Research Article

Statistics
in Medicine

Received 18 February 2010, Accepted 1 December 2010 Published online 17 January 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4178

Time-dependent ROC analysis under diverse censoring patterns

Jialiang Li[a] and Shuangge Ma[b*†]

In biomedical studies, statistical approaches based on the Receiver Operating Characteristic (ROC) analysis have been extensively used in the evaluation of classification performance of markers and construction of classifiers. In this article, we investigate time-dependent ROC approaches for censored survival data. While most existing studies have been focused on uncensored and right-censored data, insufficient attention has been paid to other censoring schemes. This study advances from existing studies by investigating more diverse censoring schemes and developing ROC measurements under such censoring. Both estimation and inference are investigated. We conduct simulation and find satisfactory performance of the proposed approaches. We apply the proposed approaches to two real data sets, compare the prognostic power of markers, and investigate whether their linear combinations have better prognostic performance. We also explore graphical tools that can assist diagnostics and efficiently monitor the classification performance. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: time-dependent ROC; marker evaluation; survival analysis

1. Introduction

ROC (Receiver Operating Characteristic)-based approaches were first developed for classification studies with categorical outcomes. They have been extensively used in biomedical studies because of their flexibility and robustness [1, 2]. Consider a study that has the binary disease status (e.g. healthy and diseased) as outcome and diagnostic marker(s). As a classifier evaluation tool, the ROC curve displays the specificity and sensitivity for the whole range of cutoff of marker values. For a classifier which can be a marker or combination of markers, the AUC (Area under the ROC Curve) provides a much more comprehensive description of the classification performance than the simple classification error. It also provides an objective tool for comparing different classifiers. As a classifier construction tool, the ROC does not make strong assumptions on the link function and can be more robust than for example the logistic regression [3].

In series of recent studies, ROC approaches have been extended to accommodate survival data [4–6]. At each time point, survival data is equivalent to binary data with the status of being alive or dead. Thus, ROC approaches can be applied at each time point. An overall accuracy measure can be obtained by integrating the AUC over time. Note that the time-dependent ROC is usually much more complicated than an integrated ROC because of censoring and the time-dependent nature of study cohort. Censoring may cause the status of censored subjects to be not well-defined. Thus, at some time points, not all observations are equally informative. In addition, unlike in cross-sectional studies and categorical outcomes, the study cohort changes as time passes. Thus, the reliability of AUC also changes over time. Instead of a simple integration, a weighted integration with time- and data-dependent weights is desirable. For uncensored and right censored data, the time-dependent ROC and related measures have been investigated in [5] and others.

[a] *Department of Statistics and Applied Probability, Duke-NUS Graduate Medical School, National University of Singapore, Singapore City, Singapore*
[b] *School of Public Health, Yale University, CT, U.S.A.*
[*] *Correspondence to: Shuangge Ma, School of Public Health, Yale University.*
[†] *E-mail: shuangge.ma@yale.edu*

In biomedical studies, censoring patterns much more complicated than right censoring are routinely encountered. A large number of examples are presented in [7, 8] and will not be repeated here. Consider for example a special type of interval-censored data where the event time is never accurately observed and is only known to lie in a finite interval. Compared with uncensored and right-censored data, intuitively, such interval-censored data is much less informative. Quantities that have simple formulations for uncensored and right-censored data (e.g. the Kaplan–Meier (KM) estimates), do not have closed analytic forms. For ROC, we have examined the formulations in [5] and others and found that they cannot be directly extended to the data under interval censoring and other complex censoring schemes.

Two biomedical studies that have motivated our research are described in Section 5. Consider for example the calcification study. The calcification of hydrogel intraocular lenses is a complication of cataract treatment and should be treated as early as possible. In the calcification study, the values of three markers were measured. From a scientific point of view, it is of interest to identify which marker has more prognostic power. More importantly, identifying prognostic marker(s) may have important clinical implications. If markers have prognostic power, then it is possible to identify patients with high risks. Such patients should be examined more frequently so as to get treated earlier. Thus, a time-dependent ROC analysis is called upon. As the survival times were either right or interval censored, new ROC techniques are needed beyond [5].

In this article, we develop time-dependent ROC tools under complex censoring patterns. This study is warranted considering the superior performance of ROC tools with categorical, uncensored and right-censored data, and the prevalence of data under complex censoring patterns. It shares a similar spirit but significantly advances from [5], which develops the ROC for uncensored and right-censored data. In the literature, there are a large number of studies on data analysis under complex censoring patterns. Parametric modeling has been comprehensively discussed in [7]. Groeneboom and Wellner [9] investigated nonparametric modeling and established the consistency and weak convergence properties. Semiparametric modeling studies include [10, 11] among many others. The aforementioned studies have been mainly focused on the estimation and inference aspects. The main findings include the consistency and slower than $\sqrt{n}$ convergence rate of the estimates of nonparametric parameters, $\sqrt{n}$ consistency and asymptotic normality of the estimates of parametric parameters, and the necessity of special inference techniques [12]. Despite the extensive research on estimation and inference, there is a lack of study on evaluating the prognostic power of markers. In most published studies, markers have been evaluated using the estimation significance (p-value), which is related to but cannot replace the evaluation of prognostic power.

The remainder of the article is organized as follows. In Section 2, we introduce several time-dependent diagnostic accuracy measurements. Some of them have been defined in previous studies. We repeat them here for the integrity of this study. In Section 3, we describe various censoring patterns and corresponding definitions of time-dependent accuracy measurements. In Section 4, we describe the estimation and inference procedures for measurements defined in Section 2. In Section 5, we conduct numerical studies, including simulation and analysis of two practical studies, to gain further insights into the proposed approaches. Beyond estimation and inference, we also explore combining multiple markers and graphical tools for diagnostics. The article concludes with a discussion in Section 6.

2. Time-dependent accuracy measurements

Consider a survival outcome T, which measures the time to onset of disease D. Here, D may also represent death. Let D and $\bar{D}$ denote the presence and absence of disease. Let X be a marker or a linear combination of markers, and let X_D and $X_{\bar{D}}$ be the markers of a diseased and a healthy subject, respectively. With cutoff c, the True Positivity Fraction (TPF) and False Positive Fraction (FPF) are both functions of time and defined as

$$\mathtt{TPF}_t(c)=P_t\{X_D>c\}, \quad \mathtt{FPF}_t(c)=P_t\{X_{\bar{D}}>c\}.$$

Here, we use the subscript t to emphasize the dependence on time. In the literature, TPF and 1-FPF have also been referred to as sensitivity and specificity [1].

