

Published in final edited form as:

Biometrics. 2015 June; 71(2): 439-449. doi:10.1111/biom.12293.

A direct method to evaluate the time-dependent predictive accuracy for biomarkers

Weining Shen*, Jing Ning**, and Ying Yuan***

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, U.S.A

Summary

Time-dependent receiver operating characteristic (ROC) curves and their area under the curve (AUC) are important measures to evaluate the prediction accuracy of biomarkers for time-to-event endpoints (e.g., time to disease progression or death). In this paper, we propose a direct method to estimate AUC(t) as a function of time t using a flexible fractional polynomials model, without the middle step of modeling the time-dependent ROC. We develop a pseudo partial-likelihood procedure for parameter estimation and provide a test procedure to compare the predictive performance between biomarkers. We establish the asymptotic properties of the proposed estimator and test statistics. A major advantage of the proposed method is its ease to make inference and to compare the prediction accuracy across biomarkers, rendering our method particularly appealing for studies that require comparing and screening a large number of candidate biomarkers. We evaluate the finite-sample performance of the proposed method through simulation studies and illustrate our method in an application to AIDS Clinical Trials Group 175 data.

Keywords

Biomarker evaluation; pseudo partial-likelihood; time-dependent AUC; time-dependent ROC

1. Introduction

Recent developments in genomic and medical research have generated a large number of candidate biomarkers that have potential use in predicting an individual's future health status. These biomarkers are often biochemical or genetic measurements on a continuous scale that calibrate the progression of patients' responses to the treatment of a disease, e.g., gene expression profiling and protein mass spectrometry (Cai and Moskowitz, 2004) and molecular biomarkers (Ransohoff, 2004). Before adopting biomarkers into clinical use, it is crucial to evaluate their accuracy (i.e., sensitivity and specificity) in predicting the likelihood

^{*}wshen@mdanderson.org. **jning@mdanderson.org. ***yyuan@mdanderson.org.

This paper has been submitted for consideration for publication in Biometrics

^{6.} Supplementary Materials

Additional simulation results, code/data and the proof of theorems, referenced in Sections 2 and 3, are available with this paper at the Biometrics website on Wiley Online Library.

of a patient experiencing a future event of interest (e.g., disease onset, disease progression and death).

One major challenge in evaluating the prediction accuracy of biomarkers arises from the fact that in many applications, the primary endpoint is the time to an event, and the risk of a patient developing the event of interest often changes over time given the baseline measure of biomarkers. For example, in a prostate cancer study, investigators analyzed data collected from samples of prostate-specific antigen and found that the test with the longest lead time did not necessarily correspond to a better diagnostic score (Pearson et al., 1994, 1996; Etzioni et al., 1999). Hence, they recommended evaluating the prediction performance of tests over a range of time rather than at the time of diagnosis. In this situation, the use of the conventional binary outcome-based receiver operating characteristic (ROC) methods and the area under the ROC curve (AUC) is not suitable and may lead to misleading results.

To address this issue, several time-dependent measures have been proposed in the literature, including time-dependent sensitivity, specificity, ROC and AUC (Heagerty et al., 2000; Heagerty and Zheng, 2005). In this article, we consider the incident/dynamic (I/D) time-dependent ROC, which was first defined by Heagerty and Zheng (2005), and focus on estimating its time-dependent AUC. One major advantage of the I/D definition over other types of time-dependent ROCs (e.g., cumulative/dynamic) is that it gives a natural way to define a weighted time-averaged summary of AUC, which is called IAUC. Moreover, Heagerty and Zheng (2005) showed that IAUC is equivalent to the global concordance summary with a proper weight function.

Two types of methods have been proposed to estimate time-dependent AUC and IAUC. The first type of methods, which might be called "indirect" methods, focus on modeling timedependent ROC curves, and obtain the estimate of the AUC as a byproduct. Heagerty and Zheng (2005) considered time-dependent ROC estimation based on a Cox model for survival time. Cai et al. (2006) used generalized linear models to estimate time-dependent sensitivity and specificity. In Song and Zhou (2008), a covariate-specific ROC estimator was constructed by including baseline covariates in the modeling of hazard functions. Zheng et al. (2008, 2010) considered positive predictive value curve estimation based on flexible regression-type method, which was extended thereafter in the presence of competing risks (Zheng et al., 2012). An extension of the time-dependent ROC for competing risks has been considered by Saha and Heagerty (2010). For the purpose of estimating the time-dependent AUC and IAUC, the "indirect" methods have some drawbacks. For example, most ROCestimation procedures rely on strong parametric assumptions such as proportional hazards, which may lead to potential bias due to the misspecification of the time-to-event model. In addition, because these methods are "indirect", they are not particularly efficient and require unnecessary intensive computation.

The second type of methods, which are called "direct" methods, bypass the step of estimating ROC curves, and directly estimate the time-dependent AUC. Song et al. (2012) considered a flexible nonparametric estimation procedures based on the proportion of concordance pairs, and adopted the inverse-probability-of-censoring weighting technique (Robins et al., 1994) to handle the censored outcomes. Recently, Saha-chaudhuri and

Heagerty (2013) proposed estimating the AUC nonparametrically by a locally weighted mean rank (WMR) kernel-based method with a bandwidth selected by cross-validation or other data-driven methods. The resulting WMR estimate is essentially a smoothed version of the proportion of concordant case-control pairs in the risk set and hence bypasses the need to model the distribution of the event time, and handles censoring in a simple way.

In modern biomarker studies, we often face the challenge of comparing and screening a large number of candidate biomarkers generated by high-throughput methods. It is highly desirable to have a direct method that is computationally easy, so that we can perform fast screening and comparison of the predictive accuracy of candidate biomarkers based on time-dependent AUC and IAUC. The WMR method of Saha-chaudhuri and Heagerty (2013) may not be particularly suitable for this task because the choice of bandwidth is often computationally intensive. In addition, it is not obvious how one should estimate and make inference of the IAUC under the WMR method, which mainly focuses on point-wise estimates of AUC(t) at a given time point t.

The goal of this paper is to propose a new direct method that provides straightforward inference for AUC(t) and IAUC. Our method directly models AUC(t) as a function of time t using a flexible and yet parsimonious parametric model, namely, the fractional polynomials model proposed by Royston and Altman (1994). We construct and maximize a pseudo partial-likelihood function to obtain the estimates of model parameters. We show that the resulting estimating equations can be viewed as U-statistics and derive asymptotic variance of the estimates. Under the proposed approach, it is convenient to obtain the confidence bands for the AUC(t), and estimate and compare the IAUC across biomarkers. These characteristics render our method particularly appealing for studies that require comparing and screening a large number of candidate biomarkers.

The rest of the paper is organized as follows. In Section 2, we introduce a pseudo partial-likelihood approach to estimate the time-dependent AUC and derive its large-sample properties. We construct a testing procedure to compare multiple biomarkers in Section 2.4. We report some simulation results in Section 3 that illustrate our methodology. We also discuss a real data example in Section 4, and provide technical proofs and additional numerical results in the online supplementary file.

