Biometrics 61, 92–105
March 2005

Survival Model Predictive Accuracy and ROC Curves

Patrick J. Heagerty

Department of Biostatistics, University of Washington, P.O. Box 357232, Seattle, Washington 98195-7232, U.S.A.
email: heagerty@u.washington.edu

and

Yingye Zheng

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, MP 702, P.O. Box 19024, Seattle, Washington 98109-1024, U.S.A.

Summary. The predictive accuracy of a survival model can be summarized using extensions of the proportion of variation explained by the model, or R^2, commonly used for continuous response models, or using extensions of sensitivity and specificity, which are commonly used for binary response models. In this article we propose new time-dependent accuracy summaries based on time-specific versions of sensitivity and specificity calculated over risk sets. We connect the accuracy summaries to a previously proposed global concordance measure, which is a variant of Kendall's tau. In addition, we show how standard Cox regression output can be used to obtain estimates of time-dependent sensitivity and specificity, and time-dependent receiver operating characteristic (ROC) curves. Semiparametric estimation methods appropriate for both proportional and nonproportional hazards data are introduced, evaluated in simulations, and illustrated using two familiar survival data sets.

Key words: Cox regression; Discrimination; Prediction; Sensitivity; Specificity.

1. Introduction

In this article we propose a new method for characterizing the predictive accuracy of a regression model when the outcome of interest is a censored survival time. We focus on data obtained from a prospective study in which a continuous follow-up time is observed for each participant, but where follow-up can be terminated either by the occurrence of the event of interest or by censoring. Thus the essential outcome information is the combination of the status at the end of follow-up (binary) and the length of follow-up (continuous). Because censored data share features of both continuous response data and binary data, the accuracy concepts that are standard for either response type may be extended for survival outcomes. Previous research has focused on extending the proportion of variation explained by the covariates, or R^2, to censored data models (Schemper and Henderson, 2000; O'Quigley and Xu, 2001). In addition, limited work has explored the use of familiar binary outcome methods such as receiver operating characteristic (ROC) curves for application in the longitudinal setting (Etzioni et al., 1999; Heagerty, Lumley, and Pepe, 2000; Slate and Turnbull, 2000). Time-dependent ROC curves offer an alternative to the use of R^2 extensions for survival data. However, the goal of an ROC analysis is to characterize the prognostic potential of a marker (or model) by focusing on the correct classification rates. Methods that summarize the proportion of variation explained by covariates require a different estimation approach, and have a different ultimate objective. The goals of this article are to introduce new time-dependent sensitivity, specificity, and ROC concepts appropriate for survival regression models; to demonstrate the connection between time-dependent ROC methods and classical concordance summaries such as Kendall's tau or the "c index" (Harrell, Lee, and Mark, 1996); and to show how standard Cox regression estimation methods directly provide the ingredients needed to calculate the proposed time-dependent accuracy summaries.

1.1 *Notation*

Let T_i be the survival time for subject i, and assume that we only observe the minimum of T_i and C_i, where C_i represents an independent censoring time. Define the follow-up time $X_i = \min(T_i, C_i)$, and let $\Delta_i = \mathbf{1}(T_i \leq C_i)$ denote the censoring indicator. The survival time T_i can also be represented through the counting process, $N_i^*(t) = \mathbf{1}(T_i \leq t)$, or the corresponding increment, $dN_i^*(t) = N_i^*(t) - N_i^*(t-)$. Note that we focus on the counting process $N_i^*(t)$ which is defined solely in terms of the survival time T_i rather than the more common notation $N_i(t) = \mathbf{1}(X_i \leq t, \Delta_i = 1)$, which depends on the censoring time (Fleming and Harrington, 1991). Let $R_i(t) = \mathbf{1}(X_i \geq t)$ denote the at-risk indicator. We also assume that for each subject we have a collection of time-invariant covariates, $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$.

We focus here on using Cox model methods to both generate a model score and to evaluate the prognostic potential of the model score. However, the evaluation methods that we propose can be used to summarize the accuracy of a prognostic score generated through any alternative regression or predictive method, and in this case varying coefficient methods (Hastie and Tibshirani, 1993) such as locally weighted partial likelihood estimation (Cai and Sun, 2003) provide a convenient approach for estimating key accuracy summaries. Therefore, we briefly introduce the relevant aspects of partial likelihood estimation. Under the proportional hazards assumption, $\lambda(t \mid \boldsymbol{Z}_i) = \lambda_0(t)\exp(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta})$, where $\lambda(t \mid \boldsymbol{Z}_i) = \lim_{\delta \to 0}\delta^{-1}P[T_i \in [t,\, t+\delta) \mid \boldsymbol{Z}_i,\, T_i \geq t]$. The partial likelihood score equations can be written as

$$\mathbf{0} = \sum_i \Delta_i \left[\boldsymbol{Z}_i - \left(\sum_k \pi_k(\boldsymbol{\beta}, X_i)\boldsymbol{Z}_k \right) \right],$$

where $\pi_k(\boldsymbol{\beta}, t) = R_k(t) \cdot \exp(\boldsymbol{Z}_k^{\mathrm{T}}\boldsymbol{\beta})/W(t)$, with $W(t) = \sum_j R_j(t) \cdot \exp(\boldsymbol{Z}_j^{\mathrm{T}}\boldsymbol{\beta})$. Solving these equations yields the consistent and asymptotically normal maximum partial likelihood estimator (MPLE) $\hat{\boldsymbol{\beta}}$ (Cox, 1972).

1.2 *Proportion of Variance Approaches*

Two main approaches exist for characterizing the proportion of variation explained by a survival model. Schemper and Henderson (2000) overview an approach where the survival time is characterized by a counting process representation, $N_i^*(t) = \mathbf{1}(T_i \leq t)$, and time-integrated variances are used to form the summary measure. Alternatively, O'Quigley and Xu (2001) consider the proportion of variation in the covariate, $\boldsymbol{Z}_i$, that is explained by the survival time T_i.

Schemper and Henderson (2000) build on earlier work that extends R^2 to Cox regression. Their approach focuses on using the counting process, $N_i^*(t)$, and marginal and conditional expectations given by the survival functions $S(t) = E[1 - N_i^*(t)]$ and $S(t \mid \boldsymbol{Z}_i) = E[1 - N_i^*(t) \mid \boldsymbol{Z}_i]$, respectively. Because the vital status indicator $N_i^*(t)$ is a binary variable, Schemper and Henderson (2000) propose using the marginal variance $S(t)[1 - S(t)]$ and the conditional variance $S(t \mid \boldsymbol{Z}_i)[1 - S(t \mid \boldsymbol{Z}_i)]$ to characterize the proportion of variation explained by the covariates $\boldsymbol{Z}_i$. In particular, a finite time range $(0,\, \tau)$ is considered and time-average variances are formed:

$$D(\tau) = \int_0^\tau S(t)[1 - S(t)] \cdot f(t)\, dt \Big/ \int_0^\tau f(t)\, dt$$

$$D_{\boldsymbol{Z}}(\tau) = \int_0^\tau E_{\boldsymbol{Z}}\{S(t \mid \boldsymbol{Z})[1 - S(t \mid \boldsymbol{Z})]\} \cdot f(t)\, dt \Big/ \int_0^\tau f(t)\, dt,$$

where $f(t)$ is the marginal density of T_i. Our representation above differs by a factor of 2 from the proposal of Schemper and Henderson (2000) as they also consider the mean absolute deviation, $E[|N_i^*(t) - S(t)|] = 2S(t)[1 - S(t)]$. Finally, the summary $V(\tau) = \lceil D(\tau) - D_{\boldsymbol{Z}}(\tau) \rceil / D(\tau)$ is proposed as the proportion of variation explained by covariates. Similarly, our approach views survival data through the counting process representation, $N_i^*(t)$, but because $N_i^*(t)$ is a binary outcome we explore the extension of standard binary response accuracy summaries such as ROC curves rather than considering an extension of R^2.

O'Quigley and Xu (2001) also develop R^2 summaries for Cox regression. In their approach the role of survival time and covariate are reversed, and the proportion of variation in the covariate that is explained by survival is proposed. The authors exploit partial likelihood estimation methods because the methods provide model-based estimates of the distribution of covariates conditional on survival time. Focusing on a scalar covariate, Xu and O'Quigley (2000) show that $\pi_i(\beta,\ t) = R_i(t)\exp(\boldsymbol{Z}_i\beta)/W(t)$ can be used to estimate the distribution of the covariate, $\boldsymbol{Z}_i$, conditional on the event occurring at time t, $\hat{P}(\boldsymbol{Z}_i \leq z \mid T_i = t) = \sum_j \pi_j(\beta, t) \cdot \mathbf{1}(Z_j \leq z)$. O'Quigley and Xu (2001) obtain estimates of the conditional variance $\mathrm{var}(\boldsymbol{Z}_i \mid T_i = t)$ and propose a global summary by integrating estimates of the marginal and conditional variance over the survival distribution. Our approach is similar in that we also use $\pi_i(\boldsymbol{\beta}, t)$ to estimate conditional distributions, but rather than computing variances we estimate time-dependent versions of sensitivity and specificity defined in the following section.

1.3 *Overview*

In Section 2 we briefly review ROC methods proposed for summarizing the accuracy of a prognostic marker or model when the outcome of interest is a survival time. We then develop new definitions of time-dependent sensitivity and specificity that are strongly connected to partial likelihood concepts. Time-dependent accuracy measures can be used to calculate time-specific ROC curves, and time-specific area under the curve (AUC) summaries. We show that a global concordance measure is the integral, or weighted average, of time-specific AUC measures. In Section 3 we discuss the estimation of time-dependent ROC and AUC summaries and provide a method that is applicable to a proportional hazards model, and a more general method that can be used to characterize any scalar prognostic score even if proportional hazards do not obtain. Finally, in Section 4 we analyze two well-known data sets. We conclude the article with a brief discussion.

