Biostatistics (2013), **14**, 1, *pp.* 42–59
doi:10.1093/biostatistics/kxs021
Advance Access publication on June 25, 2012

Non-parametric estimation of a time-dependent predictive accuracy curve

P. SAHA-CHAUDHURI*

Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, USA
paramita.sahachaudhuri@duke.edu

P. J. HEAGERTY

Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

Summary

A major biomedical goal associated with evaluating a candidate biomarker or developing a predictive model score for event-time outcomes is to accurately distinguish between incident cases from the *controls* surviving beyond t throughout the entire study period. Extensions of standard binary classification measures like time-dependent sensitivity, specificity, and receiver operating characteristic (ROC) curves have been developed in this context (Heagerty, P. J., *and others*, 2000. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344). We propose a direct, non-parametric method to estimate the time-dependent Area under the curve (AUC) which we refer to as the weighted mean rank (WMR) estimator. The proposed estimator performs well relative to the semi-parametric AUC curve estimator of Heagerty and Zheng (2005. Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105). We establish the asymptotic properties of the proposed estimator and show that the accuracy of markers can be compared very simply using the difference in the WMR statistics. Estimators of pointwise standard errors are provided.

Keywords: AUC curve; Survival analysis; Time-dependent ROC.

1. Introduction

In order to choose among candidate management options and to make medical decisions, the prediction of an individual's future health status is often necessary. The predictive objective can frequently be evaluated by quantifying how well a set of potential biomarkers or a model score can predict those subjects who subsequently experience a primary transition of health status such as the onset of disease or progression to death. One area of clinical management where predictive models or markers are used to guide treatment decisions is the general area of organ transplantation. For example, in staging patients for lung transplantation, the lung allocation score is used to prioritize transplantation candidates (Gries *and others*, 2010). In this setting, a good predictive model or marker would accurately identify those subjects who are

*To whom correspondence should be addressed.

still under study but otherwise likely to die in the near future (imminent "cases"), yet not falsely identify subjects who will not have an urgent need for transplantation (near term "controls"). A second example comes from liver disease where the model for end-stage liver disease (MELD) score is used for treatment decision making. As reviewed by Coombes and Trotter (2005, p. 87) "The MELD score has emerged as an excellent predictor of short-term mortality in patients with advanced liver disease, and patients listed for liver transplantation are now ranked on their respective MELD scores." One useful descriptive summary of the potential performance of the MELD score would be to use cohort data and show where incident death cases actually rank among their corresponding risk-set members. If the MELD score were an ideal marker, then incident cases at any time would always rank higher than all of their corresponding risk-set members. The goal of this manuscript is to show that such an empirical risk-set ranking of incident cases is directly linked to previously proposed time-dependent accuracy concepts, and also to show that risk-set ranking provides the basis for non-parametric estimation of time-dependent accuracy summary measures.

To incorporate time into predictive classification criterion, time-dependent versions of sensitivity and specificity have been proposed (Heagerty *and others*, 2000; Heagerty and Zheng, 2005). Time-dependent receiver operating characteristic (ROC) curves and area under the ROC curve (AUC) functions can characterize how well candidate markers can distinguish between those cases who experience a clinical transition from those subjects who remain event-free. Time-dependent measures describe the accuracy of a model score sequentially over time as opposed to simply providing an overall summary measure of the predictive accuracy like the C-index (Harrell *and others*, 1982) or some of its recent extensions (Antolini *and others*, 2005; Gonen and Heller, 2005). Heagerty and Zheng (2005) proposed a semi-parametric estimator for the time-dependent summary curve AUC(t). However, their estimation was "indirect" in that first estimates of time-dependent ROC curves for all observed event times are obtained. Each time-dependent ROC curve can then be integrated to provide an estimate of AUC(t).

The goal of this manuscript is to propose an alternative direct non-parametric estimator of the time-dependent accuracy curve, AUC(t). Relative to the semi-parametric estimators proposed by Heagerty and Zheng (2005), a non-parametric estimator has the major advantage of not requiring specification of a conditional hazard regression model linking the marker to the event time, and therefore the proposed methods provide valid inference for AUC(t) under minimal assumptions. In addition, we provide analytic approximations for point-wise standard errors and thus permit inference without needing to employ resampling methods.

To motivate the proposed non-parametric approach, we first note that an estimator of AUC that is commonly used in case–control studies is the empirical proportion of concordant case–control pairs (C-index). Specifically, concordance counts the pairs whose marker orderings correctly reflect the ordering of their outcomes. Such a non-parametric approach can be considered for each riskset in a time-to-event setting with incident cases. However, frequently only a small number of subjects experience an event at any given time, t, and therefore some smoothing is warranted to provide function estimation with desirable statistical properties. In this article, we propose the local rank-based concordance estimation using weighted U-statistics. We introduce the notation and motivate the estimation of the time-dependent AUC curve as a locally weighted concordance measure. We detail the large-sample properties of the proposed non-parametric estimator and discuss the estimation of asymptotic standard errors in Section 2. We also propose methods for comparison of correlated markers using difference in AUC curves. Finally, we examine the finite sample properties of the proposed AUC function estimator as well as the difference estimator via simulation studies in Section 4. In Section 5, we apply the method to Mayo PBC data and Breast cancer mortality data. We conclude the article with a brief discussion (Section 6).

2. ESTIMATION OF A TIME-DEPENDENT ACCURACY FUNCTION

2.1 *Notation*

Let n denote the total number of subjects in the study. Let T_i and C_i denote the survival time and censoring time for subject i. We assume that $T_1, T_2, \ldots, T_n$ are independent and identically distributed and so are $C_1, C_2, \ldots, C_n$. For each subject, T_i is assumed to be independent with C_i. Let $\mathbb{1}\{\cdot\}$ denote the indicator function. We observe the follow-up time $Z_i = \min\{T_i, C_i\}$ and the censoring indicator $\delta_i = \mathbb{1}\{T_i \leqslant C_i\}$. Let $R_i(t) = \mathbb{1}\{Z_i \geqslant t\}$ denote the at-risk indicator. Let $M_1, M_2, \ldots, M_n$ denote independently distributed baseline markers for the subjects. Note that the marker may be a single covariate X or based on a set of (time-independent or time-dependent) covariates $X_1, X_2, \ldots, X_p$, via a possibly time-dependent score generated through any regression or predictive model such as the proportional hazard model. Higher marker values are assumed to be more indicative of disease and therefore shorter survival time. Let $\mathcal{R}_t = \{i : R_i(t) = 1\}$ denotes the subjects that are in the riskset at time t. Among the subjects in $\mathcal{R}_t$, we denote by $\mathcal{R}_t^1$ the subjects who had an event at t or cases: $\mathcal{R}_t^1 = \{i : T_i = t\}$. The subjects who did not have an event by t are the controls: $\mathcal{R}_t^0 = \{i : T_i > t\}$. We denote by n_t the size of the control set at time t: $n_t = |\mathcal{R}_t^0|$ and $d_t = |\mathcal{R}_t^1|$. Thus, cases are incident ($\mathbb{I}$) cases and controls are dynamic ($\mathbb{D}$) controls (Heagerty and Zheng, 2005). In the next subsection, we introduce non-parametric, time-dependent, concordance measures for a baseline marker and motivate the modification for a longitudinal marker. The following definitions of time-dependent accuracy are adopted:

$$\begin{aligned}
\mathrm{TP}_t^{\mathbb{I}}(c) &= \mathrm{pr}(M_i > c | T_i = t),\\
\mathrm{FP}_t^{\mathbb{D}}(c) &= \mathrm{pr}(M_i > c | T_i > t),\\
\mathrm{ROC}_t^{\mathbb{I}/\mathbb{D}}(p) &= \mathrm{TP}_t^{\mathbb{I}}\{[\mathrm{FP}_t^{\mathbb{D}}]^{-1}(p)\},\\
\mathrm{AUC}(t) &= \mathrm{pr}(M_i > M_j | T_i = t, T_j > t).
\end{aligned}$$

