Methods for describing the time-varying predictive performance of survival models

Chao-Kang Jason Liang

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Patrick Heagerty, Chair

Michael LeBlanc

Kenneth Rice

Program Authorized to Offer Degree:
School of Public Health, Department of Biostatistics

University of Washington

Abstract

Methods for describing the time-varying predictive performance of survival models

Chao-Kang Jason Liang

Chair of the Supervisory Committee:
Professor and Chair Patrick Heagerty
Biostatistics

In this dissertation we develop new methods for quantifying the predictive performance of a survival model at different times. We broadly categorize predictive performance into either calibration or discrimination, and propose new methods for measuring time-varying discrimination that complement existing methods such as time-varying AUC. Specifically, we introduce the hazard discrimination summary, $HDS(t)$, a measure that characterizes the ability of a survival model to discriminate between incident events and survivors at each time point. We first motivate $HDS(t)$ as an incident extension of the discrimination slope, and propose a semiparametric estimator along with a study of its asymptotic properties. Second, we show that $HDS(t)$ is amenable to evaluating time-varying covariates, propose corresponding semiparametric estimators, and outline inferential procedures. Finally, we propose an alternative interpretation and nonparametric estimators for $HDS(t)$, both of which illuminate connections between $HDS(t)$ and fundamental information theoretic concepts.

TABLE OF CONTENTS

Page

LIST OF FIGURES

ACKNOWLEDGMENTS

Sincere thanks go to: Patrick Heagerty for being a special teacher, advisor, and role model; I am lucky to have learned from you and look forward to continued collaboration. My committee members Ken Rice, Lianne Sheppard, Michael LeBlanc, and Nicholas Smith for their patience and insight; in particular Ken and Michael for being on my reading committee and insisting on clarity and quality. The following University of Washington Biostatistics faculty members for shaping my growth as a graduate student: Elizabeth Brown, Gary Chan, Scott Emerson, Ken Rice, Lianne Sheppard, and David Yanez. My college math professors Nitu Kitchloo, Bill Minicozzi, W. Stephen Wilson, and Stephen Zucker for being caring instructors and mentors. My boss at JHU APL Donald Duncan for patiently introducing me to the research world while providing invaluable guidance. My high school calculus teacher Michael Siegert for being an incredible teacher and sparking my interest in math. My friends and classmates for keeping me sane. My mom, dad, brother, wife, and son for keeping it real.

DEDICATION

To J. Choe. And you too RayRay now that you're here.

Chapter 1

INTRODUCTION AND BACKGROUND

Survival models are an important class of predictive methods in biomedical research, with many different areas of application. Use has concentrated in but is not restricted to cardiovascular and oncological risk prediction. Common applications that require evaluating predictive performance include: screening many candidate (possibly longitudinal) biomarkers, evaluating a single candidate (possibly longitudinal) biomarker, developing a new risk prediction model, validating existing risk prediction models on new data, and comparing two or more prediction models.

The proliferation of data availability and interesting predictive applications has outpaced the development of suitable quantitative evaluation criteria. Even within one of the above applications there are many different ways to frame the scientific question. For example, when evaluating a candidate longitudinal biomarker, are we interested in assessing its temporal qualities? Even if we only use baseline measurements, do we care about predicting 10-year survival or are we interested in how predictive performance varies over a range of times? These are all different questions that may be best evaluated with correspondingly different performance measures.

Part of the reason for this growth in predictive applications is the availability of rich survival datasets. The Framingham Heart Study, an ongoing regional cohort study of cardiovascular disease that began in 1948, was central to identifying major prognostic factors for cardiovascular disease and continues to study new factors that may provide additional prognostic value. The Multi-Ethnic Study of Atherosclerosis (Bild et al., 2002), a modern cohort study of cardiovascular disease, is notable for its scale (more than 6000 participants, 6 different geographic sites), and available biomarker measurements (longitudinal CT scans among many others). The UK Biobank study, a study of 5-year mortality in a cohort

of almost 500000 participants, compared the predictive value of 655 individual covariates (Ganna and Ingelsson, 2015). Designing novel clinical trials or performing secondary analyses of existing clinical trial data to identify subgroups that respond more or less favorably to treatment than average – sometimes described as "personalized" or "precision" medicine – have recently been of great interest (Paik et al., 2006; Collins and Varmus, 2015). The increasing ubiquity of connected devices and wearable technology will result in a vast amount of longitudinal data at very fine time intervals, presenting another potential opportunity for development and application of new methods (Kumar et al., 2013).

Many predictive accuracy measures for survival models are extensions of measures for binary outcome models. Measures of discrimination such as receiver operating characteristic (ROC) curves and discrimination slope (DS) have been extended to survival models. Extensions of model fit and calibration measures such as R^2, the Brier score, and the Hosmer-Lemeshow test to survival models have also been attempted. We will first review the existing methods for binary outcome data before overviewing the developing field of predictive survival model evaluation. The main purpose of this dissertation is to contribute new measures of time-varying predictive accuracy for survival models. The core of our contribution is to propose additional extensions of existing discrimination measures. However, we will also provide a preliminary overview of novel connections between our proposed methods and important information theoretic concepts. The rest of this chapter will be a general review of predictive accuracy measures, with an additional focus on discrimination measures.

1.1 Notation

Let T be the failure time and C be the censoring time. Let M be a p-vector of predictors. Unless otherwise noted we assume that conditional on M, T and C are independent. Let $\{(T_i, C_i, M_i), i = 1, \ldots, n\}$ be n independent copies of (T, C, M). In practice, for each i we can only observe (Y_i, M_i, δ_i), where $Y_i = min(T_i, C_i)$ and δ_i equals 1 if $Y_i = T_i$ and is 0 otherwise.

Let $\lambda(t)$ be the hazard function for T and let $\lambda(t|M)$ be the conditional hazard function. Two hazard models which we will be referring to in later sections are the Cox model (Cox,

1972) and a local-in-time version of the Cox model (Cai and Sun, 2003). The Cox model assumes $\lambda(t|M) = \lambda_0(t)\exp(\beta' M)$, where β is a vector of parameters in $\mathbb{R}^p$. The local-in-time Cox model replaces β with $\beta(t)$, where each parameter is a function of time, and assumes $\lambda(t|M) = \lambda_0(t)\exp\{\beta(t)' M\}$.

1.2 Biomedical definitions of prediction

We will consider three types of prediction: diagnostic, prognostic, and prescriptive. Diagnostic accuracy refers to how well we are able to assess a patient's current state (e.g. whether or not they have a certain disease). Prognostic accuracy refers to how well we can assess a patient's future state (e.g. what their risk of coronary event will be in five years). Prescriptive accuracy refers to how well we can assess the benefit (or detriment) of an intervention to a patient's future state (e.g. what the coronary event risk will be while taking lipids versus not taking lipids).

In clinical settings "prescriptive accuracy" as defined above is somewhat confusingly also referred to as "predictive accuracy". For the sake of consistency with statistical jargon, in this dissertation, "predictive accuracy" will always refer to the umbrella term for different types of accuracy and "prescriptive accuracy" for the more specific definition.

1.3 Statistical methods for quantifying predictive performance

While clinically it is important to distinguish among the three types of prediction, there is often overlap in the statistical methods used to assess the types of predictive accuracy. Our methods are largely motivated by prognostic applications, but may also be useful for diagnostic and prescriptive applications.

Steyerberg et al. (2010) sorts measures of predictive accuracy for binary outcomes into five categories roughly based on scientific utility: overall performance, discrimination, calibration, reclassification, and clinical usefulness. While we feel the labels are useful for an applied audience, we will give a different treatment, focusing primarily on calibration and discrimination.

1.3.1 Calibration

For binary outcome models, calibration is defined as how close the model-based risk predictions are to the actual underlying risks. Predicted risks are considered well calibrated if they are close to the true underlying risk. In practice, it is impossible to know what the true underlying risk is for a single subject since only a binary outcome is observable. Thus, calibration is typically measured within strata of the predicted risks.

To illustrate, consider data generated from the following mechanism:

$$\begin{aligned} Y|X &\sim Bern(\mu) \\ \mu &= \text{expit}\left(1 + 0.5X_1 + 0.5X_2 + X_1 \cdot X_2\right) \\ X_1 &\sim N(0,1) \\ X_2 &\sim N(0,1) \end{aligned}$$

Suppose we have a dataset ($n = 5000$) from the above setup and two models: one that exactly matches the data generating mechanism and one that does not. A common way to visually assess the calibration of model-based risk predictions is to create a scatterplot of the observed outcomes (0s and 1s) versus the risk predictions, and then overlay a nonparametric regression line (e.g. using kernel regression). The closer the regression line is to the straight line with slope 1 and intercept 0, the better calibrated the predictions. Figure 1.1 shows graphs that characterize the calibration of the risk predictions from the two models using smoothed regression lines and decile contrasts (described in more detail below).

For a logistic regression model, calibration performance can be summarized by the Hosmer-Lemeshow statistic (Hosmer and Lemeshow, 1980), which tests the calibration of said model. Subjects are first grouped by g equally spaced quantiles of their predicted risks, and the distances (in the mean-squared sense) between average predicted risk and observed event rate within each quantile are aggregated. The test statistic is more formally defined below.

$$\widehat{HL} = \sum_{k=1}^{g} \frac{(o_k - n_k\bar{p}_k)^2}{n_k\bar{p}_k(1-\bar{p}_k)}$$

Above, o_k is the number of events among those in the kth quantile of predicted risks; n_k is the number of subjects in the kth quantile; and $\bar{p}_k$ is the average predicted risk among those

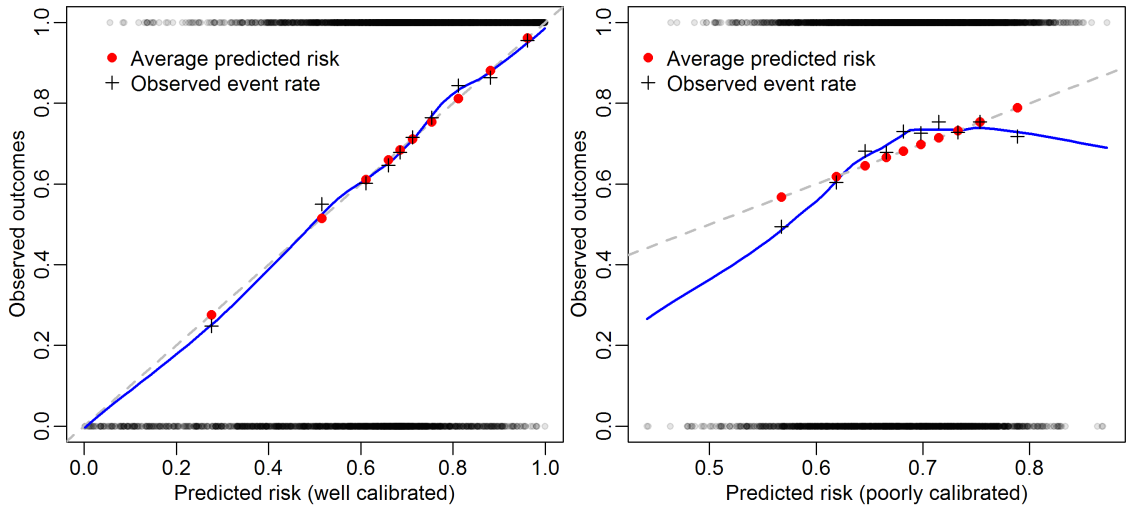

Figure 1.1: Binary outcome data was simulated using $n = 5000$. The left graph shows calibration performance when using a model that matches the data generating mechanism: $\text{logit}\{P(Y|X)\} = 1+0.5X_1+0.5X_2+X_1 \cdot X_2$. The right graph shows calibration performance when using a model that poorly matches the data generating mechanism by omitting X_2. The black dots indicate observed outcomes plotted against model-based risk predictions. The blue lines are kernel smoothed regression lines through the black dots; a well-calibrated model will have a blue line that closely resembles the dotted gray line. The red dots and black crosses are coarser measures of calibration. On each graph, the ten pairs represent deciles of predicted risk, with the dot being average predicted risk and the cross being observed event rate within that decile.

in the kth quantile. Under the null hypothesis that the logistic regression model is true, $\widehat{HL}$ asymptotically approaches a $\chi^2(g-2)$ distribution. In practice, the data is typically split into deciles, or $g = 10$. However, the test's power has been shown to be sensitive to both sample size and partition size, so care should be taken when specifying the partition size (Paul et al., 2013). Using the simulated data and models described above, average predicted risks versuses observed risks for each decile are shown graphically in Figure 1.1.

D'Agostino and Nam (2004) and Cook and Ridker (2009) proposed extensions of $\widehat{HL}$ to the Cox model which involve fixing a time of interest and comparing predicted survival probability to Kaplan-Meier estimates. Grønnesby and Borgan (1996) proposed an extension of $\widehat{HL}$ to the Cox model based on martingale residuals. An in-depth study of the above survival extensions was done by Guffey (2012).

The Brier score is another method for evaluating the calibration of binary outcome models that has been extended for use on survival models. A more extensive outline of the Brier score is given in Chapter 3.

Calibration is an important concept, but it will not be the primary focus of this dissertation. For a technical overview of calibration and proper scoring rules, see Gneiting and Raftery (2007).

1.3.2 Discrimination

Discrimination can be roughly defined as how "well separated" cases are from controls. With binary outcomes, the definition of cases and controls is unambiguous. But with survival data, there are several scientifically meaningful ways to define cases and controls, which we will outline in the following sections. Once cases and controls are defined, we then define "well separated".

ROC and AUC The receiver operating characteristic (ROC) curve was initially used for electronic signal detection applications in the 1950s (Peterson et al., 1954). It then began gaining acceptance in psychophysics and radiology studies in the 1960s (Green and Swets, 1966; Metz, 1986) before being broadly adapted by the general biomedical field for assessing diagnostic accuracy (Zweig and Campbell, 1993). Today, ROC curves have already spread

beyond diagnostic accuracy applications to being commonly used for assessing prognostic accuracy. ROC curves provide a useful graphical summary of possible sensitivity and specificity pairings for all possible marker cutpoints. Our main interest will be in the area under the ROC curve (AUC), which itself has an intuitive interpretation.

AUC measures separation between cases and controls on the rank scale - more specifically, it is the probability that a random case marker is greater than a random control's marker;

$$AUC = P(M_i > M_j | Y_i = 1, Y_j = 0).$$

An illustrative example comparing a "good" area under the ROC curve (AUC) with a "poor" AUC is shown in Figure 1.2. We generated a dataset ($n = 5000$) from the same data generating mechanism as specified in Section 1.3.1, and created two markers: M_{good} and M_{poor}. The former is defined as the predicted risks from a "good" logit model fit using $\text{logit}\{P(Y|X)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2$; the latter is defined as the predicted risks from a poor model fit using $\text{logit}\{P(Y|X)\} = \beta_0 + \beta_1 X_1$.

The AUC for M_{good} and M_{poor} are 0.73 and 0.57, respectively, suggesting that M_{good} has better discriminatory performance than M_{poor}. When comparing two models, it is also common to summarize the performance difference with ΔAUC, which in this case is $0.73 - 0.57 = 0.16$.

Several extensions of AUC for use in predictive survival models have been developed, which we will cover in more detail in the next sections.

Discrimination slope Another measure of discrimination is the discrimination slope (DS), which measures separation between cases and controls on the risk scale. It is defined as *the difference in average predicted risk between observed cases and controls*:

$$DS = \text{E}_{M|Y=1}\left\{P\left(Y = 1|M\right)|Y = 1\right\} - \text{E}_{M|Y=0}\left\{P\left(Y = 1|M\right)|Y = 0\right\}$$

DS can be traced back to at least 1982 when Yates (1982) refers to it under a different name. However, it was not until Pencina et al. (2008) popularized the integrated discrimination improvement (IDI), noting that IDI can be calculated as the difference in DS between

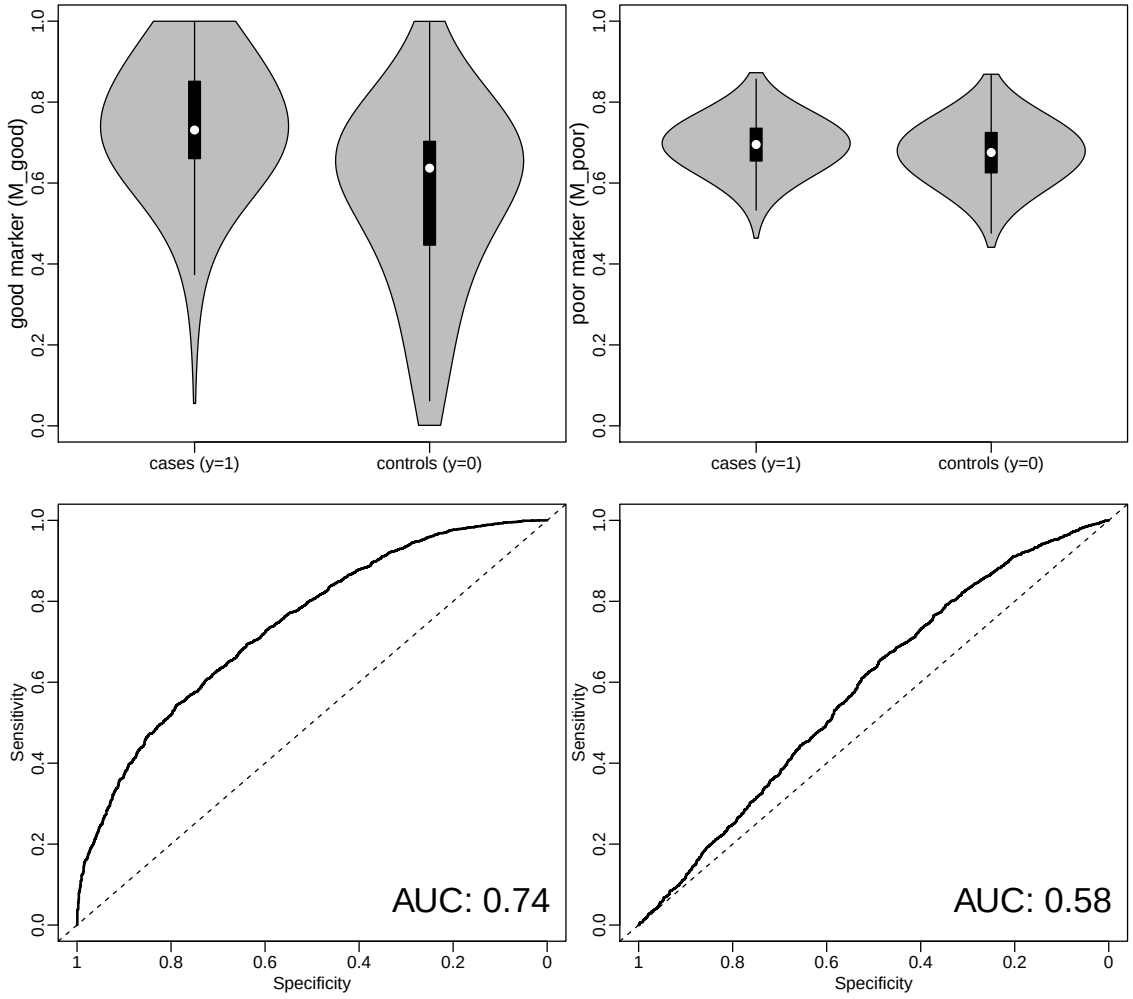

Figure 1.2: Binary outcome data was simulated using $n = 5000$. **The left graphs** show discrimination performance for M_{good}, a marker defined as the predicted risks from a model that matches the data generating mechanism: $\text{logit}\{P(Y|X)\} = 1 + 0.5X_1 + 0.5X_2 + X_1 \cdot X_2$. **The right graphs** show discrimination performance for M_{poor}, a marker defined as the predicted risks from a model that poorly matches the data generating mechanism by omitting X_2. **Each pair of violin plots** summarize the M distribution stratified by observed outcomes (cases and controls). **Each ROC curve** then summarizes the possible sensitivity and sensitivity pairs over all cutpoints. Finally, our single-number summary of discrimination is the area under the ROC curve (AUC). **Each AUC** can be interpreted as the probability that an M randomly drawn from the left violin plot (cases) is greater than one randomly drawn from the right violin plot (controls).

two models, that DS and IDI started being widely adopted for practical use. Today, DS and IDI are frequently calculated alongside AUC in applications that require evaluating predictive accuracy (Hlatky et al., 2009; Greenland et al., 2010; Goff et al., 2014). In addition to its active use in the biomedical literature, there has been an active methodological discussion of DS and IDI in the statistical literature.

Pepe et al. (2008) has shown that aside from the above definition of DS, there are several other intuitive interpretations. Specifically, DS can be mathematically expressed so as to to be 1) integrated sensitivity minus integrated one minus specificity; 2) R^2 generalized to binary regression; and 3) the difference in weighted average absolute residuals between cases and controls.

Kerr et al. (2011) found faults with the asymptotic derivations of the IDI estimator proposed by Pencina et al. (2008), and suggested that caution is needed when calculating confidence intervals for IDI or using it as a formal statistical test of risk prediction improvement.

Hilden and Gerds (2013) showed that DS is not a proper scoring rule and thus is not well suited for evaluting calibration. We agree, and consider DS to be primarily a measure of discrimination. Pencina et al. (2008) and Pepe et al. (2013) also pointed out the importance of calibration in order for DS and IDI to have meaningful interpretations, suggesting possible procedures to ensure proper calibration. In practice, researchers typically either ignore calibration entirely, or conduct an accompanying Hosmer-Lemeshow test to separately assess calibration. DS and IDI's dependence on calibration has been previosly noted (Pencina et al., 2008; Pepe et al., 2013; Hilden and Gerds, 2013), and *ad hoc* recalibration methods have been suggested (Pencina et al., 2008; Pepe et al., 2013). However, to our knowledge, no formal studies on the effectiveness of recalibration methods for DS and IDI have been done.

An illustrative example comparing a "new" DS with an "old" DS is shown in Figure 1.3. We used the same dataset ($n = 5000$) above and compare risk predictions from a "new" model and an "old" model. The former is a logit model fit using both markers plus an interaction $\text{logit}\{P(Y|X)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2$; the latter is a logit model fit using only one marker $\text{logit}\{P(Y|X)\} = \beta_0 + \beta_1 X_1$.

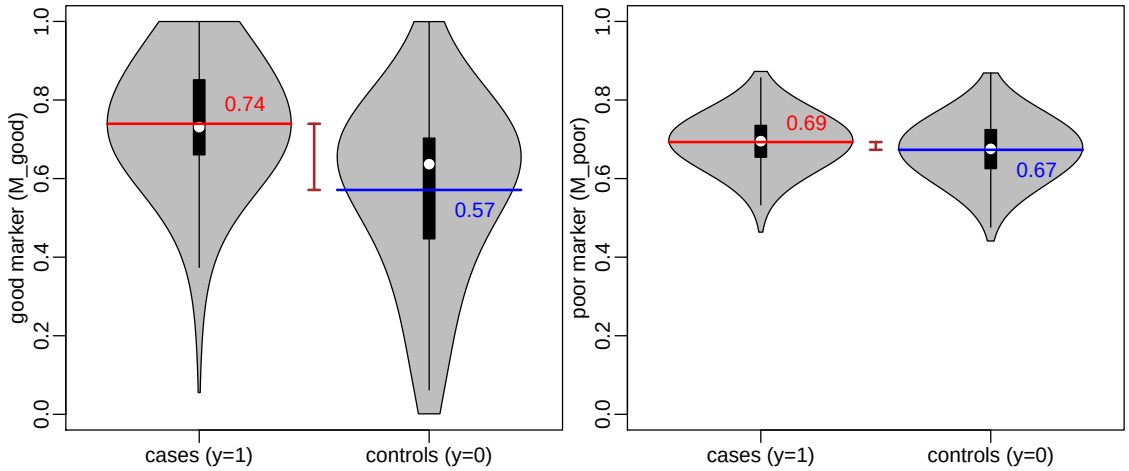

Figure 1.3: Binary outcome data was simulated using $n = 5000$. **Each pair of violin plots** summarize the distribution of predicted risks stratified by observed outcomes (cases and controls). The horizontal colored line segments represent average predicted risk for cases and controls. DS is calculated by subtracting the former from the latter. **The left graph** shows discrimination performance of the "new" model, specified to match the data generating mechanism: $\text{logit}\{P(Y|X)\} = 1 + 0.5X_1 + 0.5X_2 + X_1 \cdot X_2$. DS is equal to 0.74 - 0.58 = 0.16. **The right graph** shows discrimination performance for the "old" model, which only uses X_1. DS is equal to 0.69 - 0.68 = 0.01

The DS for new and old models are 0.16 and 0.01, respectively, suggesting that the new model has better discriminatory performance than the old model. As mentioned above, when comparing two models, it is common to summarize the performance difference with IDI, the difference between the new and old DS: $0.16 - 0.01 = 0.15$.

DS and IDI have also been partially extended for use on predictive survival models. We will cover these extensions in the next sections.

1.3.3 Other methods that do not fall within our taxonomy but are often associated with predictive accuracy

We have thus far discussed calibration and discrimination, classes of predictive accuracy measures most relevant to the contributions in the dissertation. In this section we briefly mention other methods that are useful, but fall outside the the classes of calibration and discrimination. We mention them because the methods are often associated with the general term "predictive accuracy", making it important to acknowledge their existence in order to more thoroughly frame our contribution.

R^2 extension The R^2 is a measure commonly used to assess model fit of a linear regression by quantifying how well observed values fit the predicted regression line. It is an intuitive measure in part because for ordinary linear regression, R^2 is equal to the square of the Pearson correlation coefficient, a measure of linear dependence between the predictors and outcomes.

R^2 can also be extended to measure model fit for logistic regression, but it loses much of its intuitive meaning and is more accurately interpreted as a measure of explained variation (Hosmer and Lemeshow, 2004, Section 5.2.5).

Building on the logistic regression extension, Schemper and Henderson (2000) proposed two further R^2-type extensions of predictive accuracy for survival models. With survival outcomes, for each t we can consider the variance (over study subjects) of predicted probabilities (of surviving past t) and also the variance of observed statuses (at time t). The first measure is a ratio of the weighted time-average of each of the variances, while the second is a weighted time-average of the ratio of the variances. The measures are formally defined

below.

$$V(\tau) = 1 - \frac{\int_0^\tau \mathrm{E}_M \left[S(t|M)\{1 - S(t|M)\}\right] f(t)dt}{\int_0^\tau S(t)\{1 - S(t)\} f(t)dt}$$
$$V_W(\tau) = \int_0^\tau \left(1 - \frac{\mathrm{E}_M \left[S(t|M)\{1 - S(t|M)\}\right]}{S(t)\{1 - S(t)\}}\right) f(t)dt \times \left\{\int_0^\tau f(t)dt\right\}^{-1}$$

Positive and negative predictive value Positive predictive value (PPV) and negative predictive value (NPV) are individual-level measures of predictive accuracy. This is in contrast with population-level measures such as AUC. An individual-level measure of predictive accuracy would be most useful for informing decisions that will only affect an individual, while a population-level measure would be more suitable for informing decisions that will affect an entire population. We are mostly interested in the latter, but present the following for the sake of completeness.

PPV and NPV are defined for data where both the outcome and marker are binary. Thus, it is typically used when the "marker" is the result from a thresholded test (e.g. dichotomizing blood pressure into "high" and "low"). PPV is the probability that individuals experience an outcome (e.g. myocardial infarction) given that they test positive for a marker (e.g. Framingham risk score above a certain cutoff). NPV is the probability that individuals do *not* experience an outcome given that they test negative for a marker.

$$PPV = \mathrm{P}\left(Y = 1 | M = 1\right)$$
$$NPV = \mathrm{P}\left(Y = 0 | M = 0\right)$$

Note the reversal in order of conditioning, compared to discrimination measures. Measures such as AUC and DS condition on cases and/or controls; PPV and NPV condition on the marker. This is the main reason for the distinction between individual and population measures.