At a specific time point t, the time-specific ROC curve is the two-dimensional plot $(\mathtt{FPF}_t(c), \mathtt{TPF}_t(c))$ across all possible values of c. Since both functions are time-dependent, the resulting ROC curve is also time-dependent. The ROC curve may also be written as the composition of TPF(c) and the inverse function of FPF(c)

$$\mathrm{ROC}_t(p) = \mathtt{TPF}_t\{[\mathtt{FPF}_t]^{-1}(p)\}$$

for $p \in [0, 1]$, where $[\mathtt{FPF}_t]^{-1}(p) = \inf\{u : \mathtt{FPF}_t(u) \leqslant p\}$. This formulation is especially useful in practice for selecting a proper threshold value c such that the resulting TPF and FPF are controlled at desirable levels.

At time t, the diagnostic power of marker X can be summarized with the time-dependent AUC, which is defined as

$$\mathrm{AUC}(t) = \int_0^1 \mathrm{ROC}_t(p)\,\mathrm{d}p = \int_0^1 \mathtt{TPF}_t\{[\mathtt{FPF}_t]^{-1}(p)\}\,\mathrm{d}p.$$

Another way of defining and interpreting the AUC is via

$$\mathrm{AUC}(t) = P_t\{X_D > X_{\bar{D}}\}. \tag{1}$$

That is, the AUC is equal to the probability that a randomly chosen diseased subject has a marker value greater than that of a healthy subject.

The AUC(t) can quantify a marker's prognostic power at a fixed time point. Its integration over time can measure the overall prognostic power. Consider the followup time period $[0, \tau]$. The integrated summary measure is defined as

$$C = \int_0^\tau \mathrm{AUC}(t) \times w(t)\,\mathrm{d}t, \tag{2}$$

where $w(t)$ is the weight function. Different weights can be used to emphasize and reflect the varying reliability within different time intervals. Two weight functions have attracted special attention. The first is $w(t) = 1$, under which C is simply the algebraic mean AUC(t) value over the time period $[0, \tau]$. The second is $w(t) = 2f(t)S(t)$, where f and S are the density and survival functions of the failure time distribution, respectively. Under this weight function, C is closely related to the concordance measure and can be interpreted as

$$P(X_1 > X_2, T_1 < T_2), \tag{3}$$

where X_1 and X_2 are the marker values of two randomly chosen subjects, and T_1 and T_2 are their corresponding failure times. This measure is also closely related to the one in (1) and especially useful for studying the interrelationship between X and T. Specifically, if $C > 1/2$, a subject who develops the disease earlier is more likely to have a larger marker value than a subject who develops the disease later. Unlike AUC(t), the integrated measure C is time-independent and more appropriate for comparing the overall prognostic power of different markers.

3. Diverse censoring patterns

Consider censored survival data, where the event times of some or even all subjects are not accurately observable. As can be seen in [5] and Section 2 of this article, the definition of time-dependent ROC is based on the known binary classification of disease status over time. In [5], the time-dependent TPF and FPF are defined as

$$\mathtt{TPF}_t(c) = P\{X > c | t \geqslant T\}, \quad \mathtt{FPF}_t(c) = P\{X > c | t < T\}. \tag{4}$$

Note that, in the above definition, it is necessary to assert whether or not for a subject $T > t$. However, for a specific subject, because of censoring, its status may not be well-defined at certain time points. When the censoring scenario is more complicated than right censoring, it can be even more difficult to assert the disease status. Thus, it is necessary to revise the definitions in [5] to better reflect the contribution of a subject to TPF and FPF at different time points.

Consider the scenario where the event time is only known to lie in the interval $[L, R]$ with $0 \leqslant L \leqslant R \leqslant \infty$. It includes the following special cases: (a) When $L = R < \infty$, the event time is accurately observed and uncensored. (b) When $L = 0$ and $R < \infty$, the observation is left censored, and the true event time is only known to be smaller than R. (c) When $L > 0$ and $R = \infty$, the observation is right censored, and the true event time is only known to be larger than L. (d) When $0 < L < R < \infty$, the observation is interval censored. The disease status is 0 before L, 1 after R, but undetermined between L and R.

Consider $\texttt{TPF}_t$ and $\texttt{FPF}_t$, which, as shown in (4), are defined on the marker values of diseased and healthy subjects. Below, we rewrite the definitions in (4) so that they can better reflect the interval-censored nature of data and provide a more intuitive way for estimation. When the event time is only known to lie in $[L, R]$, a subject is known *for sure* to be diseased if and only if $t \geqslant R$, and to be healthy if and only if $t \leqslant L$. Accordingly, we rewrite the definitions as

$$\texttt{TPF}_t(c) = P\{X > c | t \geqslant R\}, \quad \texttt{FPF}_t(c) = P\{X > c | t < L\}. \tag{5}$$

Thus, an uncensored observation contributes to the computation of TPF and FPF at any time t; a right-censored observation only contributes to the computation of FPF; a left censored observation only contributes to the computation of TPF; and an interval censored observation contributes to the computation of FPF only in the time interval of $[0, L)$ and TPF only in the time interval of (R, ∞). We note that in this study, the essence of the time-dependent ROC remains the same as that in [5]. The definitions in (5) have formats different from their counterparts in [5]. As discussed above, such difference has been caused by the difference in censoring schemes. We acknowledge that the old definitions may seem more intuitive. Under complex censoring, we rewrite the definitions in a way such that they can better reflect the interval-censored nature of data and provide a more intuitive way for estimation. That is, although less intuitive, the new definitions are easier to compute in practice.

To quantify the loss of information caused by censoring, we define the Undetermined Positive Fraction as

$$\texttt{UPF}_t(c) = P\{X > c | L \leqslant t \leqslant R\}.$$

Note that when $R \to L$, UPF converges to the 'incidental TPF' defined in [5]. It measures the proportion that does not provide any useful information about the classification power of marker X. When $\texttt{UPF}_t$ is non-zero, there is a loss of information about the estimation of $\texttt{TPF}_t$, $\texttt{FPF}_t$, and other diagnostic accuracy measures described in the following section. When $\texttt{UPF}_t$ increases, the accuracy of the estimates decreases, and the variances of the estimates inflate.