2. Method

2.1 Notation and estimand

Suppose that there are n subjects in the study. For subject i, we denote event time and censoring time by T_i and C_i , and assume that C_i is independent of T_i . Define the observed event time $Z_i = \min(T_i, C_i)$ and the failure indicator $\delta_i = 11(T_i - C_i)$, where $11(\cdot)$ is the indicator function. Denote $R(t) = \{j : Z_j - t\}$ as the risk set at time t. We consider independently distributed biomarkers M_1, \ldots, M_n , and assume that higher marker values are related to higher disease levels and shorter event times.

Because an ROC curve is completely determined by the set of combinations of true-positive (TP) and false-positive (FP) rates, it suffices to extend the definitions of TP and FP rates to

accommodate the time-dependent setting. Different time-dependent extensions of the TP and FP rates have been proposed in the literature. Here, we have adopted the one proposed by Heagerty and Zheng (2005):

$$TP_t(m) = P(M_i > m | T_i = t), \quad FP_t(m) = P(M_i > m | T_i > t),$$

where $TP_t(m)$ (i.e., sensitivity) measures the expected fraction of subjects with a biomarker value greater than m among the subpopulation of individuals who experience the event of interest at time t, while 1 $FP_t(m)$ (i.e., specificity) measures the fraction of subjects with a biomarker value less than or equal to m among those who survive beyond time t. Heagerty and Zheng (2005) refer to these measures as time-dependent incidence sensitivity and dynamic specificity, respectively. Under this definition, the time-dependent ROC is defined as

$$ROC_{t}\left(p\right) = TP_{t}\left\{ \left[FP_{t}\right]^{-1}\left(p\right)\right\}, \text{ for } p \in \left[0,1\right],$$

and the corresponding AUC(t) is given by

$$AUC(t) = P(M_i > M_i | T_i = t, T_i > t)$$
. (1)

2.2 Model AUC(t) using fractional polynomials

To model the AUC(t), a nonparametric method, such as the WMR estimate (Saha-chaudhuri and Heagerty, 2013), is flexible and imposes few assumptions on the functional form. However, the fitting process (e.g., selecting a bandwidth) can be computationally intensive, and estimating and making inference for the IAUC is diffcult with this type of method. In this article, we model the AUC(t) using a flexible parametric model that is parsimonious, and easy to fit and to use in making inference.

Specifically, letting $\eta(\cdot)$ denote a link function, e.g., logistic function, we directly model the AUC(t) as a parametric function of time t using fraction polynomials of K degrees as follows:

$$\eta\left(AUC\left(t\right)\right) = \sum_{k=0}^{K} \beta_k t^{(p_k)},$$

where for $k = 1, \ldots, K$,

$$t^{(p_k)} = \begin{cases} t^{p_k} & \text{if} \quad p_k \neq 0 \\ \ln(t) & \text{if} \quad p_k = 0 \end{cases}, \quad (2)$$

where $p_1 \ldots p_K$ are real-valued powers, and β_0, \ldots, β_K are unknown regression parameters. Royston and Altman (1994) showed that unlike conventional polynomials, for which lower order polynomials offer a limited family of shapes and high order polynomials

may fit poorly at the extreme values of the covariates, fractional polynomials have considerable flexibility and can mimic most function shapes we encounter in practice. The fractional polynomial regression provides an attractive alternative approach to conventional nonparametric methods, such as spline and kernel smoothers, while keeping the model parsimonious and inference straightforward given the closed-form variance formulae derived in the next subsection. According to Royston and Altman (1994), we will adopt the choice of powers p_1, \ldots, p_K in a set $\{-2, -1, -1/2, 0, 1/2, 1, 2\}$, which will be good enough for most applications. For notational simplicity, we denote our estimator of the AUC(t) as $\rho(t, \beta)$.

Our strategy of estimating $\rho(t, \beta)$ is based on the fact that the conditional event in equation (1) is the same event that has being used to construct the risk function at time t in the partial-likelihood. More specifically, let $s_1, ..., s_H$ be the ordered unique event times for $z_1, ..., z_n$. For each event time s_h , we denote its risk set by $R(s_h)$. Motivated by equation (1), we consider two types of events conditional on each risk set $R(s_h)$ derived from the observed data, denoted by

$$\begin{array}{ll} e_{1}\left(M_{i}, M_{j}, Z_{i}, Z_{j}\right) = & \left\{M_{i} {>} M_{j} | Z_{i} {=} s_{h}, \delta_{i} {=} 1, j \in R\left(s_{h}\right)\right\}, \\ e_{2}\left(M_{i}, M_{j}, Z_{i}, Z_{j}\right) = & \left\{M_{i} \leq M_{j} | Z_{i} {=} s_{h}, \delta_{i} {=} 1, j \in R\left(s_{h}\right)\right\}. \end{array}$$

Event $e_1(M_i, M_j, Z_i, Z_j)$ occurs if subject j has a smaller biomarker value compared with that of subject i, given that subject j has longer survival than subject i, so we call it a "concordant" event. Event $e_2(M_i, M_j, Z_i, Z_j)$ occurs if subject j has a larger biomarker value compared with that of subject i, given that subject j has longer survival than subject j; we similarly call this a "discordant" event.

For each event time s_h , the counts of two types of events are

$$\begin{split} n_{1}\left(s_{h}\right) &= & \sum_{j} 11\left\{j{:}M_{i}{>}M_{j}|Z_{i}{=}s_{h}, \delta_{i}{=}1, j \in R\left(s_{h}\right)\right\}, \\ n_{2}\left(s_{h}\right) &= & \sum_{j} 11\left\{j{:}M_{i} \leq M_{j}|Z_{i}{=}s_{h}, \delta_{i}{=}1, j \in R\left(s_{h}\right)\right\}. \end{split}$$

Note that $AUC(t) = P(M_i > M_j/T_i = t, T_j > t)$, thus at each event time point s_h , conditional on risk set $R(s_h)$, the count $n_1(s_h)$ follows a binomial distribution with a probability of $AUC(s_h)$. Motivated by this fact, we construct a pseudo partial-likelihood by multiplying all probabilities of observing the concordant or discordant counts over all of the risk sets from the observed event times. Synthesizing this information, we have the pseudo partial-likelihood as

$$L(\boldsymbol{\beta}) \propto \Pi_{h=1}^{H} \rho(s_h; \boldsymbol{\beta})^{n_1(s_h)} \{1 - \rho(s_h; \boldsymbol{\beta})\}^{n_2(s_h)}, \quad (3)$$

Maximizing this pseudo partial-likelihood, or alternatively solving the corresponding score equations, yields parameter estimate $\hat{\beta}$. Then, by using (2), we obtain the time-dependent AUC estimate $\rho(t, \beta)$ as a smooth function of time t and β .