Oftentimes M will not be binary but continuous, with a monotone relationship between M and risk. Moskowitz and Pepe (2004) proposed an extension of PPV and NPV that incorporates the continuous information from M by considering all possible quantiles of M.

The extensions also lend themselves well to graphical summaries, and are defined as

$$PPV(v) = \mathrm{P}\left\{Y = 1 | F_M(M) > v\right\}$$
$$NPV(v) = \mathrm{P}\left\{Y = 0 | F_M(M) \leq v\right\}.$$

Zheng et al. (2008) proposed a further extension of PPV and NPV to survival data by incorporating both continuous M and time-to-event outcomes. The extension considers all possible quantiles of M and survival time T.

$$PPV(t, v) = \mathrm{P}\left\{T \leq t | F_M(M) > v\right\}$$
$$NPV(t, v) = \mathrm{P}\left\{T > t | F_M(M) \leq v\right\}$$

Measures to assess Cox model fit Model fit typically refers to a measure that quantifies the distance between observed outcomes and predicted outcomes. For example, kernel regression residuals – defined as the difference each observed outcome and its corresponding predicted outcome – are often used to assess model fit or to aid in bandwidth selection.

Residuals are typically defined using predicted and observed *outcomes*. Such a definition for Cox regression is more difficult, given the possible censoring of time-to-event outcomes. Nevertheless, a number of residuals have been developed to assess model fit for Cox regressions. (Hosmer et al., 2008, Section 6.2) mention five different residuals: martingale, Schoenfeld, scaled Schoenfeld, score, and scaled score. Martingale residuals measure the distance between the observed censoring indicator and predicted cumulative hazard for each observation. The latter four are all related to measuring the distance between observed covariate values and expected covariate values, which rely on hazard estimates. We will briefly review just martingale and Schoenfeld residuals.

Martingale residuals are useful for checking the functional form of the covariate effects. Each residual is defined as

$$r_i = \delta_i - \hat{\Lambda}_0(t) \exp\left(m_i \hat{\beta}\right).$$

Schoenfeld residuals are useful for assessing the assumption that Cox model coefficients are time-invariant. When assessing a multivariate model, residuals are calculated for each individual predictor.

$$s_i = m_i - \frac{\sum_{j \in R_i} m_j \cdot \exp\left(m_j \hat{\beta}\right)}{\sum_{j \in R_i} \exp\left(m_j \hat{\beta}\right)}$$

Both types of residuals mentioned above are typically used by plotting them against the covariates, and visualy inspecting for systematic trends. For martingale residuals, trends would indicate deviations from the assumption of linear covariate effects; for Schoenfeld residuals trends would indicate violations of the assumption of time-invariant coefficients. We will return to Schoenfeld residuals in Chapter 4 when discussing bandwidth selection methods.

1.4 Defining cases and controls for survival data

Thus far we have provided a taxonomy of prediction "subtypes" and predictive accuracy measures in the context of binary outcome data while pointing out attempts to extend certain measures to survival data. Since the majority of the results in this dissertation extend measures of discrimination to survival data, we will first provide some background necessary for the development of our extensions.

To discuss discrimination, we must provide definitions for cases and controls. For binary outcomes, the definitions are unambiguous as they are coded into the outcomes themselves. For time-to-event outcomes, there are many ways to define cases and controls. In this section we will provide several such definitions all with meaningful but different scientific uses.

Cumulative cases, dynamic controls For a given time t_1, cumulative cases are those with events before t_1, and dynamic controls are those still event free at t_1. More formally:

$$\text{cases} : \{i : T_i \leq t_1\}$$

$$\text{controls} : \{i : T_i > t_1\}$$

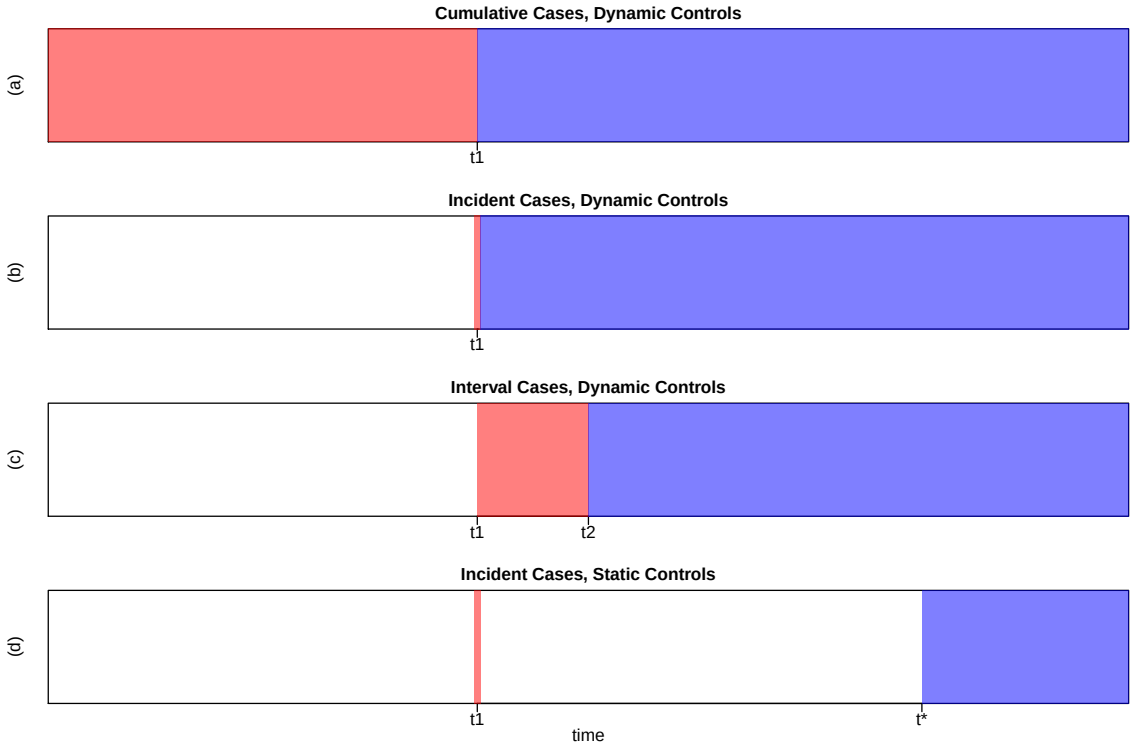

Figure 1.4: Graphical depiction of various definitions of cases and controls for survival data. Red corresponds to cases and blue to controls.

A graphical depiction is shown in Figure 1.4(a). This definition of cases and controls is convenient when one is only interested in survival past a set time. For example, cardiovascular studies sometimes have survival data but the primary interest may be in using baseline covariates to predict 10-year risk of cardiovascular disease. Such studies would be using cumulative cases and dynamic controls.

Incident cases, dynamic controls For a given time t_1, incident cases are those with events at t_1, and dynamic controls are those still event free at t_1. More formally:

$$\text{cases} : \{i : T_i = t_1\}$$
$$\text{controls} : \{i : T_i > t_1\}$$

A graphical depiction is shown in Figure 1.4(b). Predictive measures contrasting incident cases and dynamic controls are a natural fit for assessing time-varying covariates. They can also readily be converted to a time-averaged summary by performing a weighted integration over time.

Other definitions of cases and controls For a given time t_1, interval cases are those with events between t_1 and t_2, and dynamic controls are those still event free at t_2. More formally:

$$\text{cases} : \{i : T_i \in [t_1, t_2]\}$$
$$\text{controls} : \{i : T_i > t_2\}$$

It is also possible to return to incident cases but fix the controls. We will refer to such controls as static controls, or those still event free at a fixed t^*, where $t^* > t_1$. More formally:

$$\text{cases} : \{i : T_i = t_1\}$$
$$\text{controls} : \{i : T_i > t^*\}$$

Graphical depictions are shown in Figure 1.4(c)-(d). Measures based on these definitons describe how well covariate information measured up until t_1 discriminates between cases

and controls, where "case" means experiencing an event between t_1 and some later time t_2 (or exactly at t_1) and "control" means being event free past t_2 (or t^*).

Consider the scenario where during a medical visit, a patient's clinical history is updated and the physician uses the information determine whether the patient is "high risk" (case; will experience an event within the next 5 years or imminently) or "low risk" (control; will still be event-free at 5 years or some fixed time). A prognostic measure based on the above definitions would be useful in assessing how well the doctor's assessment discriminates between future high and low risk group membership.

1.5 Existing methods

In this section we will overview two types of existing methods for evaluating the discrimination of survival models: 1) various extensions of AUC, constituting a collection of rank-based methods; and 2) an extension of DS, a risk-based method. The AUC methods have been extended to cover all of the case/control definitions in Figure 1.4, while DS can be considered a measure of discrimination for *cumulative* cases and *dynamic* controls.

1.5.1 AUC extensions

Heagerty et al. (2000) developed an extension of AUC that quantifies the discrimination of cumulative cases and dynamic controls at each time t:

$$AUC^{C/D}(t) = P(M_i > M_j | T_i \leq t, T_j > t)$$

An illustrative example using the PBC Mayo data (detailed in Section 1.5.3) is shown in Figure 1.5. The example provides some intuition for how to interpret $AUC^{C/D}(t)$ at a single t of interest.

Heagerty et al. (2000) also provide an efficient nonparametric estimator of $AUC^{C/D}$ based in part on an estimator of the bivariate survival function $S(t, m)$ developed by Akritas (1994). Prior to the development of $AUC^{C/D}(t)$, investigators would often bifurcate time-to-event endpoints at a single time and proceed with a binary outcome method such as AUC. This would often require ignoring a significant number of censored outcomes, effectively throwing away data.

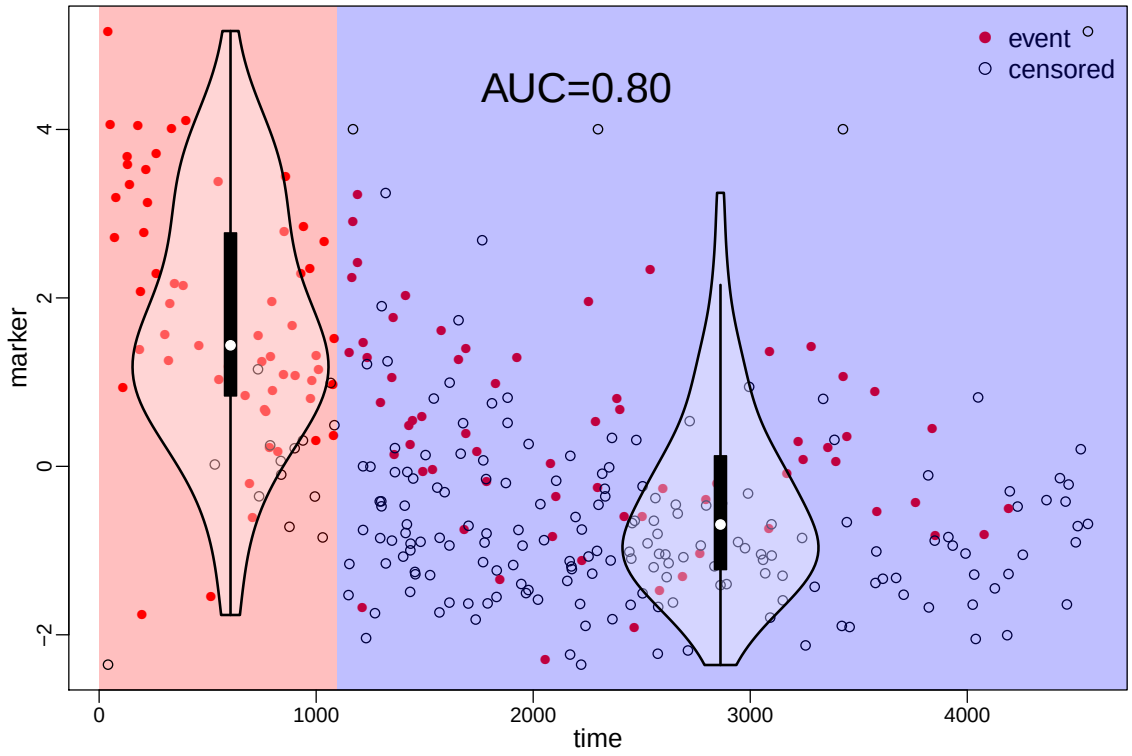

Figure 1.5: A graphical depiction of $AUC^{C/D}(t)$ for $t = 1100$, using PBC Mayo data (detailed in Section 1.5.3). Setting $t = 1100$ defines cumulative cases as those who died prior to 1100 days (red) and dynamic controls as those still alive at 1100 days (blue). The violin plots correspond to the estimated marker distributions for cases and controls. $AUC^{C/D}(1100) = 0.80$ is the probability that a random draw from the red violin plot is greater than one from the blue violin plot.

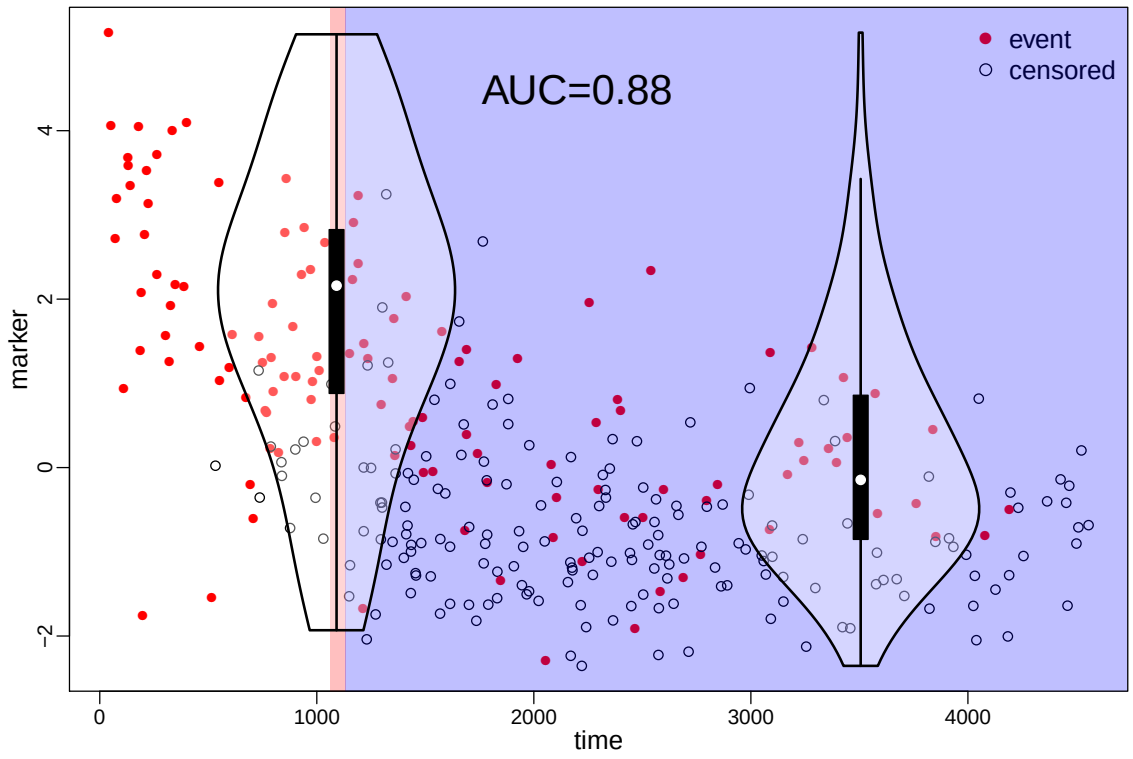

Figure 1.6: A graphical depiction of $AUC^{I/D}(t)$ for $t = 1100$, using PBC Mayo data (detailed in Section 1.5.3). Setting $t = 1100$ defines cumulative cases as those who died at 1100 days (red) and dynamic controls as those still alive at 1100 days (blue). The violin plots correspond to the estimated marker distributions for cases and controls. $AUC^{I/D}(1100) = 0.88$ is the probability that a random draw from the red violin plot is greater than one from the blue violin plot.

Heagerty and Zheng (2005) developed an extension of AUC that quantifies the discrimination of incident cases and dynamic controls at each time t:

$$AUC^{I/D}(t) = P(M_i > M_j | T_i = t, T_j > t)$$

An illustrative example using the PBC Mayo data (detailed in Section 1.5.3) is shown in Figure 1.6. The example provides some intuition for how to interpret $AUC^{I/D}(t)$ at a single t of interest.

$AUC^{I/D}(t)$ is also easily extended to accomodate time-varying covariates. Two semi-parametric estimators, based on the Cox model and a relaxed local-in-time version (Cai and Sun, 2003), were provided. Saha-Chaudhuri and Heagerty (2013) developed a nonpara-

metric estimator of $AUC^{I/D}(t)$ that is based on an intuitive graphical smoothing method. As mentioned in Section 1.4, measures for incident cases and dynamic controls are conducive to being integrated over time to achieve a global performance measure. Integrating $AUC^{I/D}(t)$ over time, weighted by $2 \cdot f(t) \cdot S(t)$ results in the intuitive global concordance measure $\mathrm{P}\left(M_i > M_j | T_i < T_j\right)$.

Additional extensions quantifying discrimination for the other definitions of cases and controls described in Section 1.4 have also been developed. Zheng and Heagerty (2004) developed an extension for incident cases and static controls; Zheng and Heagerty (2007) developed an extension for interval cases and dynamic controls. Semiparametric estimators were provided for both extensions.

1.5.2 DS and IDI extensions

The discrimination slope can be extended to survival data such that it becomes a measure of discimination for cumulative cases and dynamic controls:

$$DS(t) = \mathrm{E}_{M|T \leq t}\left\{P\left(T \leq t | M\right) | T \leq t\right\} - \mathrm{E}_{M|T>t}\left\{P\left(T \leq t | M\right) | T > t\right\}$$

An illustrative example using the PBC Mayo data (detailed in Section 1.5.3) is shown in Figure 1.7. The example provides some intuition for how $DS(t)$ is calculated at a single t of interest.

Chambless et al. (2011); Uno et al. (2012) introduced estimators for $DS(t)$ and $IDI(t)$. Their estimators are semiparametric, as they assume the Cox model to be true or use it as a working model. To our knowledge, a nonparametric estimator of $DS(t)$ does not yet exist, but we posit that M would need to be univariate, or at most bivariate, for such an estimator to be tractable. This is because while $P\left(T \leq t\right)$ is well-defined for M of any dimension, the curse of dimensionality would make it very difficult to estimate well for if M is greater than one or two dimensions.

1.5.3 Data examples

To illustrate the methods throughout this dissertation we will use data sets from three studies, which we will outline here. One is the well-known Mayo primary biliary cirrhosis

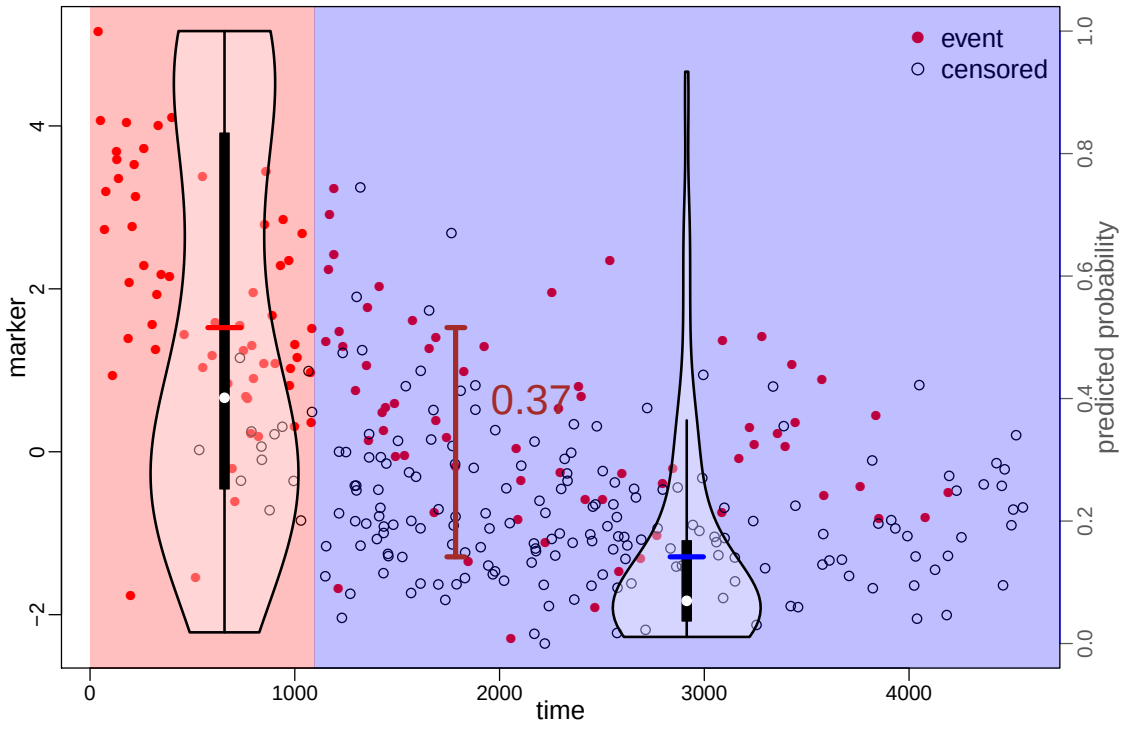

Figure 1.7: A graphical depiction of $DS(t)$ for $t = 1100$, using PBC Mayo data (detailed in Section 1.5.3). Setting $t = 1100$ defines cumulative cases as those who died at 1100 days (red) and dynamic controls as those still alive at 1100 days (blue). The violin plots correspond to the estimated predicted risks for cases and controls (right y-axis). $DS(1100) = 0.37$ is the average predicted risk for cases minus the average predicted risk for controls.

Table 1.1: Hazard ratio estimates from multivariate Cox regression using PBC data

Covariate	Hazard ratio	95% CI
log(bilirubin)	2.404	(1.981, 2.919)
log(prothrombin time)	20.352	(2.728, 151.843)
edema	2.192	(1.218, 3.944)
albumin	0.389	(0.244, 0.619)
age	1.034	(1.017, 1.052)

(PBC) study, widely regarded as a benchmark survival data set. The second is from a multiple myeloma trial (S9321) conducted by the Southwest Oncology Group (SWOG).

Mayo PBC data

The dataset contains 312 subjects, for whom 125 had events. While many covariates are available, we will focus on a subset of five covariates: log(bilirubin, log(prothrombin time), edema, albumin, and age. All covariates are treated as continuous variables. The hazard ratios from a multivariate Cox regression using the five predictors are shown in Table 1.5.3.

In the next chapters we will often use the linear predictor from the preceding Cox regression. A scatterplot of the 312 subjects showing the linear predictor values against the observed censoring or death times is shown in Figure 1.8. It is can be seen that at a qualitative level, the linear predictor is predictive of survival time – those with higher values tend to have worse survival times. How to more formally define predictive performance (in particular, discriminatory performance) while also accounting for common survival data characteristics (e.g. censoring, time-varying covariates) will be a main focus of this dissertation.

Multiple myeloma trial S9321 data

The dataset comes from a multiple myeloma clinical trial conducted by the Southwest Oncology Group (SWOG). We have data on 775 subjects, with a median survival of 48 months.

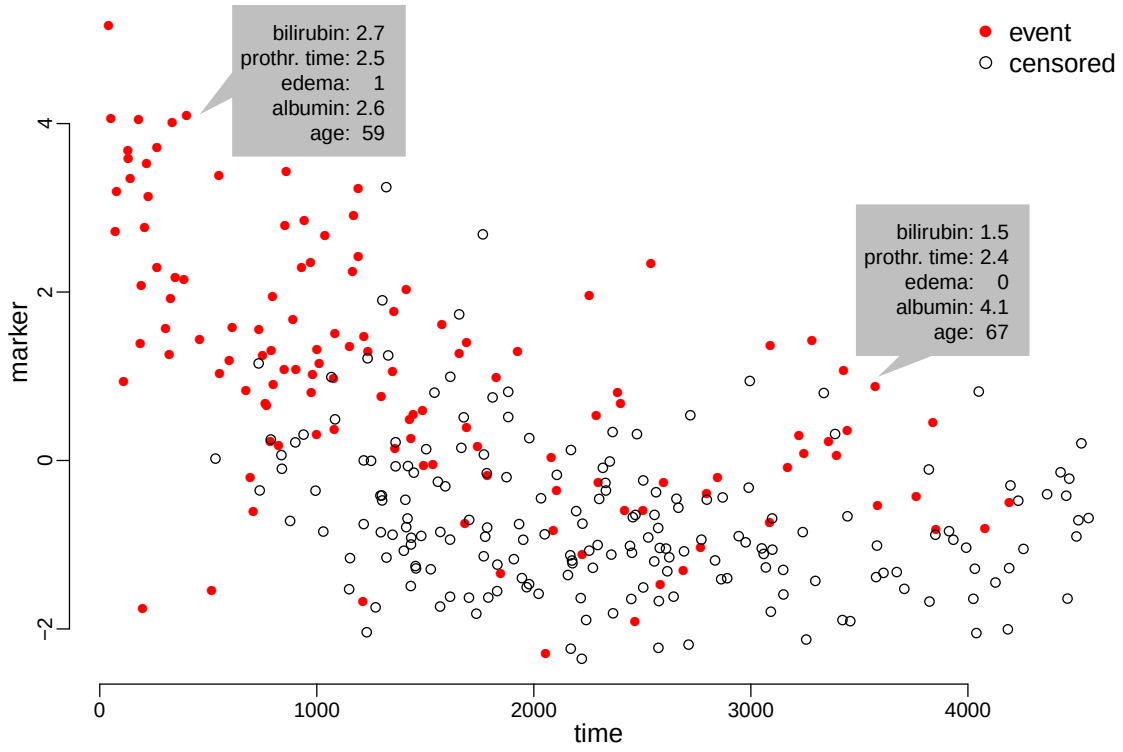

Figure 1.8: Time in days shown on the x-axis. Linear predictor values shown on the y-axis. Each dot represents one of the 312 subjects, with red dots representing those with observed deaths and hollow dots representing those who were censored. The values of the five underlying covariates for two arbitrarily chosen subjects are shown in the gray boxes.

Table 1.2: Hazard ratio estimates from multivariate Cox regression using multiple myeloma data.

Covariate	Hazard ratio	95% CI
Platelets	0.998	(0.997, 0.999)
log(SB2M)	1.189	(1.010, 1.400)
log(Creatinine)	1.024	(0.828, 1.267)
Age	1.025	(1.013, 1.036)
Albumin	0.977	(0.940, 1.016)
Calcium	1.083	(1.022, 1.149)
LDH	1.000	(1.000, 1.001)
Hemoglobin	0.952	(0.909, 0.997)

Of the subjects, 505 experienced events. We also have data on eight baseline markers: platelet count, log(SB2M), log(creatinine), age, albumin, calcium, lactacte dehydrogenase, and hemoglobin. SB2M is short for "beta-2-microglobulin".

The hazard ratios from a Cox regression with the eight predictors are shown in Table 1.2. Most of the confidence intervals include or are close to one, suggesting that the predictors are individually only weakly informative (this is in contrast with the PBC data).

Time-varying hazard ratios, calculated from local constant estimates of $\beta(t)$ (Cai and Sun, 2003), show some notable variations with time, suggesting that the proportional hazards assumption is violated for some predictors. The time-varying hazard ratios for each covariates are shown in Figure 1.9. The hazard ratios should be interpreted as having been "adjusted" for the other covariates, as the values were calculated using a multivariate model with all eight covariates.

Mayo PBC data with time varying covariates

We will also use a version of the Mayo PBC data with time varying covariates. An overview of the dataset is in Chapter 3.