2.2 *Weighted mean rank: a local concordance measure*

We first consider a baseline marker M and introduce a new approach to evaluate the predictive accuracy of this marker for a time-to-event outcome T by extending the standard binary diagnostic accuracy summaries. Note that the risk-set at time t can be represented as $\mathcal{R}_t = \{\mathcal{R}_t^1 \cup \mathcal{R}_t^0\}$. A non-parametric estimator of AUC that is frequently used in case–control studies is the proportion of concordant case–control pairs, i.e. proportion of case–control pairs where the marker value for the case is higher than the marker value for the control. This statistic can be adapted in this situation using the incident case and dynamic control definition:

$$A(t) = \frac{1}{d_t \times n_t} \sum_{i \in \mathcal{R}_t^1} \sum_{j \in \mathcal{R}_t^0} \mathbb{1}\{M_i > M_j\}.$$

Note that at a given event time t, $A(t)$ can be considered as an estimator of a time-dependent concordance measure $\mathrm{AUC}(t) = \mathrm{pr}(M_i > M_j | T_i = t, T_j > t)$, which represents the area under the *incident/dynamic* time-dependent ROC curve defined by Heagerty and Zheng (2005). However, frequently, there are only a few cases at t and often $d_t = 1$. In this situation, the information within a neighborhood around t can be utilized to estimate marker concordance at t. When $d_t = 1$, $n_t \times A(t) = \sum_{j \in \mathcal{R}_t^0} \mathbb{1}\{M^* > M_j | T^* = t, T_j > t\}$ is the rank of the case marker value among the control markers. We propose using the locally weighted mean rank (WMR) for the case markers among the available controls as an estimator of

the local concordance:

$$\mathrm{WMR}(t) := \frac{1}{|\mathcal{N}_t(h_n)|} \sum_{t_j \in \mathcal{N}_t(h_n)} A(t_j),$$

where $\mathcal{N}_t(h_n) = \{t_j : |t - t_j| < h_n\}$ denotes a neighborhood around t. This is a nearest-neighbor estimator of the AUC and can be generalized to

$$\widehat{\mathrm{AUC}}(t) = \sum_j K_{h_n}(t - t_j) \cdot A(t_j),$$

where K_{h_n} is a standardized kernel function such that $\sum_j K_{h_n}(t - t_j) = 1$.

2.3 *Asymptotic distribution of WMR estimator*

For simplicity, we assume that no two subjects fail at the same time and no subjects are censored. We denote the ordered observed failure times as $t_{(1)} < t_{(2)} < \cdots t_{(M)}$. In this section we focus on the estimation of $\mathrm{AUC}(t)$ at a fixed t using $\mathrm{WMR}(t)$. Given a bandwidth h_n, such that $h_n \to 0$ and $nh_n \to \infty$, we restrict our attention to a neighborhood around t: $\mathcal{N}_t(h_n) = \{t_j : |t - t_j| < h_n\}$ and $|\mathcal{N}_t(h_n)| = m_t$. Furthermore, we assume that, for any fixed t, the observed number of subjects at risk, $n_t + 1$, is $O(n)$ (proportional to the sample size) and therefore $\to \infty$. This assumption results when a fixed censoring distribution is assumed, and when t is within the support of the event time and censoring time distributions.

In order to detail the large sample properties of the proposed local estimator, we define the indices of ordered event times that fall within the neighborhood $\mathcal{N}_t(h_n)$. Let $L_t = |\{t_{(j)} : t_{(j)} < t - h_n\}|$ denote the number of failures that are observed before the start of the neighborhood of interest (e.g. before time $t - h_n$). The indices of the observed failure times within the neighborhood are then $L_t + 1, L_t + 2, \ldots, L_t + m_t$ and that $t_{(L_t+1)} < t_{(L_t+2)} < \cdots < t_{(L_t+m_t)}$ are the corresponding unique event times within $\mathcal{N}_t(h_n)$.

For the estimation of $\mathrm{AUC}(t)$ based on information specific to the neighborhood $\mathcal{N}_t(h_n)$, any failure time that is observed after the right boundary of the neighborhood (e.g. $> t_{(L_t+m_t)}$) is only used as a "control" for all of the observed events in the neighborhood. Therefore, we effectively censor subjects surviving past $t_{(L_t+m_t)}$ at this time.

First, we also assume that all event times are reordered such that for the ith subject with event time T_i and marker M_i this individual corresponds to the ordered time $T_i = t_{(i)}$. Secondly, for local estimation at time t, all subjects with event times beyond the right edge of the window contribute only as local "control" observations and we use only the information that they have $T_i > t_{(L_t+m_t)}$. We can therefore equivalently represent the WMR estimator in terms of the locally censored observations $T_i^* = \min[T_i, t_{(L_t+m_t+1)}]$:

$$\begin{aligned}
\mathrm{WMR}(t) &= \frac{1}{|\mathcal{N}_t(h_n)|} \sum_{t_{(i)} \in \mathcal{N}_t(h_n)} A(t_{(i)}) \\
&= \frac{1}{m_t} \sum_{i=L_t+1}^{L_t+m_t} A(t_i) \\
&= \frac{1}{m_t} \sum_{i=L_t+1}^{L_t+m_t} \frac{1}{n_{t_{(i)}}} \sum_{j=i+1}^{n} \mathbb{1}\{M_i > M_j\} \times \mathbb{1}\{T_i^* < T_j^*\} \\
&= \sum_{i \neq j} \frac{1}{2 \times m_t \times n_{t_{(i)}}} \mathbb{1}\{M_i > M_j\} \times \mathbb{1}\{T_i^* < T_j^*\}.
\end{aligned}$$

Note that for two subjects i and j with $i, j > L_t + m_t$, we cannot order the failure times, Hence, we assume that $\mathbb{1}\{T_i^* < T_j^*\} = 0$ and $\mathbb{1}\{T_j^* < T_i^*\} = 0$.

From the above representation of the WMR(t), it can be seen that WMR(t) is a linear transformation of weighted U-statistic: WMR$(t) = \frac{1}{2}(U + 1)$ defined as

$$U = \sum_{i \neq j} w_{ij} \operatorname{sgn}(j - i) \operatorname{sgn}(R_i - R_j),$$

where $\operatorname{sgn}(x) = 1$ if $x > 0$ and $\operatorname{sgn}(x) = -1$ if $x < 0$, R_i is the rank of the marker M corresponding to the ith-ordered observed failure time and $1/w_{ij} = 2 \times m_t \times n_{t_{(i)}} = 2 \times m_t \times |\mathcal{R}^0_{\min\{t_{(i)}, t_{(j)}\}}|$. Since higher marker values are more indicative of a shorter survival time, the definition of U is slightly different than the standard U-statistic.

Conditioning on the event times, we can show that standardized WMR(t) has an asymptotic normal distribution since, in this case, the weights w_{ij} can be treated as constants. We state the result in the following theorem, and explicitly denote the sample size for the proposed estimator using the notation WMR$_n(t)$:

THEOREM 1 Suppose $n - (L_t + m_t) \to \infty$ when $n \to \infty$ where $n - (L_t + m_t)$ denotes the number of subjects surviving past $t_{(L_t + m_t)}$ or equivalently, surviving past $t + h_n$ and let $h_n \to 0$, $nn_n \to \infty$. Then

$$V_n^{-1/2}[\text{WMR}_n(t) - \text{AUC}(t) - b_n(t)] \xrightarrow{\mathcal{D}} N(0, 1),$$

where $V_n = \text{var}[\text{WMR}_n(t)]$ given below, and $b_n(t)$ denotes the bias: $E[\text{WMR}_n(t)] = \text{AUC}(t) + b_n(t)$.