4. Estimation and inference

Consider data composed of subjects under various censoring schemes. As described in Section 3, all observations can be written in an 'interval censored' format. Assume a sample of n iid observations $\{S_i = (X_i, L_i, R_i): i = 1, \ldots, n\}$. In the following subsections, we investigate the estimation and inference for the quantities defined in Section 2.

4.1. TPF and FPF

Define $N_L(t) = \sum_{i=1}^n I\{L_i > t\}$ and $N_R(t) = \sum_{i=1}^n I\{R_i \leqslant t\}$, where I is the indicator function. TPF and FPF defined in (5) can be estimated by their empirical correspondences

$$\widehat{\texttt{TPF}}_t(c) = \frac{\sum_{i=1}^n I\{X_i > c, R_i \leqslant t\}}{N_R(t)}, \quad \widehat{\texttt{FPF}}_t(c) = \frac{\sum_{i=1}^n I\{X_i > c, L_i > t\}}{N_L(t)}. \tag{6}$$

Our estimates of TPF and FPF have forms different from those in [5]. With right-censored and uncensored data, the survival function can be easily computed using the KM approach. However, under general censoring schemes as considered in this study, there is no simple analytic formula for the survival function.

These estimates have variances

$$\text{var}\{\widehat{\texttt{TPF}}_t(c)\} = \frac{\texttt{TPF}_t(c)(1 - \texttt{TPF}_t(c))}{E(N_R(t))}, \quad \text{var}\{\widehat{\texttt{FPF}}_t(c)\} = \frac{\texttt{FPF}_t(c)(1 - \texttt{FPF}_t(c))}{E(N_L(t))},$$

which can be estimated by

$$\widehat{\text{var}}\{\widehat{\texttt{TPF}}_t(c)\} = \frac{\widehat{\texttt{TPF}}_t(c)(1 - \widehat{\texttt{TPF}}_t(c))}{N_R(t)}, \quad \widehat{\text{var}}\{\widehat{\texttt{FPF}}_t(c)\} = \frac{\widehat{\texttt{FPF}}_t(c)(1 - \widehat{\texttt{FPF}}_t(c))}{N_L(t)}.$$

Consider $t\in[0,\tau]$, where $P(L>t), P(R\leqslant t)>\eta_0>0$ for a fixed constant η_0. From the definition, $\mathtt{TPF}_t(c)=P(X>c,t\geqslant R)/P(t\geqslant R)$. For any t, as $n\to\infty$, $\sqrt{n}((1/n)\sum_{i=1}^{n} I\{X_i>c,R_i\leqslant t\}-P(X>c,R\leqslant t))\to N(0,\sigma_1^2)$ and $\sqrt{n}((1/n)N_R(t)-P(R\leqslant t))\to N(0,\sigma_2^2)$ with $\sigma_1,\sigma_2>0$. Such results follow from Theorem 1.14 of [13]. The consistency of the variance estimates follows from similar arguments and the Binomial distribution.

We note that the estimates of $\mathtt{TPF}_t$ and $\mathtt{FPF}_t$ in (6) have forms different from their counterparts in [5], which were designed for right-censored and uncensored data. With right-censored data, the survival function can be easily estimated using the KM approach and realized using the existing software. With interval-censored data, a few approaches for estimating the survival function are described in [7, 9]. In theory, it is possible to first estimate the survival function and then adopt the approach in [5]. However, unless the data have very special properties, for example when all observations are case I interval censored [9], estimation of the survival function can be difficult, and there is a lack of user-friendly software. We have conducted some simulation studies and found that when the sample size is moderate to large, the two estimates, namely (6) and the one in [5], are very close (results available upon request from the authors and omitted here). When the sample size is small, the two estimates may differ significantly, which may be caused by the lack of reliability of the estimate of the survival function [7].

4.2. *Time-specific AUC*

Consider a fixed time point t, where $\min_i R_i<t<\max_i L_i$. At this time point, the ROC curve can be estimated with

$$\widehat{\mathrm{ROC}}_t(p)=\widehat{\mathtt{TPF}}_t\{[\widehat{\mathtt{FPF}}_t]^{-1}(p)\}.$$

For any specific p as $n\to\infty$, the consistency of $[\widehat{\mathtt{FPF}}_t]^{-1}$ (as an estimate of $[\mathtt{FPF}_t]^{-1}$) follows from the consistency of $\widehat{\mathtt{FPF}}_t$ and monotonicity and tightness of the estimate. Combining with the consistency of $\widehat{\mathtt{TPF}}_t$, we can conclude the consistency of $\widehat{\mathrm{ROC}}_t$. Note that we focus on $\min_i R_i<t<\max_i L_i$ as there is no information elsewhere. Accordingly, $\mathrm{AUC}(t)$ can be estimated with

$$\widehat{\mathrm{AUC}}(t)=\int_0^1 \widehat{\mathrm{ROC}}_t(p)\,\mathrm{d}p.$$

The above estimator is conceptually straightforward. However, it involves the functional estimation of TPF and FPF and inversion of FPF. A computationally more feasible estimator can be based on the probabilistic interpretation of the AUC in (1). More specifically, consider the following Wilcoxon rank sum type estimator

$$\widehat{\mathrm{AUC}}(t)=\frac{\sum_i\sum_j I\{X_i>X_j,t\geqslant R_i,t<L_j\}}{\sum_i\sum_j I\{t\geqslant R_i,t<L_j\}}.$$

As $n\to\infty$, it follows from [14, 15] that

$$\frac{1}{n(n-1)}\sum_i\sum_j I\{X_i<X_j,t\geqslant R_i,t<L_j\}\to_{a.s.} P(X_1<X_2,T_1>t>T_2)=P(X_1<X_2,S_1\in\bar{D}(t),S_2\in D(t)),$$

$$\frac{1}{n(n-1)}\sum_i\sum_j I\{t\geqslant R_i,t<L_j\}\to_{a.s.} P(S_1\in\bar{D}(t),S_2\in D(t))).$$

As a result, $\widehat{\mathrm{AUC}}(t)\to_{a.s.} P(X_1<X_2|S_1\in\bar{D}(t),S_2\in D(t))$. Here, $\bar{D}(t)$ and $D(t)$ denote the index sets for healthy and diseased subjects at time t, respectively.