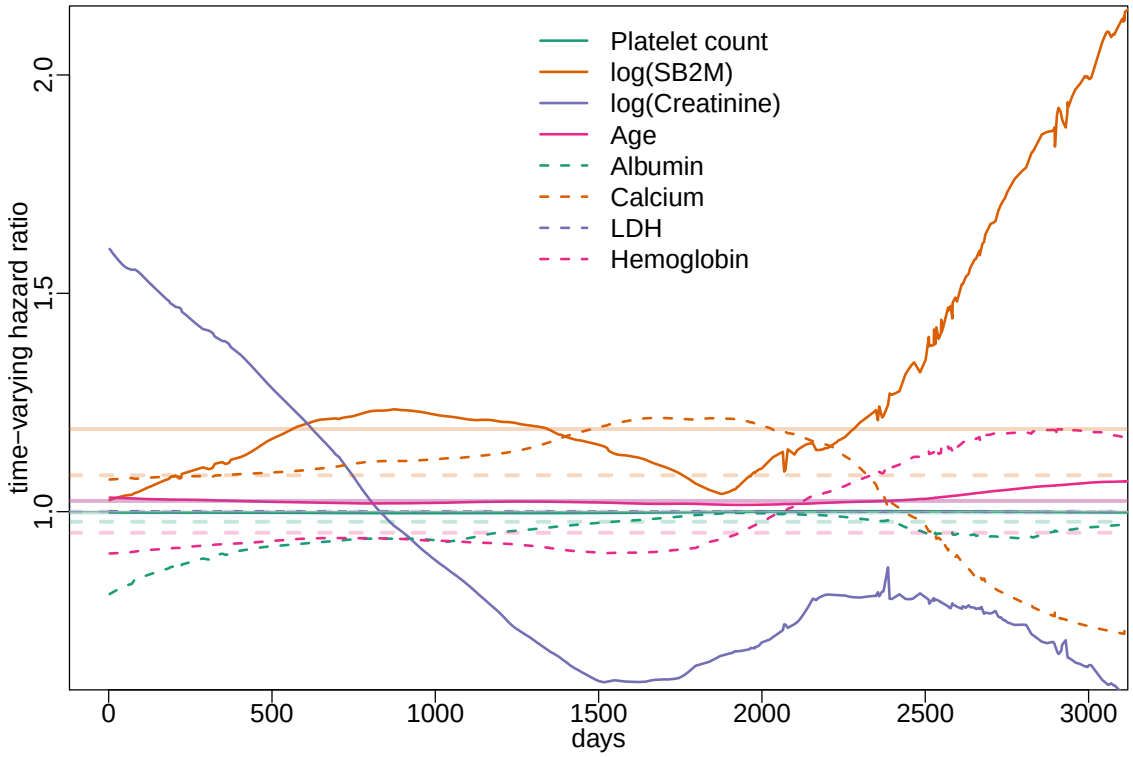

Figure 1.9: Adjusted time-varying hazard ratio estimates for each predictor. $\beta(t)$ values were estimated with a "local constant" method (Cai and Sun, 2003). Transparent lines correspond to β estimates from multivariate Cox regression. The time-trends for the $\beta(t)$ estimates suggest that the proportional hazards assumption does not hold for some predictors.

Chapter 2

HAZARD DISCRIMINATION SUMMARY

Integrated discrimination improvement (IDI) has recently been recommended as a summary of predictive model performance for binary outcomes that can complement traditional accuracy measures (Pencina et al., 2008; Steyerberg et al., 2010). Key attractive aspects of IDI include: 1) IDI measures impact based on change in absolute risk, which directly facilitates clinical interpretation; 2) IDI has several interpretations as a statistical meausure, most notably as the change in discrimination slope (DS), which summarizes the mean risk placed on cumulative cases and dynamic controls respectively (Yates, 1982; Pepe et al., 2008); 3) under certain parametric assumptions, point estimates and valid standard errors are computationally simple, allowing inference on the incremental value associated with a new marker (Kerr et al., 2011). In biomedical applications, IDI is now commonly recommended for use in conjunction with other summary measures such as AUC and net reclassification improvement (NRI) (Hlatky et al., 2009; Greenland et al., 2010; Goff et al., 2014).

Survival models are an important class of risk prediction methods, and a recent literature has been developed to generalize simple binary outcome prognostic measures to the event time setting. For example, time-dependent ROC curves and C-index methods (Heagerty et al., 2000; Heagerty and Zheng, 2005) have been developed, and partial extensions of NRI and IDI to survival outcomes have recently been detailed (Chambless et al., 2011; Uno et al., 2012). However, neither the discrimination slope nor IDI have been generalized to models for incident risk that are naturally appropriate for event-time outcomes, including standard Cox regression models. We will herein refer to survival extensions of IDI and DS as $IDI(t)$ and $DS(t)$, respectively.

Just as Heagerty and Zheng (2005) proposed an incident extension of $AUC^{C/D}(t)$, we will propose an incident extension of $DS(t)$ that is natural for hazard models. To do so we

consider three modifications to $DS(t)$: first we use the incident/dynamic definition of cases and controls (instead of cumulative/dynamic); second we judge how well cases are separated from controls by contrasting their mean hazards (instead of mean cumulative risks); and third we use the ratio as the contrast of choice (instead of the difference).

We call the resulting measure the hazard discrimination summary, or $HDS(t)$. In the next sections we will formally define $HDS(t)$; present a useful alternative interpretation; derive two estimators and study their asymptotic behaviors; and illustrate $HDS(t)$ using simulated data, data from the Mayo PBC study, and data from a multiple myeloma study.

2.1 Parameter of interest: $HDS(t)$

In Chapter 1 we reviewed $DS(t)$ – a cumulative risk-based measure of survival model discrimination. We now propose an incident alternative to $DS(t)$ that is a risk-based measure natural for hazard models. To generalize discrimination to incident events we consider two modifications of $DS(t)$: first we summarize expected hazards since this is the natural incident risk measure; and second we compare the ratio of mean hazards (risks) to reflect relative risk rather than the risk difference.

2.1.1 Parameter of interest as a generalization of discrimination slope

For any given time t, we focus on incident cases and associated dynamic controls. If a marker or set of markers are informative, we would expect a model using the marker(s) to place a higher hazard upon the cases as compared to the controls for each time t. In particular, the ratio of expected hazards would be greater than one:

$$\frac{\text{mean case hazard}}{\text{mean control hazard}} = \frac{\mathrm{E}_{M|T=t}\left\{\lambda(t|M)|T=t\right\}}{\mathrm{E}_{M|T>t}\left\{\lambda(t|M)|T>t\right\}} > 1.$$

This time-varying ratio of expected hazards is our measure of incident predictive performance, which we call the "hazard discrimination summary":

$$HDS(t) = \frac{\mathrm{E}_{M|T=t}\left\{\lambda(t|M)|T=t\right\}}{\mathrm{E}_{M|T>t}\left\{\lambda(t|M)|T>t\right\}}.$$

As a ratio measure $HDS(t)$ is not bounded, and can be interpreted relative to a value of one that is consistent with no discriminatory performance, while increasingly better performance is indicated as $HDS(t)$ approaches infinity.

2.1.2 Parameter of interest as prognostic value of a marker

Due to a simple mathematical property of hazards, the denominator of $HDS(t)$ can be shown to also equal the marginal hazard:

$$\mathrm{E}_{M|T>t}\left\{\lambda(t|M)|T>t\right\} \;=\; \lambda(t)\;, \tag{2.1}$$

providing a second attractive interpretation of $HDS(t)$. Specifically, the discrimination measure can then be shown to be equivalently expressed as

$$HDS(t) = \mathrm{E}_{M|T=t}\left[\;\frac{\lambda(t|M)}{\lambda(t)}\;\middle|\;T=t\;\right],$$

leading to the interpretation as the increase in the average risk assigned to incident cases at time t associated with using the marker, M, as compared to the marginal risk associated with not using the marker. Therefore, $HDS(t)$ also directly measures the information content, or incremental value, of the marker relative to having no prognostic data.

The derivation for relation (2.1) is as follows:

$$\begin{aligned}
&\mathrm{E}_{M|T>t}\left\{\lambda(t|M)|T>t\right\}\\
&\qquad = \int \lambda(t|m) f(m|T>t) dm\\
&\qquad = \int \frac{f(t|m)}{S(t|m)} \cdot \frac{S(t|m) f(m)}{S(t)} dm\\
&\qquad = \int \frac{f(t|m) f(m)}{S(t)} dm\\
&\qquad = \frac{1}{S(t)} \int f(t,m) dm\\
&\qquad = \frac{1}{S(t)} f(t)\\
&\qquad = \lambda(t).
\end{aligned}$$

To illustrate $HDS(t)$ under different known scenarios we consider situations where the joint distribution of M and $log(T)$ is bivariate normal with varying degrees of correlation. Under these assumptions, we can express $HDS(t)$ analytically and plot the true parameter function. The left graph in Figure 2.1 shows such iterations of $HDS(t)$ for correlations of $-0.9, -0.8, -0.7, -0.6$, with greater magnitude indicating a more predictive marker. For

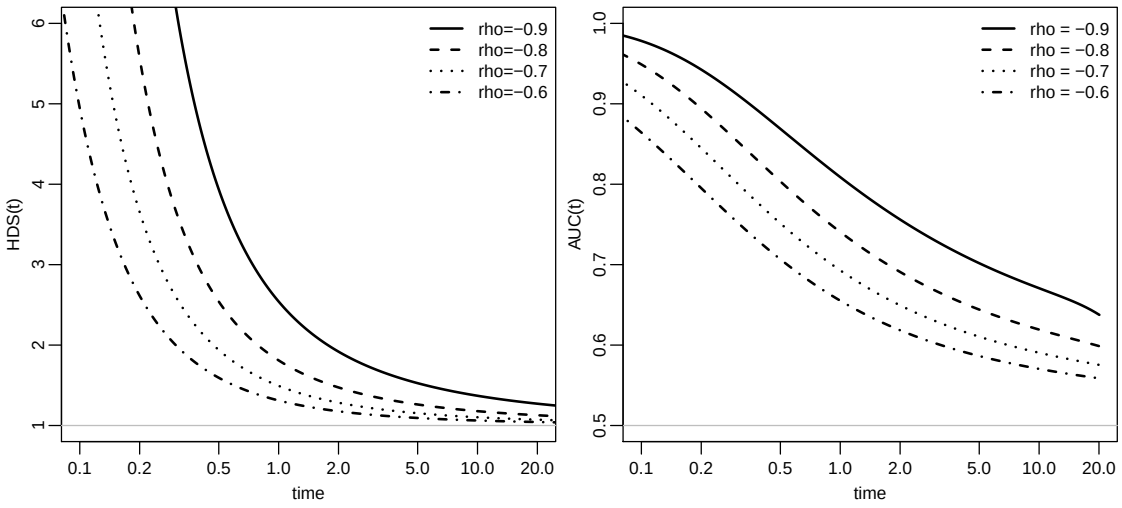

Figure 2.1: **Left**: $HDS(t)$ for varying correlations of $(M, log(T))$ using a bivariate normal distribution. The y-axis shows the ratio of the expected hazard comparing cases and controls. **Right**: $AUC^{I/D}(t)$ for varying correlations of $(M, log(T))$.

comparison, analogous plots of $AUC^{I/D}(t)$ (Heagerty and Zheng, 2005) are also shown. We see that $HDS(t)$ and $AUC^{I/D}(t)$ both have monotone trends over time and with respect to the underlying correlation, but the curves have different shapes and interpretations. For example, with $\rho = -0.9$ and at $t = 1.0$ we see that $HDS(t) \approx 2.5$ indicating that cases have more than twice the risk of controls, or equivalently that using the marker increases the risk assigned to cases by a factor of 2.5 relative to not using marker information. For the same time and correlation, $AUC^{I/D}(t)$ is approximately 0.85 which can be interpreted as the probability that a case marker exceeds a control marker. We believe that each measure is useful, and expresses the relationship between the marker and the outcome simply, although using different scales.

2.1.3 Comparing time-dependent performance of two models

While $HDS(t)$ computed for a particular survival marker or model provides a time-varying measure of the model's discriminatory performance, it is natural to compare two models via their respective $HDS(t)$. In such situations, we recommend calculating $HDS(t)$ for each of the models and then taking the ratio $HDS_{new}(t)/HDS_{old}(t)$ as a measure summarizing

the percent improvement in discrimination associated with the new model/marker.

To illustrate, suppose $(T, M1, M2)$ is trivariate normal where the correlation between T and $M1$ is -0.8 while the correlation between T and $M2$ is -0.6. Calculating $HDS(t)$ individually for $M1$ and $M2$ results in the corresponding values shown in Figure 2.1 where it is clear that $M1$ is more predictive than $M2$ across all times, as measured by $HDS(t)$. At times 0.5 and 4, $HDS(t)$ for $M1$ is 2.54 and 1.30, while for $M2$ is 1.59 and 1.11. The resulting $HDS(t)$ ratios are 1.59 and 1.17, signifying that while $M1$ is the superior marker across all times, it is more so at earlier times compared to later.

For conciseness in the following presentation, we define the following notation for the numerator and denominator of $HDS(t)$:

$$\bar{H}_1(t) = \mathrm{E}_{M|T=t}\left\{\lambda(t|M)|T=t\right\},$$
$$\bar{H}_0(t) = \mathrm{E}_{M|T>t}\left\{\lambda(t|M)|T>t\right\} \;=\; \lambda(t).$$

In section (2.1.2) we stated that $\bar{H}_0(t)$ equals the marginal hazard, $\lambda(t)$, and this relationship holds for any marker M. Therefore the marginal hazard $\lambda(t)$ is not dependent on M. The implication is that the ratio of two $HDS(t)$ values for two different markers simplifies:

$$\begin{aligned} HDS(t;M1)/HDS(t;M2) &= \frac{\bar{H}_1(t;M1)/\bar{H}_0(t;M1)}{\bar{H}_1(t;M2)/\bar{H}_0(t;M2)} = \frac{\bar{H}_1(t;M1)/\lambda(t)}{\bar{H}_1(t;M2)/\lambda(t)} \\ &= \frac{\bar{H}_1(t;M1)}{\bar{H}_1(t;M2)} = \frac{\mathrm{E}_{M1|T=t}\left\{\lambda(t|M1)|T=t\right\}}{\mathrm{E}_{M2|T=t}\left\{\lambda(t|M2)|T=t\right\}}. \end{aligned}$$

In other words, the ratio of $HDS(t)$ for two markers can be interpreted as the ratio of the expected risk predicted by M1 to the expected risk predicted by M2 for time-specific incident events. A key use of this comparison would be for nested models where $M1 = (X1, X2)$, while $M2 = (X2)$ allowing a comparison of the incremental value of $X1$ in addition to $X2$.

2.2 Estimation

In this section we outline estimation of $HDS(t)$ under both the Cox model, and a local-in-time relaxation of Cox model. Both estimators require first expressing $HDS(t)$ such that it is a function of the empirical marker distribution $\mathbf{F}_M$, and the model-based coefficient β (or $\beta(t)$) and cumulative baseline hazard $\Lambda_0(t)$.

2.2.1 $HDS(t)$ *in terms of marginal expectations*

As defined in Section 2.1, $HDS(t)$ is the ratio of expected hazards among cases to expected hazards among controls, $HDS(t) = \bar{H}_1(t)/\bar{H}_0(t)$, and we now show that $\bar{H}_1(t)$ and $\bar{H}_0(t)$ can be rewritten to facilitate estimation:

$$\begin{aligned}
\bar{H}_1(t) &= \mathrm{E}_{M|T=t}\left\{\lambda(t|M)|T=t\right\} \\
&= \frac{\mathrm{E}_M\left[\lambda^2(t|M)\exp\{-\Lambda(t|M)\}\right]}{\mathrm{E}_M\left[\lambda(t|M)\exp\{-\Lambda(t|M)\}\right]}, \\
\bar{H}_0(t) &= \mathrm{E}_{M|T>t}\left[\lambda(t|M)|T>t\right] \\
&= \frac{\mathrm{E}_M\left[\lambda(t|M)\exp\{-\Lambda(t|M)\}\right]}{\mathrm{E}_M\left[\exp\{-\Lambda(t|M)\}\right]}.
\end{aligned}$$

The derivation is as follows. Recall that $HDS(t) = \frac{\mathrm{E}_{M|T=t}\{\lambda(t|M)|T=t\}}{\mathrm{E}_{M|T>t}\{\lambda(t|M)|T>t\}}$. We will show the identities

$$\begin{aligned}
\mathrm{E}_{M|T=t}\left\{\lambda(t|M)|T=t\right\} &= \frac{1}{f(t)}\mathrm{E}_M\left[\lambda^2(t|M)\exp\left\{-\Lambda(t|M)\right\}\right], \\
\mathrm{E}_{M|T>t}\left\{\lambda(t|M)|T=t\right\} &= \frac{1}{S(t)}\mathrm{E}_M\left[\lambda(t|M)\exp\left\{-\Lambda(t|M)\right\}\right], \\
S(t) &= \mathrm{E}_M\left[\exp\left\{-\Lambda(t|M)\right\}\right].
\end{aligned}$$

For the first:

$$\begin{aligned}
&\mathrm{E}_{M|T=t}\left\{\lambda(t|M)|T=t\right\} \\
&\quad= \int \lambda(t|m)f(m|t)dm \\
&\quad= \int \lambda(t|m)\frac{f(t|m)f(m)}{f(t)}dm \\
&\quad= \frac{1}{f(t)}\int \lambda(t|m)\frac{f(t|m)}{S(t|m)}S(t|m)f(m)dm \\
&\quad= \frac{1}{f(t)}\int \lambda^2(t|m)\exp\left\{-\Lambda(t|m)\right\}f(m)dm \\
&\quad= \frac{1}{f(t)}\mathrm{E}_M\left[\lambda^2(t|M)\exp\left\{-\Lambda(t|M)\right\}\right].
\end{aligned}$$

The second:

$$\begin{aligned}
&\mathrm{E}_{M|T>t}\left\{\lambda(t|M)|T>t\right\} \\
&\quad = \int \lambda(t|m) f(m|T>t) dm \\
&\quad = \int \lambda(t|m) \frac{S(t|m) f(m)}{S(t)} dm \\
&\quad = \frac{1}{S(t)} \int \lambda(t|m) \exp\left\{-\Lambda(t|m)\right\} f(m) dm \\
&\quad = \frac{1}{S(t)} \mathrm{E}_M\left[\lambda(t|M) \exp\left\{-\Lambda(t|M)\right\}\right].
\end{aligned}$$

The third:

$$\begin{aligned}
&\mathrm{E}_M\left[\exp\left\{-\Lambda(t|M)\right\}\right] \\
&\quad = \int \exp\left\{-\Lambda(t|m)\right\} f(m) dm \\
&\quad = \int S(t|m) f(m) dm \\
&\quad = \int \frac{P(T>t, M=m)}{f(m)} f(m) dm \\
&\quad = \int P(T>t, M=m) dm \\
&\quad = S(t).
\end{aligned}$$

Combining the above three identities along with $\mathrm{E}_{M|T>t}\left\{\lambda(t|M)|T>t\right\} = \lambda(t)$ provides the desired result for facilitating estimation:

$$\begin{aligned}
HDS(t) &= \frac{\mathrm{E}_{M|T=t}\left\{\lambda(t|M)|T=t\right\}}{\mathrm{E}_{M|T>t}\left\{\lambda(t|M)|T>t\right\}} \\
&= \frac{\mathrm{E}_M\left[\lambda^2(t|M) \exp\left\{-\Lambda(t|M)\right\}\right] \mathrm{E}_M\left[\exp\left\{-\Lambda(t|M)\right\}\right]}{\mathrm{E}_M^2\left[\lambda(t|M) \exp\left\{-\Lambda(t|M)\right\}\right]}
\end{aligned}$$

We have re-expressed $\bar{H}_1(t)$ and $\bar{H}_0(t)$ in terms of marginal expectations of conditional hazard and cumulative hazard functions. Thus if we have estimators for the functions $\lambda(t|M)$ and $\Lambda(t|M)$, then natural estimators for the components of $HDS(t)$ are as follows:

$$\begin{aligned}
\hat{\bar{H}}_1(t) &= \frac{\sum_{i=1}^n \hat{\lambda}^2(t|m_i) \exp\{-\hat{\Lambda}(t|m_i)\}}{\sum_{i=1}^n \hat{\lambda}(t|m_i) \exp\{-\hat{\Lambda}(t|m_i)\}} \\
\hat{\bar{H}}_0(t) &= \frac{\sum_{i=1}^n \hat{\lambda}(t|m_i) \exp\{-\hat{\Lambda}(t|m_i)\}}{\sum_{i=1}^n \exp\{-\hat{\Lambda}(t|m_i)\}}
\end{aligned}$$

In the following sections we adopt estimators for $\hat{\lambda}(t|M)$ and $\hat{\Lambda}(t|M)$ based on either a proportional hazards model, or a non-proportional hazards relaxation using a time-varying coefficient function, $\beta(t)$, based on methods from Cai and Sun (2003).

2.2.2 Estimator for $HDS(t)$ under proportional hazards

If we assume proportional hazards, then we have $\lambda(t|M) = \lambda_0(t)\exp(\beta' M)$. Recalling that $HDS(t) = \bar{H}_1(t)/\bar{H}_0(t)$, we derive the following estimator of $HDS(t)$:

$$\begin{aligned}\widehat{HDS}(t) &= \frac{\hat{\bar{H}}_1(t)}{\hat{\bar{H}}_0(t)} \\ &= \frac{\sum_{i=1}^n \exp\left\{2\hat{\beta}' m_i - e^{\hat{\beta}' m_i}\hat{\Lambda}_0(t)\right\}\sum_{i=1}^n \exp\left\{-e^{\hat{\beta}' m_i}\hat{\Lambda}_0(t)\right\}}{\left[\sum_{i=1}^n \exp\left\{\hat{\beta}' m_i - e^{\hat{\beta}' m_i}\hat{\Lambda}_0(t)\right\}\right]^2}\end{aligned} \tag{2.2}$$

$\hat{\beta}$ is the usual estimator from the Cox proportional hazards model and $\hat{\Lambda}_0(t)$ the Breslow estimator for the cumulative baseline hazard.

2.2.3 Estimator for $HDS(t)$ under smooth non-proportional hazards

We can relax the proportional hazards assumption with a local-in-time version of the natural estimator, motivated by the local partial likelihood method proposed by Hastie and Tibshirani (1993), and also used by Cai and Sun (2003) for estimating time-varying Cox regression coefficients. Instead of assuming $\lambda(t|M) = \lambda_0(t)\exp(\beta' M)$, we now assume a less restrictive model, $\lambda(t|M) = \lambda_0(t)\exp\{\beta'(t)M\}$, with an estimator:

$$\widehat{HDS}^{LC}(t) = \frac{\sum_{i=1}^n \exp\left\{2\hat{\beta}_h'(t) m_i - e^{\hat{\beta}_h'(t) m_i}\hat{\Lambda}_h(t)\right\}\sum_{i=1}^n \exp\left\{-e^{\hat{\beta}_h'(t) m_i}\hat{\Lambda}_h(t)\right\}}{\left[\sum_{i=1}^n \exp\left\{\hat{\beta}_h'(t) m_i - e^{\hat{\beta}_h'(t) m_i}\hat{\Lambda}_{a_n}(t)\right\}\right]^2}$$

$\hat{\beta}_h(t)$ is the local constant estimate of $\beta(t)$ as proposed by Cai and Sun (2003), and $\hat{\Lambda}_h(t)$ is the corresponding estimate of $\Lambda_0(t)$ as detailed by Tian et al. (2005). Above, h is a smoothing parameter and thus varies with n. Further assumptions and properties of h are detailed in Cai and Sun (2003) and Tian et al. (2005). In practice, h can be chosen via K-fold cross-validation.

Briefly, the K-fold cross-validation procedure chooses searches for the h that minimizes a prediction error $PE(h) = \sum_{i=1}^{k} PE_i(h)$. $PE_i(h)$ is defined as the minus log partial likelihood function calculated using marker and outcome values from the ith fold of the data and $\hat{\beta}(t)$ from the full data excluding the ith fold. More formally,

$$PE_i(h) = \sum_{j \in D_i} \delta_j \left(\hat{\beta}^{-i}(t_j)' m_j - \log \left[\sum_{k:k \in D_i, t_k > t_j} \exp \left\{ \hat{\beta}^{-i}(t_j) m_k \right\} \right] \right),$$

where D_i is the set of observations in the ith fold and $\hat{\beta}^{-i}(u)$ is the estimate for $\beta(t)$ using all of the data excluding the ith fold and evaluated at u.

2.2.4 A graphical connection of $HDS(t)$ to the partial likelihood

The discrimination summary $HDS(t)$ also has a strong relationship with the partial likelihood function considered as a function of time. As shown in equation (2.1), $\bar{H}_0(t)$ is equivalent to the marginal hazard $\lambda(t)$. We can also exploit this equivalence to rewrite $HDS(t)$, under the proportional hazards assumption:

$$\begin{aligned} HDS(t) &= \frac{\mathrm{E}_{M|T=t}\{\lambda(t|M)|T=t\}}{\mathrm{E}_{M|T>t}\{\lambda(t|M)|T>t\}} = \frac{\mathrm{E}_{M|T=t}\{\lambda(t|M)|T=t\}}{\lambda(t)} \\ &= \mathrm{E}_{M|T=t}\left[\frac{\lambda(t|M)}{\mathrm{E}\{\lambda(t|M)|T>t\}} \,\middle|\, T=t \right] = \mathrm{E}_{M|T=t}\left[\frac{\exp(\beta' M)}{\mathrm{E}\{\exp(\beta' M)|T>t\}} \,\middle|\, T=t \right]. \end{aligned}$$

The above form suggests that $HDS(t)$ can also be approximated as the function estimated from regressing scaled partial likelihood contributions on time. Specifically, for each of the observations i that experienced an event at y_i, consider a plot of the points $\left\{ \left(y_i, \frac{\exp(\hat{\beta}' m_i)}{(1/n_i)\sum_{j \in R_i} \exp(\hat{\beta}' m_j)} \right) : t_i < c_i \right\}$, where R_i represents the risk set at time y_i and $n_i = \sum_{j=1}^{n} I(y_j \geq y_i)$. A simple estimator of $HDS(t)$ would then be obtained from a smoothed scatterplot. While this is not an efficient or formal estimator, it has value by directly connecting the partial likelihood contributions over time and the target discrimination function, $HDS(t)$. In Section 2.5 we use the myeloma data (see Section 1.5.3) set to illustrate the connection between $HDS(t)$ and the partial likelihood.