Proof. See Appendix for details. □

Since the WMR using a nearest-neighbor kernel is equivalent to the Nadaraya–Watson estimator, at an interior point t WMR$_n(t)$ has a bias of $b_n(t) = \frac{2}{3}h_n^2 B(t)$ for a given bandwidth h_n (Hardle, 1992). Here $B(t) = \frac{1}{2}\text{AUC}^{(2)}(t) + f^{-1}(t)\text{AUC}^{(1)}(t) f^{(1)}(t)$, where AUC$^{(1)}(t)$ (AUC$^{(2)}(t)$) denotes the first (second) derivative of the AUC, $f(\cdot)$ denotes the probability density function of the survival time and $f^{(1)}(\cdot)$ denotes the first derivative. Characterization of the finite-sample bias can be used to correct confidence interval estimates to ensure proper coverage rates. Each of the elements of bias can be estimated using standard smoothing techniques. It can be shown that for a bandwidth h_n satisfying $h_n = o(n^{-2/5})$, the bias will be asymptotically negligible relative to $V_n^{-1/2}$, but using a bandwidth $O(n^{-1/5})$ is optimal for minimizing integrated mean squared error (IMSE), yet then requires the bias correction for proper confidence interval construction.

2.4 *Estimation of variance*

Because WMR(t) is an average of random variables, we can write

$$\text{var}\{\text{WMR}(t)\} = \frac{1}{m_t^2}\left[\sum_{j \in \mathcal{N}_t(h_n)} \text{var}\{A(t_j)\} + \sum_{j \neq k} \text{cov}\{A(t_j), A(t_k)\}\right].$$

We propose the following as estimators of the variance of $A(t)$ and covariance between $A(t)$ and $A(s)$, $t < s$:

$$\widehat{\text{var}}\{A(t)\} = \frac{(n_t - 1)}{n_t}\{\hat{Q}_2(t) - \hat{Q}_0(t)^2\} + \frac{1}{n_t}[\hat{Q}_0(t)\{1 - \hat{Q}_0(t)\}],$$

$$\widehat{\text{cov}}\{A(t), A(s)\} = \frac{1}{n_t}[\{\hat{Q}_3(t,s) - \hat{Q}_{3.0}(t,s)\} + \{\hat{Q}_4(t,s) - \hat{Q}_{4.0}(t,s)\}].$$

$Q_\cdot(\cdot)$ is defined in Appendix. So, $\text{var}\{\text{WMR}(t)\} = O(m_t^{-1})$. Further, $\hat{Q}_0(t)$, $\hat{Q}_2(t)$, etc. are obtained using a normal approximation for the case and control markers after a rank-based Z-score transformation and then empirically estimating the parameters of the approximating normal distributions. Ultimately, our implementation of variance estimation is based solely on the marginal ranks of the marker measurements and survival times since, as we describe in the appendix, we first rank the raw marker values and then apply a Z-score transformation to these ranks. Secondly, we use a normal approximation to the conditional distributions, $M^*|T = t$ and $M^*|T > t$, where M^* denote the transformed marker values. Although the normal approximation is not an exact characterization, we have found that the approximation, coupled with the rank-based Z-score transformation, provides accurate standard error estimates for even highly skewed marginal marker distributions. Details of our simulation evaluations are presented in the next section.

2.5 *Selection of bandwidth*

We note that estimation of AUC via WMR is similar in spirit with several smoothing techniques that are widely used in practice and selection of a bandwidth, h_n for $\mathcal{N}_t(h_n)$ is important to the estimation of WMR. The optimal bandwidth balances bias and variance, and optimizes the estimation of time-dependent concordance. Since we want to estimate the time-dependent concordance, we focus on a criterion directly related to it. We adopt a bandwidth that minimizes the asymptotic IMSE of WMR:

$$h_n^{\text{opt}} = \underset{h>0}{\text{argmin}} \int_T \text{MSE}[\text{WMR}_{h_n}(t)]\,\text{d}t.$$

In practice, cross-validation or a data-driven method can be employed to estimate the asymptotically optimal bandwidth.

Replacing the integration over time by a sum over k unique event times, the optimal bandwidth is a minimizer of

$$\sum_{i=1}^{k} (W_{h_n}^{-i}(t_i) - A(t_i))^2,$$

where

$$W_{h_n}^{-i}(t_i) = \frac{1}{|\mathcal{N}_{t_i}(h_n)| - 1} \sum_{\substack{j \neq i \\ j \in \mathcal{N}_{t_i}(h_n)}} A(t_j).$$

Similar to other smoothing techniques, alternative types of neighborhood selection can also be employed. For example, instead of a fixed bandwidth for the entire support, we can employ a nearest-neighbor method with fixed number of "neighbor" or failure times. Fixing the number of "neighbors" instead of a fixed bandwidth will especially be useful when either the events are sparse or the sizes of risksets are small. A cross-validation method based on IMSE can be employed to estimate the optimal number of "neighbors" as before.

3. COMPARISON OF MARKERS

Suppose that it is of interest to compare the two markers A (M^{A}) and B (M^{B}) for their predictive accuracy. Comparison of the AUCs for these two markers is one simple approach to determine whether marker A is more accurate than marker B in correctly classifying the incident cases and dynamic controls at t. This can be done via the difference in the WMR statistic for the respective markers. For the marker-specific AUCs at t, define $\mathrm{dAUC}(t) = \mathrm{AUC}^{\mathrm{A}}(t) - \mathrm{AUC}^{\mathrm{B}}(t)$ and $\mathrm{dWMR}(t) = \mathrm{WMR}^{\mathrm{A}}(t) - \mathrm{WMR}^{\mathrm{B}}(t)$. Then $\mathrm{dWMR}(t)$ can be used as an estimator of $\mathrm{dAUC}(t)$.

The estimation of the variance of $\mathrm{dWMR}(t)$ is straightforward for two uncorrelated markers since, in this case,

$$\mathrm{var}\{\mathrm{dWMR}(t)\} = \mathrm{var}\{\mathrm{WMR}^{\mathrm{A}}(t)\} + \mathrm{var}\{\mathrm{WMR}^{\mathrm{B}}(t)\}.$$

This is true when markers are measured on different sub-groups. However, frequently, both the markers are assessed on the same set of subjects and hence the covariance between the two WMR statistics cannot be ignored in general. For correlated markers, the variance of $\mathrm{dWMR}(t)$ can be written as

$$\mathrm{var}\{\mathrm{dWMR}(t)\} = \mathrm{var}\{\mathrm{WMR}^{\mathrm{A}}(t)\} + \mathrm{var}\{\mathrm{WMR}^{\mathrm{B}}(t)\} - 2\,\mathrm{cov}\{\mathrm{WMR}^{\mathrm{A}}(t), \mathrm{WMR}^{\mathrm{B}}(t)\}.$$

For estimating the covariance, one needs to evaluate terms of the form

$$\mathrm{cov}\{A^{\mathrm{A}}(t), A^{\mathrm{B}}(t)\}, \quad \mathrm{cov}\{A^{\mathrm{A}}(t), A^{\mathrm{B}}(s)\}, \quad \text{and} \quad \mathrm{cov}\{A^{\mathrm{A}}(s), A^{\mathrm{B}}(t)\}, \quad t < s,$$

where $A^{\mathrm{A}}(t) = (1/n_t)\sum_j \mathbb{1}\{M^{\mathrm{A}} > M_j^{\mathrm{A}} | T = t, T_j > t\}$, etc. The estimation of these covariance terms parallels the approach for variance estimation for single marker.