2. Censored Survival and Predictive Accuracy

2.1 *Background on ROC Curve Analysis*

When outcomes Y_i are binary the accuracy of a prediction or classification rule is typically summarized through correct classification rates defined as sensitivity, $P(\hat{p}_i > c \mid Y_i = 1)$, and specificity, $P(\hat{p}_i \leq c \mid Y_i = 0)$, where $\hat{p}_i$ is a prediction, and c is a criterion for classifying the prediction as positive $(\hat{p}_i > c)$ or negative $(\hat{p}_i \leq c)$. When no a priori value of c is indicated the full spectrum of sensitivities and specificities can be characterized using an ROC curve that plots the "true positive rate" (sensitivity) versus the "false positive rate" (1-specificity) for all $c \in (-\infty, +\infty)$.

An ROC curve provides complete information on the set of all possible combinations of true-positive and false-positive rates, but is also more generally useful as a graphical characterization of the magnitude of separation between the case and control marker distributions. If case measurements and control measurements have no overlap then the ROC curve takes the value 1 (perfect true-positive rate) for any false-positive rate greater than 0. In this situation the marker is perfect at discriminating between cases and controls.

Alternatively, if the case and control distributions are identical then the ROC curve lies on the 45° line indicating that the marker is useless for separating cases from controls.

The area under the ROC curve, or AUC, is known to represent a measure of concordance between the marker and the disease status indicator (Hanley and McNeil, 1982). Specifically, the AUC measures the probability that the marker value for a randomly selected case exceeds the marker value for a randomly selected control and is directly related to the Mann–Whitney U statistic (Hanley and McNeil, 1982; Pepe, 2003). Finally, ROC curves are particularly useful for comparing the discriminatory capacity of different potential biomarkers. For example, if for each value of specificity one marker always has a higher sensitivity, then this marker will be a uniformly better diagnostic measurement. See Zhou, McClish, and Obuchowski (2002) or Pepe (2003) for more discussion of ROC analysis.

In this section we first review previous proposals for generalizing the concepts of sensitivity and specificity for application to survival endpoints. Definitions of sensitivity and specificity are given in terms of the actual survival time T_i. Censoring needs to be addressed for valid estimation. We then show that a certain choice of time-dependent true-positive and false-positive definitions leads to time-dependent ROC curves and time-dependent AUC summaries that are directly related to a previously proposed concordance summary for survival data.

2.2 *Extensions of Sensitivity and Specificity*

For survival data there are several potential extensions of cross-sectional sensitivity and specificity. Rather than a simple binary outcome, $Y_i = 1$, a survival time can be viewed as a time-varying binary outcome by focusing on the counting process representation $N_i^*(t) = \mathbf{1}(T_i \le t)$. Accuracy extensions are classified according to whether the "cases" used to define time-dependent sensitivity are *incident* cases where $T_i = t$, or equivalently $dN_i^*(t) = 1$, is used to define cases for time t, or *cumulative* cases where $T_i \le t$ or $N_i^*(t) = 1$ is used. We also consider whether "controls" are *static*, defined as subjects with $T_i > t^\star$ for a fixed value of $t^\star$, or whether controls are *dynamic* and defined for time t as those subjects with $T_i > t$. We use the superscripts $\mathbb{C}$ and $\mathbb{I}$ to denote different definitions of sensitivity, and use the superscripts $\mathbb{D}$ and $\bar{\mathbb{D}}$ to denote different definitions of specificity. In this section we focus on a scalar marker value M_i that is used as a predictor of death. When our interest is in the accuracy of a regression model we will use $M_i = \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}$.

2.2.1 *Cumulative/dynamic.* For a baseline marker value, M_i, Heagerty et al. (2000) propose versions of time-dependent sensitivity and specificity using the definitions

$$\text{sensitivity}^{\mathbb{C}}(c,t) : P(M_i > c \mid T_i \le t) = P\big(M_i > c \mid N_i^*(t) = 1\big)$$

$$\text{specificity}^{\mathbb{D}}(c,t) : P(M_i \le c \mid T_i > t) = P\big(M_i \le c \mid N_i^*(t) = 0\big).$$

Using this approach, at any fixed time t the entire population is classified as either a case or a control on the basis of vital status at time t. Also, each individual plays the role of a control for times $t < T_i$, but then contributes as a case for later times, $t \ge T_i$. Cumulative/dynamic accuracy summaries are most appropriate when a specific time t' (or a small collection of times $t'_1, t'_2, \ldots, t'_m$) is important and scientific interest lies in discriminating between subjects who die prior to a given time t' and those that survive beyond t'. ROC curves are defined as $\text{ROC}_t^{\mathbb{C}/\mathbb{D}}(p) = \text{TP}_t^{\mathbb{C}}\{[\text{FP}_t^{\mathbb{D}}]^{-1}(p)\}$ where $\text{TP}_t^{\mathbb{C}}(c) = P(M_i > c \mid N_i^*(t) = 1)$, $\text{FP}_t^{\mathbb{D}}(c) = P(M_i > c \mid N_i^*(t) = 0)$, and $[\text{FP}_t^{\mathbb{D}}(p)]^{-1} = \inf_c\{c : \text{FP}_t^{\mathbb{D}}(c) \le p\}$. In the absence of censoring $\text{ROC}_t^{\mathbb{C}/\mathbb{D}}(p)$ can be estimated using the empirical distribution of the marker separately among cases and controls. With censored survival times Heagerty et al. (2000) develop a nonparametric estimator based on the nearest-neighbor bivariate distribution estimator of Akritas (1994). A substantive application that demonstrates use of cumulative/dynamic ROC curves for a Cox regression model can be found in Fan et al. (2002).

2.2.2 *Incident/static.* Etzioni et al. (1999) and Slate and Turnbull (2000) adopt an alternative definition of time-dependent sensitivity and specificity using

$$\text{sensitivity}^{\mathbb{I}}(c,t) : P(M_i > c \mid T_i = t) = P\big(M_i > c \mid dN_i^*(t) = 1\big)$$

$$\text{specificity}^{\bar{\mathbb{D}}}(c,t^\star) : P\big(M_i \le c \mid T_i > t^\star\big) = P\big(M_i \le c \mid N_i^*(t^\star) = 0\big),$$

where $dN_i^*(t) = N_i^*(t) - N_i^*(t-)$. Using this definition, each subject does not change disease status and is treated as either a case or a control. Cases are stratified according to the time at which the event occurs (incident) and controls are defined as those subjects who are event free through a fixed follow-up period, $(0, t^\star)$ (static). These definitions facilitate the use of standard regression approaches for characterizing sensitivity and specificity because the event time, T_i, can simply be used as a covariate. To estimate the quantiles of the conditional distribution of the marker, M_i, given the event time, $T_i = t$, Etzioni et al. (1999) and Slate and Turnbull (2000) consider parametric methods that assume a normal distribution, but which allow the mean and variance to be functions of the measurement time, disease status, and the event time for the cases. Cai et al. (2003) propose methods for estimating time-dependent sensitivity and specificity when the event time is censored. Recently, Zheng and Heagerty (2004) have proposed regression quantile methods, which relax the parametric distributional assumptions of previous approaches.

2.2.3 *Incident/dynamic.* In this article we focus on the following definitions of sensitivity and specificity:

$$\text{sensitivity}^{\mathbb{I}}(c,t) : P(M_i > c \mid T_i = t) = P\big(M_i > c \mid dN_i^*(t) = 1\big)$$

$$\text{specificity}^{\mathbb{D}}(c,t) : P(M_i \le c \mid T_i > t) = P\big(M_i \le c \mid N_i^*(t) = 0\big).$$

Using this approach a subject can play the role of a control for an early time, $t < T_i$, but then play the role of case when $t = T_i$. This dynamic status parallels the multiple contributions that a subject can make to the partial likelihood function. Here sensitivity measures the expected fraction of subjects with a marker greater than c among the subpopulation of individuals who die at time t, while specificity measures the fraction of subjects with a marker less than or equal to c among those who survive beyond time t. Incident sensitivity and dynamic specificity are defined by dichotomizing the risk set at time t into those observed to die (cases) and those observed to survive (controls). In Section 3 we discuss how the observed marker data among risk sets can be used to estimate time-dependent accuracy concepts.

Incident sensitivity and dynamic specificity have some appealing characteristics relative to the alternative definitions. First, incident sensitivity and dynamic specificity are based on classification of the risk set at time t into case(s) and controls, and are, therefore, a natural companion to hazard models. Second, the definitions easily allow extension to time-dependent covariates using $P[M_i(t) > c \,|\, T_i = t]$ to define incident sensitivity and $P[M_i(t) \leq c \,|\, T_i > t]$ to define dynamic specificity with a longitudinal marker $M_i(t)$. Use of cumulative sensitivity does not permit a time-varying marker. Finally, use of incident sensitivity and dynamic specificity allows both time-specific accuracy summaries and, as shown in Section 2.4, allows time-averaged summaries that directly relate to a familiar global concordance measure. In contrast, methods have not been proposed for meaningfully averaging the time-specific incident/static or cumulative/dynamic accuracy summaries.

2.3 *Time-Dependent ROC Curves*

After selecting definitions for time-dependent sensitivity and specificity, ROC curves can be computed and interpreted. In this article we focus on incident/dynamic (I/D) ROC curves defined as the function $\mathrm{ROC}_t^{\mathbb{I}/\mathbb{D}}(p)$, where p denotes the dynamic false-positive rate, and $\mathrm{ROC}_t^{\mathbb{I}/\mathbb{D}}(p)$ denotes the corresponding incident true-positive rate. Specifically, let c^p be defined as the threshold that yields a false-positive rate of p: $P(M_i > c^p \mid T_i > t) = 1 - \text{specificity}^{\mathbb{D}}(c^p, t) = p$. The true-positive rate, $\mathrm{ROC}_t^{\mathbb{I}/\mathbb{D}}(p)$, is the sensitivity that is obtained using this threshold, or $\mathrm{ROC}_t^{\mathbb{I}/\mathbb{D}}(p) = \text{sensitivity}^{\mathbb{I}}(c^p, t) = P(M_i > c^p \mid T_i = t)$. Using the true and false-positive rate functions $\mathrm{TP}_t^{\mathbb{I}}(c) = \text{sensitivity}^{\mathbb{I}}(c, t)$ and $\mathrm{TP}_t^{\mathbb{D}}(c) = 1 - \text{specificity}^{\mathbb{D}}(c, t)$ allows the ROC curve to be written as the composition of $\mathrm{TP}_t^{\mathbb{I}}(c)$ and the inverse function $[\mathrm{TP}_t^{\mathbb{D}}]^{-1}(p) = c^p$:

$$\mathrm{ROC}_t^{\mathbb{I}/\mathbb{D}}(p) = \mathrm{TP}_t^{\mathbb{I}}\left\{ \left[\mathrm{FP}_t^{\mathbb{D}}\right]^{-1}(p) \right\}$$

for $p \in [0, 1]$. We use the notation $\mathrm{AUC}(t) = \int_0^1 \mathrm{ROC}_t^{\mathbb{I}/\mathbb{D}}(p)dp$ to denote the area under the I/D ROC curve for time t.