For inference, one possibility is based on the U-statistic theories in [13]. We have examined such an approach and found that it involves the smoothed estimation of complex functionals. An alternative nonparametric bootstrap approach is proposed in [16] (Theorem 5.3) and [17]: (a) randomly sample m subjects *without replacement*. For validity, it requires that as $n\to\infty$, $m\to\infty$ and $m/n\to 0$; (b) compute $\widehat{\mathrm{AUC}}_t$ using those m subjects; (c) repeat steps (a)–(b) B (e.g. 500) times; (d) compute the sample variance, rescale, and obtain an estimate of $\mathrm{var}(\widehat{\mathrm{AUC}})$. We note that, although Theorem 5.3 of [16] provides an asymptotic rate for m, it does not suggest how to select it in practice. Our limited experience suggests that, as long as m is not too large or too small, the inference result is satisfactory. For sample size $n=100,200,400$ as in our simulation study, $m\in[25,50],[50,100],[50,200]$ lead to satisfactory results. To the best of our knowledge, there is still a lack of data-adaptive way of choosing m.

4.3. *Integrated AUC*

There are two ways to compute the integrated AUC.

4.3.1. Approach 1. The first approach is based on definition (2), where we can estimate C with $\hat{C} = \int_0^\tau \widehat{\mathrm{AUC}}(t) \times \hat{w}(t)\,\mathrm{d}t$. Here, $\hat{w}(t)$ is the estimate of the weight function $w(t)$. Consider, for example, the concordance measure in (3). It can be estimated with

$$\hat{C} = \int_0^\tau \widehat{\mathrm{AUC}}(t) \times 2\hat{f}(t)\hat{S}(t)\,\mathrm{d}t. \tag{7}$$

In (7), $\widehat{\mathrm{AUC}}(t)$ has been developed in Section 4.2. There are multiple ways of estimating f and S. We refer to [9, 18] for references. Numerical studies suggest that, when the sample size is small to moderate, the nonparametric estimate (e.g. NPMLE) without any smoothness constraint can be numerically unreliable. When it is possible to assume that f and S are smooth functions, the following estimates can be more reliable with small sample sizes. First rewrite that $f = \exp(h)$. Assume that (a) h belongs to the Sobolev space indexed by the order of derivative s_0; and (b) $|h| \leqslant M < \infty$. Then, $S = \int f = \int \exp(h)$. h can be estimated with

$$\hat{h} = \underset{h}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \log(S(L_i) - S(R_i)) \right) - \lambda_n \int (h^{(s_0)})^2\,\mathrm{d}t \tag{8}$$

$\hat{f} = \exp(\hat{h})$, and $\hat{S} = \int \hat{f}$. In (8), $h^{(s_0)}$ is the s_0th derivative of h, λ_n is the data-dependent tuning parameter and can be chosen using V-fold cross-validation. We use the penalty on smoothness, which has been commonly adopted in spline studies, to control the estimate of h. In practice, the computation of $\hat{h}$ can be realized using a 'basis expansion + Newton maximization' approach. Following [19], we can prove that (a) $\hat{f}$ and $\hat{S}$ are $n^{s_0/(2s_0+1)}$ consistent; and (b) $\int (\hat{f}^{(s_0)})^2 = O_p(1)$ and $\int (\hat{S}^{(s_0)})^2 = O_p(1)$. These results, together with the $\sqrt{n}$ consistency of $\widehat{\mathrm{AUC}}_t$, lead to the consistency of $\hat{C}$. For inference, we use the bootstrap approach developed in [20], which has an intuitive interpretation and satisfactory empirical performance.

A drawback of this approach is its high computational cost. We refer to Section 5.1 for more details. As an alternative, the following approach is computationally more feasible.

4.3.2. Approach 2. This approach has been motivated by the probabilistic interpretation in (3). Note that the concordance measure can be interpreted as

$$P(X_1 > X_2 | T_1 < T_2) = 2P(X_1 > X_2, T_1 < T_2).$$

See Appendix of [5] for a proof.

Define $\mathscr{I} = \{(i, j)$: without ambiguity, $T_i < T_j$ or $T_i > T_j\}$. Consider two event time intervals $[L_1, R_1]$ and $[L_2, R_2]$. This pair of subjects belong to $\mathscr{I}$ when for example $R_1 < L_2$ and it is for sure that $T_1 < T_2$. We propose estimating (3) with

$$\hat{C} = \frac{\sum_{(i,j)\in\mathscr{I}} I\{X_i > X_j, T_i < T_j\}}{\sum_{(i,j)\in\mathscr{I}} I\{T_i < T_j\}} = \frac{\sum_{i=1}^n \sum_{j=1}^n I\{X_i > X_j, R_i < L_j\}}{\sum_{i=1}^n \sum_{j=1}^n I\{R_i < L_j\}}. \tag{9}$$

Note that $\{R_i < L_j\} \subseteq \{T_i < T_j\}$. Thus, under the condition $P(R_i < L_j) > 0$, $P(T_i < T_j) > 0$. As $n \to \infty$,

$$\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n I\{X_i > X_j, R_i < L_j\} \to_{a.s.} P(X_1 > X_2, R_1 < L_2) \quad \text{and} \quad \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n I\{R_i < L_j\}$$
$$\to_{a.s.} P(R_1 < L_2).$$

Such results follow from Hoeffding's theorem for two-sample U-statistics (Theorem 3.5 in [13]). As a result, $\hat{C} \to_{a.s.} P(X_1 > X_2, R_1 < L_2)/P(R_1 < L_2) = P(X_1 > X_2 | R_1 < L_2)$.

There are multiple ways of conducting inference. The first is based on the U-statistic theories. The second is the 'm out of n' nonparametric bootstrap described in Section 4.2. This approach demands that $m/n \to 0$ as $n \to \infty$. In our numerical studies, we find that when n is small, m and the size of the index set $\mathscr{I}$ can be extremely small. Performance of this bootstrap approach can thus be unsatisfactory.