Notice that the above pseudo partial-likelihood $L(\beta)$ is not a true likelihood function because the probabilities of observing the concordant or discordant counts conditional on the risk sets are multiplied together without accounting for the possible correlations between observed information from different risk sets. Nevertheless, since each term in $L(\beta)$ is a legitimate conditional probability, the corresponding score equation is an unbiased estimating equation for β and the corresponding maximum pseudo partial-likelihood estimator can be shown to be consistent and asymptotically normal. In fact, the proposed pseudo partial-likelihood belongs to the family of composite likelihoods (Lindsay, 1988; Cox and Reid, 2004), for which (weighted) conditional or marginal densities are multiplied, whether or not they are independent. For more discussion on the composite likelihood methods, we refer to the excellent review paper by Varin et al. (2011) and the references therein.

In practice, it is also often of interest to estimate the IAUC, defined as

$$C(\tau) = \int_0^{\tau} w(t;\tau) AUC(t) dt$$
, (4)

where $w(t, \tau)$ is a non-negative weight function that integrates to 1 on the observation time period $[0, \tau]$. The IAUC provides a global summary of the prediction accuracy for biomarkers, and can be used to compare the prediction performance between biomarkers and screen a pool of candidate biomarkers. Heagerty and Zheng (2005) showed that when $w(t, \tau)$ is proportional to the product of the density function and the survival function of t, $C(\tau)$ is equivalent to the global dependence measure, Kendall's tau. Under the proposed framework, the estimation of $C(\tau)$ is computationally simple, and a "plug-in" estimator of $C(\tau)$ is given by

$$C\left(\hat{\boldsymbol{\beta}};\tau\right) = \int_{0}^{\tau} w\left(t;\tau\right) \rho\left(t;\hat{\boldsymbol{\beta}}\right) dt.$$
 (5)

When the weight function $w(t, \tau)$ is invariant to time, e.g., $w(t, \tau) = 1/\tau$, the IAUC $C(\tau)$ can be viewed as the global average of the AUC curve.

Computationally, our method estimates the whole curve of AUC(t) as a function of β by maximizing the proposed pseudo-partial likelihood. Then the global measures such as IAUC and its standard error can be conveniently estimated as by-products, since these quantities can be obtained as closed-form functions of β . In contrast, the nonparametric methods (Song et al., 2012; Saha-chaudhuri and Heagerty, 2013) give point-wise estimate of the AUC(t) and require repeated calculations at different grid points of time to obtain the whole curve of AUC(t). Consequently, one would need to repeat point-wise calculation many times to estimate the global measures such as IAUC. The tuning of smoothing parameters may also be cumbersome since an optimization of the IMSE is required according to Saha-chaudhuri and Heagerty (2013).

So far we have assumed that no two patients have the same failure event time in the data, which is an assumption that may be violated in practice. In the presence of such ties in failure times, without loss of generality, we assume the observed event times of the *i*th and *j*th patients are the same. Without any knowledge of the true ordering of the failure event

times of patients i and j, we have to consider all possible orderings. There are 2 possible orderings in this case. We can construct two sets of unique failure event times by adding a small amount (e.g. 10^{-5}) to Z_i or Z_j . Then the likelihood is constructed by adding two likelihoods under two sets of unique failure event times together. This method, called the exact method, can be computationally intensive. For example, suppose there are d_j tied failure event times at the s_j distinct time, then d_j ! different orderings have to be considered and the likelihood is the sum of d_j ! different terms. To handle such computational diffculties in the case of many ties, we can approximate the likelihood under the exact method by randomly choosing one possible ordering from the observed data and maximizing the corresponding likelihood under this chosen ordering.

2.3 Inference: Asymptotic results

When making inferences about the proposed estimators of the parameters, the major challenge lies in the fact that the proposed pseudo partial-likelihood cannot be treated as either a classical partial-likelihood function or a regular likelihood function. Specifically, the asymptotic variance of the estimators is not the inverse of the negative second derivative of the pseudo partial-likelihood. To evaluate the asymptotic behavior of the proposed estimators, we will further study the score equations that correspond to the proposed likelihood. We re-write the log-likelihood in the following way:

$$l(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{i} 11(Z_{j} > Z_{i}) \left\{ 11(M_{i} > M_{j}) \log (\rho(Z_{i}; \beta)) + 11(M_{i} \leq M_{j}) \log (1 - \rho(Z_{i}; \beta)) \right\}.$$

By taking the derivative, we obtain the score function as

$$U\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n} \sum_{i=1}^{n} \delta_{i} 11\left(Z_{j} > Z_{i}\right) \left\{11\left(M_{i} > M_{j}\right) \frac{\nabla_{\boldsymbol{\beta}} \rho\left(Z_{i}; \boldsymbol{\beta}\right)}{\rho\left(Z_{i}; \boldsymbol{\beta}\right)} - 11\left(M_{i} \geq M_{j}\right) \frac{\nabla_{\boldsymbol{\beta}} \rho\left(Z_{i}; \boldsymbol{\beta}\right)}{1 - \rho\left(Z_{i}; \boldsymbol{\beta}\right)}\right\},$$

where $\nabla_{\beta} f(\beta)$ is the first order derivative of a function f with respect to a vector β . Denote $\hat{\beta}$ as a solution of $U(\beta) = 0$ and β_0 as the true value of β . By Taylor's expansions of $U(\beta)$ around β_0 to the second order, we can show consistency of $\hat{\beta}$. Moreover, based on the exchangeable structure in the score equation, we obtain the conclusion that $\hat{\beta}$ is asymptotically normally distributed under certain regularity conditions using U-statistic projection theory. We summarize these results in the following theorem.

Theorem 1—Under the regularity conditions (A1) – (A3) listed in the supplementary *file*, the estimator $\hat{\beta}$, at which the local maxima of L(β) occurs, converges to β_0 . Moreover, is asymptotically normal with mean β_0 and covariance matrix $V = \Sigma_1^{-1} \Sigma_2 \Sigma_1^{-1}$, where

$$\begin{split} & \boldsymbol{\Sigma_{1}} \!=\! E\left\{ -\nabla_{\boldsymbol{\beta}} f_{12}\left(\boldsymbol{\beta}\right); \!\boldsymbol{\beta}_{0} \right\}, \quad \boldsymbol{\Sigma_{2}} \!=\! 4cov\left\{ g_{12}\left(\boldsymbol{\beta}\right), g_{13}\left(\boldsymbol{\beta}\right); \!\boldsymbol{\beta}_{0} \right\}, \quad g_{ij}\left(\boldsymbol{\beta}\right) = \left\{ f_{ij}\left(\boldsymbol{\beta}\right) + f_{ji}\left(\boldsymbol{\beta}\right) \right\} / 2, \\ & f_{ij}\left(\boldsymbol{\beta}\right) = \! \int_{0}^{\tau} \! \int_{0}^{\tau} \! 11\left(t \! > \! s\right) \left\{ 11\left(M_{i} \! > \! M_{j}\right) \frac{\nabla_{\boldsymbol{\beta}} \rho\left(s; \boldsymbol{\beta}\right)}{\rho\left(s; \boldsymbol{\beta}\right)} - 11\left(M_{i} \leq M_{j}\right) \frac{\nabla_{\boldsymbol{\beta}} \rho\left(s; \boldsymbol{\beta}\right)}{1 - \rho\left(s; \boldsymbol{\beta}\right)} \right\} dN_{i}\left(s\right) dN_{j}\left(t\right), \end{split}$$

and $\{N_i(t), t = 0\}$ is a right-continuous counting process that records the number of events occurring over [0, t].