2.4 *Time-Dependent AUC and Concordance*

In the previous subsection we discussed how ROC methods can be used to characterize the ability of a marker to distinguish cases at time t from controls at time t. However, in many applications no a priori time t is identified, and a global accuracy summary is desired. In this subsection we show how time-dependent ROC curves are related to a standard "concordance" summary. The global summary we adopt is

$$C = P[M_j > M_k \mid T_j < T_k],$$

which indicates the probability that the subject who died at the earlier time has a larger value of the marker. This is not the usual form (i.e., $P[M_j > M_k \,|\, T_j > T_k]$), but reflects the conventions for ROC analysis.

In order to understand the relationship between this discrimination summary and ROC curves we assume independence of observations (M_j, T_j) and (M_k, T_k), and assume that T_j is continuous such that $P(T_k = T_j) = 0$. We use $P(x)$ to denote probability or density depending on the context. These assumptions imply that the concordance summary C is a weighted average of the area under time-specific ROC curves,

$$\begin{aligned} &P[M_j > M_k \mid T_j < T_k] \\ &\quad = 2 \int_t P[\{M_j > M_k\} \mid \{T_j = t\} \cap \{t < T_k\}] \\ &\qquad \times P[\{T_j = t\} \cap \{t < T_k\}]\, dt \\ &\quad = \int_t \mathrm{AUC}(t) \cdot w(t)\, dt = E_T[\mathrm{AUC}(T) \times 2 \times S(T)] \\ &\qquad \text{with} \quad w(t) = 2 \cdot f(t) \cdot S(t). \end{aligned}$$

In this notation $\mathrm{AUC}(t)$ is based on the I/D definition of sensitivity and specificity, $\mathrm{AUC}(t) = P(M_j > M_k \,|\, T_j = t,\, T_k > t)$. See the Appendix for a derivation.

In practice we would typically restrict attention to a fixed follow-up period $(0,\, \tau)$. The concordance summary can be modified to account for finite follow-up:

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

where $w^\tau(t) = 2 \cdot f(t) \cdot S(t)/W^\tau$, $W^\tau = \int_0^\tau 2 \cdot f(t) \cdot S(t)\, dt = 1 - S^2(\tau)$. The restricted concordance summary remains a weighted average of the time-specific AUCs with the weights rescaled such that they integrate to 1.0 over the range $(0,\, \tau)$. The interpretation of C^τ is a slight modification of the original concordance, where $C^\tau = P[M_j > M_k \,|\, T_j < T_k,\, T_j < \tau\,]$. Thus C^τ is the probability that the predictions for a random pair of subjects are concordant with their outcomes, given that the smaller event time occurs in $(0,\, \tau)$.

The concordance summary C is directly related to Kendall's tau. Specifically, $C = K/2 + 1/2$, where K denotes Kendall's tau (see Agresti, 2002, p. 60 for definition). Korn and Simon (1990) and Harrell et al. (1996) discuss the use of Kendall's tau (K or τ_a) with survival data and propose modifications to account for censored observations.

2.5 *Example: Gaussian Marker and Log-Normal Disease Time*

To illustrate time-dependent accuracy concepts we consider a simple example where the marker M_i and the log of survival time $\log(T_i)$ follow a bivariate normal distribution. By convention we consider a higher marker value as indicative of earlier disease onset and, therefore, explore bivariate distributions with a negative correlation between the marker and log(time).

If $[M_i,\, \log(T_i)]$ has a bivariate normal distribution with mean $(0, 0)$ and unit standard deviations then time-dependent incident sensitivity and cumulative 1-specificity are

$$P\big(M_i > c \mid dN_i^*(t) = 1\big) = \mathrm{TP}_t^{\mathbb{I}}(c) = \Phi\left[\frac{\rho \cdot \log(t) - c}{\sqrt{(1-\rho^2)}}\right]$$

$$P\big(M_i > c \mid N_i^*(t) = 0\big) = \mathrm{FP}_t^{\mathbb{D}}(c) = \frac{S_2^{\mathcal{N}}[c, \log(t); \rho]}{\Phi[-\log(t)]},$$

where $\Phi(x) = P(X < x)$ for $X \sim \mathcal{N}(0, 1)$ and $S_2^{\mathcal{N}}[x, y; \rho] = P(X > x, Y > y)$ for $(X,\ Y)$ bivariate mean 0 unit normal with correlation ρ.

Figure 1a shows I/D ROC curves for $\rho = -0.8$. The solid line corresponds to $t = \exp(-2)$ and has an AUC of 0.923

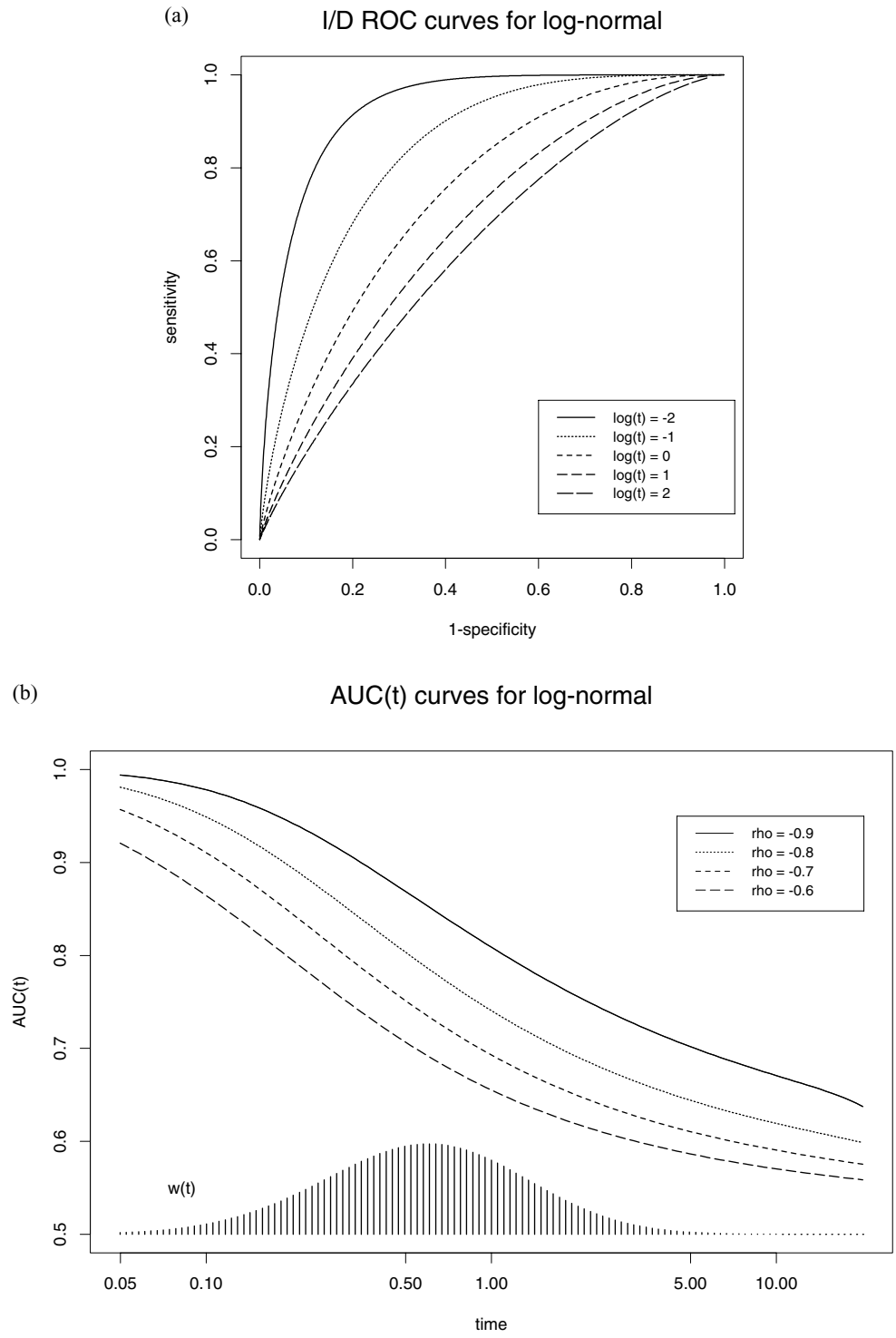

Figure 1. Incident/dynamic ROC and AUC plots for a bivariate (log) normal distribution. (a) Incident/dynamic ROC curves for a scalar marker and a disease time where $\{M_i, \log(T_i)\}$ is bivariate normal with $\rho = -0.8$. (b) Plots of AUC(t) for a scalar marker and a disease time where $\{M_i, \log(T_i)\}$ is bivariate normal with ρ taking the values $(-0.9, -0.8, -0.7, -0.6)$.

indicating very good separation between the distribution for M_i among subjects with $T_i = \exp(-2)$ as compared to the marker distribution for subjects with $T_i > \exp(-2)$. Furthermore, if the threshold value $c^{10\%} = 1.19$ were used to indicate a positive test, then by definition, only 10% of the controls (i.e., $\log(T_i) > -2$) would have a value of M_i greater than 1.19. The ROC plot shows that for this false-positive rate of 10% a sensitivity, or true-positive rate, of 75% can be obtained:

$\text{TP}_t^{\mathbb{I}}(1.19) = 0.752$. If we consider a later time such as $\log(t) = 0$ we find less overall discrimination with an AUC of 0.741. Again, specific operating points can be identified; for example, the ROC curve shows that if the false-positive rate is again controlled at 10% then a true-positive rate of only 30% is now obtained (here $c^{10\%} = 0.320$). One of the key advantages of an ROC curve is that it facilitates comparisons across different conditions in terms of the sensitivity of a marker where the specificity is controlled at a fixed level for each condition. Here we have evaluated the temporal variation in sensitivity while controlling 1-specificity at 10%.