2.3 Inference for $HDS(t)$

Under the Cox model we show that $\sqrt{n}\left\{\widehat{HDS}(t) - HDS(t)\right\}$ is asymptotically normal, pointwise for each t. Since $\lambda(t|M) = \lambda_0(t)\exp(\beta' M)$ we can write $HDS(t)$ and $\widehat{HDS}(t)$ as follows:

$$HDS(t) = \frac{\mathrm{E}_M\left[\exp\left\{2\beta' M - e^{\beta' M}\Lambda_0(t)\right\}\right]\mathrm{E}_M\left[\exp\left\{-e^{\beta' M}\Lambda_0(t)\right\}\right]}{\mathrm{E}_M\left[\exp\left\{\beta' M - e^{\beta' M}\Lambda_0(t)\right\}\right]^2},$$

$$\widehat{HDS}(t) = \frac{\sum_{i=1}^n \exp\left\{2\hat{\beta}' m_i - e^{\hat{\beta}' m_i}\hat{\Lambda}_0(t)\right\}\sum_{i=1}^n \exp\left\{-e^{\hat{\beta}' m_i}\hat{\Lambda}_0(t)\right\}}{\left[\sum_{i=1}^n \exp\left\{\hat{\beta}' m_i - e^{\hat{\beta}' m_i}\hat{\Lambda}_0(t)\right\}\right]^2}.$$

Following notation as in van der Vaart and Wellner (2007), we define the following:

$$\begin{aligned} Pf &= \int f dP, \\ \mathbb{P}_n f &= \frac{1}{n}\sum_{i=1}^n f(M_i), \\ \mathbb{G}_n f &= \sqrt{n}\left(\mathbb{P}_n - P\right) f, \\ \Theta &= \{0, 1, 2\}, \\ H &= \left(\beta, \Lambda_0(t)\right). \end{aligned}$$

Above, P is the probability measure for the i.i.d. markers $M_1, \ldots, M_n$; β is the regression parameter from the proportional hazards model and $\hat{\beta}$ is its estimate from the partial likelihood; and $\Lambda_0(t)$ is the cumulative baseline hazard from the same model and $\hat{\Lambda}_0(t)$ is the Breslow estimator. As shorthand we will use η_0 and η_n in place of $(\beta, \Lambda_0(t))$ and $\left(\hat{\beta}, \hat{\Lambda}_0(t)\right)$, respectively. We further define

$$\begin{aligned} f_{\theta,\eta_0}(m) &= \exp\left(\theta\beta m - e^{\beta m}\Lambda_0(t)\right), \\ f_{\theta,\eta_n}(m) &= \exp\left(\theta\hat{\beta} m - e^{\hat{\beta} m}\hat{\Lambda}_0(t)\right). \end{aligned}$$

We know that $\hat{\beta}$ and $\hat{\Lambda}_0(t)$ are consistent for β and $\Lambda_0(t)$ (Tsiatis 1981). Also, since Θ is finite, the class of functions indexed by Θ and H is P-Donsker. The conditions of Theorem

2.1 in van der Vaart and Wellner (2007) are thus satisfied, and we have

$$\sup_{\theta \in \Theta} \left| \mathbb{G}_n \left(f_{\theta,\eta_n} - f_{\theta,\eta_0} \right) \right| \to_p 0.$$

Using the decomposition outlined in van der Vaart and Wellner (2007), this then implies that

$$\sqrt{n} \left(\mathbb{P}_n f_{\theta,\eta_n} - P f_{\theta,\eta_0} \right) = o_p(1) + \mathbb{G}_n f_{\theta,\eta_0} + \sqrt{n} P \left(f_{\theta,\eta_n} - f_{\theta,\eta_0} \right).$$

The second term converges to a Gaussian process (or equivalently, a trivariate normal distribution, since Θ only has three terms) by the central limit theorem. Since $(\beta, \Lambda_0(t))$ is jointly normal as shown by Tsiatis (1981), the third term also converges via the delta method applied to the map $(\beta, \Lambda_0(t)) \mapsto (P f_{\theta,\eta} : \theta \in \Theta)$.

Furthermore, we know that the second term is a function of the empirical marker distribution and the third term is a function of $\left(\hat{\beta}, \hat{\Lambda}_0(t) \right)$. Since the empirical marker distribution $\mathbb{F}_M$ is asymptotically independent from $\left(\hat{\beta}, \hat{\Lambda}_0(t) \right)$, the two terms are asymptotically independent.

More explicit representations of the two terms and their limiting covariance matrices are outlined below.

$$\sqrt{n} \begin{pmatrix} \mathbb{P}_n f_{0,\eta_0} & - & P f_{0,\eta_0} \\ \mathbb{P}_n f_{1,\eta_0} & - & P f_{1,\eta_0} \\ \mathbb{P}_n f_{2,\eta_0} & - & P f_{2,\eta_0} \end{pmatrix} \to N\left(\mathbf{0}_3, \Sigma_1 \right),$$

$$\sqrt{n} \begin{pmatrix} P f_{0,\eta_n} & - & P f_{0,\eta_0} \\ P f_{1,\eta_n} & - & P f_{1,\eta_0} \\ P f_{2,\eta_n} & - & P f_{2,\eta_0} \end{pmatrix} \to N\left(\mathbf{0}_3, \Sigma_2 \right),$$

where Σ_1 and Σ_2 are defined as follows

$$\Sigma_1 = \begin{pmatrix} \text{Var}\left[f_{0,\eta_0}(M_i)\right] & \text{Cov}\left[f_{0,\eta_0}(M_i), f_{1,\eta_0}(M_j)\right] & \text{Cov}\left[f_{0,\eta_0}(M_i), f_{2,\eta_0}(M_j)\right] \\ \text{Cov}\left[f_{0,\eta_0}(M_i), f_{1,\eta_0}(M_j)\right] & \text{Var}\left[f_{1,\eta_0}(M_i)\right] & \text{Cov}\left[f_{1,\eta_0}(M_i), f_{2,\eta_0}(M_j)\right] \\ \text{Cov}\left[f_{0,\eta_0}(M_i), f_{2,\eta_0}(M_j)\right] & \text{Cov}\left[f_{1,\eta_0}(M_i), f_{2,\eta_0}(M_j)\right] & \text{Var}\left[f_{2,\eta_0}(M_i)\right] \end{pmatrix},$$

$$\Sigma_2 = B\Sigma_{\eta_0}B',$$

$$B = \begin{pmatrix} \frac{\partial}{\partial\beta}Pf_{0,\eta_0} & \frac{\partial}{\partial\Lambda_0(t)}Pf_{0,\eta_0} \\ \frac{\partial}{\partial\beta}Pf_{1,\eta_0} & \frac{\partial}{\partial\Lambda_0(t)}Pf_{1,\eta_0} \\ \frac{\partial}{\partial\beta}Pf_{2,\eta_0} & \frac{\partial}{\partial\Lambda_0(t)}Pf_{2,\eta_0} \end{pmatrix}.$$

Above, Σ_{η_0} is the asymptotic covariance matrix for $(\beta, \Lambda_0(t))$ as derived by Tsiatis (1981), and B is the Jacobian of the map $(\beta, \Lambda_0(t)) \mapsto (Pf_{\theta,\eta} : \theta \in \Theta)$ evaluated at $(\beta, \Lambda_0(t))$. The elements of B are more explicitly shown below:

$$\frac{\partial}{\partial\beta}Pf_{\theta,\eta_0} = \int \exp\left\{\theta\beta m - e^{\beta m}\Lambda_0(t)\right\}\left\{\theta m - e^{\beta m}\Lambda_0(t)m\right\}dF_m,$$
$$\frac{\partial}{\partial\Lambda_0(t)}Pf_{\theta,\eta_0} = -\int \exp\left\{\theta\beta m - e^{\beta m}\Lambda_0(t)\right\}e^{\beta m}dF_m.$$

By asymptotic independence, we thus have

$$\sqrt{n}\begin{pmatrix} \mathbb{P}_n f_{0,\eta_n} & - & Pf_{0,\eta_0} \\ \mathbb{P}_n f_{1,\eta_n} & - & Pf_{1,\eta_0} \\ \mathbb{P}_n f_{2,\eta_n} & - & Pf_{2,\eta_0} \end{pmatrix} \to N\left(\mathbf{0}_3, \Sigma_1 + \Sigma_2\right).$$

Since $HDS(t) = \frac{Pf_{2,\eta_0}Pf_{0,\eta_0}}{\left(Pf_{1,\eta_0}\right)^2}$ and $\widehat{HDS}(t) = \frac{\mathbb{P}_n f_{2,\eta_n}\mathbb{P}_n f_{0,\eta_n}}{\left(\mathbb{P}_n f_{1,\eta_n}\right)^2}$, using the delta method we have

$$\sqrt{n}\left\{\widehat{HDS}(t) - HDS(t)\right\} \to N\left(0, A\left(\Sigma_1 + \Sigma_2\right)A'\right),$$

where A is the Jacobian of the map $(x, y, z) \mapsto \frac{zx}{y}$, evaluated at $(Pf_{0,\eta_0}, Pf_{1,\eta_0}, Pf_{2,\eta_0})$:

$$A = \left(\frac{Pf_{2,\eta_0}}{\left(Pf_{1,\eta_0}\right)^2}, -\frac{2Pf_{2,\eta_0}Pf_{0,\eta_0}}{\left(Pf_{1,\eta_0}\right)^3}, \frac{Pf_{0,\eta_0}}{\left(Pf_{1,\eta_0}\right)}\right).$$

To estimate $A\left(\Sigma_1 + \Sigma_2\right)A'$, we propose using a plug-in estimator $\hat{A}\left(\hat{\Sigma}_1 + \hat{\Sigma}_2\right)\hat{A}'$. We calculate $\hat{\Sigma}_1$ by replacing, in Σ_1, the covariances with sample covariances and η_0 with η_n. Likewise, $\hat{A}$ is just A, with P and η_0 replaced by $\mathbb{P}$ and η_n. We define $\hat{\Sigma}_2 = \hat{B}\hat{\Sigma}_{\eta_0}\hat{B}'$, where

$\hat{\Sigma}_{\eta_0}$ is the covariance estimator proposed by Tsiatis (1981), and $\hat{B}$ is just B with P and η_0 replaced by $\mathbb{P}$ and η_n.

Again using the decomposition in van der Vaart and Wellner (2007), we know that $\mathbb{P}_n f_{\theta,\eta_n} - P f_{\theta,\eta_0} \to 0$, for general f. Since the estimator for the asymptotic variance consists of components that can all be represented in the preceding form (e.g., $\widehat{\mathrm{Cov}}\left[f_{0,\eta_0}(M_i), f_{1,\eta_0}(M_j)\right] = \mathbb{P}_n\left(f_{0,\eta_n} f_{1,\eta_n}\right) - \mathbb{P}_n f_{0,\eta_n} \mathbb{P}_n f_{1,\eta_n}$), by Slutsky's theorem, the estimator for the asymptotic variance is consistent.

2.3.1 Inference for $\widehat{HDS}^{LC}(t)$

If we relax the proportional hazards assumption to be local in time so that $\lambda(t|M) = \lambda_0(t)\exp\{\beta'(t)M\}$, asymptotic normality of $\widehat{HDS}^{LC}(t)$ is shown using arguments similar to those for $\widehat{HDS}(t)$. Briefly, we define f_{θ,η_0} and f_{θ,η_n} using the local constant versions of β and $\Lambda_0(t)$ as follows:

$$f_{\theta,\eta_0}(m) = \exp\left\{\theta\beta(t)m - e^{\beta m}\Lambda_0(t)\right\},$$
$$f_{\theta,\eta_n}(m) = \exp\left\{\theta\hat{\beta}_h(t)m - e^{\hat{\beta}_h(t)m}\hat{\Lambda}_h(t)\right\}.$$

Above, $\theta \in \{0,1,2\}$. $\hat{\beta}_h(t)$ and $\hat{\Lambda}_h(t)$ are local constant estimators of β and $\Lambda_0(t)$ proposed by Cai and Sun (2003) and Tian et al. (2005), respectively. We let h be a smoothing parameter that is needed for estimating $\hat{\beta}_h(t)$ and $\hat{\Lambda}_h(t)$. As shown by Tian et al. (2005), we must have $h = O(n^{-v})$ with $1/4 < v < 1/2$. While Cai and Sun (2003) suggest $h = O(n^{-1/5})$ for purposes of minizing the mean integrated square error of $\hat{\beta}_h(t)$, Tian et al. (2005) show that "undersmoothing" is necessary to ensure $\sqrt{n}$-consistency of $\hat{\Lambda}_h(t)$.

Using notation from van der Vaart and Wellner (2007), and letting $\widehat{HDS}^{LC}(t) = \frac{f_{2,\eta_n} f_{0,\eta_n}}{f_{1,\eta_n}^2}$ and $HDS(t) = \frac{f_{2,\eta_0} f_{0,\eta_0}}{f_{1,\eta_0}^2}$, we have the following decomposition for each θ:

$$\sqrt{nh}\left(\mathbb{P}_n f_{\theta,\eta_n} - P f_{\theta,\eta_0}\right) = o_p(1) + \sqrt{h}\mathbb{G}_n f_{\theta,\eta_0} + \sqrt{nh} P\left(f_{\theta,\eta_n} - f_{\theta,\eta_0}\right).$$

The variability from the second term above comes from the empirical distribution of M and thus the second term is $o_p(1)$. The variance from the third term comes from $\hat{\beta}_h(t)$ and $\hat{\Lambda}_h(t)$; the former is $\sqrt{nh}$-consistent (Cai and Sun, 2003) while the latter is $\sqrt{n}$-consistent (Tian et al., 2005). Thus the asymptotic variance for the third term is dominated by the

variation in $\hat{\beta}_h(t)$. Applying the delta method in turn shows that $\widehat{HDS}^{LC}(t)$ is asymptotically normal at $\sqrt{nh}$ rate and its asymptotic standard errors can be approximated using a function of the standard error for $\hat{\beta}_h(t)$.

2.3.2 Standard errors for ratio of two $\widehat{HDS}(t)$ functions

Assuming $HDS_{new}(t) \neq HDS_{old}(t)$, it is straightforward to show that $\widehat{HDS}_{new}(t)/\widehat{HDS}_{old}(t)$ is asymptotically normal. However, deriving the asymptotic covariance is more complicated. Fortunately, Hjort (1985) showed that the bootstrap is asymptotically valid for $\hat{\beta}$ and $\hat{\Lambda}_0(t)$ from the Cox model. Since $\widehat{HDS}(t)$ is a function of $\hat{\beta}$, $\hat{\Lambda}_0(t)$, and the empirical distribution function of M, the result from Hjort (1985) combined with results from van der Vaart and Wellner (1996, Theorems 3.6.2 and 3.9.11) imply that the bootstrap can be used to obtain asymptotically valid standard errors for $\widehat{HDS}_{new}(t)/\widehat{HDS}_{old}(t)$.

2.4 Simulations

To assess $\widehat{HDS}(t)$ and $\widehat{HDS}^{LC}(t)$, and their standard error estimators, we will generate data under two scenarios: 1) when a Cox proportional hazards model is true; and 2) when a Cox proportional hazards model with time-varying $\beta(t)$ is true.

2.4.1 Coverage under Cox model

For our example where the data is generated from a Cox proportional hazards model, we used a Weibull model with two covariates. The covariates were each drawn from independent $Unif(0,2)$ distributions, and there is uninformtive censoring of roughly 25% of the outcomes for each dataset. It is also straightforward to derive a closed form expression for the true underlying $HDS(t)$, and calculate it numerically. The hazard function and covariate distributions are specified as

$$\begin{aligned} \lambda(t|M) &= 0.5 \cdot \exp\left(\beta_1 \cdot M_1 + \beta_2 \cdot M_2\right), \\ M_1 &\sim Unif(0,2), \\ M_2 &\sim Unif(0,2). \end{aligned}$$

Table 2.1: Pointwise coverage results for $\widehat{HDS}(t)$ 95% confidence intervals using analytic estimator. Results are shown for two different Cox model data generators: one where $\beta_1 = \beta_2 = 1$ and a second where $\beta_1 = 0.5, \beta_2 = 1.5$. Row one results are shown in Figure 2.2.

	time											
(β_1, β_2)	0.025	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$(1.0, 1.0)$	0.928	0.924	0.916	0.916	0.925	0.933	0.938	0.944	0.952	0.957	0.963	0.960
$(0.5, 1.5)$	0.948	0.945	0.935	0.915	0.928	0.935	0.948	0.963	0.967	0.970	0.968	0.981

We considered two different models: one where $\beta_1 = \beta_2 = 1$ and second where $\beta_1 = 0.5, \beta_2 = 1.5$.

For each model we simulated 1000 datasets (each $n = 500$) and for each dataset calculated $\widehat{HDS}(t)$ and the 95% confidence interval estimate outlined in Section 2.3. The pointwise coverage rates of nominal 95% confidence intervals are reported in Table 2.1. Results from the model where $\beta_1 = \beta_2 = 1$ is shown in Figure 2.2. The graph in Figure 2.2 also shows the true $HDS(t)$ along with $\widehat{HDS}(t)$ from 250 datasets to provide some empirical intuition of the estimator's small sample behavior.

Overall, for both models, the coverage rates are quite satisfactory, with under-coverage early on and slight over-coverage at later times.

Bootstrap confidence intervals tended to outperform analytic confidence intervals by achieving coverage rates closer to 95%, though were much more computationally intensive. Thus if computational cost is not an issue, we recommend using bootstrap confidence intervals. Otherwise, the analytic standard errors still offer reasonable confidence intervals at a much lower computational cost. Pointwise coverage rates of percentile bootstrap 95% confidence intervals (using 1000 datasets, each $n = 500$ and 500 bootstrap replications) are reported in Table 2.2.

2.4.2 Coverage under local constant Cox model

For our example where the data is generated from a Cox proportional hazards model with time-varying covariates $\beta(t)$, we used a slightly modified version of a model specified by Cai

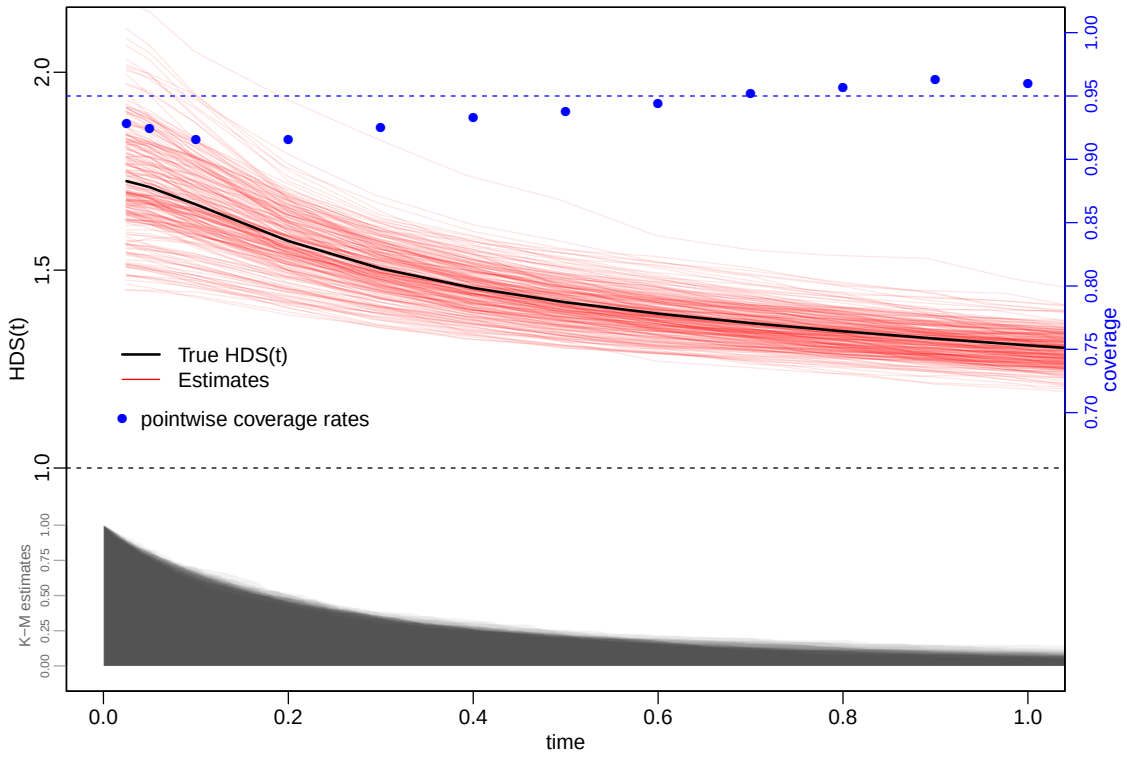

Figure 2.2: Results from 1000 simulations (each $n = 500$) are shown above. Data for each simulation is generated from a Cox model. The true underlying $HDS(t)$ is shown by the black line. A random sample of 250 $\widehat{HDS}(t)$ estimates (without confidence intervals) are overlayed in red. The pointwise coverage rates at various time points for the 95% confidence intervals are shown in blue. At the bottom of the graph are overplotted K-M estimates from 20 random datasets, to illustrate roughly how much data is available for estimation at each time.

Table 2.2: Pointwise coverage results for $\widehat{HDS}(t)$ 95% confidence intervals using percentile bootstrap estimator. Results are shown for two different Cox model data generators: one where $\beta_1 = \beta_2 = 1$ and a second where $\beta_1 = 0.5, \beta_2 = 1.5$. Row one results are shown in Figure 2.2.

	time											
(β_1, β_2)	0.025	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$(1.0, 1.0)$	0.944	0.945	0.938	0.935	0.932	0.934	0.935	0.939	0.945	0.940	0.942	0.943
$(0.5, 1.5)$	0.948	0.946	0.948	0.942	0.946	0.942	0.938	0.942	0.942	0.940	0.945	0.944

Table 2.3: Pointwise coverage results for $\widehat{HDS}^{LC}(t)$ and $\widehat{HDS}(t)$ 95% confidence intervals using analytic estimators. Data generated from proportional hazards model with time-varying coefficients. A visual aide corresponding to the results is shown in Figure 2.3.

	time											
Estimator	0.025	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	1.0	1.25	1.5
$\widehat{HDS}^{LC}(t)$	0.910	0.908	0.918	0.924	0.933	0.934	0.927	0.907	0.899	0.929	0.944	0.958
$\widehat{HDS}(t)$	0.971	0.968	0.963	0.962	0.955	0.930	0.811	0.565	0.265	0.005	0.013	0.103

and Sun (2003) in one of their simulated examples. The model consists of two independent covariates, where the effect of one varies linearly with time.

$$\lambda(t|M) = \exp\left(t \cdot M_1 + 0.5 \cdot M_2\right)$$
$$M_1 \sim Unif(1,3)$$
$$M_2 \sim N(0,1)$$

We simulated 1000 datasets using the above model (each $n = 500$) and for each dataset calculated $\widehat{HDS}^{LC}(t)$ (a fixed bandwidth of 0.4 was used for all calculations) and the 95% confidence interval estimate outlined in Section 2.3.1. For comparison, we did the same with $\widehat{HDS}(t)$, which is based on a misspecified model assuming time-invariant coefficients.

The pointwise coverage rates of nominal 95% confidence intervals are reported in Table 2.3, and in Figure 2.3. The graph in Figure 2.3 also shows the true $HDS(t)$ along with $\widehat{HDS}^{LC}(t)$ from 100 datasets to provide some empirical intuition of the estimator's small sample behavior. Samples of the misspecified $\widehat{HDS}(t)$ are also shown for comparison.

Overall, coverage rates for $\widehat{HDS}^{LC}(t)$ are satisfactory across most of the support times. In contrast, while $\widehat{HDS}(t)$ estimates display less variance (unsurprising as it is a more parametric estimator), coverages rates drop steeply below 80% after $t = 0.5$, largely due to the heavily biased estimates from model misspecification.

Just as in Section 2.4.1, bootstrap confidence intervals tended to outperform analytic confidence intervals by achieving coverage rates closer to 95%, though were much more computationally intensive. Thus if computational cost is not an issue, we recommend using

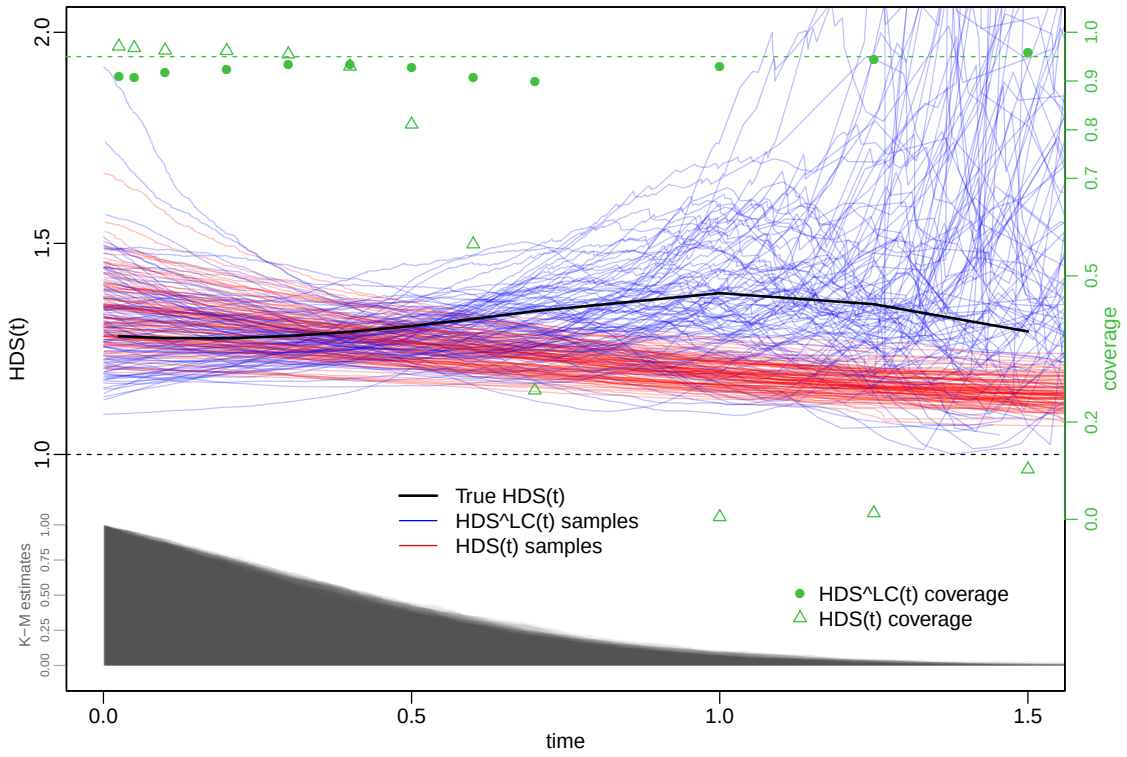

Figure 2.3: Results from 1000 simulations (each $n = 500$) are shown above. Data for each simulation is generated from a proportional hazards model with time-varying coefficients. The true underlying $HDS(t)$ is shown by the black line. A random sample of 100 $\widehat{HDS}^{LC}(t)$ estimates (without confidence intervals) are overlayed in blue. For comparison, a random sample of 100 $\widehat{HDS}(t)$ estimates are shown in red. The pointwise coverage rates at various time points for the 95% confidence intervals are shown in green (circles for $\widehat{HDS}^{LC}(t)$ and triangles for $\widehat{HDS}(t)$. At the bottom of the graph are overplotted K-M estimates from 20 random datasets, to illustrate roughly how much data is available for estimation at each time.

Table 2.4: Pointwise coverage results for $\widehat{HDS}^{LC}(t)$ 95% confidence intervals using percentile bootstrap estimators. Data generated from proportional hazards model with time-varying coefficients.

	time											
Estimator	0.025	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	1.0	1.25	1.5
$\widehat{HDS}^{LC}(t)$	0.943	0.942	0.941	0.951	0.952	0.956	0.949	0.952	0.943	0.932	0.960	0.976

bootstrap confidence intervals. Otherwise, the analytic standard errors still offer reasonable confidence intervals at a much lower computational cost. Pointwise coverage rates of percentile bootstrap 95% confidence intervals (using 1000 datasets, each $n = 500$ and 500 bootstrap replications) are reported in Table 2.4.

2.5 Examples

Mayo PBC data We first use the Mayo PBC data to illustrate $HDS(t)$ on a benchmark dataset. See Section 1.5.3 for a detailed overview of the dataset.

Using all five predictors and assuming proportional hazards, we can estimate $HDS(t)$ using methods detailed in Section 2.2.2. We can also relax the proportional hazards to be local-in-time (bandwidth of 730 days) and again estimate $HDS(t)$. The results for both are shown in Figure 2.4. The discriminatory ability of the five predictors is strong for the duration of follow-up, but particularly so during the first two years.

In particular, for the first two years, the five predictors assign 5-10 times more risk on cases than controls. For later times the ratio decreases but is still substantial and greater than 2. We also see that in this example, qualitatively, we would likely draw similar conclusions using either the Cox model estimator or the local estimator.

Multiple myeloma trial S9321 data To illustrate $HDS(t)$ as a tool for comparing time-varying performance, we use the multiple myeloma dataset from the Southwest Oncology Group (SWOG) trial S9321. See Section 1.5.3 for a detailed overview of the dataset.