The covariance matrix V can be consistently estimated by $\hat{\Sigma}_1^{-1}\hat{\Sigma}_2\hat{\Sigma}_1^{-1}$, where

$$\hat{\Sigma}_{1} = -\frac{1}{n^{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{i} 11(Z_{j} \geq Z_{i}) \nabla_{\beta} \left\{ 11(M_{i} > M_{j}) \frac{\nabla_{\beta} \rho(Z_{i}; \hat{\beta})}{\rho(Z_{i}; \hat{\beta})} - 11(M_{i} \leq M_{j}) \frac{\nabla_{\beta} \rho(Z_{i}; \hat{\beta})}{1 - \rho(Z_{i}; \hat{\beta})} \right\},$$
and
$$\hat{\Sigma}_{2} = \frac{4}{n(n-1)(n-2)} \sum_{i=1}^{n} \sum_{j \neq j', j, j' \neq i} g_{ij} (\hat{\beta}) g_{ij'} (\hat{\beta})^{T}.$$

The asymptotic behavior of the estimated AUC curve and its integration, which follows that of $\hat{\beta}$, are summarized in the next corollary.

Corollary 1—*Under the regularity conditions (A1)–(A3),* $\rho\left(t;\hat{\boldsymbol{\beta}}\right)$ is a consistent estimator of AUC(t). Given that the weight function w is continuous on $[0, \tau]$, it follows that

$$\sqrt{n}\left(C\left(\hat{\pmb{\beta}};\tau\right)-C\left(\pmb{\beta_0};\tau\right)\right)\xrightarrow{d}N\left(0,V\right),$$

The variance V can be consistently estimated by

$$\hat{V} = \frac{4}{n(n-1)(n-2)} \sum_{i=1}^{n} \sum_{j \neq j', j, j' \neq i} s_{ij} (\hat{\beta}) s_{ij'} (\hat{\beta}).$$

2.4 Testing: comparing the predictive accuracy of biomarkers

Rapid progress in biomedical technology has allowed the scientific community to identify numerous biomarkers, such as those for cancer diagnosis and prognosis. With such a large number of candidate biomarkers, it is highly important to select the most useful biomarkers based on their prediction performances. In this section, we use AUC curves and their integration measures to address the scientific question, "How do we compare the prediction accuracies of two biomarkers, M^A and M^B ?" . We consider $\rho(t, \beta^A) - \rho(t, \beta^B)$ as a quantitative measure of the prediction difference between M^A and M^B and derive its asymptotic distribution, as presented in Corollary 2. Moreover, we construct a test statistic based on a global measure of the prediction difference between two markers across the entire study period. In this subsection, we will use the respective superscripts A and B to denote the corresponding asymptotic quantities of M^A and M^B derived in the previous subsection.

Corollary 2—Under the regularity conditions (A1) – (A3) listed in the supplementary file, $\rho\left(t;\hat{\boldsymbol{\beta}}^{A}\right) - \rho\left(t;\hat{\boldsymbol{\beta}}^{B}\right)$ is asymptotically normal with mean AUC^A(t) AUC^B(t) and variance V ^{AB}(t), where

$$\begin{split} V^{AB}\left(t\right) = & 4cov\left\{h_{12}\left(\boldsymbol{\beta}^{A}, \boldsymbol{\beta}^{B}, t\right), h_{13}\left(\boldsymbol{\beta}^{A}, \boldsymbol{\beta}^{B}, t\right); \boldsymbol{\beta_{0}^{A}}, \boldsymbol{\beta_{0}^{B}}\right\}, \\ and \quad & h_{ij}\left(\boldsymbol{\beta_{0}^{A}}, \boldsymbol{\beta_{0}^{B}}, t\right) = & \nabla_{\boldsymbol{\beta}}^{T}\rho\left(t; \boldsymbol{\beta_{0}^{A}}\right)\left(\boldsymbol{\Sigma_{1}^{A}}\right)^{-1}g_{ij}^{A}\left(\boldsymbol{\beta_{0}}\right) - \nabla_{\boldsymbol{\beta}}^{T}\rho\left(t; \boldsymbol{\beta_{0}^{B}}\right)\left(\boldsymbol{\Sigma_{1}^{B}}\right)^{-1}g_{ij}^{B}\left(\boldsymbol{\beta_{0}}\right). \end{split}$$

Moreover, $C\left(\hat{\boldsymbol{\beta}}^A;\tau\right) - C\left(\hat{\boldsymbol{\beta}}^B;\tau\right)$ is asymptotically normal with mean $C\left(\boldsymbol{\beta_0^A};\tau\right) - C\left(\boldsymbol{\beta_0^B};\tau\right)$ and variance V^{AB} , where

$$V^{AB}{=}4cov\left\{r_{12}\left(\boldsymbol{\beta}^{A},\boldsymbol{\beta}^{B}\right),r_{13}\left(\boldsymbol{\beta}^{A},\boldsymbol{\beta}^{B}\right);\boldsymbol{\beta}_{\boldsymbol{0}}^{A},\boldsymbol{\beta}^{B}\right\}\quad and\\ r_{ij}\left(\boldsymbol{\beta}^{A},\boldsymbol{\beta}^{B}\right){=}\int_{0}^{\tau}w\left(t;\tau\right)\nabla_{\boldsymbol{\beta}}^{T}\rho\left(t;\boldsymbol{\beta}^{A}\right)dt\left(\boldsymbol{\Sigma}_{1}^{A}\right)^{-1}g_{ij}^{A}\left(\boldsymbol{\beta}^{A}\right)-\int_{0}^{\tau}w\left(t;\tau\right)\nabla_{\boldsymbol{\beta}}^{T}\rho\left(t;\boldsymbol{\beta}^{B}\right)dt\left(\boldsymbol{\Sigma}_{1}^{B}\right)^{-1}g_{ij}^{B}\left(\boldsymbol{\beta}^{B}\right).$$

The variance $V^{AB}(t)$ of $\rho\left(t;\hat{\boldsymbol{\beta}}^{A}\right)-\rho\left(t;\hat{\boldsymbol{\beta}}^{B}\right)$ can be consistently estimated by $V^{\hat{A}B}(t)$, which is given by

$$\frac{4}{n(n-1)(n-2)} \sum_{i=1}^{n} \sum_{j \neq j', j, j' \neq i} h_{ij} \left(\hat{\boldsymbol{\beta}}^{A}, \hat{\boldsymbol{\beta}}^{B}, t \right) h_{ij'} \left(\hat{\boldsymbol{\beta}}^{A}, \hat{\boldsymbol{\beta}}^{B}, t \right).$$