In Figure 1b we show the AUC(t) functions for different values of ρ. For each value of ρ we find a decreasing AUC(t) with increasing time. In addition, with decreasing correlation between the marker and the disease time we find uniformly decreasing values for AUC(t). A global accuracy summary can be obtained using C, which integrates AUC(t) using the weight function proportional to $2 \cdot f(t) \cdot S(t)$. Figure 1b also displays the weight function, which for this example is $w(t) = 2 \cdot \phi(t)[\,1 - \Phi(t)]$, where $\phi(x)$ and $\Phi(x)$ are the standard normal density and distribution functions, respectively. In this bivariate normal situation there exists an analytical solution for the concordance: $C = \sin^{-1}(-\rho)/\pi + 0.5$. For $\rho = -0.9$ we find $C = 0.827$, while with $\rho = -0.6$ we find $C = 0.703$. Therefore, when the marker M_i and log-survival time have a correlation of -0.9 there is a 82.7% chance that for a random pair of observations the marker value for the earlier survival time is greater than the marker value for the larger survival time. This concordance probability is reduced to 70.3% when $\rho = -0.6$.

3. Estimation of Incident/Dynamic Time-Dependent Accuracy

In this section we propose methods for the estimation of time-dependent accuracy summaries using a single scalar marker M_i. When interest is in the accuracy of a survival regression model we propose using the linear predictor as a scalar marker, $M_i = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}$, and then using nonparametric or semiparametric methods to characterize the time-dependent sensitivity and specificity of the model score. In particular, we discuss how the Cox model and partial likelihood concepts can be conveniently used to provide semiparametric estimates of I/D accuracy. However, the methods that we propose do not require the model score, M_i, to be derived from a proportional hazards model and are potentially applicable for any prognostic scale.

3.1 *Estimation:* $\text{TP}_t^{\mathbb{I}}(c)$ *and* $\text{FP}_t^{\mathbb{D}}(c)$ *under Proportional Hazards*

Properties of the partial likelihood function make estimation of I/D ROC curves a natural companion to Cox regression. Here we assume that the censoring time C_i is independent of the failure time T_i and marker M_i. To clearly distinguish between the general model score, $M_i = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}$, and a Cox model that uses this score, we denote γ as the proportional hazards regression parameter $\lambda(t \mid M_i) = \lambda_0(t)\exp(M_i\gamma)$. It is well known that under a proportional hazards model the weights, $\pi_i(\gamma,\, t) = R_i(t) \cdot \exp(M_i\gamma)/W(t)$ introduced in Section 1.1, are used to compute an estimate of the expected value of the marker given failure: $\hat{E}(M_i \mid T_i = t) = \sum_k M_k \cdot \pi_k(\gamma, t)$. However, Xu and O'Quigley (2000) show that these weights can also be used to estimate the distribution of the covariate conditional on death at time t:

$$\widehat{\text{TP}}_t^{\mathbb{I}}(c) = \hat{P}(M_i > c \mid T_i = t) = \sum_k 1(M_k > c) \cdot \pi_k(\gamma, t), \quad (1)$$

where the estimate $\hat{P}(M_i > c \mid T_i = t)$ is a consistent estimator when the Cox model for M_i holds. Estimation of γ using partial likelihood provides a semiparametric estimate for $\text{TP}_t^{\mathbb{I}}(c)$. An empirical estimator can be used for $\text{FP}_t^{\mathbb{D}}(c)$:

$$\begin{aligned}\widehat{\text{FP}}_t^{\mathbb{D}}(c) &= \hat{P}(M_i > c \mid T_i > t) \\ &= \sum_k 1(M_k > c) \cdot R_k(t+)/W^R(t+), \end{aligned} \quad (2)$$

where $R_k(t+) = \lim_{\delta \to 0} R_k(t + |\delta|)$, and $W^R(t+) = \sum_k R_k(t+)$. The term $W^R(t+)$ denotes the size of the "control set" at time t, where we define the control set as the risk set minus subjects who fail at time t. Essentially, $\widehat{\text{FP}}_t^{\mathbb{D}}(c)$ is the empirical distribution function for marker values among the control set, and $\widehat{\text{TP}}_t^{\mathbb{I}}(c)$ is an exponential tilt of the empirical distribution function for the marker among risk set subjects (Anderson, 1979).

3.2 *Estimation:* $\text{TP}_t^{\mathbb{I}}(c)$ *and* $\text{FP}_t^{\mathbb{D}}(c)$ *under Nonproportional Hazards*

In order to use equation (1) to estimate incident sensitivity the proportional hazards assumption must be satisfied. However, this aspect can be relaxed by adopting a varying-coefficient model of the form $\lambda(t \mid M_i) = \lambda_0(t)\exp[M_i\gamma(t)]$. The time-varying coefficient function $\gamma(t)$ can be estimated either in a one-step fashion based on routine Cox model residuals, or through locally weighted partial likelihood methods. Note that if proportional hazards do obtain then $\gamma(t) \equiv 1$ when $M_i = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}$.

Grambsch and Therneau (1994) describe residual-based methods for assessing the proportional hazards model that can also be used to obtain estimates of time-varying coefficient functions. In order to define the residuals we adopt the following notation: $S^{(p)}(\boldsymbol{\beta}, t) = \sum_k R_k(t)\exp(\boldsymbol{Z}_k^{\mathrm{T}}\boldsymbol{\beta}) \cdot \boldsymbol{Z}_k^{\otimes p}$, where $\boldsymbol{Z}_k^{\otimes p}$ refers to $\mathbf{1}$, $\boldsymbol{Z}_k$, and $\boldsymbol{Z}_k\boldsymbol{Z}_k^{\mathrm{T}}$ for $p = 0$, 1, 2, respectively. The "scaled Schoenfeld residuals" are defined for each observed ordered failure time, $t_{(j)}$, as the vector

$$r_j^*(\boldsymbol{\beta}) = V^{-1}[\boldsymbol{\beta}, t_{(j)}]\{\boldsymbol{Z}_{(j)} - e[\boldsymbol{\beta}, t_{(j)}]\},$$

where $e[\boldsymbol{\beta},\, t_{(j)}] = S^{(1)}[\boldsymbol{\beta},\, t_{(j)}]/S^{(0)}[\boldsymbol{\beta},\, t_{(j)}]$, $V[\boldsymbol{\beta},\, t_{(j)}] = S^{(2)}[\boldsymbol{\beta},\, t_{(j)}]/S^{(0)}[\boldsymbol{\beta},\, t_{(j)}] - e[\boldsymbol{\beta},\, t_{(j)}]e[\boldsymbol{\beta},\, t_{(j)}]^{\mathrm{T}}$, and $\boldsymbol{Z}_{(j)}$ denotes the covariate for the subject observed to die at time $t_{(j)}$. Grambsch and Therneau (1994) show that $E\{r_j^* \mid \mathcal{F}[t_{(j)}]\} \approx [\boldsymbol{\beta}(t) - \boldsymbol{\beta}_0]$, where $\boldsymbol{\beta}_0$ is the time-averaged coefficient and $\mathcal{F}(t)$ is the right-continuous filtration specifying the survival process history. This property is used to obtain focused tests of proportionality, and to obtain estimates of the time-varying coefficient function, $\beta_k(t)$ corresponding to covariate $Z_{i,k}$. As a graphical diagnostic tool standard regression-smoothing techniques are now commonly applied to the points $[t_{(j)}, \hat{\beta}_k + r_{j,k}^*(\hat{\boldsymbol{\beta}})]$ following a Cox model fit in order to obtain estimates of time-dependent coefficient functions, $\beta_k(t)$.

For the evaluation of the accuracy of a marker, M_i, the smoothing of Schoenfeld residuals can be used to obtain a simple estimate of I/D AUC(t) by exploiting standard Cox model output. First a Cox model of the form $\lambda_0(t)\exp(M_i\gamma)$ is fit, followed by use of regression-smoothing methods to obtain $\hat{\gamma}(t)$. Second, equation (2) can still be used to obtain estimates of false-positive rates, and (1) can now be evaluated using $\gamma(t)$ rather than a constant value γ:

$$\widehat{\mathrm{TP}}_t^{\mathbb{I}}(c) = \hat{P}(M_i > c \mid T_i = t) = \sum_k 1(M_k > c)\cdot\pi_k[\hat{\gamma}(t), t]. \quad (3)$$

By using equation (3) we are adopting the flexible semiparametric hazard model, $\lambda_0(t)\exp[M_i\gamma(t)]$, which no longer assumes proportionality, but rather only assumes smoothly varying hazard ratios over time.

More formal flexible semiparametric statistical methods can be used to estimate a varying-coefficient hazard model and subsequently produce time-dependent accuracy summaries based on minimal model assumptions. For example, Hastie and Tibshirani (1993) discuss both smooth parametric methods and nonparametric penalized likelihood methods for estimating the function $\gamma(t)$ in the model $\lambda_i(t) = \lambda_0(t)\exp[M_i\gamma(t)]$. More recently Cai and Sun (2003) characterize the properties of locally weighted partial likelihood methods used to obtain varying coefficient estimates. Using kernel weights that are specified as a function of time, t, allows use of local-linear estimation methods. Cai and Sun (2003) prove the pointwise consistency and asymptotic normality of the resulting function estimator, $\hat{\gamma}(t)$. Smooth parametric and/or nonparametric methods allow valid estimation of accuracy summaries such as AUC(t) based on the minimal model assumptions because models of the form $\lambda_i(t) = \lambda_0(t)\exp[M_i\gamma(t)]$ only assume linearity in M_i and smoothly varying hazard ratios over time. The linearity assumption can be relaxed by using a model with single or multiple transformations of M_i and a vector of time-varying coefficients.