An alternative approach, which has better empirical performance, is the semiparametric bootstrap and consists of the following steps. (a) Construct $\hat{S}_X(t)$, an estimate of the conditional distribution function (conditional on X). In our numerical studies, we adopt the nonparametric estimation in [21, 22]. (b) For $i=1,\ldots,n$, simulate survival time T_i^* from the conditional distribution $\hat{S}_{X_i}(t)$. (c) Compute $\hat{C}^*=\sum_i\sum_j I\{X_i>X_j, T_i^*<T_j^*\}/\sum_i\sum_j I\{T_i^*<T_j^*\}$. (d) Repeat steps (b)–(c) B (e.g. 500) times. (e) Compute the sample variance and obtain an estimate of var($\hat{C}$).

4.3.3. Remarks. When estimating the integrated AUC, Approach 1 is generically applicable, however at the price of a high computational cost. Approach 2 is computationally simpler, and our numerical studies suggest its better empirical performance. However, it is limited to certain weight functions where C has simple probabilistic interpretations. This limitation may not be serious, considering that the concordance measure is perhaps the only commonly used integrated AUC.

When conducting inference for $\hat{C}$ with Approach 2, both the nonparametric and semiparametric bootstrap approaches are asymptotically valid. When the sample size is large, both approaches have satisfactory performance. However, when the sample size is moderate to small, the semiparametric bootstrap is preferred. We conjecture that the unsatisfactory performance of the nonparametric bootstrap is caused by the extremely small size of $\mathscr{I}$.

5. Numerical studies

5.1. Simulation

We conduct simulation to gain further insights into the proposed approaches. In simulation, we examine (a) whether the prognostic accuracy measures can be consistently estimated, (b) performance of the inference procedures, and (c) whether the prognostic measures can correctly rank the markers. We set the sample size $n=100$, 200, and 400, and consider three different ways of generating markers X_1, X_2, and X_3:

- Case 1: X_1, X_2, and X_3 are independently generated from Uniform(0, 2).
- Case 2: $(X_1, X_2, X_3)^{\mathrm{T}}$ are generated from a tri-variate normal distribution with mean $\mathbf{1}=(1,1,1)^{\mathrm{T}}$ and covariance $0.33\mathbf{I}$ where $\mathbf{I}$ is the 3×3 identity matrix.
- Case 3: $(X_1, X_2, X_3)^{\mathrm{T}}$ are generated from a tri-variate normal distribution with mean $\mathbf{1}=(1,1,1)^{\mathrm{T}}$ and covariance $\boldsymbol{\Sigma}$. Here, $\boldsymbol{\Sigma}$ follows a heterogeneous compound symmetry structure with the diagonal elements equal to 0.32, 0.25, and 0.35 and correlation parameter 0.45.

We generate the failure time T from a proportional hazards model $S(t)=\exp\{-\exp(\beta_1X_1+\beta_2X_2+\beta_3X_3)\Lambda(t)\}$ with a Gompertz cumulative baseline hazard $\Lambda(t)=\frac{1}{6}\exp\{0.06t\}$. We set the regression coefficients $\beta_1=2, \beta_2=0.5$, and $\beta_3=0$. Therefore, the three markers have, respectively, strong, moderate, and weak prognostic power for the event outcome. We generate the left end of the censoring interval as $L=\omega t$, where $\omega\sim$Uniform(0.5, 1), and the right end as $R=\xi t$ with probability 0.65 where $\xi\sim$ Uniform(1, 2), and $R=\infty$ with probability 0.35. The resulted data set is thus a mixture of right- and interval-censored observations.

Although several accuracy measurements are defined in Sections 2 and 4, we focus on the integrated AUC, which is defined on other accuracy measurements. The accuracy of the integrated AUC can partly reflect the accuracy of other measurements. In Section 4.3, we propose two approaches for estimating the integrated AUC. We find that they lead to similar estimates under all simulation scenarios. For example, under simulation Case 1 and sample size $n=400$, the average biases over 1000 simulations for marker X_1 are 0.021 and 0.015 under Approaches 1 and 2, respectively. Comparable results are obtained under other simulation settings. We also find that the computational cost of Approach 1 is much higher than that of Approach 2. For example, with code written in R and run on a PC with Pentium 4 2.8 GHz processor, one estimation of the integrated AUC along with 500 bootstrap estimates take 10 and 2 min using Approaches 1 and 2, respectively. Thus, in what follows, we only report results using Approach 2.

We simulate 1000 replicates and present the summary statistics in Table I. We can see that when the sample size is reasonably large, the estimated AUCs are close to the true values. The discrepancy decreases as the sample size increases. However, when the sample size is equal to 100, the estimated

Table I. Estimation of the integrated AUC: summary statistics based on 1000 replicates.

Case	Marker	True	Sample size	Est.	SD	SE
1	X_1	0.75	100	0.8289	0.0371	0.0406
			200	0.8005	0.0270	0.0295
			400	0.7648	0.0211	0.0225
	X_2	0.55	100	0.5427	0.0608	0.0626
			200	0.5554	0.0387	0.0394
			400	0.5530	0.0274	0.0288
	X_3	0.50	100	0.5263	0.0640	0.0672
			200	0.4934	0.0418	0.0462
			400	0.5031	0.0295	0.0308
2	X_1	0.75	100	0.8294	0.0354	0.0415
			200	0.8230	0.0261	0.0351
			400	0.7740	0.0189	0.0253
	X_2	0.55	100	0.5804	0.0623	0.0659
			200	0.5689	0.0405	0.0442
			400	0.5537	0.0318	0.0385
	X_3	0.50	100	0.5058	0.0585	0.0640
			200	0.4902	0.0393	0.0429
			400	0.4981	0.0261	0.0279
3	X_1	0.76	100	0.8257	0.0348	0.0465
			200	0.8059	0.0276	0.0350
			400	0.7668	0.0179	0.0266
	X_2	0.64	100	0.6892	0.0545	0.0704
			200	0.6700	0.0327	0.0425
			400	0.6448	0.0250	0.0293
	X_3	0.61	100	0.6467	0.0591	0.0715
			200	0.6270	0.0368	0.0417
			400	0.6155	0.0252	0.0286

‘True’ is the true AUC values; ‘Est.’ is the mean of the estimated AUC values; ‘SD’ is the standard deviation of the estimated AUC values; ‘SE’ is the mean of the estimated semiparametric bootstrap standard errors.