To more formally assess the time-varying predictive performance of each marker, we calculate $HDS(t)$ using each individual marker. We also assess the combined time-varying

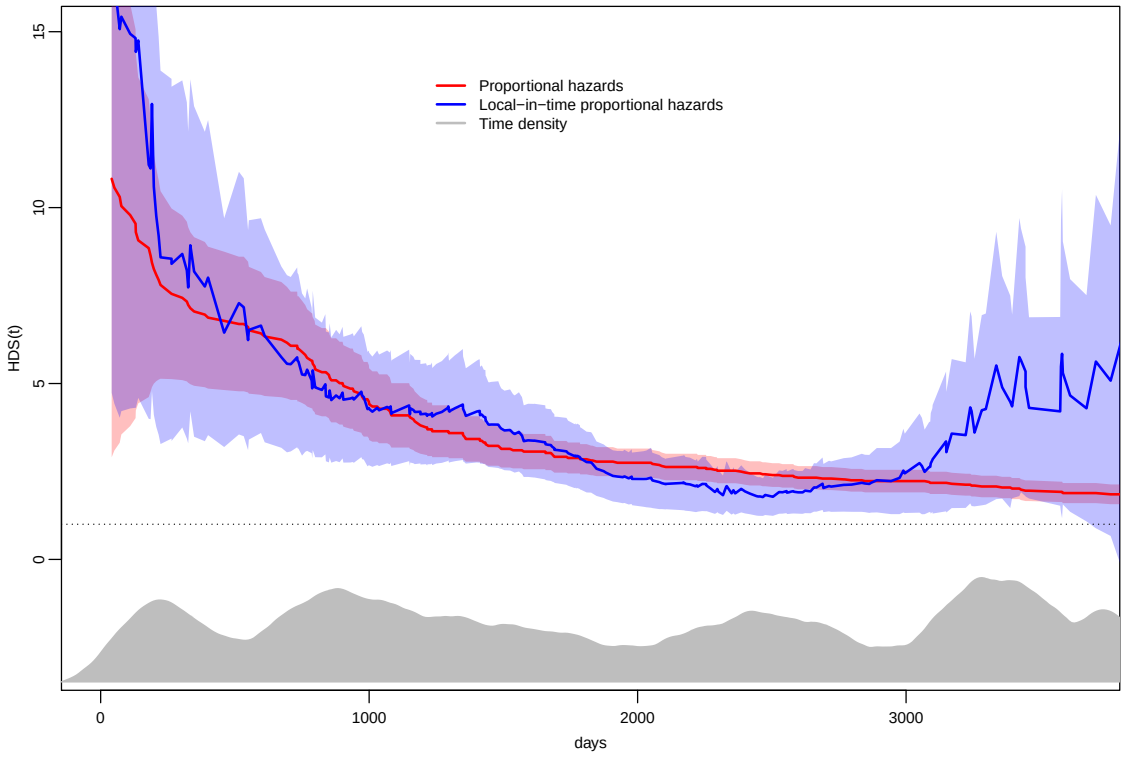

Figure 2.4: $HDS(t)$ estimates using the proportional hazards assumption and the relaxed local-in-time (bandwidth = 730 days) proportional hazards assumption, with pointwise 95% confidence intervals. Truncated marginal time density (estimated by using a kernel-smoothed Kaplan-Meier estimator) shown at the bottom.

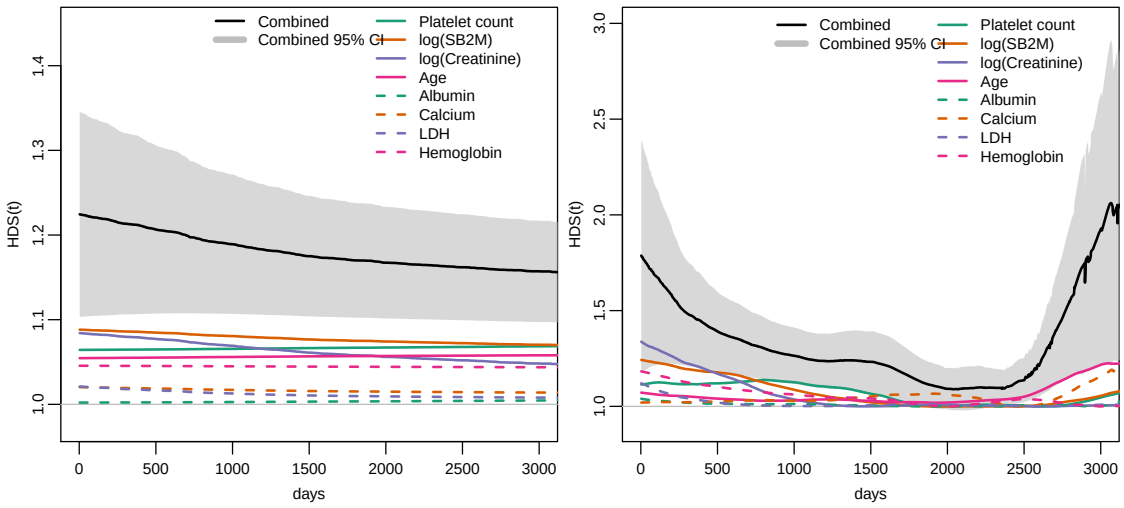

Figure 2.5: $HDS(t)$ for eight markers - individually and then combined. The grey areas represents pointwise 95% confidence interval for $HDS(t)$ using all eight markers. **Left**: under proportional hazards assumption. **Right**: under local-in-time proportional hazards assumption (window width: 730 days). Note the different scales on the y-axes.

predictive performance of all eight markers. We performed two sets of analyses - one under the proportional hazards assumption, and one under the local-in-time proportional hazards assumption. The results are shown in Figure 2.5.

Estimates of $HDS(t)$ under the proportional hazards assumption for individual markers are all fairly close to one. The markers log(SB2M), log(creatinine), and platelet count appear the most predictive relative to others but still do not offer substantial time-varying discriminatory ability. However, combining the multiple weakly predictive markers yields noticeably improved discrimination across all times as $HDS(t)$ is approximately 1.2 over time. Nevertheless, even the multivariate model only assigns about 20% more risk on cases as compared to controls.

As suggested by the estimates of $\beta(t)$, however, the proportional hazards assumption likely does not hold. If we relax the proportional hazards assumption to be local, the $HDS(t)$ estimates suggest that while some individual markers are still weak, log(SB2M) and log(creatinine) offer a moderate amount of discriminatory value for earlier events. The latter two markers have $HDS(t)$ values closer to around 1.3 early on, before dropping to

be closer to one at around 1000 days. Furthermore, $HDS(t)$ for all markers starts off at 2 before dropping off to around 1.5. Thus we see, unlike in the Mayo PBC example, that estimating $HDS(t)$ after relaxing the proportional hazards assumption can lead to qualitatively different results.

Finally, we use the myeloma data set to illustrate the graphical connection between $HDS(t)$ and the partial likelihood outlined in Section 2.2.4. Figure 2.6 shows a scatterplot of scaled partial likelihood contributions versus time, $\left\{\left(y_i, \frac{\exp(\hat{\beta}m_i)}{(1/n_i)\sum_{j\in R_i}\exp(\hat{\beta}m_j)}\right) : t_i < c_i\right\}$, where m is the vector of all eight markers. The smoothed scatterplot line and $\widehat{HDS}(t)$ estimated using formula (2.2) are both shown as well, and the two summary curves are in close agreement.

2.6 Discussion

We have proposed a time-varying measure of discrimination for survival models, $HDS(t)$, that naturally generalizes the discrimination slope for binary data for application to incident events. For any time t, $HDS(t)$ measures, on the incident risk scale, how far apart those subjects with an event are from those who remain event-free. $HDS(t)$ is thus a natural complement to existing measures of discrimination for survival models and is particularly suited for use with hazard models. With our incident risk-based method, it is now possible to choose between the two common survival risk scales of interest; cumulative risk and incident risk. Also, while rank-based methods such as time-varying AUC curves (Heagerty et al., 2000; Heagerty and Zheng, 2005) are invariant to monotone transformations of the marker, they do not provide an assessment of the magnitude of the difference in risk between cases and controls, and the use of a risk-based measure may provide more information than simply using ranks.

Although we mention properties of $HDS(t)$ relative to rank-based methods, there is an additional difference between the standard approaches to estimation for these summaries. When measuring the prognostic accuracy of multiple predictors with time-dependent AUC curves, there is an explicit "decoupling" of model generation from model evaluation. Specifically, since the AUC is only defined when M is univariate, a model must first be used to

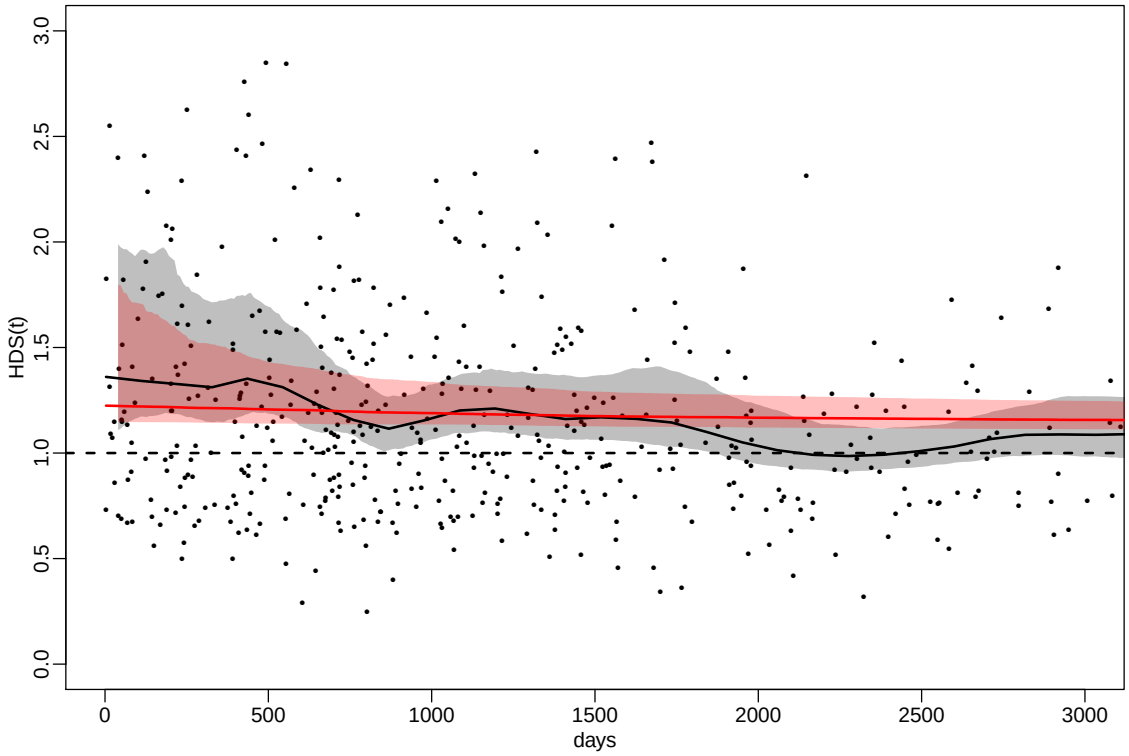

Figure 2.6: Each dot corresponds to an event i and is equal to $\frac{\exp(\hat{\beta}m_i)}{(1/n_i)\sum_{j\in R_i}\exp(\hat{\beta}m_j)}$, which is the ith partial likelihood contribution (evaluated at $\hat{\beta}$) scaled by the size of the risk set. The red line is $\widehat{HDS}(t)$, calculated using (2.2). The black line is a smoother through the scaled partial likelihood contributions. The bands are 95% confidence intervals using the percentile bootstrap estimator.

combine the predictors into a univariate prognostic score, and only then is the prognostic value of the score evaluated using AUC. However, $HDS(t)$ is well-defined even when M is multidimensional, so one can forego the task of using a model to first generate a prognostic score from the multiple predictors. Of course, this requires either the proportional hazards assumption (or its relaxed version) to hold if the subsequent estimate of $HDS(t)$ is to have a meaningful interpretation. If one does not believe either assumption holds, it is then advisable to use a separate model to combine the multiple predictors into a prognostic score, and then evaluate the prognostic accuracy of the score with $HDS(t)$.

Chapter 3

HAZARD DISCRIMINATION SUMMARY WITH LONGITUDINAL MARKERS

In the previous chapter we introduced $HDS(t)$ as a measure of prognostic accuracy, and gave examples where it could be used to assess baseline markers. In practice there are often situations where markers are updated over time (e.g. blood work being updated at each doctor's visit or wearable devices continuously taking biometric measurements) and thus it can be important to quantify the predictive value of time-varying markers.

Existing measures of predictive performance for time-varying markers can roughly be put in two categories: 1) partially conditional measures; and 2) incident measures. In this chapter we define "partially conditional" and give an overview of the existing measures of predictive performance. The main contribution of this chapter is to show that the incident measure $HDS(t)$ is still well-defined and has a useful scientific interpretation when replacing M with $M(t)$. We will also detail the technical steps needed to extend $HDS(t)$ to accomodate time-varying markers, and illustrate $HDS(t)$ using both simulated and study data.

3.1 Background

In this section we review three partially conditional measures and one incident measure of predictive performance. The partially conditional measures are an R^2-type extension, a Brier score extension, and an $AUC(t)$ extension. The incident measure is $AUC^{I/D}(t)$, which was covered in Section 1.5.1, and is amenable to evaluating time-varying markers.

The partially conditional measures are all extensions of existing methods for evaluating survival models with time-invariant covariates, which are in turn extensions of methods for evaluting binary outcome models. Additionally, the partially conditional measures incorporate the temporal aspects of the covariates in similar ways: an additional earlier time t_1 is specified; comparisons are then made conditional on survivors past t_1, and on covariate

measurements available through t_1. The conditioning procedure is analogous to a scenario where we have observed a cohort up until a specified time t_1, at which point we want to assess future predictions made using currently available data.

In contrast to partially conditional measures, which require the conditioning procedure described above to accomodate time-varying markers, the incident measure $AUC^{I/D}(t)$ is already well-defined (and interpretable) for both time-varying and time-invariant markers. As an incident measure, $HDS(t)$ shares the same property and is the focus of this chapter. For the sake of completeness we review both partially conditional and incident measures.

R^2-type measure Schemper and Henderson (2000) proposed two R^2-type measures of predictive accuracy for survival models with time-invariant covariates. For binary outcomes, R^2 can be seen as the ratio of the variance of the predicted values to the variance of the outcomes. With survival outcomes, for each t we can consider the variance of predicted probabilities (of surviving past t) and also the variance of observed statuses (at time t). The first measure is a ratio of the weighted time-average of each of the variances, while the second is a weighted time-average of the ratio of the variances. The measures are formally defined as

$$V(\tau) = 1 - \frac{\int_0^\tau \mathrm{E}_M \left[S(t|M)\{1 - S(t|M)\}\right] f(t)dt}{\int_0^\tau S(t)\{1 - S(t)\} f(t)dt},$$
$$V_W(\tau) = \int_0^\tau \left(1 - \frac{\mathrm{E}_X \left[S(t|M)\{1 - S(t|M)\}\right]}{S(t)\{1 - S(t)\}}\right) f(t)dt \times \left\{\int_0^\tau f(t)dt\right\}^{-1}.$$

Henderson et al. (2002) then proposed partially conditional extensions of $V(\tau)$ and $V_W(\tau)$ to allow for time-varying covariates by conditioning on a second time τ^*, where $\tau > \tau^*$. The new measures still compare predicted probabilities with observed statuses, with the following modifications: the predicted probabilities are based on covariate history up through τ^* and conditional on surviving past τ^*; and observed statuses are just among those still surviving at τ^*. While Henderson et al. (2002) do not explicitly define the parameters of interest, we

assume the following definitions:

$$V(\tau;\tau^*) = 1 - \frac{\int_{\tau^*}^{\tau} \mathrm{E}_{M_{\tau^*}}\left[S(t|M_{\tau^*}, T>\tau^*)\{1-S(t|M_{\tau^*}, T>\tau^*)\}\right] f(t|T>\tau^*)dt}{\int_{\tau^*}^{\tau} S(t|T>\tau^*)\{1-S(t|T>\tau^*)\} f(t|T>\tau^*)dt},$$

$$V_W(\tau;\tau^*) = \int_{\tau^*}^{\tau} \left(1 - \frac{\mathrm{E}_{M_{\tau^*}}\left[S(t|M_{\tau^*}, T>\tau^*)\{1-S(t|M_{\tau^*}, T>\tau^*)\}\right]}{S(t|T>\tau^*)\{1-S(t|T>\tau^*)\}}\right) f(t|T>\tau^*)dt$$
$$\times \left\{\int_{\tau^*}^{\tau} f(t|T>\tau^*)dt\right\}^{-1}.$$

There also exists a connection between $DS(t)$ and $V_W(t)$. Specifically, $DS(t)$ can be formulated as a generalization of R^2 for binary outcomes (where the binary outcome is defined as experiencing an event before t). Pepe et al. (2008) and Chambless et al. (2011) showed that

$$DS(t) = \frac{Var\{S(t|M)\}}{S(t)\{1-S(t)\}}.$$

Thus $V_W(t)$ can also be interpreted as a weighted integral of $DS(t)$ over time.

Brier score (quadratic loss) The Brier score (Brier, 1950) is a common method for assessing risk predictions for binary outcomes. It is the average squared deviation of the observed outcome (0 or 1) from the risk prediction, and is formally defined as

$$BS = \mathrm{E}\left[\left\{Y - \hat{P}\left(Y=1|M\right)\right\}^2\right].$$

The Brier score can be adapted to survival models by instead defining it as, for any time t, the average squared deviation between true survival status at t and the predicted probability (Graf et al., 1999; Gerds and Schumacher, 2006):

$$BS(t) = \mathrm{E}\left[\left\{1\left(T>t\right) - \hat{P}\left(T>t|M\right)\right\}^2\right].$$

Schoop et al. (2008) adapted the above loss function for time-varying covariates by conditioning on an additional time s, which is similar to the partially conditional approach Henderson et al. (2002) took to extend R^2 for time-varying covariates:

$$BS(t;s) = \mathrm{E}\left[\left\{1\left(T>t\right) - \hat{P}\left(T>t|T>s, M_s\right)\right\}^2 \;\middle|\; T>s\right].$$

Above, $t > s$, and M_s is all the historical covariate information up until time s. With time-invariant covariates, $s = 0$. This extension is similar to the R^2 extension in the sense that we are conditioning on a second earlier time (here, s instead of τ^*), and predicted probabilities are calculated based on marker information up through this earlier time s. In practice, the predicted probabilities are difficult to calculate. Schoop et al. (2008) assumes that predicted probabilities are already available, and point the reader to Tsiatis and Davidian (2004) for joint modeling approaches and Zheng and Heagerty (2005) for a semiparametric partly conditional approach.

ROC methods In Chapter 1 we reviewed ROC and AUC methods for binary outcome models. There are two extensions into survival models that can also accomodate time-varying covariates. The first is $AUC^{I/D}(t)$ (Heagerty and Zheng, 2005) , or a method for measuring how well incident cases are separated from dynamic controls (see Figure 1.4(b) for a visual definition). Sensitivity and specificity at time t and cutpoint c and corresponding AUC are defined as

$$\begin{aligned}
\text{Sensitivity}^I(c \mid t) &= P\{M(s) > c \mid T = t\}, \\
\text{Specificity}^D(c \mid t) &= P\{M(s) \leq c \mid T > t\}, \\
AUC^{I/D}(t) &= P\{M_i(t) > M_j(t) | T_i = t, T_j > t\}.
\end{aligned}$$

Another extension, which we will refer to as $AUC^{Int/D}(t)$ (Zheng and Heagerty, 2007), measures how well interval cases are separated from dynamic controls (see Figure 1.4(c) for a visual definition). It can also be thought of as a partially conditional generalization of $AUC^{C/D}(t)$. The definitions for interval/dynamic sensitivity, specificity and AUC are shown below.

$$\begin{aligned}
\text{Sensitivity}^{Int}(c \mid \text{start} = t_1, \text{stop} = t_2) &= P\{M(t_1) > c \mid T \geq t_1, T \leq t_2\}, \\
\text{Specificity}^D(c \mid \text{start} = t_1, \text{stop} = t_2) &= P\{M(t_1) \leq c \mid T \geq t_1, T > t_2\}, \\
AUC^{Int/D}(t) &= P\left\{M_i(t_1) > M_j(t_1) | T_i \in [t_1, t_2], T_j > t_2\right\}.
\end{aligned}$$

Above, $t_1 < t_2$. The interval/dynamic generalization of AUC is similar to the R^2 and Brier score in the sense that a second earlier time t_1 (τ^* and s in the other examples) is

used to specify the amount of covariate history used for evaluating prognostic performance. We will see that our generalization of $HDS(t)$ is dissimilar in this regard, and more comparable to $AUC^{I/D}(t)$. A common feature of $AUC^{I/D}(t)$ and $HDS(t)$ is that they are more amenable to being averaged over time to create more concise performance summaries.

3.2 Parameter of interest

Our parameter of interest is similar to $HDS(t)$ for baseline covariates, with the main difference being that M is replaced with $M(t)$:

$$HDS(t) = \frac{\mathrm{E}\left[\lambda\left\{t|M(t)\right\} \mid T = t\right]}{\mathrm{E}\left[\lambda\left\{t|M(t)\right\} \mid T > t\right]}.$$

$HDS(t)$ as the ratio of mean case risk to mean control risk We interpret $\lambda\{t|M(t)\}$ as "the hazard rate at t, conditional on the marker information available from baseline through time t". Note that the Cox model assumes that of the historical information available, only the measurement at t has an impact on the risk.

We can in turn interpret $HDS(t)$ as a measure of how well covariate information up through time t discriminates between incident cases ($T = t$) and dynamic controls ($T > t$) at time t.

$HDS(t)$ as the prognostic value of $M(t)$ Analogously to Section 2.1.2, the denominator of $HDS(t)$ can be shown to be equal to the marginal hazard:

$$\mathrm{E}\left[\ \lambda\left\{t|M(t)\right\} \mid T > t\ \right] = \lambda(t). \tag{3.1}$$

The above relation then implies the following mathematically equivalent representation of $HDS(t)$:

$$HDS(t) = \mathrm{E}\left[\ \frac{\lambda\{t|M(t)\}}{\lambda(t)}\ \middle|\ T = t\ \right]. \tag{3.2}$$

Thus an alternative interpretation for $HDS(t)$ is as a measure of how much better (as a multiplicative factor) our risk predictions become if we choose to use the covariate information available to us through time t.

3.3 Estimation

The $HDS(t)$ estimator for baseline covariates involves estimating the conditional survival function. This can be problematic when time-varying covariate are involved (Fisher and Lin, 1999), so we propose an estimation method that does not require estimating the conditional survival function. However, one downside to this method is that it requires the stronger assumption of random censoring (whereas our esimator from Section 2.2 for baseline covariates only requires that T be conditionally independent of M given C). Further, we also make the assumption that $M(t)$ is an external covariate, in the sense that the value of $M(t)$ is independent of events occurring prior to t.

Derivation of estimator We first note that $HDS(t)$ can be rewritten as a function of just $\lambda\{t|M(t)\}$ and the marker distribution in the risk set at t:

$$HDS(t) = \frac{\mathrm{E}\left[\lambda^2\left\{t|M(t)\right\}|T>t\right]}{\mathrm{E}^2\left[\lambda\left\{t|M(t)\right\}|T>t\right]}. \tag{3.3}$$

The derivation of (3.3) relies on property (3.1) and the following property of the numerator of $HDS(t)$:

$$\mathrm{E}\left[\lambda\left\{t|M(t)\right\}|T=t\right] = \frac{1}{\lambda(t)}\mathrm{E}\left[\lambda\left\{t|M(t)\right\}|T>t\right].$$

Details of this property are derived below, where m_t runs through the support of $M(t)$ at time t:

$$\begin{aligned}
\mathrm{E}\left[\lambda\left\{t|M(t)\right\}|T=t\right] &= \int \lambda\left\{t|m(t)\right\} f\left\{m(t)|T=t\right\} dm_t \\
&= \int \lambda\left\{t|m(t)\right\} \frac{f\left\{t|m(t)\right\} f\left\{m(t)\right\}}{f(t)} dm_t \\
&= \int \lambda\left\{t|m(t)\right\} \frac{f\left\{t|m(t)\right\}}{S\left\{t|m(t)\right\}} \frac{f\left\{m(t)\right\}}{f(t)} S\left\{t|m(t)\right\} dm_t \\
&= \int \lambda^2\left\{t|m(t)\right\} \frac{S(t)}{f(t)} f\left\{m(t)|T>t\right\} dm_t \\
&= \frac{1}{\lambda(t)} \int \lambda^2\left\{t|m(t)\right\} f\left\{m(t)|T>t\right\} dm_t \\
&= \frac{1}{\lambda(t)}\mathrm{E}\left[\lambda\left\{t|M(t)\right\}|T>t\right].
\end{aligned}$$

From (3.3) we can further expand $HDS(t)$ as follows:

$$\begin{aligned} HDS(t) &= \frac{\mathrm{E}\left[\lambda^2\left\{t|M(t)\right\}|t \geq T\right]}{\mathrm{E}^2\left[\lambda\left\{t|M(t)\right\}|t \geq T\right]} \\ &= \frac{\int \lambda^2\left\{t|M(t)\right\} \frac{1}{P(T>t)} dH\left\{M(t),t\right\}}{\left[\int \lambda\left\{t|M(t)\right\} \frac{1}{P(T>t)} dH\left\{M(t),t\right\}\right]^2}, \end{aligned}$$

where $H(M(t),t)$ is the joint probability function $P\left\{M(t) \leq m(t), T > t\right\}$. Define C as the censoring variable and $Y = min(T,C)$. Note that $P(Y > t) = P(T > t)P(C > t)$. Thus under independent censoring and plugging into the above we have

$$\begin{aligned} HDS(t) &= \frac{\frac{1}{P(Y>t)} \int \lambda^2\left\{t|M(t)\right\} dG\left\{M(t),t\right\}}{\left[\frac{1}{P(Y>t)} \int \lambda\left\{t|M(t)\right\} dG\left\{M(t),t\right\}\right]^2} \\ &= \frac{\frac{1}{P(Y>t)} \int \exp\left\{2\beta M(t)\right\} dG\left\{M(t),t\right\}}{\left[\frac{1}{P(Y>t)} \int \exp\left\{\beta M(t)\right\} dG\left\{M(t),t\right\}\right]^2}, \end{aligned}$$

where $G(M(t),t)$ is the joint probability function $P(M(t) \leq m(t), Y > t)$. The latter equality comes from applying the proportional hazards assumption, which then leads to the estimator

$$\begin{aligned} \widehat{HDS}^R(t) &= \frac{\frac{1}{R(t)} \sum_{i:t_i>t} \hat{\lambda}^2\left\{t|m_i(t)\right\}}{\left[\frac{1}{R(t)} \sum_{i:t_i>t} \hat{\lambda}\left\{t|m_i(t)\right\}\right]^2} \\ &= \hat{P}\left(Y > t\right) \frac{\frac{1}{n} \sum_{i:t_i>t} \exp\left\{2\hat{\beta} m_i(t)\right\}}{\left[\frac{1}{n} \sum_{i:t_i>t} \exp\left\{\hat{\beta} m_i(t)\right\}\right]^2}. \end{aligned} \tag{3.4}$$

$\hat{\beta}$ is the usual estimator from maximizing the partial likelihood, and $\hat{P}(Y > t)$ is the empirical estimator.

Using this method, large jumps may occur as t changes. As $\widehat{HDS}^R(t)$ is calculated by just using values in the risk set at t, a large jump may occur when a large linear predictor (relative to the other in the risk set) enters the risk set. Similarly, a large drop may occur when a large linear predictor exits the risk set. This can be particularly apparent in situations where covariates jump by a large amount after being updated. An example of this occurring will be shown in section 3.6. If one is more interested in broader time-trends than local variations, a solution is to pre-process the covariates by performing a simple linear

interpolation of the covariates before calculating $\widehat{HDS}^R(t)$. Another is to use the following alternate estimator based on smoothing scaled partial likelihood contributions.