3.3 *Estimation:* $\mathrm{ROC}_t^{\mathbb{I}/\mathbb{D}}(p), \mathrm{AUC}(t)$, *and* C^τ

Given estimates of $\mathrm{TP}_t^{\mathbb{I}}(c)$ and $\mathrm{FP}_t^{\mathbb{D}}(c)$ the area under the ROC curve at time t, AUC(t), and the integrated area, C^τ, can be calculated. The estimated ROC curve is given as

$$\widehat{\mathrm{ROC}}_t^{\mathbb{I}/\mathbb{D}}(p) = \widehat{\mathrm{TP}}_t^{\mathbb{I}}\left\{\left[\widehat{\mathrm{FP}}_t^{\mathbb{D}}\right]^{-1}(p)\right\},$$

where $[\widehat{\mathrm{FP}}_t^{\mathbb{D}}]^{-1}(p) = \inf_c\{c : \widehat{\mathrm{FP}}_t(c) \le p\}$. The estimated AUC($t$) is simply $\widehat{\mathrm{AUC}}(t) = \int \widehat{\mathrm{ROC}}_t^{\mathbb{I}/\mathbb{D}}(p)\,dp$ estimated using standard numerical integration methods such as the trapezoid rule. Finally, the estimated concordance is given by

$$\hat{C}^\tau = \int \widehat{\mathrm{AUC}}(t)\cdot\hat{w}^\tau(t)\,dt,$$

where $\widehat{\mathrm{AUC}}(t)$ is given above and $\hat{w}^\tau(t) = 2\cdot\hat{f}(t)\cdot\hat{S}(t)/[1-\hat{S}^2(\tau)]$. The Kaplan–Meier estimator can be used for $\hat{S}(t)$, and a discrete approximation to $\hat{f}(t)$ can be used based on the increments in the Kaplan–Meier estimator. If Kaplan–Meier is used to estimate $f(t)$ and $S(t)$ then $\widehat{\mathrm{AUC}}(t)$ only needs to be evaluated at the observed failure times in order to calculate $\hat{C}^\tau$.

3.4 *Inference for Incident/Dynamic Accuracy Summaries*

Xu and O'Quigley (2000) show that the estimator $\widehat{\mathrm{TP}}_t^{\mathbb{I}}(c)$ given in equation (1) is consistent provided that the proportional hazards model obtains, and provided the independent observations are subject to independent censoring. Parallel arguments apply for the estimator obtained using a varying-coefficient model given in equation (3) whenever a consistent estimator of $\gamma(t)$ is used. Cai and Sun (2003) show that the locally weighted MPLE is consistent under standard regularity conditions. In addition, because $\widehat{\mathrm{FP}}_t^{\mathbb{D}}(c)$ is an empirical distribution function calculated over the control set (i.e., the risk set minus the case), consistency obtains provided the control set represents an unbiased sample (i.e., independent censoring). Therefore, consistent estimates of time-dependent sensitivity and specificity and corresponding AUC(t) and C^τ summaries are obtained under the proportional hazards assumption using equations (1) and (2), and under more general nonproportional hazards assumptions using equation (3). Finally, because the accuracy summaries are defined over the joint distribution of the marker M_i and the survival time T_i, the nonparametric bootstrap of Efron (1979) based on resampling of observations (M_i, X_i, Δ_i) may be used to compute standard errors or to provide confidence intervals.

3.5 *Discrete Times and General Hazard Models*

Our motivation for developing tools to summarize predictive accuracy stems from interest in characterizing the prognostic potential of Cox models for continuous survival times. However, the basic time-dependent accuracy concepts and the estimation method outlined in Section 3.2 generalizes to discrete survival times and/or alternative hazard regression models.

The key to estimation of $\mathrm{TP}_t^{\mathbb{I}}(c)$ presented in Sections 3.1 and 3.2 is that a hazard model can be used to reweight the empirical distribution of M_i calculated over the risk set at time t. Equations (1) and (3) show specific details for Cox models. More generally, let $P(T_i = t \mid T_i \ge t,\, M_i)$ denote the hazard, where $P(t)$ represents either density for continuous survival times or probability for discrete times. A hazard regression model can be formulated as $g[P(T_i = t \mid T_i \ge t,\, M_i)] = \alpha(t) + M_i\beta(t)$, where $g(x)$ is a link function. The Cox model is a special case where a log link is used; $\alpha(t) = \log\lambda_0(t)$; and $\beta(t) \equiv \beta$ under the proportional hazards assumption. Following arguments given in Xu and O'Quigley (2000) the general model implies:

$$P(M_i = m \mid T_i = t) \propto g^{-1}[\alpha(t) + m\cdot\beta(t)] \times P(M_i = m \mid T_i \ge t), \quad (4)$$

where $P(M_i = m \mid T_i \ge t)$ denotes either the marker density or probability depending on whether a continuous or discrete marker distribution is assumed. See the Appendix for a derivation. Equation (4) shows that $P(M_i = m \mid T_i = t)$ can be estimated from separate estimates of the hazard model and the distribution of the marker conditional on $T_i \ge t$. Therefore, the general estimation approach outlined in Section 3.2 can be adopted for either discrete survival times or for general hazard regression models provided that consistent estimates of $[\alpha(t),\, \beta(t)]$ and $P(M_i = m \mid T_i \ge t)$ are available. Tied survival times impact choice of a method for estimating the hazard model parameters. In addition, with

discrete survival times calculation of the concordance summary $C = \int \mathrm{AUC}(t) \cdot w(t)\, dt$ requires modification to account for the fact that $P(T_j = T_k) \neq 0$ and, therefore, the constant 2 in the weight $w(t) = 2 \cdot f(t) \cdot S(t)$ needs to be computed as $1/P(T_j < T_k)$. Finally, Cox models are convenient because the baseline hazard, $\alpha(t) = \log \lambda_0(t)$, drops out of (4), and is thus not required for estimation of $\mathrm{TP}^{\mathbb{I}}_t(c)$.

3.6 *Simulations to Evaluate Incident/Dynamic Estimation*

In order to demonstrate the feasibility of using Cox regression methods and the marker distribution among risk sets for estimating I/D ROC curves and global concordance we conducted a set of simulation studies.

For each of $m = 500$ simulated data sets a sample of $n = 200$ marker values, M_i, and survival times, T_i, were generated such that $(M_i, \log T_i)$ is bivariate normal with a correlation of $\rho = -0.7$. An independent log-normal censoring time was generated to yield a fixed expected fraction of censored observations (either 20% or 40% censored). For each simulated data set we estimated the I/D AUC(t) function and the concordance summary C^τ using the largest observed survival time to truncate follow-up time. We applied four methods of estimation to the censored data: maximum likelihood assuming a bivariate normal distribution for the survival time and the marker; maximum partial likelihood using the Cox model, which for this example incorrectly assumes proportional hazards; locally weighted maximum partial likelihood (MPL) estimation for the model $\lambda_0(t) \exp[M_i \gamma(t)]$ using the method of Cai and Sun (2003); and simple local linear smoothing of the scaled Schoenfeld residuals. For local MPL estimation and local linear smoothing we used an Epanechnikov kernel with a span of $n^{-1/5}$ where n is the number of observations.

In order to estimate AUC(t) and C^τ using semiparametric methods the model for the survival time conditional on the marker, $\lambda_0(t) \exp[M_i \gamma(t)]$, is combined with the observed marker distribution within each risk set according to the methods described in Section 3.2. We have adopted a survival model that assumes that the log hazard increases linearly in M_i for each time t. The true data-generating model is actually nonlinear with a concave risk function. Therefore, for this simulation our estimation used a first-order approximation to the true conditional hazard surface.

Table 1 displays the mean and standard deviation for the estimate of AUC(t) at various values of t when data are generated with 20% and with 40% censoring. When 20% of the observations are censored we find that the MLE for AUC(t) has minimal bias for $\log(t)$ between -2 and 2. Estimates based on the locally weighted MPLE and the residual smoother yield approximately unbiased estimates for all but the most extreme values of time with some negative bias observed for both the semiparametric estimators. For example, at $\log(t) = -2$ the mean $\widehat{\mathrm{AUC}}(t)$ using the locally weighted MPLE is 0.860 (relative bias of $1 - 0.860/0.884 = -3\%$) and using the residual smoother the average is 0.881 (relative bias of

Table 1

Simulation results for estimation of I/D accuracy. Data $(M_i, \log T_i)$ were generated as bivariate normal with a correlation of $\rho = -0.7$. The sample size for each simulated data set was $N = 200$. The AUC(t) *curve and the integrated curve, C^τ, were estimated using: maximum likelihood assuming a bivariate normal model; Cox model, which assumes proportional hazards; local maximum partial likelihood for the varying-coefficient model $\lambda(t) = \lambda_0(t) \exp[\gamma(t) M_i]$; and a local linear smooth of the scaled Schoenfeld residuals to estimate the varying-coefficient model.*

		MLE		Cox model		Local MPLE		Residual smooth	
Log time	AUC(t)	Mean	SD	Mean	SD	Mean	SD	Mean	SD
				20% censoring					
-2.0	0.884	0.884	0.018	0.743	0.028	0.860	0.052	0.881	0.044
-1.5	0.833	0.834	0.019	0.734	0.026	0.817	0.033	0.829	0.035
-1.0	0.782	0.782	0.019	0.725	0.024	0.768	0.031	0.771	0.033
-0.5	0.734	0.734	0.019	0.716	0.023	0.722	0.032	0.720	0.033
0.0	0.693	0.693	0.018	0.707	0.021	0.688	0.034	0.686	0.034
0.5	0.660	0.660	0.016	0.700	0.023	0.655	0.041	0.657	0.040
1.0	0.634	0.634	0.015	0.691	0.028	0.633	0.044	0.637	0.041
1.5	0.614	0.614	0.013	0.670	0.044	0.621	0.064	0.622	0.048
2.0	0.598	0.598	0.012	0.600	0.075	0.579	0.076	0.573	0.060
C^τ	0.741	0.741	0.016	0.720	0.020	0.737	0.018	0.740	0.018
				40% censoring					
-2.0	0.884	0.884	0.019	0.749	0.031	0.859	0.054	0.875	0.048
-1.5	0.833	0.834	0.021	0.742	0.029	0.818	0.035	0.827	0.037
-1.0	0.782	0.782	0.021	0.732	0.026	0.770	0.035	0.772	0.035
-0.5	0.734	0.734	0.020	0.722	0.024	0.724	0.038	0.722	0.039
0.0	0.693	0.693	0.019	0.712	0.024	0.689	0.042	0.687	0.041
0.5	0.660	0.660	0.018	0.702	0.026	0.654	0.045	0.655	0.043
1.0	0.634	0.635	0.016	0.689	0.035	0.633	0.057	0.637	0.048
1.5	0.614	0.614	0.015	0.653	0.055	0.617	0.075	0.614	0.051
2.0	0.598	0.599	0.013	0.560	0.073	0.555	0.075	0.546	0.058
C^τ	0.741	0.741	0.017	0.727	0.022	0.740	0.021	0.742	0.021

$1 - 0.881/0.884 < -1\%$), while at $\log(t) = 2$ the locally weighted MPLE mean estimate is 0.579 (relative bias $= 1 - 0.579/0.598 = -3\%$) and for the residual smoother the mean is 0.573 (relative bias $= 1 - 0.573/0.598 = -4\%$). As expected for local regression methods Table 1 shows that the nonparametric methods yield substantially greater variances for specific values of t compared to the MLE.