An optimal bandwidth may be selected for marker comparison as before. In particular, an optimal bandwidth for marker comparison is a minimizer of

$$\sum_{i=1}^{k} \{\mathrm{d}W_{h_n}^{-i}(t_i) - \mathrm{d}A(t_i)\}^2,$$

where

$$\mathrm{d}W_{h_n}^{-i}(t_i) = \frac{1}{|\mathcal{N}_{t_i}(h_n)| - 1} \sum_{\substack{j \neq i \\ j \in \mathcal{N}_{t_i}(h_n)}} \mathrm{d}A(t_j),$$

$$\mathrm{d}A(t_i) = \frac{1}{n_{t_i}} \sum_j \mathbb{1}\{{M_i}^{\mathrm{A}} > {M_j}^{\mathrm{A}} | T_i = t_i, T_j > t_i\} - \frac{1}{n_{t_i}} \sum_j \mathbb{1}\{{M_i}^{\mathrm{B}} > {M_j}^{\mathrm{B}} | T_i = t_i, T_j > t_i\}.$$

4. SIMULATION STUDY

4.1 *Single marker*

To demonstrate the use of the WMR to estimate time-dependent concordance and the rank-based variance estimator, we conducted a set of simulation studies. Suppose that T denotes the (log) time until failure and M denotes the marker. We assumed that (T, M) jointly follows a bivariate normal distribution with means 0, variances 1, and correlation $\rho = -0.7$. For each of $N = 1000$ simulated datasets, a sample of $n = 200$ marker values M and survival times T were generated. An additional independent censoring (log) time was

Table 1. *Simulation results for WMR and comparison with semi-parametric method. We assumed* $(log(T), M) \sim N_2(0, 0, 1, 1, -0.7)$. *An independent censoring time was generated such that either* 20% *or* 40% *subjects were censored. The results are based on* 1000 *such simulations. The MLE and semi-parametric estimates are cited from Heagerty and Zheng (2005). MCSD denotes the Monte-Carlo SD. The nominal coverage is* 95.0. *Estimate of integrated AUC* (*iAUC*) *and its variance is based on* 10 *equi-spaced quantiles of time. The variance of integrated AUC is estimated assuming a constant weight*

		MLE		Semi-parametric		WMR ($n = 200$)			
Log time	AUC(t)	Mean	SD	Mean	SD	Mean	MCSD	EstSD	Coverage
20% censoring									
−2.0	0.884	0.884	0.018	0.881	0.044	0.876	0.055	0.050	90.2
−1.5	0.833	0.834	0.019	0.829	0.035	0.828	0.040	0.041	93.4
−1.0	0.782	0.782	0.019	0.771	0.033	0.780	0.037	0.036	93.2
−0.5	0.734	0.734	0.019	0.720	0.033	0.734	0.035	0.034	94.1
0.0	0.693	0.693	0.018	0.686	0.034	0.695	0.037	0.037	94.0
0.5	0.660	0.660	0.016	0.657	0.040	0.664	0.047	0.045	94.9
1.0	0.634	0.634	0.015	0.637	0.041	0.638	0.066	0.064	93.0
iAUC	0.741	0.741	0.016	0.740	0.018	0.738	0.020	0.017	89.6
40% censoring									
−2.0	0.884	0.884	0.019	0.875	0.048	0.876	0.056	0.050	89.9
−1.5	0.833	0.834	0.021	0.827	0.037	0.828	0.041	0.042	92.9
−1.0	0.782	0.782	0.021	0.772	0.035	0.781	0.038	0.037	92.7
−0.5	0.734	0.734	0.020	0.722	0.039	0.735	0.039	0.038	94.3
0.0	0.693	0.693	0.019	0.687	0.041	0.696	0.044	0.043	93.5
0.5	0.660	0.660	0.018	0.655	0.043	0.666	0.061	0.059	92.6
1.0	0.634	0.635	0.016	0.637	0.048	0.640	0.108	0.100	91.8
iAUC	0.741	0.741	0.017	0.742	0.021	0.739	0.023	0.019	88.3

generated such that either 20% or 40% subjects are censored. To estimate that WMR, a bandwidth of $n^{-1/5}$ ($b = 1$) was considered. For each simulated dataset, we estimated that WMR at $t = -2.0, -1.5, \ldots, 1.0$. The average WMR over 1000 simulations, the Monte-Carlo standard deviation (MCSD), the estimated SD (EstSD) using the proposed variance estimator and the coverage can be found in Table 1. For comparison purposes, we also cite the maximum likelihood estimate (MLE) and semi-parametric estimates of time-dependent AUC as in Heagerty and Zheng (2005). Additionally, simulation results for comparison of two markers can be found in Table 2.

We find that when 20% observations are censored, the relative bias of WMR is $< 1\%$. For example, at $\log(t) = 1.0$, the mean WMR is 0.638 while the true concordance is 0.634 with a relative bias of 0.6%. The relative bias is reduced with an increasing sample size. The estimated standard deviation is comparable with the MCSD with a coverage close to the nominal level except for the edges. This kind of edge behavior is typical of many other scatter plot smoothing techniques like LOWESS. Thus, the variance estimator can be used in practice for inference instead of a resampling-based variance estimator since the variance estimator may be less computationally intensive than, say, bootstrap.

4.2 *Comparison between semi-parametric and non-parametric approaches*

To demonstrate the potential robustness advantage of the proposed non-parametric rank-based estimators over the semi-parametric estimator of Heagerty and Zheng (2005) (henceforth HZ2005), we carried out additional simulations. Here we focus on an underlying heterogeneous population where one subset has a

Table 2. *Simulation results for comparison of two markers via the difference in respective WMR estimators. We assumed* $(\log T, M_A, M_B)$ *follows a multivariate normal distribution with* $\text{cor}(\log T, M_A) = -0.7$, $\text{cor}(\log T, M_B) = -0.5$, *and* $\text{cor}(M_A, M_B) = 0.8$. *An independent censoring time was generated such that either* 20% *or* 40% *subjects were censored. We show the difference between the theoretical AUC for marker A versus marker B* (dAUC(t)), *the difference between the corresponding estimated WMR statistics* (dWMR(t)), *the Monte-Carlo SD* (*MCSD*) *and the SD estimated using the proposed variance estimator* (*EstSD*) *and the coverage* (*nominal*: 95.0)

Log time	dAUC(t)	dWMR(t)	MCSD	EstSD	Coverage
20% censoring					
−2.0	0.101	0.099	0.062	0.054	88.8
−1.5	0.098	0.096	0.040	0.040	94.5
−1.0	0.091	0.090	0.032	0.032	94.6
−0.5	0.080	0.078	0.029	0.029	94.8
0.0	0.070	0.069	0.030	0.029	94.9
0.5	0.060	0.060	0.037	0.035	93.2
1.0	0.051	0.053	0.054	0.051	93.3
40% censoring					
−2.0	0.101	0.099	0.062	0.054	88.8
−1.5	0.098	0.096	0.041	0.040	94.7
−1.0	0.091	0.090	0.033	0.033	95.0
−0.5	0.080	0.079	0.031	0.031	94.8
0.0	0.070	0.071	0.035	0.035	94.5
0.5	0.060	0.061	0.048	0.047	94.8
1.0	0.051	0.053	0.100	0.089	93.3

normal (low) value of the marker, yet is likely to die relatively quickly, while a second major subset of the population does have the marker predictive of time until death. Biologically this scenario is plausible in situations where the disease group under study is not homogeneous, and the marker of interest correlates only with the outcome for one disease subgroup. Specifically, we assumed that the (log) time until failure T and the marker M comes from a mixture distribution where

$$(T, M) = \begin{cases} (T^{(I)}, M^{(I)}) & \text{if } Z = 1, \\ (T^{(N)}, M^{(N)}) & \text{if } Z = 0, \end{cases}$$

with $T^{(I)} \sim N(-1.5, 1)$, independent of $M^{(I)} \sim N(-1.5, 1)$, $(T^{(N)}, M^{(N)}) \sim N_2(0, 0, 1, 1, \rho)$, and $Z \sim$ Bernoulli(p).

Further, $\rho = -0.8$ and $p = 0.2$. An independent censoring time was generated such that 20% of the subjects were censored. For each of the $N = 1000$ simulated datasets, a sample of $n = 1000$ marker values M and survival time T were generated. For both HZ2005 and WMR, a bandwidth of 0.1 was considered. For each simulated dataset, we estimated WMR and HZ2005 at $t = -2.5, -2.0, \ldots, 1.0$. The average WMR, HZ2005, MCSD for both the estimators and EstSD and coverage of WMR can be found in Table 3. We found that, for such a scenario, the relative bias for WMR is $< 1\%$ at all times while the relative bias for HZ2005 can be as high as 54% and the relative absolute bias for HZ2005 was greater than that of WMR at all times. The WMR estimator in general was less efficient than HZ2005, but not terribly so. However, the bias in HZ2005 remained substantial when compared with WMR at all times.