Table II. Proportion of correct ranking of the three markers over 1000 replicates.

	Sample size		
Case	100 (percent)	200 (per cent)	400 (per cent)
1	98.7	100.0	100.0
2	95.4	98.1	99.6
3	92.0	96.5	97.7

AUCs can be inflated. This has been caused by the small denominators. A similar phenomenon has been observed with AUC for categorical data and uncensored and right-censored survival data.

For inference, we have experimented with both the nonparametric and semiparametric bootstrap approaches. When the sample size is equal to 400, both approaches can generate satisfactory results. However, when the sample size is equal to 100, the nonparametric bootstrap is less satisfactory. Thus, we only report results using the semiparametric bootstrap. We can see that the bootstrap standard errors are close to the empirical standard deviations of the estimates, suggesting satisfactory performance.

We also examine whether the integrated AUCs can correctly rank the prognostic power of markers, with the proper rank being $X_1 > X_2 > X_3$. With the 1000 replicates, we compute the proportions of correct rankings and present the results in Table II. We find that a very high proportion of replicates are properly ranked. The performance is best when the markers are independent and bounded (Case 1). We also note that, when the sample size is small, the AUC of X_3 may be over estimated. In a few occasions, X_3 is mistakenly ranked higher than X_2.

5.2. *Analysis of HDSD*

The Hypobaric Decompression Sickness Data (HDSD) study was conducted by NASA to investigate the risk of decompression sickness in hypobaric environments. The outcome of interest is the time to onset of grade IV venous gas emboli, which was mixed-case interval censored because of measurement limitations. The study was first described in [23]. The HDSD contains 549 records, among which 124 were interval censored and the rest were right censored. One record has missing measurements and is removed from downstream analysis. Besides censoring times and event indicators, values of the following markers were recorded: (a) AGE, which ranges from 20 to 54 with median 30; (b) BMI, body mass index; and (c) TR360, which measures the decompression stress. It is the ratio of the partial pressure of nitrogen to ambient pressure at the final altitude. It is an experimental variable and is related to the particular pre-breathe protocol being tested.

For the three markers, we use the approach described in Section 4.3.2 and compute their integrated AUCs as:

$$\text{Age}: 0.583[0.581, 0.586]; \quad \text{BMI}: 0.530[0.527, 0.532]; \quad \text{TR360}: 0.518[0.516, 0.520],$$

where the numbers in '[ ]' are the 95 per cent bootstrap confidence intervals. Although their integrated AUCs are statistically significantly higher than 0.5, all three markers have very weak discriminatory ability for the development of grade IV venous gas emboli.

In the literature, it has been noted that the prognostic power of single markers may not be sufficient, and that the combinations of multiple markers may be needed [24]. The approach for combining multiple markers proposed in [3], although may have better asymptotic properties, is computationally expensive. As an alternative, we use a simple step-up approach. At each step, we enter an additional marker and search for regression coefficient(s) so that the resulting AUC is maximized. Under related but different settings, the stepwise approach has been shown to be satisfactory when the number of markers is not too large. For better interpretability, we focus on the linear combinations of markers. For identifiability, we keep the coefficient of TR360 to be 1 in all steps. This stepwise procedure leads to the following results:

Step	Marker	AUC
(I)	TR360	0.583[0.581, 0.586]
(II)	$0.001 \times \text{AGE} + \text{TR360}$	0.642[0.639, 0.644]
(III)	$-0.0299 \times \text{BMI} + 0.0152 \times \text{AGE} + \text{TR360}$	0.669[0.666, 0.671]

The AUC increases as we introduce more markers. The linear combination obtained in Step III has an AUC significantly larger than that of any single marker. The prognostic accuracy improves as we combine information from different markers.

As the same data set is used for deriving the combination rule and for calculating the prognostic accuracy, the above AUCs may be over estimated. To more objectively gauge the prognostic performance, we consider the following cross-validation-based procedure. We first randomly partition the data into two sets with equal sizes. The first set is the training set and used to derive the combination rule. The second set is the testing set and used to evaluate the prognostic power. To avoid an extreme partition, we repeat this process 1000 times. The average AUCs are 0.636 and 0.658 at Steps II and III, respectively, which lead to similar conclusions as made above.

The integrated AUC reflects the 'integrated' or 'overall' prognostic performance. To gain further insights, we explore the ROC curve as a function of time. In Figure 1, for the three markers and their linear combination obtained above, we plot their time-dependent ROC curves. We note that, at different time points, the effective sample sizes may differ, with a larger sample size leading to more reliable estimates. Not all the estimated ROC curves in the spanned surface deserve the same attention due to the varying degree of efficiency. To reflect the variation of sample size, we use different colors in Figure 1, with cold color (for example blue) corresponding to small sample sizes and warm color (for example red) corresponding to large sample sizes. The change of color from blue to red corresponds to the increase in sample size. Such a plot can be obtained using existing software such as MATLAB or R. Sample code is available from the authors. Figure 1 suggests that, for all markers considered, the prognostic performance improves as time passes.

5.3. Analysis of calcification study

The study investigated the calcification of hydrogel intraocular lenses, which is an infrequently reported complication of cataract treatment [25]. In this study, patients were examined by an ophthalmologist to