Incorrectly assuming proportional hazards lead to biased estimates. Table 1 shows that the estimated AUC(t) obtained using equation (1) with an estimated Cox model coefficient is negatively biased for $\log(t) < 0$. For example, at $\log(t) = -2$ we obtain a negative bias of $1 - 0.743/0.884 = -16\%$. For $\log(t) > 0$ the estimates obtained using the Cox model and equation (1) are positively biased indicating that direct use of the proportional hazards assumption produces an estimated AUC(t) curve that is flatter than the target with early underestimation and late overestimation.

When censoring is increased to 40% similar patterns are found for all estimators. Table 1 shows that the bias in $\widehat{\text{AUC}}(t)$ is slightly larger with increased censoring. For example, at $\log(t) = 2$ the mean estimate for the locally weighted MPLE is 0.555 (relative bias of $1 - 0.555/0.598 = -7\%$) and for the residual smoother it is 0.546 (relative bias of $1 - 0.546/0.598 = -9\%$). Therefore, even with 40% censoring the smooth semiparametric methods appear to perform adequately.

Finally, Table 1 also shows the results for the estimation of the global concordance summary C^τ. In the simulations we estimate C using the analytical results for the MLE: $\hat{C} = \sin^{-1}(-\hat{\rho})/2 + 1/2$. For the methods that adopt a varying-coefficient hazard model we set τ equal to the largest uncensored survival time in the observed data and, therefore, truncate follow-up at slightly different times for each simulated data set. However, even with 40% censoring the largest uncensored time had a median value of exp(2.30) with an interquartile range of exp(2.04) to exp(2.65), and thus typically very little mass in the survival distribution is lost because $S[\exp(2.30)] = 1 - \Phi(2.30) = 0.01$. With 20% censoring the mean estimate for the MLE, locally weighted MPLE, and residual smoother are 0.741 (SD = 0.016), 0.737 (SD = 0.018), and 0.740 (SD = 0.018), respectively. In contrast the estimate obtained naively assuming proportional hazards is negatively biased with an average estimate of 0.720 (relative bias $= 1 - 0.720/0.741 = -3\%$). These results suggest that the smooth semiparametric methods yield little bias, and for this example exhibit high efficiency relative to the MLE. A similar pattern is seen with 40% censoring where slightly increased standard deviations are observed relative to results obtained with 20% censoring.

4. Examples

In this section we illustrate the proposed methods using two well-studied data sets.

4.1 *VA Lung Cancer Data*

Kalbfleisch and Prentice (2002) present and analyze Veteran's Administration (VA) lung cancer data from a clinical trial in which males with inoperable cancer were randomized to a standard treatment or a test therapy. Baseline covariates that were considered important predictors of mortality include: patient age, histological type of tumor, and a performance status measure known as the Karnofsky score. Schemper and Henderson (2000) use these covariates plus a treatment indicator and report an R^2 of $\hat{V} = 0.24$. This would suggest that the covariates explain only 24% of the time-integrated variance in survival status.

Table 2
Cox regression estimates for the VA lung cancer data where follow-up is truncated at 500 *days. The reference category for cell type is squamous.*

Covariate	Estimate	SE	Z
Treatment	−0.323	0.206	−1.566
Age/10	−0.086	0.093	−0.937
Karnofsky score	−0.032	0.005	−5.931
Cell type (small)	0.841	0.270	3.116
Cell type (adeno)	1.151	0.295	3.896
Cell type (large)	0.350	0.285	1.231

For comparison we use the same covariates and Cox regression to create estimates of $\text{ROC}_t^{\mathbb{I}/\mathbb{D}}(p)$ for select t, the AUC(t) function, and the concordance summary C^τ. For our analysis we terminate follow-up at 500 days. Estimated model coefficients and standard errors are given in Table 2. Using the proportional hazards assumption we can employ equations (1) and (2) to estimate time-specific I/D ROC curves, and then integrate the ROC curve to obtain $\widehat{\text{AUC}}(t)$. Estimates of AUC(t) and pointwise 90% confidence intervals are displayed in Figure 2a. Over the first 60 days of follow-up the AUC(t) ranges between 0.66 and 0.73. The substantive interpretation is: on any day, t, between 0 and 60, the probability that a subject who dies on day t having a model score greater than a subject who survives beyond day t is at least 0.66. The accuracy summaries suggest good short-term discriminatory potential of the model score. The estimated AUC(t) function tends to decline over time to approximately 0.65 for $100 < t < 300$. Estimates of AUC(t) also become increasingly variable over time due to the diminishing size of the risk set. Using a follow-up of $\tau = 365$ days yields a concordance estimate of $\int_0^\tau \widehat{\text{AUC}}(t) \cdot \hat{w}^\tau(t)/dt = 0.713$ with a standard error of 0.026. This implies that conditional on one event occurring within the first year, the probability that the model score is larger for the subject with the smaller event time is 71.3%. The concordance estimate $\hat{C}^\tau$ is relatively modest in magnitude, but is significantly different from the null value of 0.50 (95% CI for C^τ: 0.661, 0.765).

To characterize the model score, $M_i = \boldsymbol{Z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, using fewer assumptions we relax the proportional hazards assumption for M_i by using a varying coefficient model: $\lambda_0(t)\exp[M_i\gamma(t)]$. Note that we are still focusing on use of the Cox model with a proportional hazards assumption to generate the model score, but are relaxing the assumptions needed to characterize model accuracy. This highlights the fact that different methods can be used for generating and evaluating a survival regression model score. For the VA lung cancer data we simply use a kernel smooth of the scaled Schoenfeld residuals to estimate $\gamma(t)$. The estimate of $\gamma(t)$ suggests a decreasing log-relative hazard with increasing time (not shown).

Figure 2b shows estimates of AUC(t) based on equations (2) and (3), which relax the proportional hazards assumption.

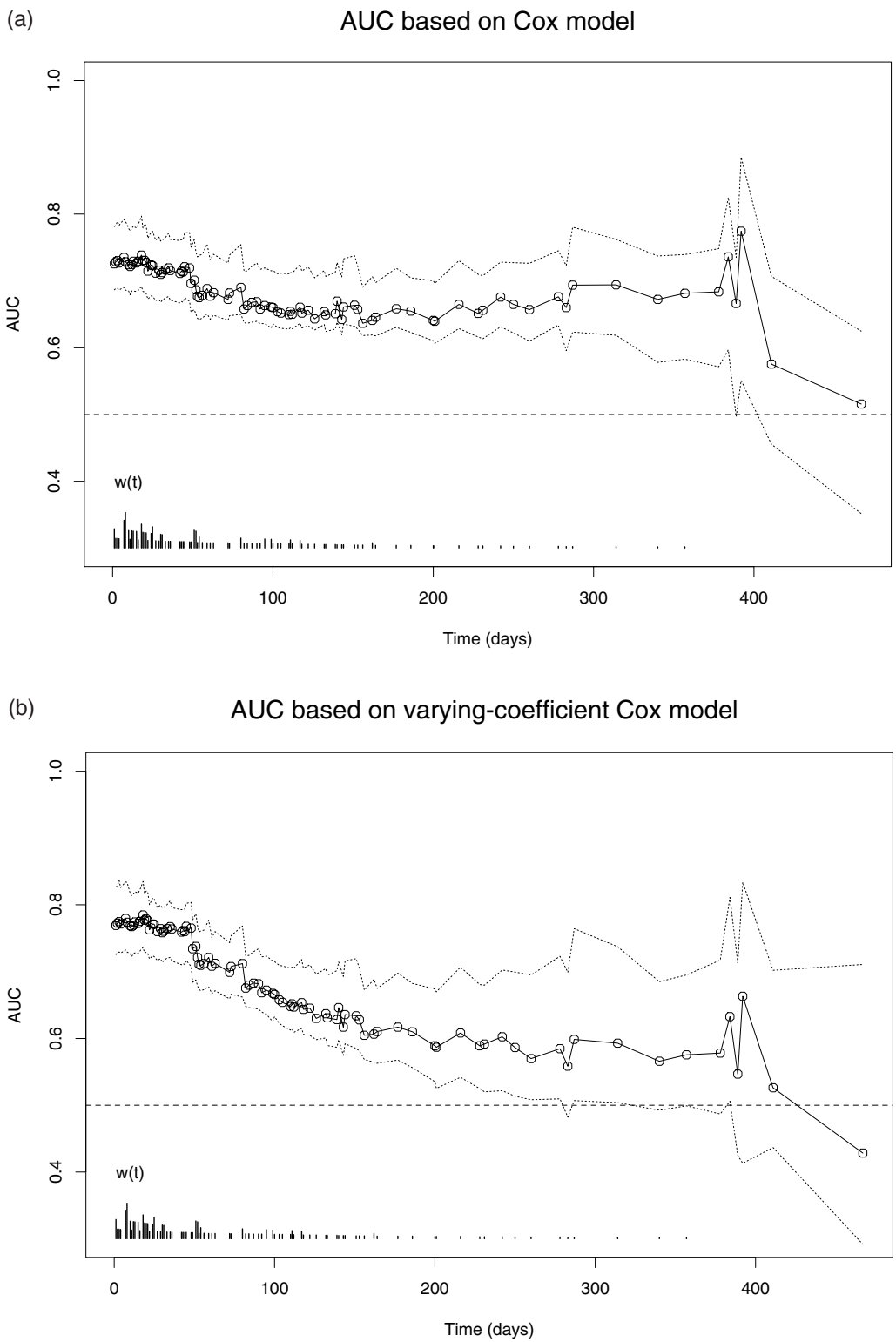

Figure 2. Incident/dynamic AUC plots for the VA lung cancer data. (a) Accuracy of the model score (linear predictor) under the assumption of proportional hazards. Estimates of I/D AUC(t) versus time with pointwise 90% confidence intervals. Using $\tau = 365$ we obtain $\hat{C}^\tau = \int_0^\tau \widehat{\mathrm{AUC}}(t) \cdot \hat{w}^\tau(t)\, dt = 0.713$ (SE $= 0.026$). (b) Accuracy of the model score (linear predictor) based on a varying-coefficient multiplicative hazard model. Estimates of I/D AUC(t) versus time with pointwise 90% confidence intervals. Using $\tau = 365$ we obtain $\hat{C}^\tau = \int_0^\tau \widehat{\mathrm{AUC}}(t) \cdot \hat{w}^\tau(t)\, dt = 0.738$ (SE $= 0.022$).