Table 3. *Simulation results for comparison of semi-parametric approach of Heagerty and Zheng (2005) (HZ2005) and non-parametric approach WMR. We assumed* $(\log T, M)$ *follows a mixture of two multivariate normal distributions:* $(\log T^{(I)}, M^{(I)}) \sim N_2(-1.5, -1.5, 1, 1, 0)$ *if* $Z = 1$ *and* $(\log T^{(N)}, M^{(N)}) \sim N_2(0, 0, 1, 1, -0.8)$ *if* $Z = 0$*, where* $Z \sim$ *Bernoulli* (0.2)*. An independent censoring time was generated such that 20% of the subjects were censored. We show the estimated HZ2005, WMR, MCSDs, and the SD estimated using the proposed variance estimator (EstSD) and the coverage (nominal: 95.0) for WMR*

Log time	AUC(t)	HZ2005	MCSD	WMR	MCSD	EstSD	Coverage
−2.5	0.378	0.173	0.069	0.376	0.109	0.097	87.7
−2.0	0.481	0.346	0.087	0.477	0.083	0.075	91.1
−1.5	0.591	0.551	0.071	0.595	0.062	0.057	92.0
−1.0	0.673	0.660	0.048	0.674	0.048	0.043	91.6
−0.5	0.709	0.689	0.032	0.708	0.035	0.034	93.7
0.0	0.709	0.684	0.023	0.710	0.030	0.030	94.8
0.5	0.691	0.666	0.022	0.692	0.034	0.033	92.5
1.0	0.669	0.646	0.022	0.669	0.042	0.041	93.9

5. Examples

5.1 *Example 1*

We now illustrate the proposed methods using data from 312 randomized subjects from the Mayo primary biliary cirrhosis (PBC) study, a randomized, placebo-controlled trial of the drug D-penicillamine (DPCA) for the treatment of PBC conducted at the Mayo clinic between 1974 and 1984 (Fleming and Harrington, 1991). Among these subjects, 125 died by the end of the follow-up. We demonstrate AUC curves associated with four-covariate and five-covariate model, their difference and the associated point-wise 95% confidence interval.

We consider a score from Cox model with covariates: log(bilirubin), log(prothrombin time), edema, albumin, and age. The log hazard ratios under PH assumption can be found in (Heagerty and Zheng, 2005, Table 3). The prognostic score from this model has been used previously (Costa and Shaw, 2009; Fleming and Harrington, 1991). Here, we address the question: how well does this score discriminate between subjects who are at risk of an imminent failure from those who are not. We plot the WMR curve and associated point-wise 95% confidence interval (Figure 1(a)). A bandwidth of 504 days was used such that the IMSE is minimized ($h = 0.8$). Another WMR curve based on a four-covariate model (bandwidth = 504 days) (log(prothrombin time), edema, albumin, and age) with point-wise 95% confidence interval is plotted in Figure 1(b). This modified Mayo score is less predictive than the score from the five-covariate model and the time-specific accuracy decreases steadily over time. We further compared the predictive accuracy of the five-covariate and four-covariate model scores using the difference in the respective WMR curves. We plotted the difference in the WMRs in Figure 2 along with the point-wise 95% confidence interval for the difference using the proposed variance estimator of the dWMR(t) (see Section 3). Throughout the entire time period considered here, the predictive accuracy of the five-covariate model is generally better than the four-covariate model.

In Figure 1(a), we show the proposed estimator, and show the result of using a simple LOWESS smoother (solid, blue line) that tends to dampen the fluctuation observed after 2000 days. We do not have any biological explanation for the late increase, and expect this to likely reflect variability and/or be the result of cross-validation choosing a relatively small bandwidth. One additional analysis that could explain the observed pattern would be to consider bandwidth estimation based on IMSE separately over the first 2000 days and the later times—this would likely allow a larger bandwidth at later times and therefore

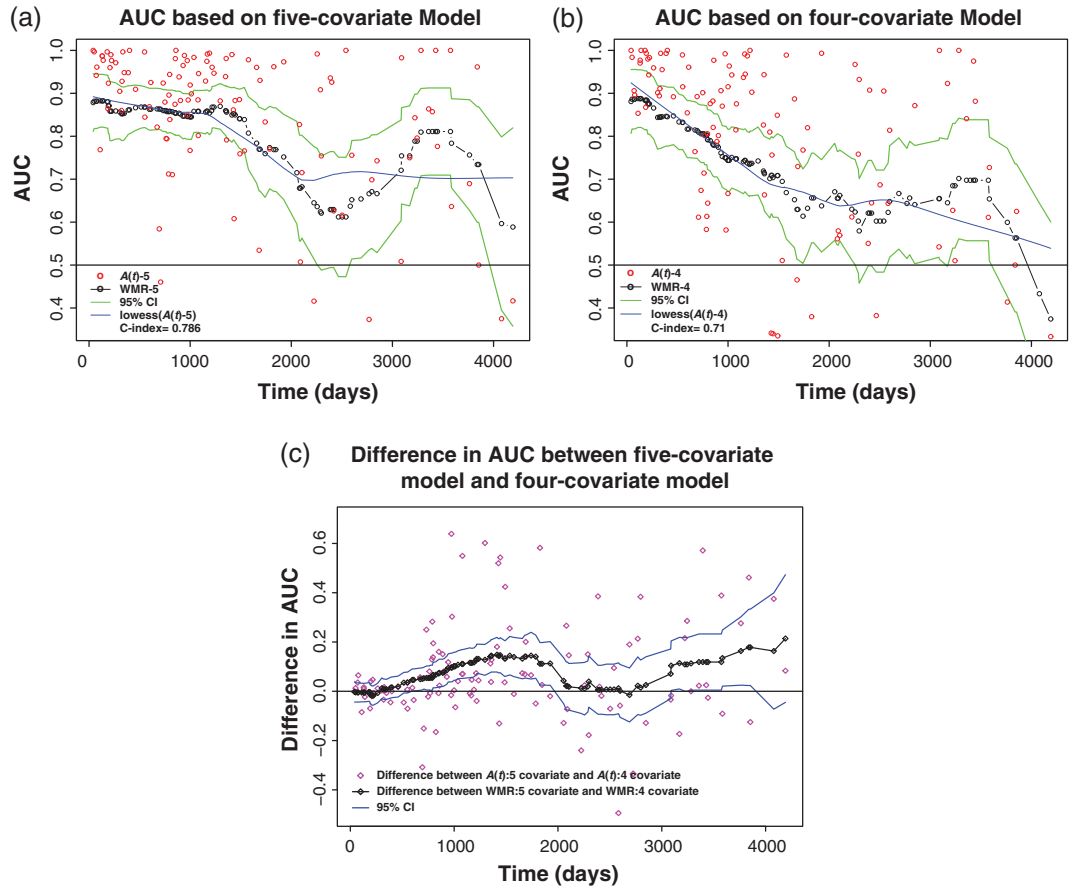

Fig. 1. (a) and (b) The WMR curve and point-wise confidence interval for Mayo PBC data (bandwidth = 504 days). (c) The dWMR curve for the difference in AUC from five-covariate and four-covariate models and point-wise confidence interval for Mayo PBC data (bandwidth = 504 days). (a) Five-covariate model. (b) Four-covariate model. (c) Difference between five- and four-covariate models.

tend to smooth this part of the curve in more agreement with the LOWESS estimator. We have introduced IMSE as one objective method for choosing the bandwidth, but other options are worth consideration. Figure 1(b) uses a different composite marker (excludes bilirubin) and shows a similar yet dampened pattern after 2000 days. The WMR curves, dWMR curve and the associated confidence intervals based on leave-one-out cross-validation were very similar to those estimated without cross-validation.