Similarly, the variance V^{AB} of $C\left(\hat{\boldsymbol{\beta}}^{A};\tau\right) - C\left(\hat{\boldsymbol{\beta}}^{B};\tau\right)$ can be consistently estimated by

$$\hat{V} = \frac{4}{n(n-1)(n-2)} \sum_{i=1}^{n} \sum_{j \neq i', j, j' \neq i} r_{ij} \left(\hat{\boldsymbol{\beta}}^{A}, \hat{\boldsymbol{\beta}}^{B} \right) r_{ij'} \left(\hat{\boldsymbol{\beta}}^{A}, \hat{\boldsymbol{\beta}}^{B} \right).$$

One hypothesis of interest for the overall predictive equivalence of biomarkers M^A and M^B can be formulated as $H_0:C\left(\boldsymbol{\beta}_0^A;\tau\right)=C\left(\boldsymbol{\beta}_0^B;\tau\right)$. Following Corollary 2, we can accordingly construct a Wald-type test statistic, termed T^{AB} , as

$$T^{AB} = \frac{C\left(\hat{\boldsymbol{\beta}}^{A}; \tau\right) - C\left(\hat{\boldsymbol{\beta}}^{B}; \tau\right)}{\sqrt{\hat{v}^{AB}}}, \quad (6)$$

which converges to a standard normal distribution under the null hypothesis H_0 .

It is straightforward to generalize the proposed two-biomarker test statistic (6) for simultaneous testing of multiple markers $\{M^1,...,M^K\}$. Specifically, we wish to test for

$$H_0^K: C\left(\nabla \boldsymbol{\beta_0^1}; \tau\right) = C\left(\boldsymbol{\beta_0^2}; \tau\right) = \cdots = C\left(\boldsymbol{\beta_0^K}; \tau\right).$$

Accordingly, we can extend the test statistic (6) by considering $T^K = C_K^T (\hat{V}^K)^{-1} C_K$, where C_K is a (K-1)-dimensional vector:

$$C_{K} = \left(C\left(\hat{\boldsymbol{\beta}}^{1};\tau\right) - C\left(\hat{\boldsymbol{\beta}}^{2};\tau\right), C\left(\hat{\boldsymbol{\beta}}^{1};\tau\right) - C\left(\hat{\boldsymbol{\beta}}^{3};\tau\right), \dots, C\left(\hat{\boldsymbol{\beta}}^{1};\tau\right) - C\left(\hat{\boldsymbol{\beta}}^{K};\tau\right)\right)^{T},$$

and $V^{\hat{K}}$ is a $(K-1) \times (K-1)$ matrix, the (s, t)-th element of which is

$$v_{s,t} = \frac{4}{n(n-1)(n-2)} \sum_{i=1}^{n} \sum_{j \neq j', j, j' \neq i} r_{ij} \left(\hat{\boldsymbol{\beta}}^{1}, \hat{\boldsymbol{\beta}}^{s+1} \right) r_{ij'} \left(\hat{\boldsymbol{\beta}}^{1}, \hat{\boldsymbol{\beta}}^{t+1} \right).$$

Under the null hypothesis H_0^K , the test statistic T^K converges to a standard $\chi^2_{_{K-1}}$ distribution.

3. Simulation

To explore the theoretical results empirically, we conducted several simulation studies for the proposed estimators of the AUC(t) and the test that compares the predictive accuracy of biomarkers. In particular, we considered a logistic link function and chose the power of the polynomials as $\{-2, -1, -1/2, 0, 1/2, 1, 2\}$.

3.1 Estimating the AUC(t) of a single biomarker

We first considered estimating the AUC(t) of a single biomarker. Following the settings of Heagerty and Zheng (2005), we generated 200 pairs of survival time T and marker M, in which (log T, M) follows a bivariate normal distribution with means 0, variances 1 and correlation -0.7. In addition, we generated an independent censoring time uniformly from (0, τ^*) such that either 20% or 40% of the subjects were censored. We calculated the estimated values of the AUC(t) at various time points based on t = 1000 replications and compared the fitting results obtained by the nonparametric kernel estimator (WMR) with bandwidth of t t as reported by Saha-chaudhuri and Heagerty (2013).

Table 1 lists the empirical bias (Bias), empirical standard error (SD), average of the model-based standard error estimates (SE) and the coverage probability (CP) of the 95% confidence intervals for the estimates. Under both 20% and 40% censoring rates, we nd good estimation of the AUC(t) by the proposed method in that the biases of the estimates are small and the empirical standard errors are close to the averages of the model-based standard error estimates. The empirical coverage probabilities based on the 95% estimated confidence intervals are close to the nominal value of 0.95, especially when we have sufficient data around the given time points, i.e., -1.5 log T 0.5. Compared to the WMR, the proposed method has smaller biases and empirical standard deviations (i.e., about 10% smaller), except at the tails (log T = 1), due partly to the use of a regression-type function, ρ , in our method.

We considered another example in which the data (log T, M) were generated from a mixture of normal distributions ZN(-1.5, -1.5, 1, 1, 0) + (1 - Z)N(0, 0, 1, 1, -.8), where Z follows a Bernoulli distribution with success probability 0.2. An independent censoring time was generated such that 20% of the patients were censored. We summarize the estimated AUC(t) and compare the results with those of the WMR method in Table 2. It can be seen that the proposed method has comparable estimation bias and more accurate coverage probabilities than the WMR method.

Note that our method is computationally easy and straightforward in producing confidence bands by using the delta method. In Figure 1, we plot the estimated AUCs and the corresponding 95% confidence bands under several data generating mechanisms. For the first two subplots, (log T, M) follows a bivariate normal distribution $N(0, 0, 1, 1, \gamma)$ with $\gamma = .5$ and -.9. For the third subplot, we generate (log T, M) from a mixture of normal distributions ZN(-1.5, -1.5, 1, 1, 0) + (1 - Z)N(0, 0, 1, 1, -.6), where Z follows a Bernoulli distribution with success probability 0.2. For the fourth subplot, we generate M from an exponential distribution with mean one, and let T follow another exponential distribution with mean (1 + M) $^{-1}$ given fixed M. It can be seen that the estimated curve shows good agreement with the true curve in each of the four scenarios.