Alternate estimator Equation (3.2) suggests that regressing scaled partial likelihood contributions on time, as suggested in Section 2.2.4, still gives a reasonable estimator for $HDS(t)$. The kernel regression line through the partial likelihood contributions is defined as

$$\widehat{HDS}^{PL}(t) = \frac{1}{nh} \sum_{i:\delta_i=1} S_i K\left(\frac{t - y_i}{h}\right), \tag{3.5}$$

where δ_i is the censoring indicator for observation i, where 1 indicates no censoring; $K(u)$ is a kernel function such as the Epanechnikov or Gaussian kernel. S_i is the scaled partial likehood contribution defined as

$$S_i = \frac{\exp\left\{\hat{\beta}' m_i(t_i)\right\}}{(1/n_i)\sum_{j \in R_i} \exp\left\{\hat{\beta}' m_j(t_i)\right\}},$$

where R_i is the risk set at time y_i and n_i is the number of observations in R_i. Note that if n_i were omitted, S_i would be exactly the partial likelihood contribution for observation i.

3.4 Inference

We can show that $\widehat{HDS}^R(t)$ is asymptotically normal using a procedure similar to the asymptotic derivations for $\widehat{HDS}(t)$ in Section 2.3. In particular, we use results from van der Vaart and Wellner (2007) to show asymptotic normality in an analogous way. While analytic standard error estimates do appear feasible in theory, we have found the coverage in finite samples to be poor. The analytic standard error estimators are based on approximations of asymptotic results. Thus, the parts of the decomposition which are asymptotically zero are not estimated. It may be the case that they do not approach zero quickly enough to be practically ignorable for smaller samples. It will require further investigation to find an analytic confidence interval estimator with better small sample performance.

We thus suggest using the percentile bootstrap to generate confidence intervals. Validity of the percentile bootstrap follow in a way analogous to the procedure from Section 2.3.2.

Specifically, results from Hjort (1985) and van der Vaart and Wellner (1996, Theorems 3.6.2 and 3.9.11), combined with the asymptotic normality of $\widehat{HDS}^R(t)$ allows us to conclude that bootstrap confidence intervals will be asymptotically valid. The results from a simulation study of pointwise percentile bootstrap confidence intervals are shown in the next section.

3.5 Simulations

To illustrate the performance of bootstrap confidence intervals, we simulated data from a model where the true $HDS(t)$ can be calculated analytically. The model assumes the hazard function

$$\lambda\{t|M(t)\} = 0.5 \cdot \exp\{0.8M(t)\}.$$

Each realization of the time-varying covariate $M(t)$ is a step-function approximation (with jumps every 0.05) of a straight line with a random intercept and slope, defined as

$$\begin{aligned} M(t) &= 2 + a_1 + a_2 \left\lfloor \frac{t}{0.05} \right\rfloor \cdot 0.05, \\ a_1 &\sim Unif(-1, 1), \\ a_2 &\sim Unif(-0.25, 0.25). \end{aligned}$$

The step function representation is used for practical and computational convenience. Data is typically collected at discrete times (e.g. annual doctor's visit or daily blood pressure measurement). Even if continuous covariates are available, the measurements are usually discretized for computational tractability. Thus in this simulated example, we sidestep issues that may arise by defining the $M(t)$ realizations to be discrete step functions. An example of 100 draws of $M(t)$ is shown in Figure 3.1.

The true underlying $HDS(t)$ corresponding to this data generating mechanism is shown in Figure 3.2. For contrast, $HDS(t)$ calculated using only baseline markers values $M_i(0)$ for all observations is also shown. Unsurprisingly, the model using temporally updated markers shows a more sustained prognostic performance, compared to the model using only baseline markers.

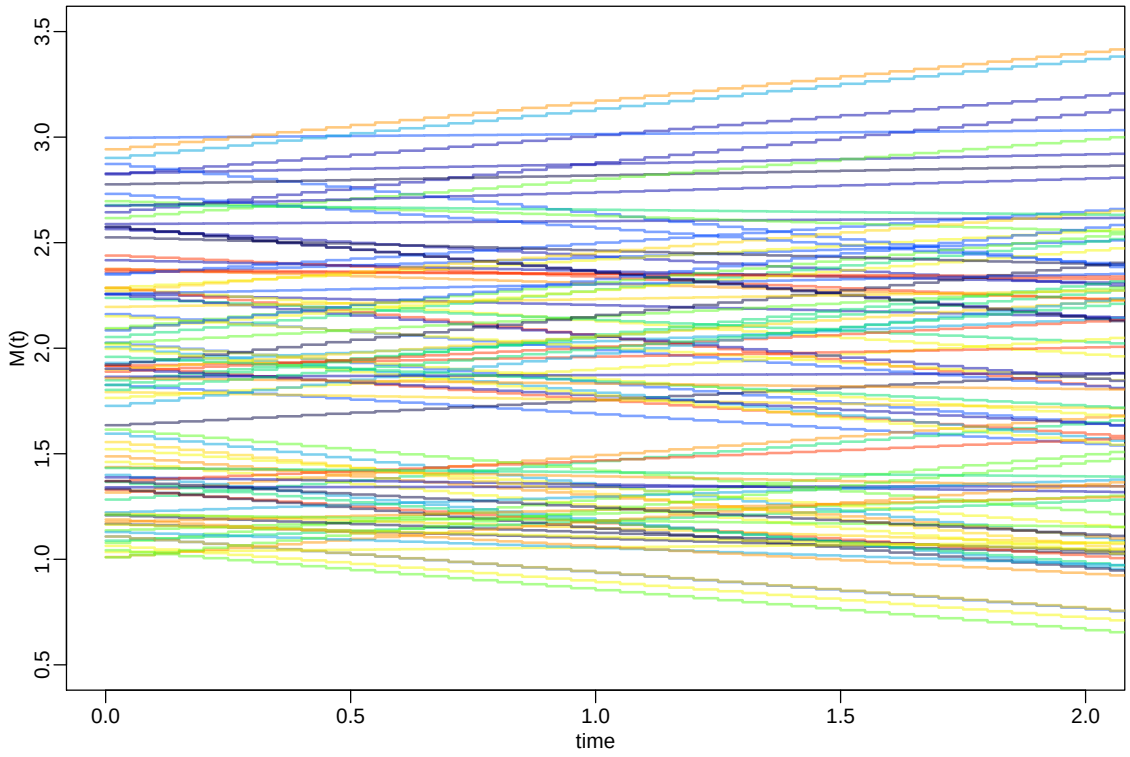

Figure 3.1: 100 draws of the time-varying covariate $M(t)$. Each $M(t)$ is a left-continuous step function with jumps every 0.05, and can be thought of as an approximation of a straight line with a random intercept and slope.

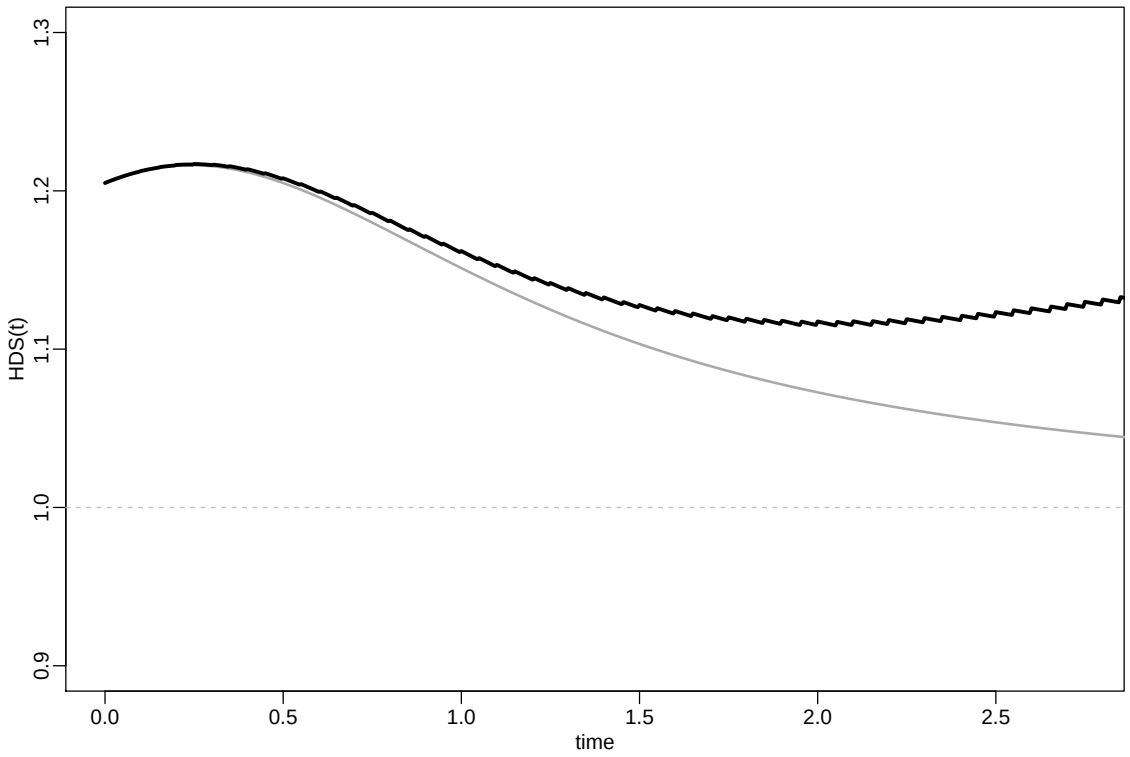

Figure 3.2: The black line is the true $HDS(t)$ for the data generating mechanism when correctly using all the temporal information from $M(t)$. The gray line is the true $HDS(t)$ (as defined in Chapter 1) when only using $M(0)$. A plot showing 100 random draws of $M(t)$ is shown in Figure 3.1.

Table 3.1: Coverages for bootstrap confidence intervals at different times and various sample sizes. Each coverage rate based on 1000 confidence intervals. Each confidence interval generated using 500 bootstrap replications.

sample size	0.055	0.105	0.205	0.305	0.405	0.505	0.755	1.005	1.505	2.005
n=500	0.937	0.935	0.933	0.935	0.935	0.937	0.929	0.874	0.754	0.692
n=2000	0.943	0.943	0.940	0.942	0.947	0.944	0.933	0.930	0.868	0.774
n=5000	0.947	0.951	0.949	0.950	0.953	0.953	0.953	0.946	0.895	0.847

To assess the coverage of percentile bootstrap confidence intervals for different sample sizes ($n = 500, n = 2000, n = 5000$), we used the data generating mechanism above to simulate 1000 sets of data for each sample size and calculated pointwise 95% confidence intervals for each set. Each confidence interval was generated using 500 bootstrap replications. The specific coverage rates are detailed in Table 3.1. Overall, the coverage rates deteriorate as t increases and the data gets sparser (only about 5% of the data lies beyond $t = 1.5$). Also, the coverage rates tend to improve as n increases; though they are still quite satisfactory for $n = 500$.

Based on our simulation results, it is important to note the number of observations in the risk set for determining when confidence intervals are still valid. Specifically, according to our simulations, a rule of thumb is that the confidence intervals are likely valid at t if there are at least 250 observations in the risk set at t.

3.6 Examples

To further illustrate $HDS(t)$, we use the time-varying covariates available from the Mayo PBC dataset. We use the time-varying versions of the same five covariates studied in Chapter 1: log(bilirubin), log(prothrombin time), edema, albumin, and age. A Cox regression with the five covariates produces hazard ratio estimates similar to when only baseline covariates are used. The results are shown in Table 3.2.

If we treat the Cox regession as a way to compress the five covariates into a time-varying risk score (referred to in this section as $M(t)$), it is possible to create a visualization of the

Table 3.2: Hazard ratio estimates from multivariate Cox regression using time-varying covariates from the Mayo PBC data

Covariate	Hazard ratio	95% CI
log(bilirubin)	2.912	(2.336, 3.630)
log(prothrombin time)	17.352	(5.225, 61.415)
edema	2.143	(1.357, 3.385)
albumin	0.253	(0.170, 0.375)
age	1.046	(1.028, 10.64)

data that reflects the prognostic performance of $M(t)$ at different times. This is shown in Figure 3.3.

Broadly, for each t, the better separated the red points around t are from the line segments in the same vertical space, the more informative $M(t)$ is at t. Measures like $HDS(t)$ and $AUC^{I/D}(t)$ are different ways of quantifying the separation at each t, and will be demonstrated later in this section.

Another way to visualize the same data is shown in a similar plot in Figure 3.4. In reality each observed $M(t)$ is a step function; however, since covariate updates happen at fairly common times, linear interpolations between $M(t)$ updates were used to prevent overplotting at such times.

Actual estimates of $HDS(t)$ using various estimators are shown in Figure 3.5. Comparing $\widehat{HDS}^{R}(t)$ versus $\widehat{HDS}(t)$ reveals that using temporally updated covariate values provide more sustained prognostic performance compared to just using baseline values. The same conclusion holds when comparing $\widehat{HDS}^{PL}(t)$ for both baseline and time-varying covariates.

For reference, estimates of $AUC^{I/D}(t)$ using the weighted mean rank estimator (Saha-Chaudhuri and Heagerty, 2013) are shown in Figure 3.6. The time-varying linear predictor from a Cox regression with the five time-varying covariates and the linear predictor from a Cox regression with the five baseline covariates were used for the calculations. Qualitatively, we would draw the same conclusions – that using temporal marker information greatly

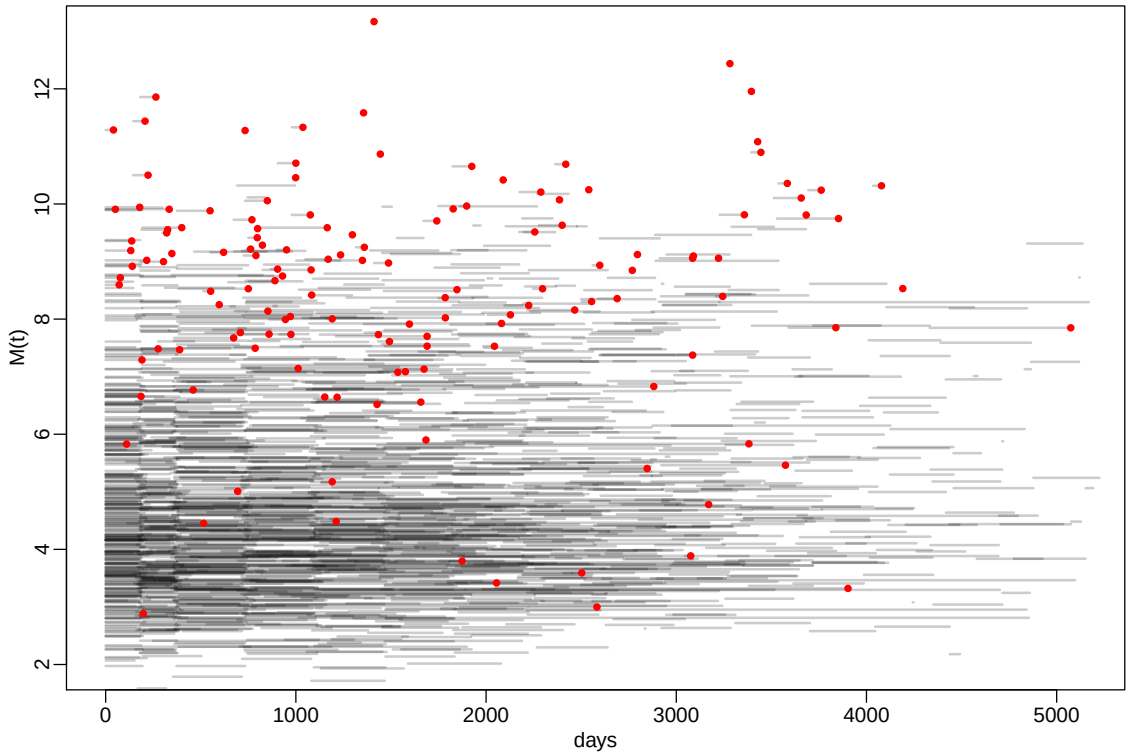

Figure 3.3: Visualization of a Cox model time-varying linear predictor $M(t)$ using the Mayo PBC data. The time-varying covariates used were log(bilirubin), log(prothrombin time), edema, albumin, and age. There are 312 lines, and 140 red points; each line represents an observation and each point represents a time of death. It is possible to visually compare $M(t)$ values for cases versus controls at each t with this graph (compared to 3.4), where each $M(t)$ is plotted as a step function represented by time segments. One drawback is that information about an individual's $M(t)$ trajectory is lost, since to avoid overplotting, the segments for each observation are not connected.

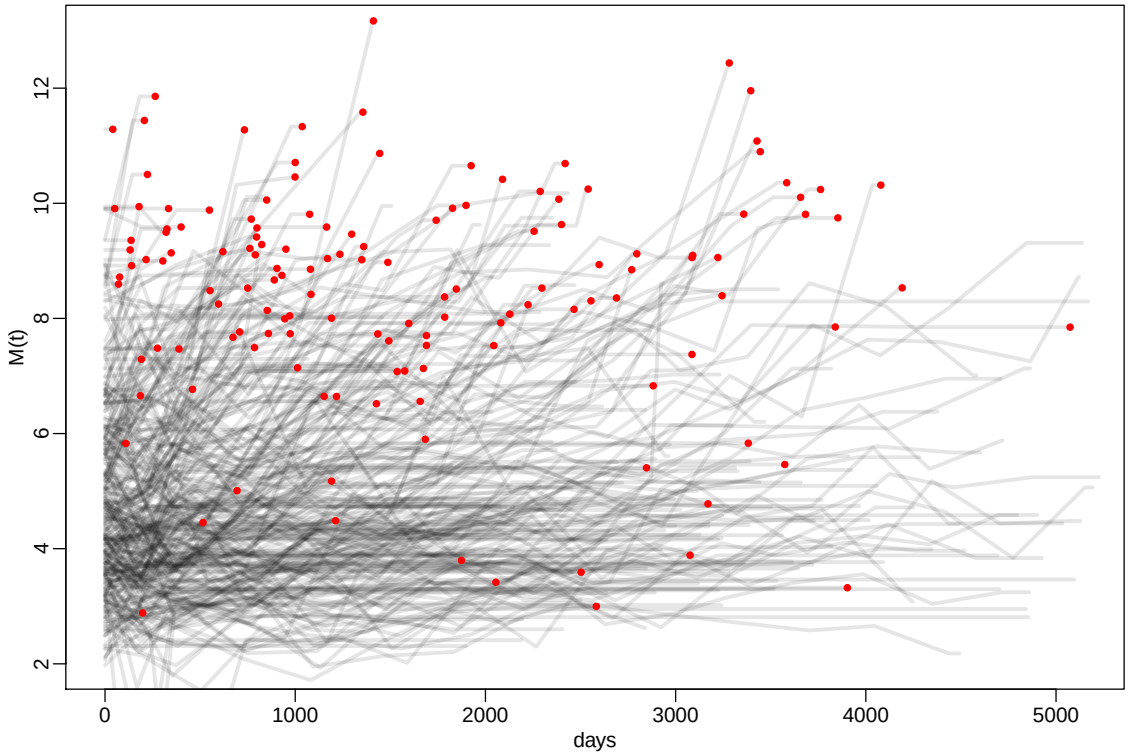

Figure 3.4: There are 312 lines, and 140 red points; each line represents an observation and each point represents a time of death. Each $M(t)$ is in reality a step function; however, since covariate updates happen at fairly common times, linear interpolations between $M(t)$ updates were used to prevent overplotting at such times. Broadly, for each t, the better separated the red points around t are from the line segments in the same vertical space, the more informative $M(t)$ is at t.

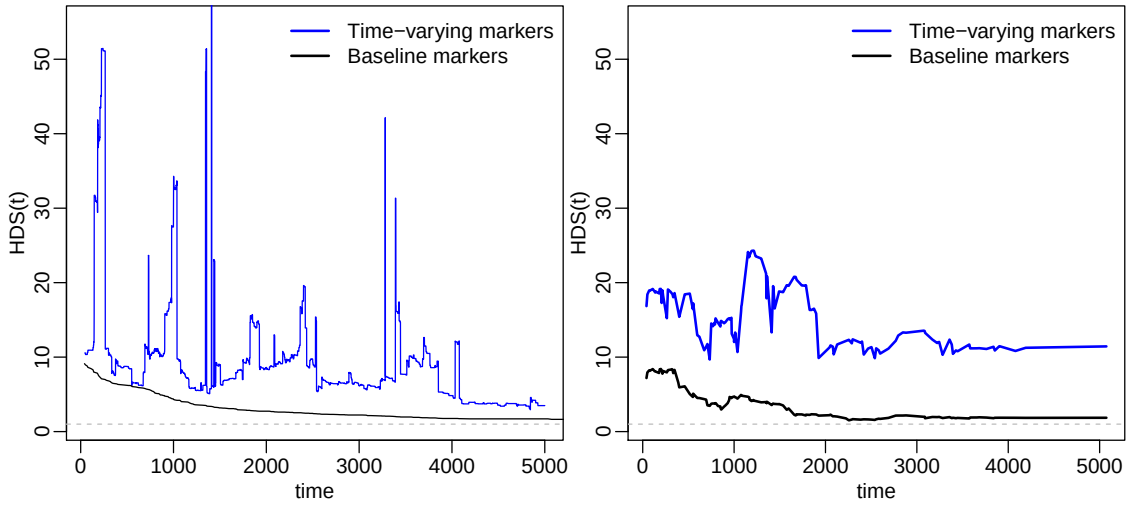

Figure 3.5: **Left**: $HDS(t)$ using the estimators $\widehat{HDS}(t)$ (2.2) and $\widehat{HDS}^{R}(t)$ (3.4) for baseline and time-varying covariates. Overall the latter one shows more sustained performance but also has large jumps due to some covariates having large jumps when they are updated. **Right**: $HDS(t)$ using the scaled partial likelihood estimator $\widehat{HDS}^{PL}(t)$ (3.5) for both baseline and time-varying covariates. The estimate using time-varying covariates again shows better sustained performance. Lines were smoothed using 25 nearest neighbors and a box kernel for each point.

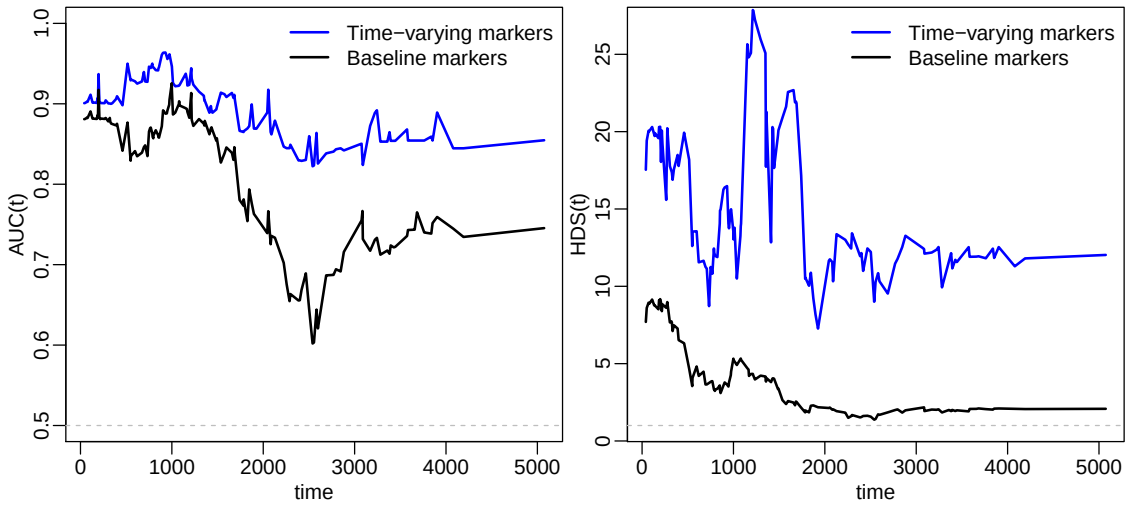

Figure 3.6: **Left**: $AUC^{I/D}(t)$ estimates for baseline and time-varying covariates to contrast with $HDS(t)$. **Right**: $HDS(t)$ using the scaled partial likelihood estimator $\widehat{HDS}^{PL}(t)$ (3.5) for both baseline and time-varying covariates. This is identical to the right plot for Figure 3.5 with the exception of a rescaled y-axis. Lines were smoothed using 30 nearest neighbors and a box kernel for each point.

improves prognostic performance across all times.

3.7 Discussion

Time-dependent covariates bring both a great opportunity for improved prognostic performance and challenging technical hurdles for modeling and evaluation. In this chapter we have extended $HDS(t)$ to be compatible with time-dependent covariates, joining the thin field of existing evaluation measures that can accomodate temporal covariate updates. $HDS(t)$ also continues to be a risk-based complement to the rank-based $AUC^{I/D}(t)$ for assessing time-varying markers.

Intuitively, one would expect covariates that are updated over time to have better prognostic performance than just baseline measurements alone. Both our results on simulated data and Mayo PBC data reflect this. While the prognostic performance of baseline measurements tends to weaken over time, when the covariates are temporally updated the performance tends to be more sustained. This qualitative behavior is displayed whether we

use $HDS(t)$ or $AUC^{I/D}(t)$ to quantify the performance.

A more practical issue concerning our primary estimator $\widehat{HDS}(t)$ is its sensitivity to coarsely sampled (in time) covariates that display great variation between samples. $\widehat{HDS}(t)$ only uses covariate measurements from those in the risk set at t. Thus, if a small change in t means a subject's risk score was updated to be very high, the estimate would be highly influenced and would increase sharply. Conversely, if the same subject then had an event (thus exiting the risk set), the estimate would sharply decrease.

While this is not necessarily a negative feature of $\widehat{HDS}(t)$, it can make for jarring visualizations when the estimator is plotted over time. If such local variations are not of scientific interest, pre-processing $M(t)$ to be smoother or choosing $\widehat{HDS}^{PL}(t)$ based on smoothing partial likelihood contributions may be attractive alternatives. Yet another option is to use a logarithmic transformation of the risk or risk ratios, though a more formal justification for this procedure is still lacking.

Future work The estimator $\widehat{HDS}^{R}(t)$ relies on a working Cox model with time-invariant coefficients. It seems likely that extending the estimator to be less parametric by allowing for time-varying coefficients as in Section 2.2.3 should be straightforward, though a more thorough investigation with simulated and study data is needed.

We showed that some datasets may result in large local deviations of $HDS(t)$ estimates. To mitigate such behavior we suggested either pre-processing $M(t)$ to be smoother in time or using $\widehat{HDS}^{PL}(t)$. While the latter suggestion is a practical *ad hoc* solution, it remains to more formally investigate the asymptotic properties of $\widehat{HDS}^{PL}(t)$. Intuitively, since $\widehat{HDS}^{PL}(t)$ involves kernel smoothing, one would expect it to be less efficient than $\widehat{HDS}^{R}(t)$. Further work is required to more precisely quantify the relative efficiencies to assess the viability of $\widehat{HDS}^{PL}(t)$ as an alternative.

Finally, just as $AUC^{Int/D}(t)$ is a generalization of $AUC^{C/D}(t)$, it may be worth exploring whether an analogous generalization for $HDS(t)$ and $DS(t)$ is of practical interest and can be feasibly estimated. A generalization may look something like

$$DS^{Int}(t_2;t_1) = \frac{\mathrm{E}\left[P\{t_1 \leq T \leq t_2 \mid M_{t_1}, T > t_1\} \mid M_{t_1}, t_1 \leq T \leq t_2\right]}{\mathrm{E}\left[P\{t_1 \leq T \leq t_2 \mid M_{t_1}, T > t_1\} \mid M_{t_1}, T > t_2\right]},$$

where $t_1 < t_2$, and M_{t_1} refers to the covariate history up through t_1.