5.2 *Example 2*

In the second example, we focus on comparison of a new pathology measurement with a standard one as predictors of mortality among young breast cancer patients. A cohort of young (onset before age 45) subjects had primary tumor samples analyzed for the percent of cells in the S-phase of the replication cycle. A standard measurement of S-phase (in %) was obtained in addition to a new measurement (in %) based exclusively on epithelial cells that were sorted using flow cytometry. Survival subsequent to disease diagnosis was recorded on $n = 253$ subjects with a median observed follow-up time of 62 months. A prior

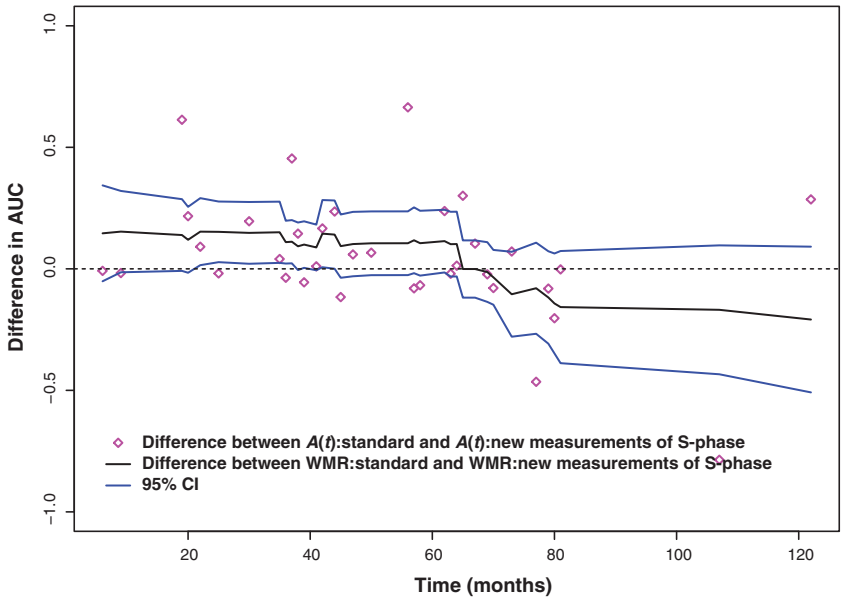

Fig. 2. The dWMR curve for the difference between AUC (solid, middle line) from the new and standard percent S-phase measurements and point-wise confidence interval (solid, top and bottom lines) among young-onset breast cancer patients. The raw differences are indicated by diamonds.

analysis of the data (Heagerty *and others*, 2000) using cumulative events inferred that the new marker was only superior during the first 60 months. Our proposed WMR methods allow us to directly compare the riskset rank based on the standard measure versus the new measure to see whether improvement is obtained, and to see whether improvement is uniform across time. Figure 2 shows the difference in the WMR as a function of time and clearly shows an advantage for the new marker but only during the first 60 months where cases ranked using the new measurement had an average percentile that was 12% higher. Thus, the use of the new graphical method allows a direct evaluation of both how much of an improvement is obtained, and of *when* improvement is obtained.

6. Discussion

In this article, we introduce a non-parametric estimator of a time-dependent predictive accuracy function that is useful for characterizing the accuracy of a predictive score or scalar marker for a survival time that may be censored. For the entire study duration, we look at the ability of the score to discriminate between subjects who are at risk of an impending failure (incident cases) from those who are not (dynamic controls). We use AUC(t) as a measure of predictive accuracy at t and show that a non-parametric estimator of AUC, namely the locally weighted average of the proportion of concordant case–control pairs, can be used to estimate this time-dependent concordance. The proposed estimator, WMR at t, is a smoothed version of time-specific AUCs within a neighborhood around each t and is similar to other smoothing approaches, yet is based on local U-statistic summaries. Furthermore, we showed that under certain conditions on the size of the control set when compared with the number of failures, the standardized estimator WMR(t) asymptotically follows a normal distribution. We provide an estimator of the asymptotic variance. Both the estimator and the variance can be easily estimated. We examined the estimator WMR(t) and the variance estimator

via simulation studies and showed that both perform satisfactorily under realistic censoring proportions. We introduced a simple approach for comparing the predictive accuracy of two correlated markers using the difference in the respective WMR estimates. We proposed a variance estimator for this difference and evaluated the properties of the estimator via simulation. We applied the methods to the well-known Mayo PBC dataset, and compared the results with a semi-parametric estimator and drew similar conclusions.

For any given time point t, construction of a confidence interval for WMR(t) is straight-forward using the asymptotic properties presented here. In many applications, however, it may be desirable to obtain simultaneous confidence bands for the function WMR(t), so that a single probability statement can be made regarding the predictive accuracy curve over the entire time span of interest. While estimation of simultaneous confidence band (e.g. for cumulative hazard, survival curve, etc.) has received a wide attention in the survival analysis literature (Hall and Wellner, 1980; Parzen *and others*, 1997), the associated theory may not be applicable in our case (since the limiting process of WMR($\cdot$) may not possess an independent increment structure). When an estimator in such situations does not possess independent incremental structure, a well-known approach is to simulate a Gaussian process while keeping the observations fixed to approximate the distribution of the process and to estimate the appropriate quantiles based on the perturbed distribution. This approach has been used in the literature for the estimation of subject-specific cumulative hazard and survival function (Lin *and others*, 1994), the difference in two survival functions (Parzen *and others*, 1997), and contrast in two hazard functions (Gilbert *and others*, 2002) and may be applicable for the estimation of simultaneous confidence bands for the WMR curve. Computationally simpler Bonferroni-type bands may also be useful in our setting. A third approach, stemming from literature on kernel smoothing (Wu *and others*, 1998), first establishes simultaneous confidence intervals for a set of grid points and then bridges the gap between the grid points via smoothness conditions of the curve. However, additional careful work is needed to extend simultaneous methods to our application and is an important direction for future research.

For the basic development, we assume that the censoring time, C_i, is independent of the survival time, T_i. Estimation that allows this assumption to be relaxed to permit the weaker assumption of conditional independence between the failure and censoring times given the marker would be appropriate in certain settings and therefore warrants further development.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (PSC) and by grants R01 HL072966 and U54 RR024379.

APPENDIX A

A.1 *Asymptotic Normality of WMR*

Suppose that we are interested in characterizing the predictive accuracy of the marker at t using $(1/|\mathcal{N}|)\Sigma_{t_j\in\mathcal{N}}A(t_j)$, where $\mathcal{N}=\{t_j:|t-t_j|<h_n\}$. We assume that $s_{L_t+1}<s_{L_t+2}<\cdots<s_{L_t+m_t}\in\mathcal{N}$ denotes the unique failure times with $|\mathcal{N}|=m_t$. We first assume that no subjects were censored in $(s_{L_t+1}, s_{L_t+m_t})$ and the subjects surviving $s_{L_t+m_t}$ are considered censored at $s_{L_t+m_t}$, so that the event times of these subjects cannot be ordered.

We first restate the following theorem from Lee (1990).

THEOREM A1 Let $W_n = \sum_{(n.k)} w(S)\psi(S)$ be a weighted U-statistic of order $k = 2$ and define

$$w_{i,n} = \sum_{S:i\in S} w(S),$$

$$W_n^* = \sum_{i=1}^{n} w_{i,n}\psi_i(X_i).$$

Suppose that the following conditions hold when $n \to \infty$:

(i)
$$\max_{1\leqslant i\leqslant n} \frac{|w_{i,n}|^2}{\sum_{i=1}^n w_{i,n}^2 \to 0}.$$

(ii)
$$\frac{\sum_{(n,2)} w^2(S)}{\sum_{i=1}^n w_{i,n}^2 \to 0}.$$

(iii) $E|\psi_1(X_1)|^{2+\delta} < \infty$ for some $\delta > 0$.

Then $(W_n - \theta)/(\text{var } W_n)^{1/2} \underrightarrow{\mathcal{D}} N(0, 1)$.

Here $\psi_i(x) = E(\psi(X_i, X_j)|X_i = x)$ is the conditional expectation given $X_i = x$.