3.2 Testing predictive performance between two biomarkers

We next considered estimating the difference in the AUC(t) between two markers, M^A and M^B . For each of N = 1000 replications, we generated n = 200 realizations of $(\log T, M^A, M^B)$ from a three-dimensional Gaussian distribution with means zero, variances one and corr(log T, M^A) = -0.7, corr(log T, M^B) = -0.5, corr(M^A, M^B) = 0.8. An independent censoring time was generated uniformly from $(0, \tau^*)$ such that either 20% or 40% of the observations were censored. As summarized in Table 3, in estimating the AUC(t), the performance of the proposed method is comparable to that of the WMR; whereas the empirical standard deviation of the proposed estimator is 10% smaller than that of the WMR. The estimated coverage probabilities based on 95% confidence intervals are close to the nominal value of 95% (range: 94.4% - 96.9%), except when the time point t is close to the boundary of the study period due to the limited observed information. The lengths of the confidence intervals increased with time, and sharply increased at the tails, which was due to the decreasing sizes of the risk sets over time. We also calculate the difference in IAUC of two biomarkers. For the 20% censoring rate, the estimation bias is .001, with an estimated standard error of .020, an empirical standard error of .021 and a coverage probability of 93.5%. For 40% censoring, the estimation bias is .001, with an estimated standard error of .023, an empirical standard error of .025 and a coverage probability of 90.8%.

We also consider another example where M^A and M^B follow a bivariate normal distribution and the survival time T is generated from a Cox model. More details and relevant results are presented in the supplementary material.

We then evaluated the performance of the test statistic proposed in (6), including the size and power. We generated (log T, M^A , M^B) from a normal distribution with means zero, variances one and correlations $\operatorname{corr}(\log T, M^A) = \gamma_A$, $\operatorname{corr}(\log T, M^B) = \gamma_B$ and $\operatorname{corr}(M^A, M^B) = 0$. We fixed $\gamma_B = -0.4$ and then let γ_A take a set of different values in $\{-0.4, -0.5, \ldots, \}$

-0.9}. We calculated the probabilities of rejecting the null hypothesis

 $H_0:C\left(\beta_0^A;\tau\right)=C\left(\beta_0^B;\tau\right)$ based on percentiles (0.025 and 0.975) of the standard normal distribution. We based the calculation of the size of the test on 5000 Monte Carlo replications, and that of the power on 1000 replications. When the null hypothesis is true (i.e., $\gamma_A=-0.4$), the sizes of the proposed test were in a reasonable range: 0.052-0.083 under all scenarios considered. This suggests that the asymptotic approximation for the test statistic (6) is adequate for moderate sample sizes. As expected, the powers of the test increased with decreasing values of γ_A . The power decreased with an increase in the degree of censoring, indicating that the censoring percentage has an impact on the power.

4. Real data application

We applied the proposed methodology to the data collected from the AIDS Clinical Trials Group (ACTG) 175 study (Hammer et al., 1996). This study enrolled 2139 individuals and equally randomized them into one of four treatment arms that provided either single drug or combination treatments involving zidovudine (AZT), didanosine (ddI) and zalcitabine (ddC). Here, we focus on the combination treatment arm of AZT and ddC. Among 524 patients randomized to that treatment arm, 109 patients developed AIDS or died before the end of the study. The outcome of interest is the time to the development of AIDS. The follow-up time ranged from 14 days to 1231 days, with a median time of 997 days. Our goal here is to assess the ability of two biomarkers, CD4 and CD8 T cell counts (Catalfamo et al., 2011), to predict the risk of developing AIDS. Specifically, we are interested in two assessments: how well do the two biomarkers discriminate between patients who are at high risk of the progression of HIV infection to AIDS or death and those who are not; and which biomarker better discriminates the patients at higher risk.

We made the first assessment by using the proposed method to estimate the respective AUC(t) of two biomarkers. We applied a logistic link function with the power of the polynomials as $\{-2, -1, -1/2, 0, 1/2, 1, 2\}$. We present two estimated AUC curves and their 95% confidence bands for the two biomarkers in the first two subplots of Figure 2. These two AUC curves decrease steadily over time. Specifically, over the first 400 days of follow-up, the value of the AUC(t) of CD4 is above 0.7, implying that for any time t before day 400, the probability of a patient who has a progression of AIDS on day t having a CD4 counts larger than a patient who survives beyond day t is at least 0.7. In contrast, the estimated AUC values of CD8 are between 0.5 and 0.6, implying that the predictive probability is close to non-informative value (0.5). It can been seen that estimates of the two AUC curves become increasingly variable with increasingly wider confidence intervals at the tails of the study period. Specifically, 10 out of 524 (1.9%) subjects died or were censored before day 200, and only 22.7% of the patients remained in the study at day 1100.

We performed the second assessment by testing the difference between the two estimated AUC curves. We present our estimates of the difference in the AUCs in the last subplot of Figure 2. It can be seen that the difference is always positive, indicating that CD4 has a better predictive performance than CD8 most of the time, although the lower confidence bound is below 0 around days 150 and 750. To determine whether such advantage is

statistically significant when accounting for the estimation variation, we tested the difference in the IAUCs between the two markers. We chose the cut-off endpoint τ to be the 95th percentile of the observed survival time to avoid unnecessary variability. The estimated difference in the IAUC is 0.110, with a standard error of 0.037, which leads to the conclusion that CD4 has a significantly better prediction accuracy than that of CD8, with a p-value of 0.002. We performed the same analysis for the other three study arms and found a similar prediction superiority for CD4 compared to CD8. These results agree with the previous ndings of Catalfamo et al. (2011).

5. Discussion

In this paper, we have proposed a semi-parametric estimator of time-dependent AUC curve and have constructed a test procedure to compare the performances of biomarkers in terms of risk prediction. The pseudo partial-likelihood approach we proposed for estimation and inference readily and advantageously handles censoring. Compared to the time-dependent ROC estimation methods available in the literature, our approach is robust as it does not require any assumptions on the underlying survival time. One major difference between our method and existing nonparametric methods is that our method estimates the entire curve as a function of β and t through the fractional polynomial function, while the non-parametric methods use a point-wise approach to estimate the AUC. Consequently, our method is computationally very convenient for obtaining global statistics such as IAUC and the confidence band plot (e.g., Figure 1) through an analytic expression. In contrast, nonparametric methods would have to repeat the kernel-based estimation procedure N times, including the selection of bandwidths, in order to estimate AUC(t) at N different time points t. In addition, estimating the summary statistics would be computationally intensive. For example, to decrease the computational burden, Saha-chaudhuri and Heagerty (2013) calculated IAUC as the average of AUC(t) at 10 time points, which inevitably will lead to some approximation errors. Conceptually, our method may be easily understandable or interpretable for practitioners, since it is a "regression-type" method, with covariates being functions of time. Note that the inference results we developed apply to other types of basis functions as well, such as wavelets and local polynomials, which may be more appealing when modeling non-smooth AUC curves with local changes. In the future, it might be of interest to develop methods to nd optimal choice of orders of fractional polynomials and the related link functions.

In this paper, our focus is to develop a computationally convenient evaluation of the predictive ability of a single biomarker, and deal with the challenge of comparing and screening a large number of candidate biomarkers generated by high-throughput methods. A more clinically meaningful question is how to build a scoring system based on multiple biomarkers and covariates selected from our evaluation. Developing rigorous statistical tools for such scoring systems is beyond the scope of this paper, and is a worthy objective for future research.