Chapter 4

NONPARAMETRIC $HDS(t)$ ESTIMATION AND INFORMATION THEORY

In Section 2.6, we mentioned that the definitions of risk-based measures such as $HDS(t)$ and $DS(t)$ can allow for ambiguity in how they are used to assess multivariate marker panels. We will illustrate the ambiguity by first considering how AUC is used.

Suppose we are interested in measuring the prognostic accuracy of multiple predictors. When using AUC methods, there is an explicit "decoupling" of model generation from model evaluation. Recall the definition of $AUC^{I/D}(t)$:

$$AUC^{I/D}(t) = \mathrm{P}(M_i > M_j | T_i = t, T_j > t).$$

Since the above requires that M be ordered, M can not be multivariate and must be univariate. Thus, to assess a multivariate panel of predictors, a model must first be used to combine the predictors into a univariate score. Only then is the prognostic value of the score evaluated using AUC. In contrast, consider the definition of $HDS(t)$:

$$HDS(t) = \frac{\mathrm{E}\{\lambda(t|M)|T = t\}}{\mathrm{E}\{\lambda(t|M)|T > t\}}.$$

Note that $HDS(t)$ is still well-defined even when M is multidimensional. Thus, one need not use a separate model to first generate a prognostic score from the multiple predictors. However, as shown in the preceding chapters, the proportional hazards assumption (or its relaxed version) needs to hold if the subsequent estimate of $HDS(t)$ is to have a meaningful interpretation as a *risk-based* measure of prognostic accuracy. If one does not believe that the assumption holds, it is advisable to use a separate model to combine the multiple predictors into a prognostic score, and then evaluate the prognostic accuracy of the score with $HDS(t)$.

In this chapter we will propose an $HDS(t)$ estimator suited for the scenario above where one is interested in evaluating a univariate prognostic score. The prognostic score

could be, for example: risk scores from a multivariate model (e.g. the linear predictor from a multivariate Cox regression); or risk scores from an existing risk model (e.g. outputs from the Framingham risk calculator).

We will also describe some novel connections between risk-based measures of discrimination and long-standing information theoretic concepts such as mutual information and f-divergences.

4.1 Parameter of interest

Mathematically, we are still interested in $HDS(t)$ as defined in Section 2.1.1, though we restrict the marker M to be univariate. Recall,

$$HDS(t) = \frac{\mathrm{E}_{M|T=t}\left\{\lambda(t|M)|T=t\right\}}{\mathrm{E}_{M|T>t}\left\{\lambda(t|M)|T>t\right\}}.$$

By restricting M to be univariate, $HDS(t)$ becomes more comparable to $AUC^{I/D}(t)$ in the sense that model building and model evaluation are explicitly decoupled. In other words, when evaluating a panel of predictors, we must first combine them into a univariate prognostic score M.

A mathematical consequence of M being univariate is that nonparametric estimation of $\lambda(t|M)$ becomes tractable. This in turn facilitates a nonparametric estimator for $HDS(t)$, details of which will be shown in next sections.

In Section 2.1.2 we presented an additional formulation of $HDS(t)$ that implies an attractive alternative interpretation as the "prognostic value of a marker". Formally,

$$HDS(t) = \mathrm{E}_{M|T=t}\left[\left.\frac{\lambda(t|M)}{\lambda(t)}\right| T=t\right].$$

The above formulation is particularly compelling when M is univariate, as it begins to resemble mutual information (MI). Mutual information is an important information theoretic concept (Cover and Thomas, 2012, Chapter 2) that measures the dependence between two random variables.

4.2 Mutual information connection

In this section we will briefly review mutual information and its connection with $HDS(t)$. A complete overview of mutual information is beyond the scope of this dissertation, but

Cover and Thomas (2012) provides much useful background.

Mutual information is often expressed as the expected log ratio of the joint distribution to the product of the marginal distributions,

$$MI(M,T) = \mathrm{E}_{M,T}\left\{\log\frac{p(T,M)}{p(T)p(M)}\right\},$$

where $p(T,M)$ is the joint density function of T and M, while $p(T)$ and $p(M)$ are the marginal density functions. Through the above formulation it is easy to see that if T and M are independent, their mutual information is zero. Mathematically, mutual information is also equivalent to

$$MI(M,T) = \mathrm{E}_{M,T}\left\{\log\frac{p(T|M)}{p(T)}\right\}, \tag{4.1}$$

where $p(T|M)$ is the conditional density function of $T|M$. Using the above formulation we can see mutual information as a measure of the increase in precision of a variable due to knowledge of the second variable. A greater increase in precision would be represented by a "sharper" conditional density and thus greater value of $MI(M,T)$.

Instead of the density-based definitions above, it is possible to derive another formulation of mutual information based on hazards, which to our knowledge has not been shown before. The derivation is as follows:

$$\begin{aligned}
MI(T,M) &= E_{T,M}\left\{\log\frac{p(T|M)}{p(T)}\right\} \\
&= E_{T,M}\left\{\log\frac{\frac{p(T|M)}{S(T|M)}S(T|M)}{\frac{p(T)}{S(T)}S(T)}\right\} \\
&= E_{T,M}\left\{\log\frac{\lambda(T|M)}{\lambda(T)}\right\} + E_{T,M}\left\{\log S(T|M)\right\} - E_{T,M}\left\{\log S(T)\right\} \\
&= E_{T,M}\left\{\log\frac{\lambda(T|M)}{\lambda(T)}\right\} - 1 + 1 \\
&= E_{T,M}\left\{\log\frac{\lambda(T|M)}{\lambda(T)}\right\}.
\end{aligned}$$

The derivations for the equalities $E_{T,M}\left\{\log S(T|M)\right\} = -1$ and $E_{T,M}\left\{\log S(T)\right\} = -1$

used above are shown below:

$$\begin{aligned}
E_{T,M}\left\{\log S(T)\right\} &= E_T\left\{\log S(T)\right\} \\
&= E_T\left\{-\Lambda(T)\right\} \\
&= -\int\int_0^T \frac{f(u)}{S(u)}du \cdot f(T)dT \\
&= S(T)\int_0^t \frac{f(u)}{S(u)}du\Big|_0^\infty - \int f(T)dT \\
&= -\,S(T)\cdot\log\left\{S(T)\right\}|_0^\infty - 1 \\
&= -1.
\end{aligned}$$

The equality between the third and fourth lines follows via integration by parts. The proof showing that $E_{T,M}\left[\log S(T|M)\right] = -1$ follows in a similar fashion.

Having established that mutual information can be interpreted as log hazard ratios, we can apply conditional expectations to show

$$MI(T,M) = \mathrm{E}_T\mathrm{E}_{M|T=t}\left\{\log\frac{\lambda(t|M)}{\lambda(t)}\middle| T=t\right\}.$$

Recall that $HDS(t) = \mathrm{E}_{M|T=t}\left\{\frac{\lambda(t|M)}{\lambda(t)}\middle| T=t\right\}$, which makes apparent that the inner expectation of $MI(T,M)$ is equivalent to $HDS(t)$ on the log scale. To our knowledge, this interpretation of mutual information is not commonly used. However, since we are studying survival data and time-varying prognostic performance, we find this connection between a fundamental information theory concept and temporal prognostic performance to be of great interest. Specifically, we see that a "longitudinal deconstruction" of MI can serve as a measure of time-varying incident prognostic performance. Furthermore, note that a first-order Taylor expansion of $log(x)$ around $x = 1$ shows that $log(x) \approx x - 1$. Thus, when $\frac{\lambda(t|M)}{\lambda(t)}$ is not too far from 1 (i.e. no greater than 2), $\log\frac{\lambda(t|M)}{\lambda(t)}$ can be intuitively interpreted as "percent *improvement* in risk prediction".

Another consequence of the above formulation is that it may facilitate an estimator for $MI(T,M)$ with censored survival data. This in turn may be useful for applications where one has survival endpoints and is interested in filtering a large number of candidate univariate predictors.

Table 4.1: A nonexhaustive list of distance measures and their corresponding f-divergence subtypes.

Distance measure	Choice of $f(t)$
Kullback-Leibler divergence	-log(t)
Hellinger distance	$\left(\sqrt{t}-1\right)^2$
Total variation distance	$\lvert t-1\rvert$
χ^2 divergence	$(t-1)^2$

4.2.1 f-divergences

Mutual information, particularly as defined in (4.1), can be interpreted as a measure of distance between two probability measures: a marginal density and a conditional density. However, this is by no means the only method of quantifying distance between two probability measures, as many other measures have been thoroughly studied. For example, the Kullback-Leibler divergence of two continuous distrubutions with densities $p(x)$ and $q(x)$ is familiar to statisticians:

$$KL\{p(x)||q(x)\} = \int_{-\infty}^{\infty} p(x) \cdot \log\frac{p(x)}{q(x)} dx.$$

More generally, Morimoto (1963) and Ali and Silvey (1966) independently studied a class of divergence measures called f-divergences that include well known distance measures such as the Kullback-Leibler divergence and Hellinger distance. An f-divergence is defined as

$$D_f\{p(x)||q(x)\} = \int f\left\{\frac{p(x)}{q(x)}q(x)\right\} dx,$$

where f is any convex function such that $f(1) = 0$. Table 4.1 is a nonexhaustive list of choices of f that correspond to some well-known divergences.

Mutual information and discrimination slope are f-divergences Though the majority of this dissertation has focused on measures of time-varying discrimination, we have found that some of these measures have interesting mathematical connections with information theory. Specifically, it can be shown that mutual information and discrimination slope are both f-divergences.

By noting that MI is equal to the Kullback-Leibler divergence of $p(t,m)$ from $p(t)p(m)$, or a measure of the distance between the joint distribution of time and marker from the product of the marginal distributions, it is straightforward to see that MI is an f-divergence:

$$MI(T,M) = KL\{p(t,m)||p(t)p(m)\}.$$

We can also show that the discrimination slope, DS, for a binary outcome Y and marker M is equivalent to the f-divergence subtype known as the χ^2 divergence by choosing $f(t) = (t-1)^2$:

$$\begin{aligned}
D_{\chi^2}(f(y,m)||f(y)f(m)) &= \int \left\{ \frac{f(y,m)}{f(y)f(m)} - 1 \right\}^2 f(y)f(m)dydm \\
&= \int \left\{ \frac{f^2(y,m)}{f^2(y)f^2(m)} - 2\frac{f(y,m)}{f(y)f(m)} + 1 \right\} f(y)f(m)dydm \\
&= \int \frac{f(y,m)}{f(y)f(m)} f(y,m)dydm - 2\int f(y,m)dydm + \int f(y)f(m)dydm \\
&= \{DS(M)+1\} - 2 + 1 \\
&= DS(M).
\end{aligned}$$

The derivation for the equality $\int \frac{f(y,m)}{f(y)f(m)} f(y,m)dydm = DS(M)+1$ is shown below:

$$\begin{aligned}
\int \frac{f(y,m)}{f(y)f(m)} f(y,m)dydm &= \int \frac{f(y|m)}{f(y)} f(y,m)dydm \\
&= \int \frac{f(y=1|m)}{P(y=1)} f(y=1,m)dm + \int \frac{f(y=0|m)}{P(y=0)} f(Y=0,m)dm \\
&= \mathrm{E}\left\{P(Y=1|M)|Y=1\right\} \\
&\quad + \int \frac{1-f(y=0|m)}{P(y=0)} f(y=0,m)dm \\
&= \mathrm{E}\left\{P(Y=1|M)|Y=1\right\} \\
&\quad - \int \frac{f(y=0|m)}{P(y=0)} f(y=0,m) + \int \frac{1}{P(y=0)} f(y=0,m) \\
&= \mathrm{E}\left\{P(Y=1|M)|Y=1\right\} - \mathrm{E}\left\{P(Y=1|M)|Y=0\right\} + 1 \\
&= DS(M)+1.
\end{aligned}$$

4.3 Estimation

Recall that $HDS(t)$ can be written as a function of $\lambda(t|M)$, $\Lambda(t|M)$, and the empirical marker distribution:

$$HDS(t) = \frac{\mathrm{E}\left[\lambda^2(t|M)\exp\{-\Lambda(t|M)\}\right]\mathrm{E}\left[\exp\{-\Lambda(t|M)\}\right]}{\mathrm{E}^2\left[\lambda(t|M)\exp\{-\Lambda(t|M)\}\right]}. \tag{4.2}$$

As with the semiparametric estimator described in Section 2.2, our general strategy for creating an estimator is to replace $\lambda(t|M)$ and $\Lambda(t|M)$ with estimators and the marginal expectations with empirical means.

Unlike Section 2.2, instead of using $\hat{\lambda}(t|m)$ and $\hat{\Lambda}(t|m)$ from the Cox model, we use conditional versions of nonparametric estimates of the marginal $\hat{\lambda}(t)$ and $\hat{\Lambda}(t)$. We will first outline nonparametric hazard estimation methods before returning to nonparametric $HDS(t)$ estimation.

4.3.1 Nonparametric hazard estimation

Broadly, nonparametric marginal hazard estimation can be grouped into kernel, spline, and wavelet methods. Wang (2005) has a broad review of the various hazard estimation methods. Kernel methods appear to be the most thoroughly studied, perhaps due to their relative simplicity. Since our primary interest is in applying hazard estimation to a different estimator of our own, we choose to focus on the most popular kernel methods.

Watson and Leadbetter (1964a,b) introduced nonparametric estimation of $\lambda(t)$ for uncensored outcomes. Ramlau-Hansen (1983), Tanner and Wong (1983), and Yandell (1983) studied different kernel-based methods that allow for right-censoring. Muller and Wang (1994), Patil (1993), and González-Manteiga et al. (1996) studied bandwidth selection for kernel estimators that allow for right-censoring.

Beran (1981) introduced an estimator for $\Lambda(t|M)$ based on a conditional version of the Nelson-Aalen estimator, while Dabrowska (1987) studied its asymptotic properties. McKeague and Utikal (1990) proposed a nonparametric estimator for $\lambda(t|M)$ based on kernel-smoothing the aforementioned estimator for $\Lambda(t|M)$.

The estimators we will use for $\Lambda(t|M)$ and $\lambda(t|M)$ are

$$\hat{\Lambda}(t|m) = \sum_{i:y_i \leq t} \delta_i \frac{1}{\sum_{j:y_j > y_i} I\left(\frac{m_j - m}{b_m}\right)}, \tag{4.3}$$

$$\hat{\lambda}(t|m) = \sum_{i=1}^{n} \delta_i \frac{\frac{1}{b_t} K\left(\frac{y_i - t}{b_t}\right)}{\sum_{j:y_j > y_i} I\left(\frac{m_j - m}{b_m}\right)} I\left(\frac{m_i - m}{b_m}\right), \tag{4.4}$$

where $K(u)$ is a kernel such as the Epanechnikov kernel; $I(u)$ is the uniform kernel $1_{[-0.5,0.5]}(u)$, and is used along with the marker bandwidth b_m (which approaches 0 at a rate dependent on n) to specify a subset of observations. $\hat{\Lambda}(t|M)$ is the Nelson-Aalen estimator applied to observations with marker values close to the target M. $\hat{\lambda}(t|M)$ is a kernel smoothed version of $\hat{\Lambda}(t|M)$, and depends on an additional time bandwidth b_t, which approaches 0 at a rate dependent on n. Details on how to select a two-dimensional bandwidth are in Section 4.4.

Figure 4.1 contains a pair of plots that illustrate the estimators $\hat{\Lambda}(t|M)$ and $\hat{\lambda}(t|M)$ at $M = 7$. The example uses the Mayo PBC data, and M is the Cox regression linear predictor using log(bilirubin), log(prothrombin time), edema, albumin, and age.

4.3.2 Nonparametric $HDS(t)$ estimation

Having established nonparametric cumulative hazard and hazard rate estimation methods, we can now propose a nonparametric $HDS(t)$ estimator $HDS^{NP}(t)$. Recall from Section 2.2 that $HDS(t)$ can be expressed as a function of the hazard rate, cumulative hazard, and empirical marker distribution (4.2). This formulation suggests a plug-in estimator

$$\widehat{HDS}^{NP}(t) = \frac{\sum_{i=1}^{n} \hat{\lambda}^2(t|m_i) \exp\{-\hat{\Lambda}(t|m_i)\} \sum_{i=1}^{n} \exp\{-\hat{\Lambda}(t|m_i)\}}{\left[\sum_{i=1}^{n} \hat{\lambda}(t|m_i) \exp\{-\hat{\Lambda}(t|m_i)\}\right]^2},$$

where $\hat{\lambda}(t|m_i)$ and $\hat{\Lambda}(t|m_i)$ are the nonparametric hazard estimators defined in (4.4) and (4.3).

We expect analytic results for asymptotics to be more complicated, owing to the need for two bandwidths (in time and in marker). We recommend using the percentile bootstrap for confidence intervals. While we do not provide formal verification that the bootstrap provides valid confidence interval, they appear reasonable in practice, and we expect the estimator to be asymptotically normal at a rate slower than $\sqrt{n}$. Our expectation of asymptotic

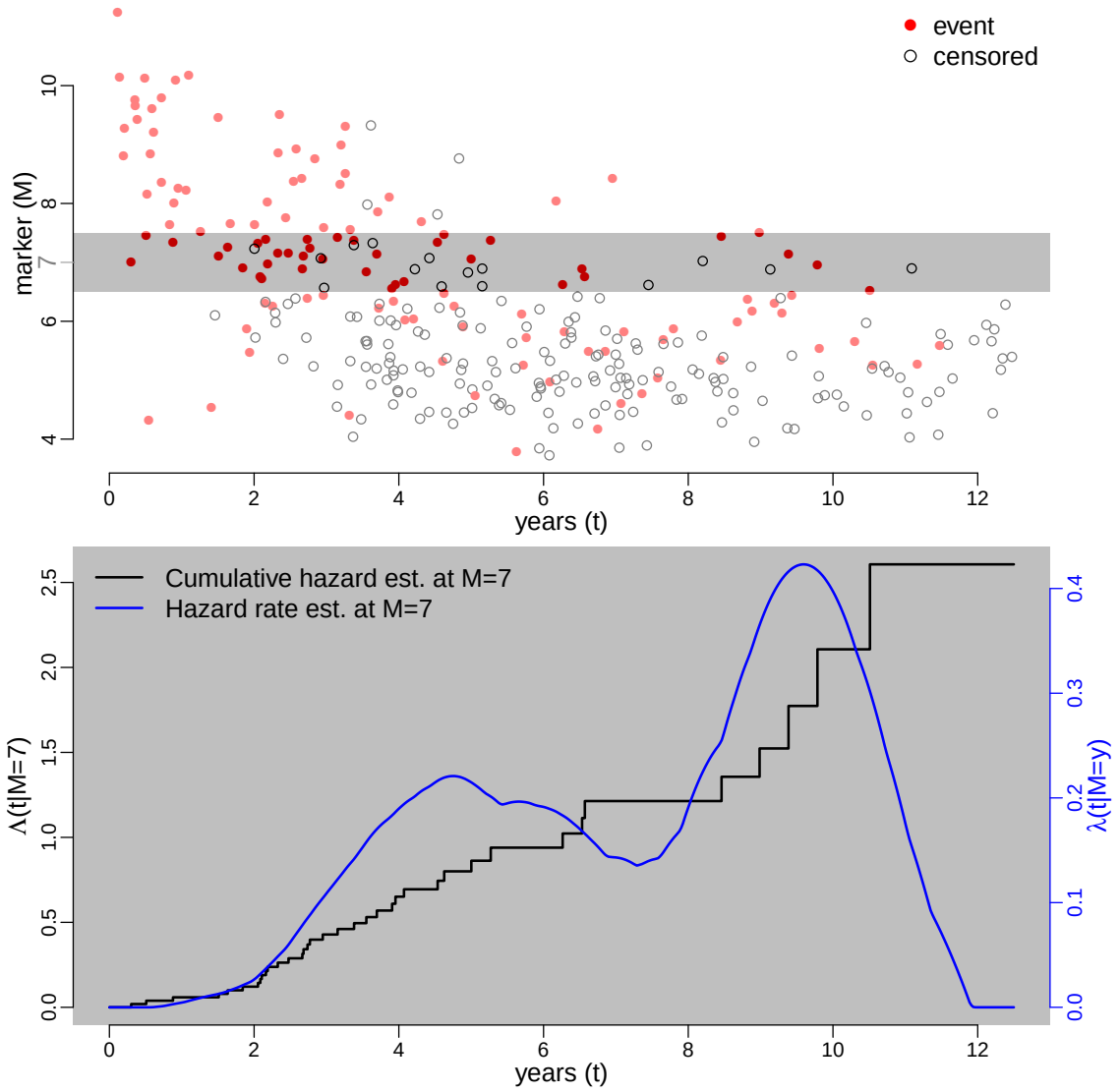

Figure 4.1: The above two plots are a visual aide for how the conditional cumulative hazard, $\Lambda(t|M)$ and hazard rate $\lambda(t|M)$ functions are nonparametrically estimated at $M = 7$. **The top plot** is a visualization of the Mayo PBC data; the y-axis is the linear predictor (of the usual 5 predictors) scale and the x-axis is the time scale. A strip of observations centered around $M = 7$ is highlighted in gray. These observations are used to calculate $\hat{\Lambda}(t|M = 7)$ and $\hat{\lambda}(t|M = 7)$, **shown in the bottom plot**. The black line is the Nelson-Aalen estimator using the points in the gray strip; note that jumps only occur at event times (red dots). The blue line is a kernel-smooth of the black line.

normality is based on existing results showing that the estimates for $\lambda(t|m)$ and $\Lambda(t|m)$ are asymptotically normal (Dabrowska, 1987; McKeague and Utikal, 1990).

4.3.3 Estimating $HDS(t)$ using a graphical smoothing technique

Our nonparametric $HDS(t)$ estimator is based on the semiparametric estimator $\widehat{HDS}(t)$ from Section 2.2. In the same section we also defined an alternative semiparametric estimator based on smoothing scaled partial likelihood contributions. In this section we construct an analogous nonparametric version based on smoothing ratios of conditional to marginal hazard rates against time.

Recall again the *prognostic value of a marker* formulation of $HDS(t)$:

$$HDS(t) = \mathrm{E}_{M|T=t}\left\{ \frac{\lambda(t|M)}{\lambda(t)} \middle| T = t \right\}.$$

The above formulation suggests that $HDS(t)$ can be estimated as the regression curve of "risk prediction improvements" against time, for those who experienced events. We define the "risk prediction improvement" for the ith individual as $\frac{\hat{\lambda}(t_i|m_i)}{\hat{\lambda}(t_i)}$, or how much an individual's (who experienced an event at t_i) risk prediction improves when we incorporate marker information.

Visually, consider each individual i who experienced at event at time t_i and plot the points $\left\{ \left(t_i, \frac{\hat{\lambda}(t_i|m_i)}{\hat{\lambda}(t_i)} \right) \right\}$. We can then estimate $HDS(t)$ by fitting a nonparametric regression curve through the points:

$$\widehat{HDS}^{G}(t) = \frac{1}{n \cdot a_n} \sum_{i:\delta_i=1} K\left(\frac{y_i - t}{a_n} \right) \frac{\hat{\lambda}(y_i|m_i)}{\hat{\lambda}(y_i)},$$

where $K(u)$ is a kernel such as the Epanechnikov kernel and a_n is bandwidth value that varies with n. We have found that heuristically, setting a_n to be equal to the time-bandwidth b_t tends to give reasonable results.

An illustrative using Mayo PBC data is provided in Section 4.7. An accompanying series of graphs illustrating the above procedure are also shown in Figure 4.10.

One drawback of our alternate nonparametric $HDS(t)$ estimator is it requires yet another bandwidth choice for estimating the regression curve. However, advantages include:

1) a close visual tie between $HDS(t)$ and an empirical representation of the underlying observations; 2) much less computationally intensive, which can be relevant for larger datasets; 3) the estimator can be easily modified to serve as a reasonable estimator for mutual information, details of which are in the next section.

4.3.4 Estimating Mutual Information

Mutual information is used extensively for continuous outcome data, and there is an extensive active literature on how to nonparametrically estimate mutual information in those situations (Kinney and Atwal, 2014). To our knowledge, there are no proposals for estimating mutual information when one has censored survival outcomes.

We have found that the graphical smoothing technique outlined in the previous section can be easily adapted to created an estimator for mutual information. To show this, recall from Section 4.2 that mutual information can be written as

$$MI(T, M) = \mathrm{E}_T \mathrm{E}_{M|T=t} \left\{ \log \frac{\lambda(t|M)}{\lambda(t)} \middle| T = t \right\}. \tag{4.5}$$

In other words, if we estimate $\frac{\lambda(t_i|m_i)}{\lambda(t_i)}$ at every i where an event is observed and then take a weighted sum of the estimates (weighted by jumps in the Kaplan-Meier estimator), the result is a reasonable estimator for MI. Formally,

$$\widehat{MI} = \int \log \frac{\hat{\lambda}(T|M)}{\hat{\lambda}(T)} d\hat{F}_T.$$

Just as with $\widehat{HDS}^{NP}(t)$ and $\widehat{HDS}^{G}(t)$, estimates of $\hat{\lambda}(t|m)$ and $\hat{\lambda}(t|m)$ require bandwidths for both marker and time. We propose using the bandwidth selection technique from Section 4.4. Results from a simulation study of the above estimator are shown in Section 4.6.

Equation (4.5) also suggests a natural MI-based measures of time-varying prognostic performance. Specifically, the inner expectation $E_{M|T=t} \left\{ \log \frac{\lambda(t|M)}{\lambda(t)} \middle| T = t \right\}$ becomes a "longitudinal deconstruction" of mutual information. While this interpretation is not as meaningful when T is just a continuous outcome, when T is a survival endpoint it becomes much more compelling.

An illustrative example using the Mayo PBC data is provided in Section 4.7.4 along with a visual aide in Figure 4.11.

4.4 2-D bandwidth selection method

Our nonparametric estimator for $HDS(t)$ requires first estimating $\Lambda(t|M)$ and $\lambda(t|M)$. The former requires a marker bandwidth, while the latter requires both a marker and time bandwidth.

Bandwidth selection for estimating the marginal hazard function $\lambda(t)$ has been extensively studied. Muller and Wang (1994) proposed choosing dynamic bandwidths that vary with time so as to minimize the mean-squared error at local time segments, rather than the usual method of choosing a fixed bandwidth to minimize integrated mean-squared error. Hess et al. (1999) conducted an extensive simulation study of different estimators and bandwidth selection methods for $\lambda(t)$, and concluded that using both dynamic bandwidth selection as suggested by Muller and Wang (1994) and boundary correction usually result in much more accurate estimates. Our own simulation studies of using cross validation to find a fixed time bandwidth that minimizes integrated squared error (ISE) as suggested by Patil (1993) gave largely unsatisfactory results.

While the bandwidth selection procedure is fully outlined in Patil (1993, Section 3.2), we briefly review it here. First consider the ISE for an estimate of $\lambda(t)$,

$$\begin{aligned} ISE(h) &= \int \left\{ \hat{\lambda}_h(x) - \lambda(x) \right\}^2 w(x) dx \\ &= \int \hat{\lambda}_h^2(x) w(x) dx - 2 \int \left\{ \frac{\hat{\lambda}_h(x)}{1 - F(x)} w(x) \right\} f(x) dx + \int \lambda^2(x) w(x) dx, \end{aligned}$$

where $\hat{\lambda}_h(x)$ is an estimate of $\lambda(x)$ using the time-bandwidth h, and $w(x)$ is a weight function that can be used to downweight variance in the tails of the estimates. In the above formulation, only the first two terms depend on h, thus it is sufficient to focus on those two terms. An estimate of the two terms as a function of h is

$$CV(h) = \int \hat{\lambda}^2(x) w(x) dx - \frac{2}{n} \sum_{i=1}^{n} \frac{\hat{\lambda}_h^{-i}(x_i)}{1 - F_n(x_i)} w(x_i) \delta_i,$$

where $\hat{\lambda}_h^{-i}$ is the leave-one-out version of $\hat{\lambda}(t)$ (with the ith observation left out), and $F_n(x)$

is the empirical distribution of the follow-up time $Y = min(T, C)$. We then select, among a pre-specified set of candidates, the h that minimizes $CV(h)$.