For asymptotic normality of WMR, consider $X = (T, M^*) = (T, -M)$. We define a U-statistic with event time T and a baseline marker M^* as

$$U = \sum_{i\neq j} w_{ij}\,\text{sgn}(i - j)\,\text{sgn}(R_i - R_j),$$

where R_i is the rank of the marker value M^* for a subject $a(i)$ who experienced the ith ordered failure: $T_{a(i)} = s_i$. Also

$$\text{sgn}(x - y) = \begin{cases} 1 & \text{if } x > y, \\ -1 & \text{if } x < y, \\ 0 & \text{otherwise.} \end{cases}$$

Note that, the event times T_{l_1}, T_{l_2} for two subjects surviving past $s_{L_t+m_t}$ cannot be ordered; we assume that $\text{sgn}(l_1 - l_2) = 0$ for those subjects. Finally, we assume that the weights are as follows:

$$w_{ij} = \frac{1}{2 \times m_t \times |\mathcal{R}^0_{\min\{s_i,s_j\}}|}$$

if s_i and s_j can be ordered and $w_{ij} = 0$ when the event times s_i and s_j cannot be ordered. We now state the asymptotic normality of WMR as in the following theorem.

THEOREM A2 Suppose $n - (L_t + m_t) \to \infty$ when $n \to \infty$, where $n - (L_t + m_t)$ denotes the number of subjects surviving past $t_{(L_t+m_t)}$ or equivalently, surviving past $t + h_n$ and let $h_n \to 0$, $nn_n \to \infty$. Then

$$V_n^{-1/2}[\text{WMR}_n(t) - \text{AUC}(t) - b_n(t)] \underrightarrow{\mathcal{D}} N(0, 1),$$

where $V_n = \text{var}[\text{WMR}_n(t)]$ and $b_n(t)$ denotes the bias: $E[\text{WMR}_n(t)] = AUC(t) + b_n(t)$.

Proof. To prove the asymptotic normality of WMR(t), we note that $\text{WMR}(t) = \frac{1}{2}(U+1)$ and simply verify conditions (i)–(iii).

Note that the assumption (iii) holds for $\psi(\mathbf{X}_i, \mathbf{X}_j) = \text{sgn}(i-j)\,\text{sgn}(R_i - R_j)$ since $|\psi(\mathbf{X}_i, \mathbf{X}_j)| \leqslant 1$.

Dropping subscript n from $w_{i,n}$ to simplify the notation, we see

$$w_i = \sum_j w_{ij} = \begin{cases} \dfrac{1}{2m_t}\left(1 + \displaystyle\sum_{j<i} \frac{1}{|\mathcal{R}^0_{s_j}|}\right) & i = L_t+1, L_t+2, \ldots, L_t+m_t, \\ \dfrac{1}{2m_t} \displaystyle\sum_{j=L_t+1}^{L_t+m_t} \frac{1}{|\mathcal{R}^0_{s_j}|} & i > L_t+m_t, \end{cases}$$

so that

$$\max_{L_t+1 \leqslant i \leqslant n} |w_i|^2 = w^2_{L_t+m_t} \leqslant \frac{1}{4m_t^2}\left(1 + \frac{m_t}{n_{L_t+m_t}}\right)^2,$$

where $|\mathcal{R}^0_{s_i}| = n_i = (n-i)$. Also,

$$\begin{aligned} 4m_t^2 \times \sum_i w_i^2 &= 1 + \left(1 + \frac{1}{n_{L_t+1}}\right)^2 + \left(1 + \frac{1}{n_{L_t+1}} + \frac{1}{n_{L_t+2}}\right)^2 \\ &\quad + \cdots + \left(1 + \frac{1}{n_{L_t+1}} + \frac{1}{n_{L_t+2}} + \cdots + \frac{1}{n_{L_t+m_t-1}}\right)^2 \\ &\quad + (n - (L_t+m_t)) \times \left(\frac{1}{n_{L_t+1}} + \frac{1}{n_{L_t+2}} + \cdots + \frac{1}{n_{L_t+m_t-1}} + \frac{1}{n_{L_t+m_t}}\right)^2 \\ &\geqslant (n - (L_t+m_t)) \times \left(\frac{m_t}{n_{L_t+1}}\right)^2 \\ \Rightarrow \frac{\max_{L_t+1 \leqslant i \leqslant n} |w_i|^2}{\sum_i w_i^2} &\leqslant \frac{(1 + m_t/(n_{L_t+m_t}))^2}{(n - (L_t+m_t)) \times (m_t/n_{L_t+1})^2} \longrightarrow 0 \quad \text{if } n - (L_t+m_t) \to \infty \text{ as } n \to \infty. \end{aligned}$$

Thus, condition (i) is satisfied. If subjects were censored between s_{L_t+1} and $s_{L_t+m_t}$, the inequality above will still hold. Finally,

$$\begin{aligned} 4m_t^2 \sum_{i \neq j} w_{ij}^2 &= \left(\underbrace{\frac{1}{n^2_{L_t+1}} + \cdots + \frac{1}{n^2_{L_t+1}}}_{n_{L_t+1} \text{ times}}\right) + \left(\frac{1}{n^2_{L_t+1}} + \underbrace{\frac{1}{n^2_{L_t+2}} + \cdots + \frac{1}{n^2_{L_t+2}}}_{n_{L_t+2} \text{ times}}\right) \\ &\quad + \cdots + \left(\frac{1}{n^2_{L_t+1}} + \frac{1}{n^2_{L_t+2}} + \cdots + \frac{1}{n^2_{L_t+m_t-1}} + \underbrace{\frac{1}{n^2_{L_t+m_t}} + \cdots + \frac{1}{n^2_{L_t+m_t}}}_{n_{L_t+m_t} \text{ times}}\right) \end{aligned}$$

$$
\begin{aligned}
&+ (n - (L_t + m_t)) \times \left(\frac{1}{n^2_{L_t+1}} + \frac{1}{n^2_{L_t+2}} + \cdots + \frac{1}{n^2_{L_t+m_t-1}} + \frac{1}{n^2_{L_t+m_t}} \right) \\
&= \frac{1}{n_{L_t+1}} + \left(\frac{1}{n^2_{L_t+1}} + \frac{1}{n_{L_t+2}} \right) + \left(\frac{1}{n^2_{L_t+1}} + \frac{1}{n^2_{L_t+2}} + \frac{1}{n_{L_t+3}} \right) \\
&\quad + \cdots + \left(\frac{1}{n^2_{L_t+1}} + \frac{1}{n^2_{L_t+2}} + \cdots + \frac{1}{n^2_{L_t+m_t-1}} + \frac{1}{n_{L_t+m}} \right) \\
&\quad + (n - (L_t + m)) \times \left(\frac{1}{n^2_{L_t+1}} + \frac{1}{n^2_{L_t+2}} + \cdots + \frac{1}{n^2_{L_t+m_t-1}} + \frac{1}{n^2_{L_t+m_t}} \right) \\
&= 2 \left(\frac{1}{n_{L_t+1}} + \frac{1}{n_{L_t+2}} + \cdots + \frac{1}{n_{L_t+m_t}} \right) \\
&\leqslant \frac{2m_t}{n_{L_t+m_t}}
\end{aligned}
$$

and hence condition (ii) is also satisfied (whether subjects were censored between s_{L_t+1} and $s_{L_t+m_t}$) if $n - (L_t + m_t) \to \infty$ as $n \to \infty$. □