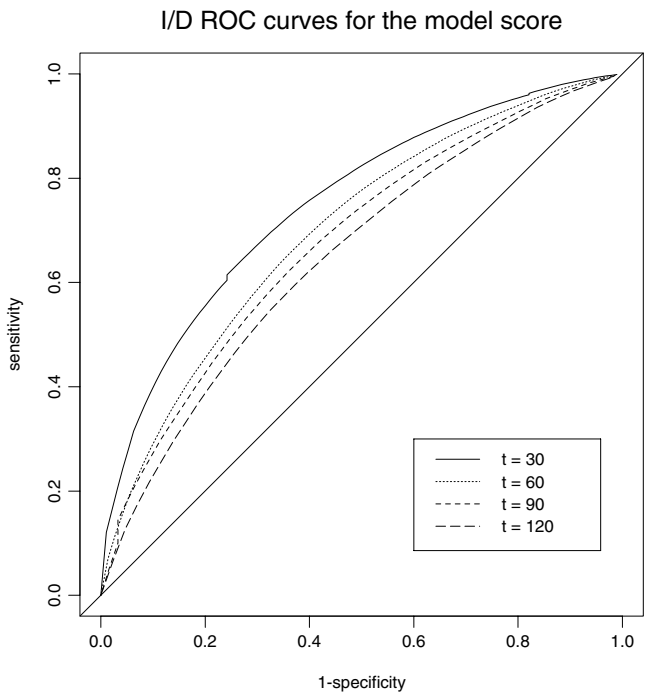

Figure 3. Incident/dynamic ROC curves for the VA lung cancer data. A model score is derived using Cox regression with Karnofsky score, age, and cell type. ROC curves are estimated using a varying-coefficient Cox model with the derived model score as the single predictor.

First, notice that the short-term accuracy of the model score remains good with $\widehat{\mathrm{AUC}}(t)$ between 0.70 and 0.78 over the first 60 days of follow-up. Second, the discriminatory ability of the model score declines substantially over time, and estimates of AUC(t) approach 0.50 after approximately 300 days, suggesting that the model score is essentially useless at discriminating incident cases from controls after 300 days. The 1-year concordance is estimated as $\hat{C}^{\tau} = 0.738$, a slight increase from the estimate obtained assuming proportional hazards. In this example the AUC(t) curve is particularly useful for displaying the fact that the baseline model score is good at discriminating early cases from early controls, but is of decreasing prognostic utility with increasing temporal distance from the baseline measurement. Declining prognostic value is not surprising, particularly because the Karnofsky score is actually a time-varying health status measure, but only the baseline value is available for the regression model. Figure 3 shows select estimates of I/D ROC curves based on the varying-coefficient model. Similar to the plot of AUC(t) the ROC curves show that predictive accuracy is uniformly decreasing with increasing time since baseline. For example, controlling the dynamic false-positive rate at 20% leads to an incident sensitivity of 56% at 30 days, decreasing to 45%, 42%, and 38% for 60, 90, and 120 days. The ROC curves also show details regarding the trade-off between sensitivity and specificity. If a stricter false-positive rate of 10% was desired then the corresponding sensitivity would only be 40% at 30 days and less than 30% for follow-up times of 60 days or greater.

Table 3
Cox regression estimates for the PBC data

Covariate	Estimate	SE	Z
	Model 1		
Log(bilirubin)	0.877	0.099	8.866
Log(prothrombin time)	3.013	1.025	2.939
Edema	0.785	0.300	2.617
Albumin	−0.944	0.237	−3.985
Age	0.033	0.009	3.881
	Model 2		
Log(prothrombin time)	4.141	0.870	4.758
Edema	1.190	0.295	4.031
Albumin	−1.314	0.223	−5.897
Age	0.024	0.009	2.660

4.2 *Mayo PBC Data*

Next, we consider data from a randomized placebo-controlled trial of the drug D-penicillamine (DPCA) for the treatment of primary biliary cirrhosis (PBC) conducted at the Mayo Clinic between 1974 and 1984 (Fleming and Harrington, 1991). Among the 312 subjects randomized to the study, 125 died by the end of the follow-up. Although the study established that DPCA is not effective for the treatment of PBC, the data have been used to develop a commonly used clinical prediction model. We use this example to illustrate how ROC curves and/or AUC(t) summaries can be used to compare different model scores.

We first consider a Cox model containing five covariates: log(bilirubin), albumin, log(prothrombin time), edema, and age. Table 3 gives the regression estimates using the proportional hazard, model with mortality as the response. Except for log(prothrombin time), all covariates are strong predictors of survival. The model has been used to create a widely used prognostic score. We now address the basic question: How well does the model score discriminate subjects who are likely to die from subjects who are likely to survive? In addition, we consider whether the accuracy of the score changes over time. Using the fitted linear predictor from the Cox model, we construct I/D time-dependent ROC curves and associated summaries for the "Mayo model." Figure 4a plots AUC(t) evaluated at each failure time. The model score has very good discriminatory capacity for distinguishing those patients who die at time t from those who live beyond time t. The accuracy is especially good for follow-up times less than 1000 days, with early AUC(t) estimates exceeding 0.85. The accuracy of the model score gradually decreases with time. Based on $\widehat{\mathrm{AUC}}(t)$ and the Kaplan–Meier estimator of the marginal survival distribution we estimate a concordance summary, C^{τ}, of 0.80, with τ fixed at 4000 days for this and subsequent analysis.

To quantify the impact of a single covariate on the accuracy of prediction we fit a second Cox regression model that does not include the covariate log(bilirubin). Table 3 displays coefficient estimates for this new four-covariate model. The estimate of C^{τ} drops from 0.80 to 0.73 when log(bilirubin) is excluded from the model. In addition, we can use the estimated AUC(t) curves shown in Figure 4a to quantify for each follow-up time t the additional predictive accuracy that

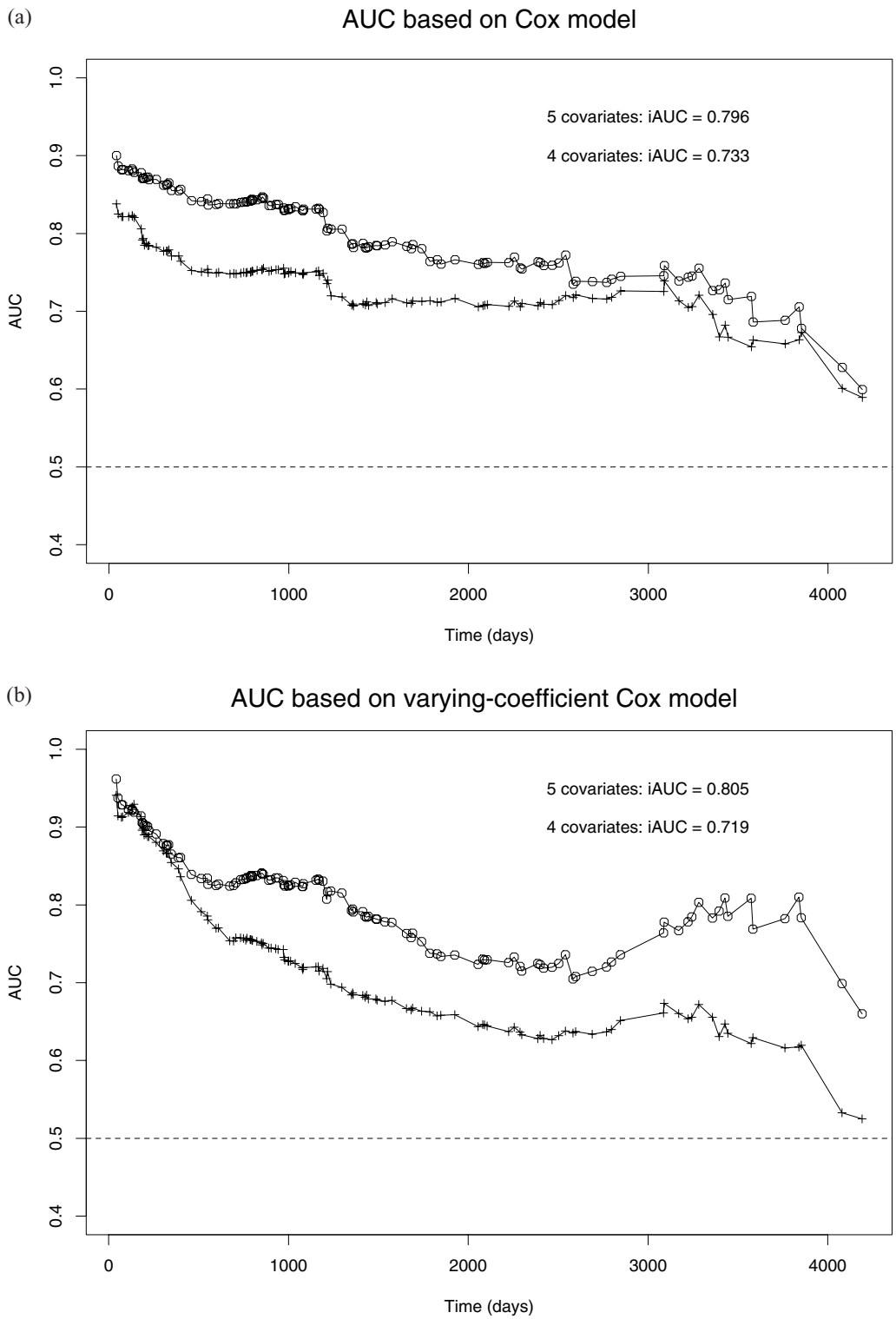

Figure 4. Incident/dynamic AUC plots for the Mayo PBC data. (a) Accuracy of the model score using five covariates (○) log(bilirubin), log(prothrombin), edema, albumin, and age, and the model score using four covariates (+), where log(bilirubin) is excluded. Lines plot the estimates of I/D AUC(t) versus time under the assumption of proportional hazards. (b) Accuracy of the model score using five covariates (○) log(bilirubin), log(prothrombin), edema, albumin, and age, and the model score using four covariates (+), where log(bilirubin) is excluded. Estimation is based on a varying-coefficient multiplicative hazard model. Lines plot the estimates of I/D AUC(t) versus time.

is obtained by using bilirubin in addition to the other model covariates. Relative to the five-covariate model the estimated AUC(t) for the four-covariate model is approximately 0.10 units below the five-covariate model $\widehat{\text{AUC}}(t)$ for t between 0 and 2000 days.