Although our semi-parametric model employs a parametric regression form of time, strong modeling assumptions are avoided by using sufficiently flexible functions of time, such as fractional polynomial functions. In the numerical studies, we conducted a sensitivity

analysis by trying a few different choices of parametric form for ρ and basis functions, including adding t^3 to the basis defined by (2) and considering a more general transformation by replacing the number 1 in the logistic function with an additional parameter. We found that the results were not sensitive to these changes. Another issue in practice is the choice of τ , which depends on the range of accurate estimation of the AUC curves. In this paper, since we specify ρ as a logistic-regression-type function, we recommend picking τ as data-dependent, for example, as the 90%th percentile of the observed survival time points. We found that our test results when comparing $C(\beta, \tau)$ did not vary much for different values of τ . For example, in the ACTG data study, when we chose τ to be the 80%th, 90%th or 95%th percentile of the observed survival time, each resulting p-value did not differ much within each treatment group.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the editor, associate editor, and two referees for comments that have led to significant improvements in the article. The research was supported in part by the National Cancer Institute grants CA154591, CA016672 and 5P50CA098258 (Yuan) and R21-HL109479 (Ning).

References

- Cai T, Moskowitz CS. Semi-parametric estimation of the binormal roc curve for a continuous diagnostic test. Biostatistics. 2004; 5:573–586. [PubMed: 15475420]
- Cai T, Pepe MS, Lumley T, Zheng Y, Jenny NS. The sensitivity and speci city of markers for event times. Biostatistics. 2006; 7:182–197. [PubMed: 16079162]
- Catalfamo M, Wilhelm C, Tcheung L, Proschan M, Friesen T, Park J-H, Adelsberger J, Baseler M, Maldarelli F, Davey R, Roby G, Rehm C, Lane C. CD4 and CD8 T Cell Immune Activation during Chronic HIV Infection: Roles of Homeostasis, HIV, Type I IFN, and IL-7. The Journal of Immunology. 2011; 186:2106–2116. [PubMed: 21257970]
- Cox DR, Reid N. A note on pseudolikelihood constructed from marginal densities. Biometrika. 2004; 91:729–737.
- Etzioni R, Pepe MS, Longton G, Hu C, Goodman G. Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. Medical Decision Making. 1999; 19:242–251. [PubMed: 10424831]
- Hammer SM, Katzenstein DA, Hughes MD, Gundacker H, Schooley RT, Haubrich RH, Henry WK, Lederman MM, Phair JP, Niu M, Hirsch MS, Merigan TC. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. The New England Journal of Medicine. 1996; 335:1081–1090. [PubMed: 8813038]
- Heagerty PJ, Lumley T, Pepe MS. Time-dependent roc curves for censored survival data and a diagnostic marker. Biometrics. 2000; 56:337–344. [PubMed: 10877287]
- Heagerty PJ, Zheng Y. Survival model predictive accuracy and roc curves. Biometrics. 2005; 61:92–105. [PubMed: 15737082]
- Lindsay BG. Composite likelihood methods. Contemporary Mathematics. 1988; 80:221–239.
- Pearson JD, Luderer AA, Metter EJ, Partin AW, Chan DW, Fozard JL, Carter HB. Longitudinal analysis of serial measurements of free and total psa among men with and without prostatic cancer. Urology. 1996; 48:4–9. [PubMed: 8973693]

Pearson JD, Morrell CH, Landis PK, Carter HB, Brant LJ. Mixed-effects regression models for studying the natural history of prostate disease. Statistics in Medicine. 1994; 13:587–601. [PubMed: 7517570]

- Ransohoff D. Rules of evidence for cancer molecularmarker discovery and validation. Nature Reviews Cancer. 2004; 4:309–314.
- Robins JM, Rotnitxky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89:846–866.
- Royston P, Altman DG. Regression using fractional polynomials of continuous covariates:

 Parsimonious parametric modelling. Journal of the Royal Statistical Society. Series C (Applied Statistics). 1994; 43:429–467.
- Saha P, Heagerty PJ. Time-dependent predictive accuracy in the presence of competing risks. Biometrics. 2010; 66:999–1011. [PubMed: 20070296]
- Saha-chaudhuri P, Heagerty PJ. Non-parametric estimation of a time-dependent predictive accuracy curve. Biostatistics. 2013; 14:42–59. [PubMed: 22734044]
- Song X, Zhou X-H. A semiparametric approach for the covariate-speci c roc curve with survival outcome. Statistica Sinica. 2008; 18:947–965.
- Song X, Zhou X-H, Ma S. Nonparametric receiver operating characteristic-based evaluation for survival outcomes. Statistics in Medicine. 2012; 31:2660–2675. [PubMed: 22987578]
- Varin C, Reid N, Firth D. An overview of composite likelihood methods. Statistica Sinica. 2011; 21:5–12.
- Zheng Y, Cai T, Jin Y, Feng Z. Evaluating prognostic accuracy of biomarkers under competing risk. Biometrics. 2012; 68:388–396. [PubMed: 22150576]
- Zheng Y, Cai T, Pepe MS, Levy WC. Time-dependent predictive values of prognostic biomarkers with failure time outcome. Journal of the American Statistical Association. 2008; 103:362–368. [PubMed: 19655041]
- Zheng Y, Cai T, Stanford JL, Feng Z. Semiparametric models of time-dependent predictive values of prognostic biomarkers. Biometrics. 2010; 66:50–60. [PubMed: 19397579]

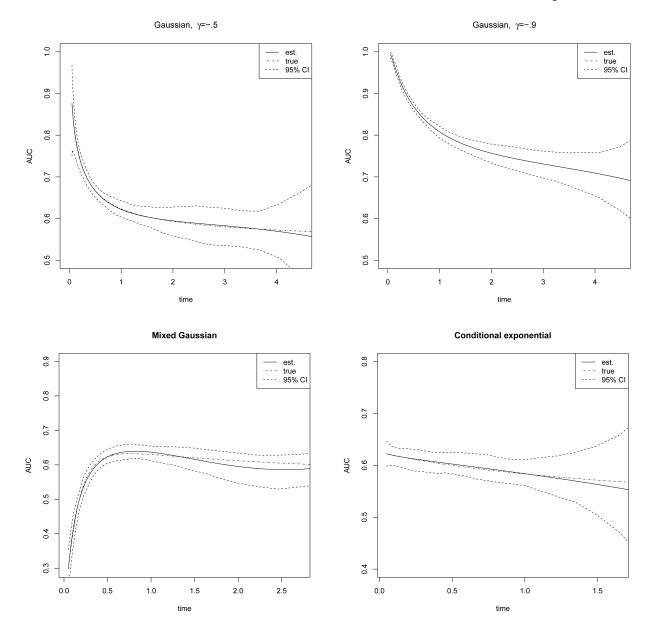
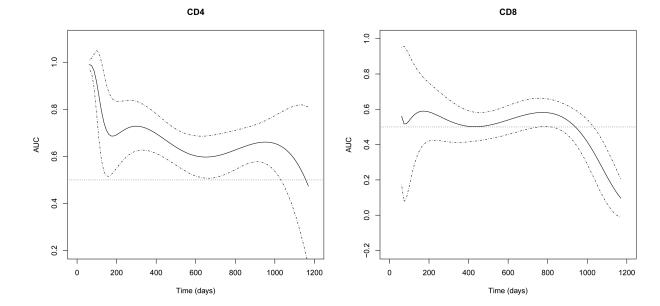


Figure 1. Estimated AUC with 95% confidence intervals.