Appendix B

B.1 *Variance of WMR*

As before, we want to characterize the predictive accuracy of the marker at t using $(1/|\mathcal{N}|)\Sigma_{t_j \in \mathcal{N}} A(t_j)$, where $\mathcal{N} = \{t_j : |t - t_j| < h_n\}$. Suppose, $t, s \in \mathcal{N}$. Note that

$$
\begin{aligned}
A(t) &= \frac{1}{n_t} \sum_{\mathcal{R}(t)} \mathbb{1}\{M_j < M_i | T_i = t, T_j > t\}, \\
\Rightarrow \mathbb{E}\{A(t)\} &\approx Q_0(t), \\
\text{var}\{A(t)\} &\approx \frac{1}{n_t}[Q_0(t)\{1 - Q_0(t)\} + (n_t - 1)\{Q_2(t) - Q_0(t)^2\}], \\
\text{cov}\{A(t), A(s)\} &\approx \frac{1}{n_t}[\{Q_3(t,s) - Q_{3.0}(t,s)\} + \{Q_4(t,s) - Q_{4.0}(t,s)\}], \quad \text{with } t < s
\end{aligned}
$$

where

$$
\begin{aligned}
Q_0(t) &= \text{pr}(M_j < M_i | T_i = t, T_j > t), \\
Q_2(t) &= \text{pr}(M_j < M_i, M_k < M_i | T_i = t, T_j > t, T_k > t), \\
Q_3(t,s) &= \text{pr}(M_j < M_i, M_j < M_l | T_i = t, T_l = s, T_j > s), \\
Q_{3.0}(t,s) &= \text{pr}(M_j < M_i | T_i = t, T_j > s) \times \text{pr}(M_j < M_l | T_l = s, T_j > s), \\
Q_4(t,s) &= \text{pr}(M_l < M_i, M_j < M_l | T_i = t, T_l = s, T_j > s), \\
Q_{4.0}(t,s) &= \text{pr}(M_l < M_i | T_i = t, T_l = s) \times \text{pr}(M_j < M_l | T_l = s, T_j > s).
\end{aligned}
$$

The variance is the same as the variance of an AUC estimator using the proportion of concordant case–control pairs as obtained from a case-control study (Pepe, 2003). When there are d_t events at t,

$$\text{var}\{A(t)\} \approx \frac{1}{n_t \times d_t}[Q_0(t)\{1 - Q_0(t)\} + (d_t - 1)\{Q_1(t) - Q_0(t)^2\} + (n_t - 1)\{Q_2(t) - Q_0(t)^2\}],$$

where

$$Q_1(t) = \text{pr}(M_j < M_i, M_j < M_l | T_i = t, T_l = t, T_k > t).$$

However, the expression for $\text{cov}\{A(t), A(s)\}$ remains the same.

To estimate the variance, we use the empirical normal quantiles of the rank of the markers instead of the markers themselves. Finally, the quantities of interest are estimated assuming normality of the transformed marker:

$$\begin{aligned}
Q_0(t) &\hat{=} \Phi\left(-\frac{\mu_0(t) - \mu_1(t)}{\sqrt{(\sigma_0^2(t) + \sigma_1^2(t))}}\right), \\
Q_2(t) &= \text{pr}(M_j < M_i, M_k < M_i | T_i = t, T_j > t, T_k > t) \\
&\hat{=} \Phi_2(\mathbf{0} | \tilde{\mu}_2, \tilde{\Sigma}_2), \\
Q_3(t,s) &= \text{pr}(M_j < M_i, M_j < M_l | T_i = t, T_l = s, T_j > s) \\
&\hat{=} \Phi_2(\mathbf{0} | \tilde{\mu}_3, \tilde{\Sigma}_3), \\
Q_{3.0}(t,s) &= \text{pr}(M_j < M_i | T_i = t, T_j > s) \times \text{pr}(M_j < M_l | T_l = s, T_j > s) \\
&\hat{=} \Phi\left(-\frac{\mu_0(s) - \mu_1(t)}{\sqrt{(\sigma_0^2(s) + \sigma_1^2(t))}}\right) \times \Phi\left(-\frac{\mu_0(s) - \mu_1(s)}{\sqrt{(\sigma_0^2(s) + \sigma_1^2(s))}}\right), \\
Q_4(t,s) &= \text{pr}(M_l < M_i, M_j < M_l | T_i = t, T_l = s, T_j > s) \\
&\hat{=} \Phi_2(\mathbf{0} | \tilde{\mu}_4, \tilde{\Sigma}_4), \\
Q_{4.0}(t,s) &= \text{pr}(M_l < M_i | T_i = t, T_l = s) \times \text{pr}(M_j < M_l | T_l = s, T_j > s) \\
&\hat{=} \Phi\left(-\frac{\mu_1(s) - \mu_1(t)}{\sqrt{(\sigma_1^2(s) + \sigma_1^2(t))}}\right) \times \Phi\left(-\frac{\mu_0(s) - \mu_1(s)}{\sqrt{(\sigma_0^2(s) + \sigma_1^2(s))}}\right), \\
\tilde{\mu}_2 &= \begin{pmatrix} \mu_0(t) - \mu_1(t) \\ \mu_0(t) - \mu_1(t) \end{pmatrix} \quad \tilde{\Sigma}_2 = \begin{pmatrix} \sigma_0^2(t) + \sigma_1^2(t) & \sigma_1^2(t) \\ \sigma_1^2(t) & \sigma_0^2(t) + \sigma_1^2(t) \end{pmatrix}, \\
\tilde{\mu}_3 &= \begin{pmatrix} \mu_0(s) - \mu_1(t) \\ \mu_0(s) - \mu_1(s) \end{pmatrix} \quad \tilde{\Sigma}_3 = \begin{pmatrix} \sigma_0^2(s) + \sigma_1^2(t) & \sigma_0^2(s) \\ \sigma_0^2(s) & \sigma_0^2(s) + \sigma_1^2(s) \end{pmatrix}, \\
\tilde{\mu}_4 &= \begin{pmatrix} \mu_1(s) - \mu_1(t) \\ \mu_0(t) - \mu_1(s) \end{pmatrix} \quad \tilde{\Sigma}_4 = \begin{pmatrix} \sigma_1^2(s) + \sigma_1^2(t) & -\sigma_0^2(s) \\ -\sigma_0^2(s) & \sigma_0^2(s) + \sigma_1^2(s) \end{pmatrix}.
\end{aligned}$$

If there are more than one case at t, $\mu_1(t)$ and $\sigma_1^2(t)$ can be estimated as the mean and variance of the (transformed) marker of the cases at t, while $\mu_0(t)$ and $\sigma_0^2(t)$ are estimated as the mean and variance of the (transformed) marker of the controls at t. However, usually there is only a single incident event at t. Hence, we assume that the mean marker values for the cases within the neighborhood of interest $\mathcal{N}$ (or $\tilde{\mathcal{N}} \subset \mathcal{N}$) is constant: $\mu_1(t) = \mu_1(s)$ and $\sigma_1^2(t) = \sigma_1^2(s)$, and use all the cases within the neighborhood to estimate $\mu_1(t)$ and $\sigma_1^2(t)$.

References

Antolini, L., Boracchi, P. and Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine* **24**, 3927–3944.

Coombes, J. M. and Trotter, J. F. (2005). Development of the allocation system for deceased donor liver transplantation. *Clinical Medicine and Research* **3**, 87–92.

Costa, M. J. and Shaw, J. E. H. (2009). Parametrization and penalties in spline models with an application to survival analysis. *Computational Statistics and Data Analysis* **53**, 657–670.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Process and Survival Analysis*. New York: John Wiley & Sons.

Gilbert, P. B., Wei, L. J., Kosorok, M. R. and Clemens, J. D. (2002). Simultaneous inferences on the contrast of two hazard functions with censored observations. *Biometrics* **58**, 773–780.

Gonen, M. and Heller, G. (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* **92**, 965–970.

Gries, C. J., Rue, T. C., Heagerty, P. J., Edelman, J. D., Mulligan, M. S. and H., Goss C. (2010). Development of a predictive model for long-term survival after lung transplantation and implications for the lung allocation score. *Journal of Heart Lung Transplant* **29**, 731–738.

Hall, W. J. and Wellner, J. A. (1980). Confidence bands for a survival curve from censored data. *Biometrika* **67**, 133–143.

Hardle, W. (1992). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. and Rosati, R. A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association* **247**, 2543–2546.

Heagerty, P. J., Lumley, T. and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105.

Lee, A. J. (1990). *U-statistics: Theory and Practice*. New York: Marcel Dekker.

Lin, D. Y., Fleming, T. R. and Wei, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**, 73–81.

Parzen, M. I., Wei, L. J. and Ying, Z. (1997). Simultaneous confidence intervals for the difference of two survival functions. *Scandanavian Journal of Statistics* **24**, 309–314.

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

Wu, C. O., Chiang, C-T. and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association* **93**, 1388–1402.

[*Received July 1, 2010; revised February 3, 2012; accepted for publication May 17, 2012*]