We then relax the proportional hazard assumption and use the time-varying coefficient models as described in Section 3.2 to characterize the accuracy of the model score $M_i = \boldsymbol{Z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. The bottom panel of Figure 4 displays the AUC function based on the estimated time-varying coefficient obtained using locally weighted MPL. Early estimates of AUC(t) now exceed 0.90 and decline sharply to approximately 0.75 at 2000 days for the five-covariate model and to less than 0.65 at 2000 days for the four-covariate model. Using the estimated AUC(t) reveals that the "Mayo model" is excellent at short-term prediction but that the predictive accuracy declines to $\widehat{\text{AUC}}(t) < 0.80$ by 1 year for the model without bilirubin, and to $\widehat{\text{AUC}}(t) < 0.80$ by 5 years for the five-covariate model. Finally, using the time-varying coefficient produces a global concordance summary of 0.80 for the five-covariate model and 0.72 for the model that excludes bilirubin.

5. Discussion

This article introduces a new version of time-dependent sensitivity, specificity, and associated ROC curves that are useful for characterizing the predictive accuracy of a scalar marker, such as a derived model score, when the outcome is a censored survival time. We show that the area under the time-specific ROC curves can be plotted as a function of time to characterize temporal changes in accuracy, and can be integrated using the marginal distribution of the failure time to provide a global concordance summary. Incident sensitivity and dynamic specificity are shown to be easily estimated using a fitted hazard model and the empirical distribution of the marker data within risk sets. Using only a routine Cox model output allows estimates of accuracy that assume proportional hazards and simple regression smoothing of scaled Schoenfeld residuals provides accuracy summaries appropriate for markers that do not satisfy proportional hazards. Simulations suggest that residual smoothing and locally weighted partial likelihood estimators both provide feasible and accurate estimates.

Our methods explicitly decouple the generation of a predictive score from the evaluation of prognostic accuracy. An investigator may use Cox regression to create a model score $M_i = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}$ that is a time-invariant linear combination of baseline covariates $\boldsymbol{Z}_i$. However, using the flexible methods proposed in Section 3.2 to evaluate the prognostic potential of M_i does not require commitment to the proportional hazards assumption. A practical advantage of using $M_i = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}$ is that a single "scoring" of the baseline covariates is conducted to generate M_i, but if proportional hazards is clearly violated then a more general model such as $\lambda_0(t)\exp[\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}(t)]$ may be appropriate, and would lead to a time-varying score $M_i(t) = \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\beta}(t)$.

A number of aspects warrant additional research. First, estimation methods proposed in Sections 3.1 and 3.2 assume that the censoring time is independent of the survival time. Relaxation to allow conditional independence given the marker, M_i, or covariates, $\boldsymbol{Z}_i$, would be useful. Second, we have proposed estimators that assume a prospective study design. Extension to case–cohort data may be important for characterizing the accuracy of markers for rare diseases. Third, development of analytical approximations that characterize the large sample distribution of the proposed estimators would facilitate approximate inference for time-dependent ROC curves, the AUC(t) curve, or the concordance summary C^{τ}. Finally, exploration of time-dependent accuracy methods with a longitudinal marker, $M_i(t)$, would be important for the common prospective medical setting in which predictive covariate information is updated over time.

Résumé

L'adéquation d'un modèle de survie peut être résumée grâce à des extensions du pourcentage de variabilité expliquée par le modèle, ou R2, utilisé habituellement pour les modèles expliquant une réponse continue, ou grâce à des extensions de la sensibilité et spécificité, utilisées habituellement pour prédire une réponse binaire. Dans cet article nous proposons une version dépendant du temps de l'adéquation, en utilisant des fonctions du temps de la sensibilité et la spécificité calculées sur les groupes à risque. Nous relions les résumés de l'adéquation à une mesure globale de la concordance, proposée auparavant, qui est une extension du tau de Kendall. De plus, nous montrons comment utiliser les résultats obtenus par un modèle de Cox afin d'obtenir les estimations de la sensibilité et la spécificité dépendant du temps ainsi que des courbes ROC (Receiver Operating Characteristic) dépendant du temps. Des méthodes d'estimation semi-paramétrique adaptées à la fois aux modèles à hasards proportionnels et non proportionnels sont présentées, évaluées par des simulations et illustrées par deux jeux de données de survie.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. New York: John Wiley & Sons.

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics* **22,** 1299–1327.

Anderson, J. A. (1979). Multivariate logistic compounds. *Biometrika* **66,** 17–26.

Cai, T., Pepe, M. S., Lumley, T., Zheng, Y., and Jenny, N. S. (2003). The sensitivity and specificity of markers for event times. *University of Washington Technical Report* **188,** 1–30.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics* **30,** 93–111.

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7,** 1–26.

Etzioni, R., Pepe, M., Longton, G., Hu, C., and Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making* **19,** 242–251.

Fan, V., Au, D., Heagerty, P., Deyo, R., McDonell, M., and Fihn, S. (2002). Validation of case-mix measures derived

from self-reports of diagnoses and health. *Journal of Clinical Epidemiology* **55,** 371–380.

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

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals (Corr: 1995, **82,** 668). *Biometrika* **81,** 515–526.

Hanley, J. A. and McNeil, B. (1982). The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36.

Harrell, F. E., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15,** 361–387.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55,** 757–796.

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.

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

Korn, E. L. and Simon, R. (1990). Measures of explained variation for survival data. *Statistics in Medicine* **9,** 487–503.

O'Quigley, J. and Xu, R. (2001). Explained variation in proportional hazards regression. In *Handbook of Statistics in Clinical Oncology*, J. Crowley (ed), 397–409. New York: Marcel Dekker.

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

Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression. *Biometrics* **56,** 249–255.

Slate, E. H. and Turnbull, B. W. (2000). Statistical models for longitudinal biomarkers of disease onset. *Statistics in Medicine* **19,** 617–637.

Xu, R. and O'Quigley, J. (2000). Proportional hazards estimate of the conditional survival function. *Journal of the Royal Statistical Society, Series B, Methodological* **62,** 667–680.

Zheng, Y. and Heagerty, P. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* **5,** 615–632.

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

Received August 2003. *Revised March* 2004.
Accepted March 2004.

Appendix

Concordance as Function of AUC(t)

Assume independent observations (M_j, T_j) and (M_k, T_k), and assume that T_j is continuous such that $P(T_k = T_j) = 0$. Let $P(x)$ denote probability or density depending on the context:

$$P[T_j < T_k] = \tfrac{1}{2} \text{ (by independence)}$$

$$\begin{aligned}
&P[M_j > M_k \mid T_j < T_k] \\
&\quad = P[\{M_j > M_k\} \cap \{T_j < T_k\}] \times 2 \\
&\quad = \int_t P[\{M_j > M_k\} \cap \{T_j = t\} \cap \{t < T_k\}] \times 2\, dt \\
&\quad = \int_t P[\{M_j > M_k\} \mid \{T_j = t\} \cap \{t < T_k\}] \times 2 \\
&\qquad \times P[\{T_j = t\} \cap \{t < T_k\}]\, dt \\
&\quad = \int_t \mathrm{AUC}(t) \times 2 \times P[T_j = t] \times P[t < T_k]\, dt \\
&\quad = \int_t \mathrm{AUC}(t) \cdot w(t)\, dt = E_T[\mathrm{AUC}(T) \times 2 \times S(T)], \\
&\qquad \text{with} \quad w(t) = 2 \cdot f(t) \cdot S(t).
\end{aligned}$$

Hazard as Bridge from $P(M_i = m \mid T_i \geq t)$ *to* $P(M_i = m \mid T_i = t)$

Let $P(x)$ denote probability or density depending on the context and specific assumptions. For either continuous or discrete survival times the conditional hazard can be defined as

$$\lambda(t \mid M_i = m) = P(T_i = t \mid M_i = m)/P(T_i \geq t \mid M_i = m).$$

Let $P(m)$ denote the marginal density or distribution of the marker M. Following Xu and O'Quigley (2000) we obtain the following general relationship:

$$\begin{aligned}
&P(M_i = m \mid T_i = t) \\
&\quad = P(T_i = t \mid M_i = m) \cdot P(M_i = m)/P(T_i = t) \\
&\quad = \lambda(t \mid M_i = m) \cdot P(T_i \geq t \mid M_i = m) \\
&\qquad \cdot P(M_i = m)/P(T_i = t) \\
&\quad = \lambda(t \mid M_i = m) \cdot P(M_i = m \mid T_i \geq t) \cdot P(T_i \geq t)/P(T_i = t)
\end{aligned}$$

$$P(M_i = m \mid T_i = t) \propto \lambda(t \mid M_i = m) \cdot P(M_i = m \mid T_i \geq t).$$