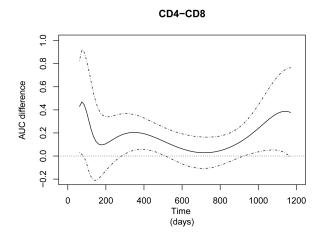


Figure 2. Estimated AUC curves/associated 95% confidence intervals of CD4, CD8 and their AUC difference.

Table 1

Simulation results for estimation of AUC of a single biomarker, comparing the proposed method with the WMR method cited from Table 1 of Saha-chaudhuri and Heagerty (2013). SD and SE are the empirical and model-based standard error estimates. CP is the calculated coverage probability of 95% confidence intervals.

			WMR			Proposed method			
Log time	AUC(t)	Bias	SD	CP(%)	Bias	SD	SE	CP (%)	
			20% се	ensoring					
-2.0	0.884	-0.008	0.055	90.2	-0.001	0.051	0.056	87.9	
-1.5	0.833	-0.005	0.040	93.4	-0.001	0.037	0.041	94.1	
-1.0	0.782	-0.002	0.037	93.2	-0.001	0.032	0.036	96.1	
-0.5	0.734	0.000	0.035	94.1	0.000	0.029	0.035	98.2	
0.0	0.693	0.002	0.037	94.0	0.000	0.032	0.037	97.0	
0.5	0.660	0.004	0.047	94.9	0.002	0.042	0.046	96.4	
1.0	0.634	0.004	0.066	93.0	0.004	0.075	0.074	93.1	
$C(\beta; \ \tau)$	0.741	-0.003	0.020	89.6	0.001	0.022	0.026	95.0	
			40% ce	ensoring					
-2.0	0.884	-0.008	0.056	89.9	-0.001	0.054	0.057	87.2	
-1.5	0.833	-0.005	0.041	92.9	-0.001	0.039	0.042	94.0	
-1.0	0.782	-0.001	0.038	92.7	-0.001	0.035	0.038	94.3	
-0.5	0.734	0.001	0.039	94.3	-0.001	0.034	0.039	97.0	
0.0	0.693	0.003	0.044	93.5	0.000	0.040	0.044	95.7	
0.5	0.660	0.006	0.061	92.6	0.006	0.062	0.065	94.3	
1.0	0.634	0.006	0.108	91.8	-0.027	0.235	0.285	85.8	
$C(\beta; \tau)$	0.741	-0.002	0.023	88.3	0.001	0.025	0.032	94.0	

Table 2

Simulation results for estimation of AUC of a single biomarker, comparing the proposed method with the WMR method cited from Table 3 of Saha-chaudhuri and Heagerty (2013). SD and SE are the empirical and model-based standard error estimates. CP is the calculated coverage probability of 95% confidence intervals.

		WMR			Proposed method			
Log time	AUC(t)	Bias	SD	CP (%)	Bias	SD	SE	CP (%)
-2.5	0.378	-0.002	0.109	87.7	0.000	0.104	0.111	91.9
-2.0	0.481	-0.004	0.083	91.1	0.006	0.077	0.085	95.2
-1.5	0.591	0.004	0.062	92.0	0.000	0.059	0.065	96.0
-1.0	0.673	0.001	0.048	91.6	-0.006	0.043	0.047	96.0
-0.5	0.709	-0.001	0.035	93.7	-0.001	0.040	0.041	95.6
0.0	0.709	0.001	0.030	94.8	0.005	0.037	0.039	95.0
0.5	0.691	0.001	0.034	92.5	-0.003	0.054	0.051	93.0
1.0	0.669	0.000	0.042	93.9	-0.002	0.088	0.102	96.2

Table 3

Simulation results for estimated difference between AUCs of two biomarkers, comparing the proposed method with the WMR method as cited from Table 2 of Saha-chaudhuri and Heagerty (2013). dAUC(t) is the true value of difference in AUCs at time point t. SD and SE are the empirical and model-based standard error estimates. CP is the calculated coverage probability of 95% confidence intervals.

		WMR			Proposed method			
Log time	dAUC(t)	Bias	SD	CP (%)	Bias	SD	SE	CP (%)
			20% се	ensoring				
-2.0	0.101	-0.002	0.062	88.8	0.001	0.060	0.063	88.7
-1.5	0.098	-0.002	0.040	94.5	0.000	0.036	0.041	96.6
-1.0	0.091	-0.001	0.032	94.6	-0.001	0.028	0.033	96.4
-0.5	0.080	-0.002	0.029	94.8	0.000	0.025	0.029	96.2
0.0	0.070	-0.001	0.030	94.9	-0.001	0.027	0.030	96.0
0.5	0.060	0.000	0.037	93.2	0.000	0.037	0.037	96.9
1.0	0.051	0.002	0.054	93.3	0.001	0.069	0.061	91.1
			40% ce	ensoring				
-2.0	0.101	-0.002	0.062	88.8	0.000	0.068	0.063	89.0
-1.5	0.098	-0.002	0.041	94.7	0.000	0.039	0.042	95.3
-1.0	0.091	-0.001	0.033	95.0	-0.001	0.031	0.034	96.0
-0.5	0.080	-0.001	0.031	94.8	-0.001	0.030	0.032	96.4
0.0	0.070	0.001	0.035	94.5	-0.002	0.032	0.036	96.9
0.5	0.060	0.001	0.048	94.8	0.003	0.055	0.053	94.4
1.0	0.051	0.002	0.100	93.3	-0.008	0.258	0.266	85.8

Table 4 Empirical rejection rates (%) for comparing biomarkers' predictive accuracy with null hypothesis $H_0:C\left(\beta_0^A;\tau\right)=C\left(\beta_0^B;\tau\right)$.

sample size	$\gamma_A =4$	$\gamma_A =5$	$\gamma_A =6$	$\gamma_A =7$	$\gamma_A =8$	$\gamma_A =9$
			20% censoring			
100	7.4	10.3	25.3	54.5	86.9	99.7
200	5.4	12.8	43.3	83.0	99.4	100
300	5.2	18.5	63.0	94.8	100	100
			40% censoring			
100	8.3	11.1	21.0	43.4	74.4	96.2
200	6.3	12.9	36.0	71.0	94.6	100
300	5.6	15.1	50.8	87.1	99.7	100