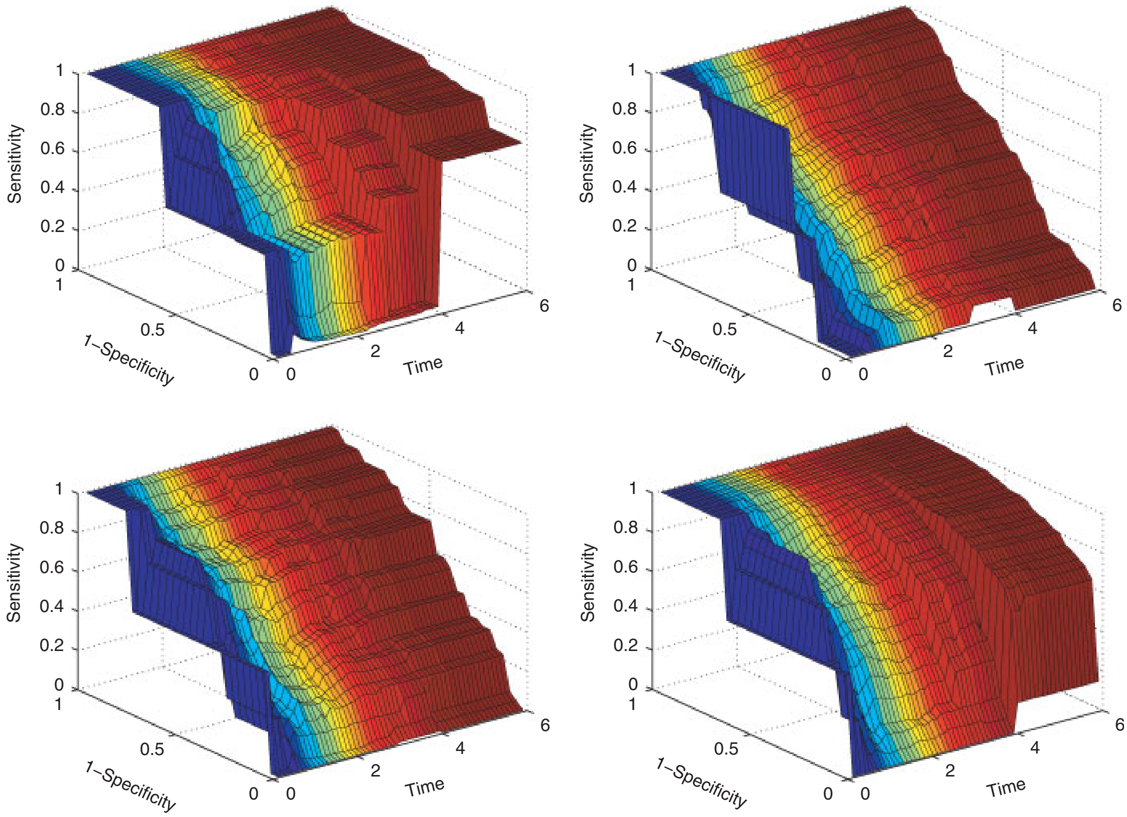

Figure 1. Analysis of HDSD: estimated time-dependent ROC curves for TR360 (upper left), BMI (upper right), AGE (lower left), and their linear combination (lower right). The change of color from blue to red as reflected in the online colored version corresponds to the increase in sample size.

determine the status of calcification at a random time ranging from 0 to 36 months after implantation of the intraocular lenses. Thus, all observations were case I interval censored [9]. At the examination, the severity of calcification was graded on a discrete scale ranging from 0 to 4, with severity $\leqslant 1$ classified as 'not calcified'. The data set contains 379 records, among which one has missing measurements. Among the 378 subjects with complete measurements, 48 experienced calcification during the followup. The markers of interest include AGE, incision width, and incision length.

For the three markers, we compute their integrated AUCs as

$$\text{Age}: 0.503[0.494, 0.511]; \quad \text{incision width}: 0.658[0.647, 0.667];$$

$$\text{incision length}: 0.805[0.795, 0.814].$$

Among the three markers, AGE has no prognostic power. The marker incision length has the largest AUC and can make around 80 per cent correct discrimination between the calcified and normal subjects.

We consider the stepwise approach described in Section 5.2 and search for linear combinations of markers that may have better discriminative power. We find that no linear combination has significantly better discriminative power than the incision length. This result indicates that for discrimination purpose, information in age and incision width is redundant given that in incision length. Thus, in future clinical practice, only the incision length needs to be measured for prognosis prediction.

We plot the ROC curves as a function of time in Figure 2. With this data set, the effective sample size is small at the two ends of the time line and is relatively large in the middle. Simply eyeballing Figure 2 suggests that the ROC curves for the incision length dominate those of incision width and age at almost all time points.

6. Discussion

The ROC analysis provides a useful framework for evaluating the diagnostic and prognostic power of classifiers. It provides a comprehensive classification accuracy measurement and can be practically

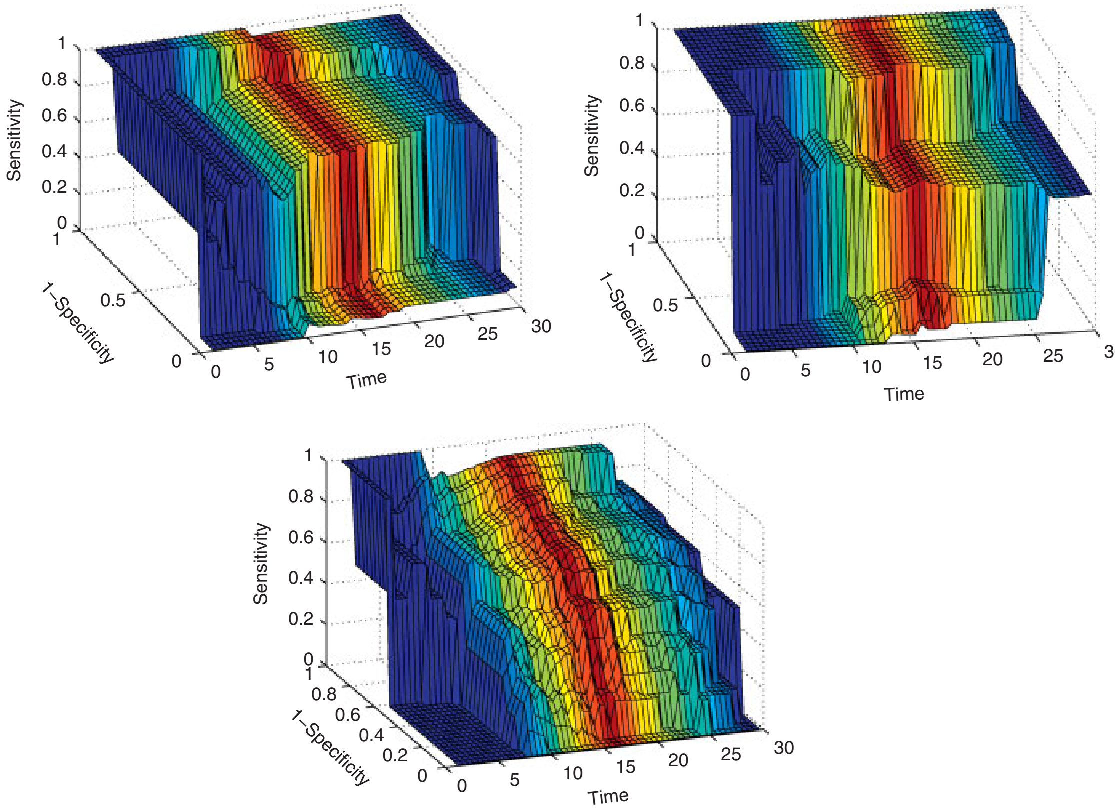

Figure 2. Analysis of calcification study: estimated time-dependent ROC curves for incision length (upper left), incision width (upper right), and AGE (lower).

more relevant than alternatives such as classification error and likelihood-based measurements. It does not make specific assumptions on, for example, the form of the link function, and thus can be more robust. There has been extensive investigation on the ROC techniques with categorical data. Recent development on ROC techniques with censored survival data has generated interesting results and has attracted considerable attention.

In this article, we investigate the time-dependent ROC under censoring patterns more complicated and more general than right censoring. The type of censoring we consider, although not as popular as right censoring, is routinely encountered. The remarkable difference in censoring patterns demands a new set of tools. For various ROC measurements, we develop the estimation and inference approaches. The proposed estimation procedures are simple, and their asymptotic consistency can be easily established using the central limit theorem and the consistency properties of U-statistics [13]. Without attempting to construct smoothed estimation of complex functionals, we propose using the bootstrap approaches for inference. Their validity can be established following [16, 17]. Simulation shows that the proposed estimation and inference approaches have satisfactory finite sample performance. Applications to two real data sets show that the proposed approach can discriminate markers with different prognostic power. For certain data, it is possible to improve prognostic performance by combining multiple markers under the ROC framework. A stepwise approach has been adopted for this purpose. In our data analysis, we use the integrated AUC as the measure of prognostic power. An alternative is the ROC curve (and corresponding AUC) at a specific time at which researchers may want to make prediction. For example in cancer research, three- or five-year survival may be of special interest. With the two data sets we analyze, there is no time point of special interest. Thus, we focus on the integrated measure.

For right-censored data, there has been some recent development closely related to the time-dependent ROC. For example, when there are a large number of markers, Song *et al*. [26] proposes an approach that can effectively combine multiple markers. Zheng and Heagerty [27] and followup studies consider longitudinal markers. Following a similar strategy, it is possible to extend the present study and accommodate more complex marker scenarios. Such an endeavor is nontrivial and will be pursued in the future studies.

Acknowledgements

The authors thank the Associate Editor and two referees for very insightful comments, which have led to significant improvement of this article. This study has been supported by grants from the Academic Research Funding (ARF R-155-000-109-112) and the National Medical Research Council (NMRC/NIG/0054/2009) in Singapore (Li) and DMS 0805984 from the NSF USA (Ma).

References

1. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.
2. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley: New York, 2002.
3. Ma S, Huang J. Combining multiple markers for classification using ROC. *Biometrics* 2007; **63**:751–757.
4. Heagerty P, Lumley T, Pepe M. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**:337–344.
5. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; **61**:92–105.
6. Schoop R, Graf E, Schumacher M. Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics* 2007; **64**:603–610.
7. Sun J. *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer: Berlin, 2006.
8. Gomez G, Calle ML, Oller R, Langohr K. Tutorial on methods for interval-censored data and their implementation in R. *Statistical Modeling* 2009; **9**:259–297.
9. Groeneboom P, Wellner JA. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhauser: Basel, 1992.
10. Huang J. Efficient estimation for the proportional hazard model with interval censoring. *Annals of Statistics* 1996; **24**:540–568.
11. Xue H, Lam KF, Li G. Sieve maximum likelihood estimator for semiparametric regression models with current status data. *JASA* 2004; **99**:346–356.
12. Ma S, Kosorok MR. Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis* 2005; **96**:190–217.
13. Shao J. *Mathematics Statistics*. Springer: New York, 1999.
14. Kowalski J, Tu XM. *Modern Applied U-Statistics*. Wiley: New York, 2007.
15. Khan S, Tamer E. Partial rank estimation and duration models with general forms of censoring. *Journal of Econometrics* 2007; **136**:251–280.
16. Gine E. Lectures on some aspects of the bootstrap. *Lecture Notes in Mathematics* 1997; **1665**:37–152.
17. Arcones MA, Gine E. On the bootstrap of U and V statistics. *Annals of Statistics* 1992; **20**:655–674.
18. Schick A, Yu Q. Consistency of the GMLE with mixed case interval censored data. *Scandinavian Journal of Statistics* 2000; **27**:45–55.
19. Ma S, Kosorok MR. Adaptive penalized M-estimation with current status data. *Annals of Institute of Statistical Mathematics* 2006; **58**:511–526.
20. Liu A, Meiring W, Wang Y. Testing generalized linear models using smoothing spline methods. *Statistica Sinica* 2005; **15**:235–256.
21. Tu XM. Nonparametric estimation of survival distributions with censored initiating time, and censored and truncated terminating time: application to transfusion data for acquired immune deficiency syndrome. *Applied Statistics* 1995; **44**:3–16.
22. Pique RO. Survival analysis issues with interval-censored data. *Ph.D. Thesis*, Universitat Politecnica de Catalunya, 2006.
23. Conkin J, Powell M. Lower body adynamia as a factor to reduce the risk of hypobaric decompression sickness. *Aviation*, *Space and Environmental Medicine* 2001; **72**:202–214.
24. Zheng Y, Cai T, Feng Z. Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics* 2006; **62**:279–287.
25. Yu AKF, Kwan KYW, Chan DHY, Fong DYT. Clinical features of 46 eyes with calcified hydrogel intraocular lenses. *Journal of Cataract and Refractive Surgery* 2001; **27**:1596–1606.
26. Song X, Ma S, Huang J, Zhou XH. A semiparametric approach for the nonparametric transformation survival model with multiple covariates. *Biostatistics* 2007; **8**:197–211.
27. Zheng Y, Heagerty PJ. Prospective accuracy for longitudinal markers. *Biometrics* 2007; **63**:332–341.