We simulated 150 sets of survival times (each $n = 500$) from a standard normal distribution, with roughly 33% random censoring. For each dataset we searched 100 candidate time-bandwidths $h \in \{0.01, 0.02, \ldots, 1\}$ using the above procedure. A summary of our cross validation results using the ISE method is shown in Figure 4.2. The cross validation results show that in general $CV(h)$ is not convex, and the ISE method tends to choose extremely narrow or extremely wide time-bandwidths.

Since we are interested in $\lambda(t|M)$, we need to select not just a time bandwidth, but an additional marker bandwidth. One way to extend the above bandwidth selection methods for $\lambda(t)$ to $\lambda(t|M)$ is to first select a marker bandwidth, then iteratively apply a marginal hazard bandwidth selection technique to estimate $\lambda(t|M)$ for each M. This would result in a single marker bandwidth and multiple (possibly dynamic) time bandwidths, which is computationally burdensome. If we were primarily interested in estimating $\lambda(t|M)$ well, such an inconvenience may be worthwhile. However, we are primarily interested in using our $\lambda(t|M)$ estimate as a means to estimate $HDS(t)$. In other words, we just need an estimate of $\lambda(t|M)$ that is "good enough" to translate into an accurate estimate of $HDS(t)$. Thus we considered a number of a bandwidth selection methods that allow us to simply perform a two-dimensional grid search of candidate bandwidths so as to select a 2-D fixed bandwidth for marker and time.

Unlike regression with continuous outcomes (such as splines or kernel based methods), where it is natural to choose a bandwidth based on mean squared distance between the observed and predicted outcomes, there is no direct analog for survival models due to the potentially censored time-to-event outcomes. Hosmer et al. (2008, Section 6.2) mention five types of residuals used to assess the adequacy of the Cox model for $\lambda(t|M)$: martingale, Schoenfeld, scaled Schoenfeld, score, and scaled score. Martingale residuals measure the distance between the observed censoring indicator and predicted cumulative hazard for each observation. The latter four are all related to measuring the distance between observed covariate values and expected covariate values, which rely on hazard estimates.

While the above residuals were specifically developed for assessing Cox model fit, the

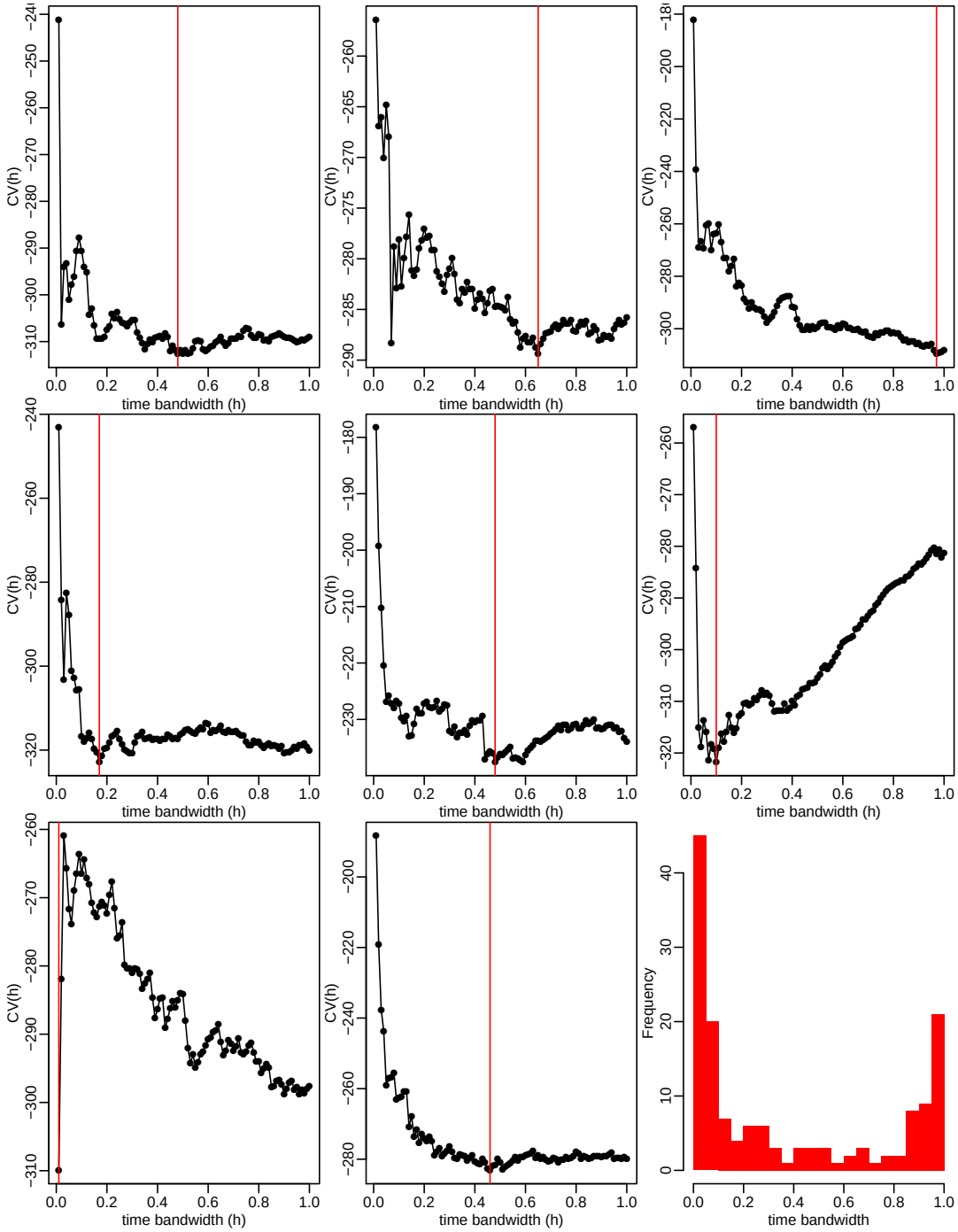

Figure 4.2: Eight sets (out of 150) of cross validation results from random data. Each line plot shows the $CV(h)$ score against candidate bandwidths for a random dataset. Data was generated with $n = 500$, standard normal survival times, and roughly 33% censoring. A summary of the selected bandwidths (the h values that minimized $CV(h)$) from all 150 random data sets is shown in the red histogram.

underlying concept of *distance between observed and predicted values* can be extended to nonparametric models of $\lambda(t|M)$. We chose Schoenfeld residuals, as the "expected covariate" estimates rely on $\lambda(t|M)$ and thus are dependent on both a time and marker bandwidth. In contrast, martingale residuals rely on estimating $\Lambda(t|M)$, which only require selecting a marker bandwidth.

Consider the scenario where we have a univariate covariate M and wish to assess model fit. The Schoenfeld residual for the ith individual experiencing an event is defined as

$$s_i = m_i - \frac{\sum_{j \in R_i} m_j \cdot \exp\left(m_j \hat{\beta}\right)}{\sum_{j \in R_i} \exp\left(m_j \hat{\beta}\right)},$$

where m_i is the *observed* covariate value; the fraction is the corresponding *predicted* or *expected* covariate value, which can be seen as a weighted average of observed marker values in the risk set R_i. Note that the weights are estimates of $\lambda(t|M)$ based on the Cox model. The s_i values are then plotted against m_i, and a smoothed regression line assessed for systematic patterns, the presence of which would indicate a violation of the proportional hazards assumption. If the proportional hazard assumption held, then the regression line is expected to vary around zero without a systemic pattern.

We first consider the following nonparametric extension of Schoenfeld residuals,

$$s_i^b = m_i - \frac{\sum_{j:t_j \in R_i} m_j \hat{\lambda}_b\left(t_i|m_j\right)}{\sum_{j:t_j \in R_i} \hat{\lambda}_b\left(t_i|m_j\right)},$$

which is similar to s_i, except that the Cox model estimates of $\lambda(t|M)$ are replaced with a nonparametric – and crucially, 2-D bandwidth b dependent – estimate $\hat{\lambda}_b(t_i|m_j)$.

Since we are not interested in assessing proportional hazards but rather selecting b from a grid of candidates, we use the residuals in the same manner that residuals from ordinary regression are used - by taking the mean of the squared residuals. Further, and again analogously to ordinary regression, we use a leave-one-out technique to calculate the predicted marker value.

Specifically, we consider the following quantity as our bandwidth score, where d_k is 1 if

observation k experienced an event and 0 if the observation was censored:

$$S(b) = \frac{1}{\sum_{j=1}^{n} d_j} \sum_{i}^{n} d_i \left\{ m_i - \frac{\sum_{j:t_j \in R_i \setminus \{i\}} m_j \hat{\lambda}_b^{-i}(t_i|m_j)}{\sum_{j:t_j \in R_i \setminus \{i\}} \hat{\lambda}_b^{-i}(t_i|m_j)} \right\}^2$$

After specifying a grid of candidate b values, we select the b that minimizes $S(b)$, and use that for calculating $\widehat{HDS}^{NP}(t)$ or estimating $MI(T, M)$.

Example bandwidth selection method We perform a study of our bandwidth selection method using data generated from a model where time T and marker M are bivariate normal. We also specify M to be a moderately strong marker, with $corr(M, T) = -0.7$, and allow for roughly 33% censoring.

Bandwidth selection scores for a random sample of nine simulated datasets (each $n = 500$) are shown in Figure 4.3. The bandwidth selection scores appear to be approximately convex, making automated bandwidth selection viable, since a minimum score will typically be readily available.

$\widehat{HDS}^{NP}(t)$ estimates using the same nine datasets and selected bandwidths are shown in Figure 4.4, along with the true underlying $HDS(t)$.

An aggregate view of the bandwidths selected using our selection technique is shown in Figure 4.5. In contrast, the bandwidths selected using the oracle minimum integrated squared error $\int \left\{ \widehat{HDS}^{NP}(t) - HDS(t) \right\}^2 dt$ are also shown. In practice it is impossible to select bandwidths in such a way, but it provides an aggregate comparison of how close to optimal our selected bandwidths are.

4.5 Inference

We expect $\widehat{HDS}^{NP}(t)$ to be asymptotically normal at a nonparametric rate, but also expect the formal derivations to be quite complicated. We thus suggest an *ad hoc* percentile bootstrap estimator for calculating confidence intervals. We have found the performance of this confidence interval estimator to be acceptable for practical usage.

One area that requires further clarification is whether to re-select the bandwidths for each bootstrap replication. We have found that re-selecting the bandwidth for each bootstrap replication to be computationally intensive, while not offering major gains in coverage

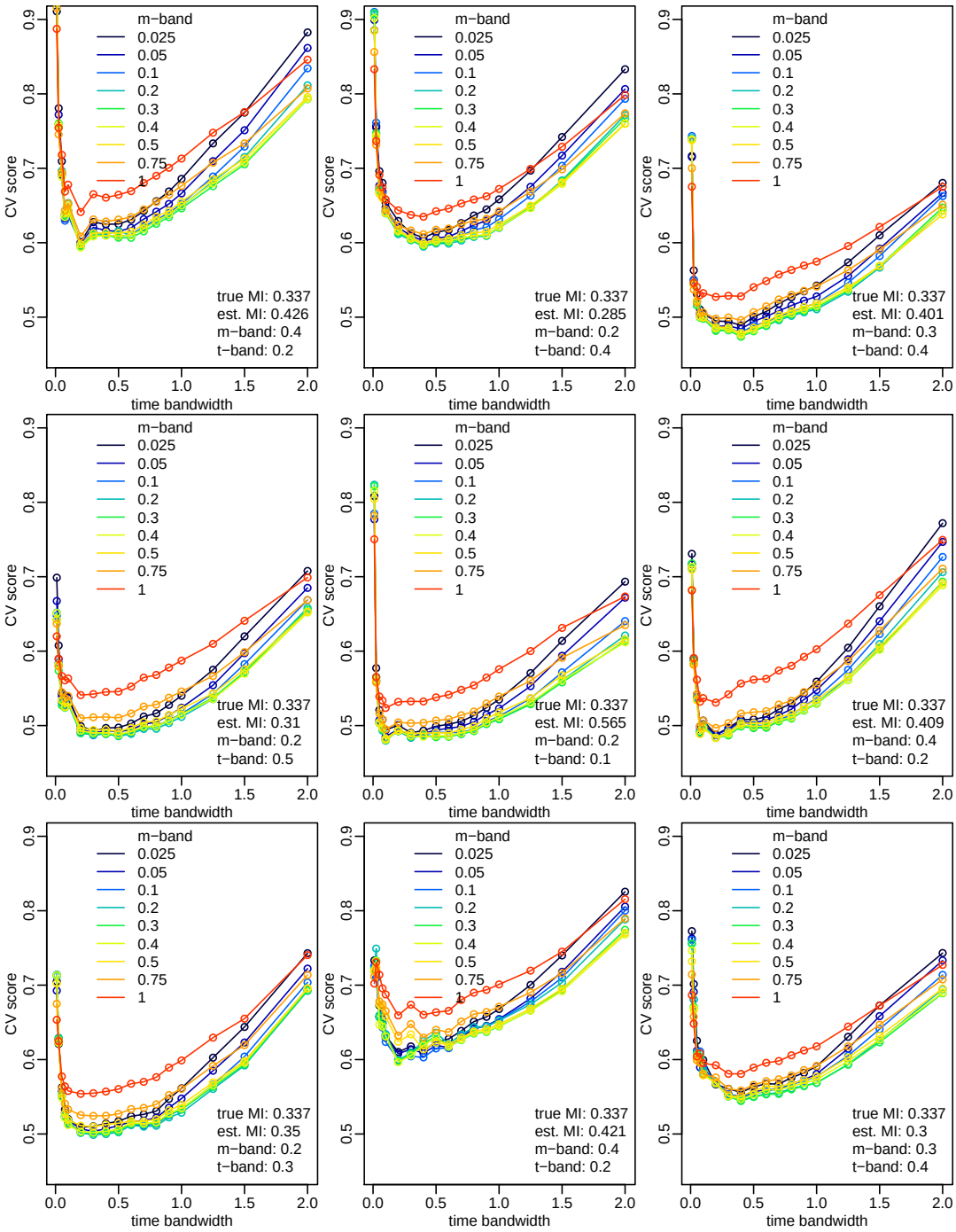

Figure 4.3: Nine sets of cross validation results from random data. Each plot summarizes the cross validation scores from a 2-D grid of marker and time bandwidths. The x-axis corresponds to time bandwidths, while the colors correspond to marker bandwidths. Data was generated with $n = 500$, where time and marker values are bivariate normal with $corr(M, T) = -0.7$, and roughly 33% censoring. The bandwidths corresponding to the minimum cross validation scores are printed in the bottom right of each plot, along with the corresponding mutual information estimate.

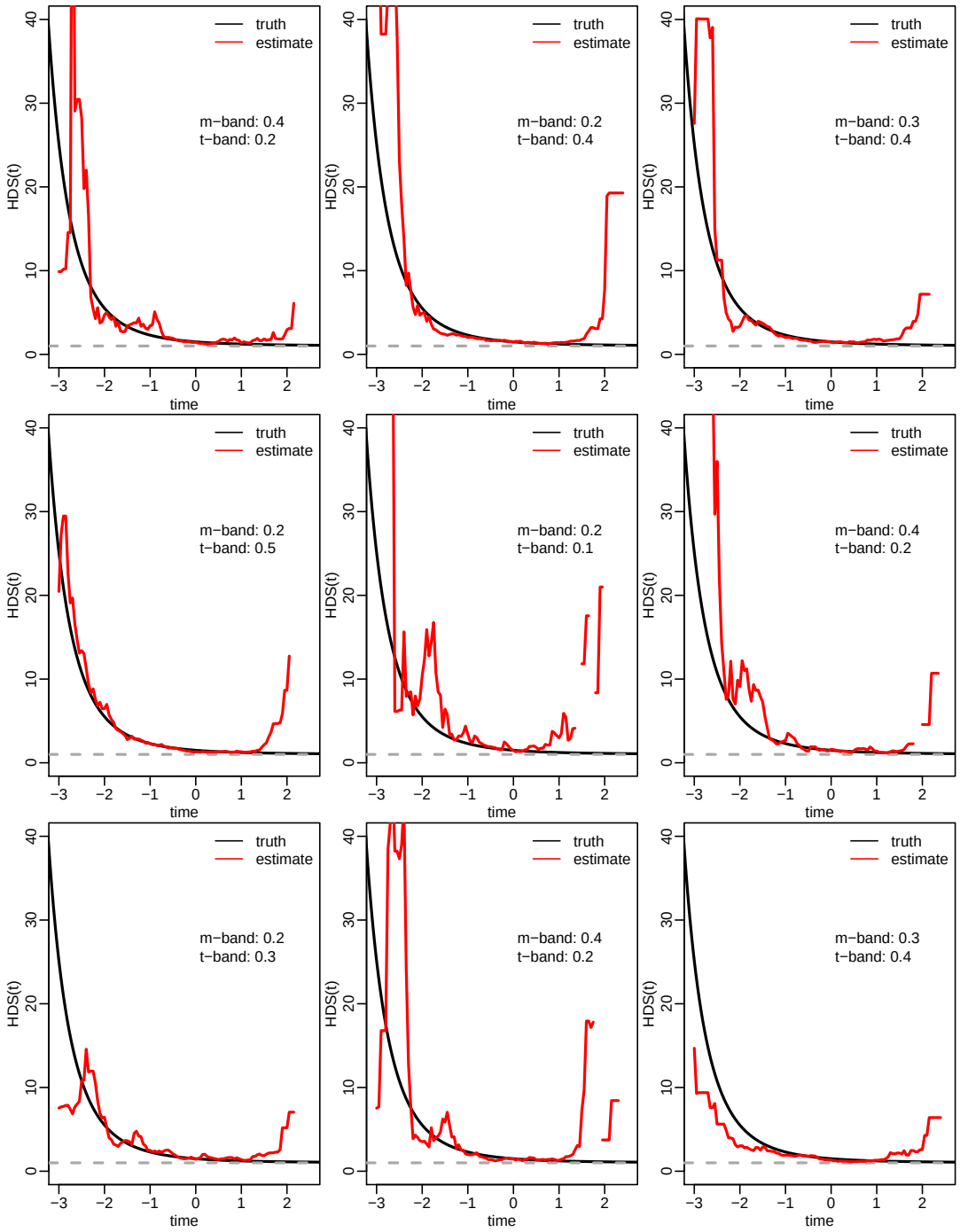

Figure 4.4: Shown are nine estimates (in red) of $HDS(t)$ along with the true $HDS(t)$ (in black). The data and bandwidths used are the same ones used for the corresponding plots in Figure 4.3. That the estimates do not deviate too far from the truth suggest that our bandwidth selection method chooses reasonable bandwidths. While the $HDS(t)$ estimates are not optimal in the integrated square error (ISE) sense (i.e. for each plot there are $HDS(t)$ estimates using bandwidths other than the "optimum" one that have better ISE), this is largely due to the boundary effects. We believe that using dynamic bandwidths (such as nearest neighbors estimates of $\lambda(t|M)$) would mitigate this behavior.

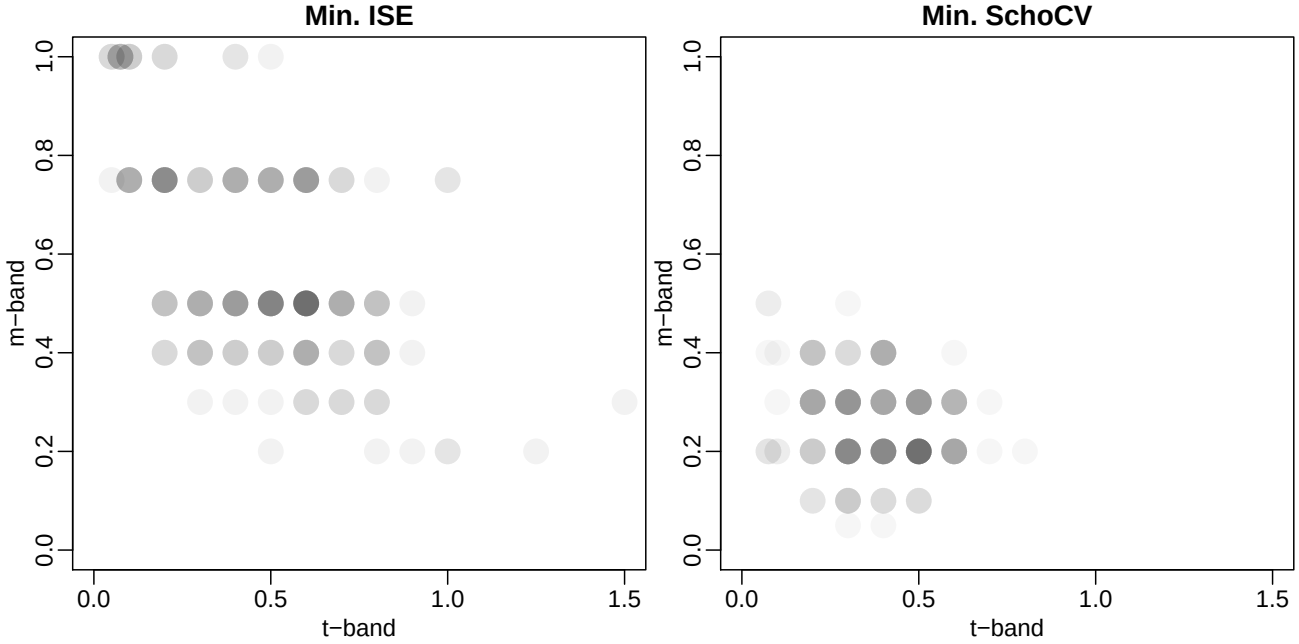

Figure 4.5: Bandwidths corresponding to minimum weighted ISE versus bandwidths selected using selection precedure from Section 4.4

accuracy. Thus we suggest using, for all bootstrap replications, the same bandwidth selected using the entire original dataset.

A more detailed investigation of the bootstrap confidence interval estimator is outlined in the next section.

4.6 Simulations

Coverage rates for $\widehat{HDS}^{NP}(t)$ Figure 4.6 shows the pointwise coverage rates of our proposed percentile bootstrap confidence intervals for $\widehat{HDS}^{NP}(t)$. The rates using $\widehat{HDS}^{LC}(t)$ from Section 2.2.3 and its corresponding standard error estimator are shown for comparison. The actual pointwise coverage rates for a selection of times are shown in Table 4.2.

Data was generated from a model where the survival time T and marker M are bivariate normal with correlation -0.7 and roughly 30% censoring. We generated 1000 datasets (each $n = 500$) and each set of confidence intervals were generated using 250 bootstrap replications. Bandwidths for $\widehat{HDS}^{NP}(t)$ were fixed for all replications (marker bandwidth: 0.2;

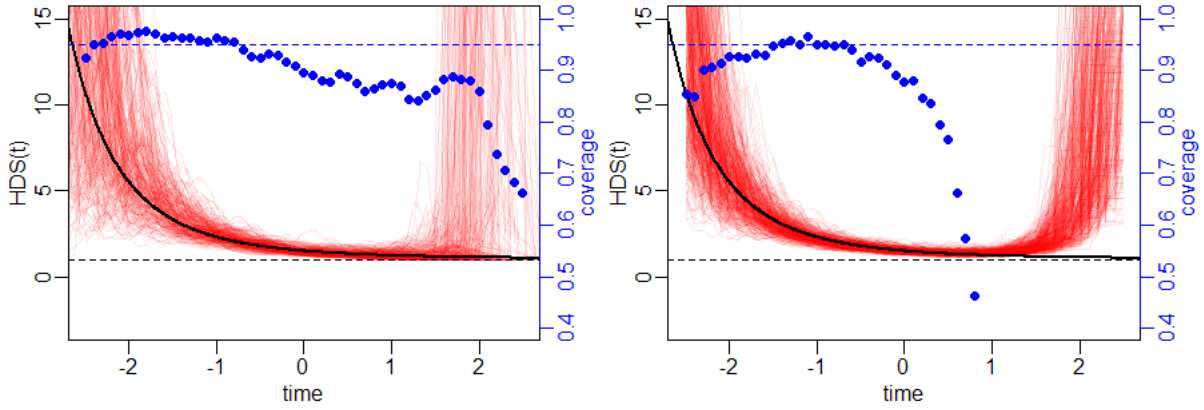

Figure 4.6: Plots comparing performance of confidence intervals for $\widehat{HDS}^{LC}(t)$ **(left)** versus $\widehat{HDS}^{NP}(t)$ **(right)**. Blue dots signify the pointwise coverage rates (above blue line signifies overcoverage, under signifies undercoverage). Each graph also has 250 replications of the respective estimators to illustrate the general distribution of each estimator. The simulated data was generated from a model where survival time and M are bivariate normal with correlation -0.7. Each replication has $n = 500$ and roughly 30% censoring.

time bandwidth: 0.4), as was the bandwidth for $\widehat{HDS}^{LC}(t)$ (time bandwidth: 0.3).

Coverage rates for $\widehat{HDS}^{NP}(t)$ begin steeply dropping around time 0.5. This is explained in part because the risk set only contains roughly 20% of remaining observations at the time, but also because of known difficulties estimating the hazard in the tails. Specifically, estimates of $\widehat{HDS}^{NP}(t)$ at each time require nonparametric hazard estimates of $\lambda(t|m_i)$ for *all* m_i in the dataset. At later times (particularly for data simulated from bivariate normal distributions), this involves an increasing number of hazard estimates in the tails. We believe another reason is that, in the example we chose, the true $HDS(t)$ approaches 1, which is a lower bound for both the parameter and our estimator. The percentile bootstrap confidence intervals are known to perform poorly for extreme values. Further study is warranted in this area.

Mutual information estimation Figure 4.7 shows simulation results for our proposed mutual information estimator under two different scenarios: 1) a moderately strong marker M; and 2) a moderately weak marker M. For the first scenario, we simulated 200 sets

Table 4.2: Pointwise coverage results for $\widehat{HDS}^{LC}(t)$ and $\widehat{HDS}^{NP}(t)$ 95% percentile bootstrap confidence intervals. A visual aide corresponding to row one results is shown in Figure 4.6.

					sup				
Estimator	-2.5	-2.0	-1.5	-1.0	-0.5	0	0.5	1	1.5
$\widehat{HDS}^{LC}(t)$	0.92	0.97	0.97	0.96	0.92	0.90	0.89	0.88	0.86
$\widehat{HDS}^{NP}(t)$	0.85	0.93	0.95	0.95	0.92	0.88	0.77	0.17	0.00

of data ($n = 500$ each) from a model where event time and M are bivariate normal with correlation -0.7, and roughly 33% censoring. For the second scenario, we simulated 200 sets of data ($n = 500$ each) from the same model but with correlation -0.4. For each MI estimate, the bandwidths were automatically selected using the technique proposed in Section 4.4.

The results show that this estimator for $MI(T, M)$ appears to have reasonable bias and variance. Mutual information estimation for continuous or quantitative variables is still considered an open problem Kinney and Atwal (2014), and to our knowledge, there are no published methods for mutual information estimation when one variable is censored. We believe that our results, while still preliminary, suggest it is possible to derive reasonable mutual information estimators with censored data.

4.7 Examples

In this section we use the Mayo PBC data to illustrate how $\widehat{HDS}^{NP}(t)$ can be used to evaluate a single biomarker, and how it compares to the more parametric versions $\widehat{HDS}(t)$ $\widehat{HDS}^{LC}(t)$. We also provide a visual guide for interpreting $HDS(t)$ as the "value of a prognostic marker", again using the Mayo PBC data.

4.7.1 Biomarker evaluation

We consider a scenario using the Mayo PBC data where we would like to evaluate the prognostic value of log(bilirubin) incremental to a default set of predictors; log(prothrombin time), edema, albumin, and age. We thus consider two univariate markers: M_{old}, defined

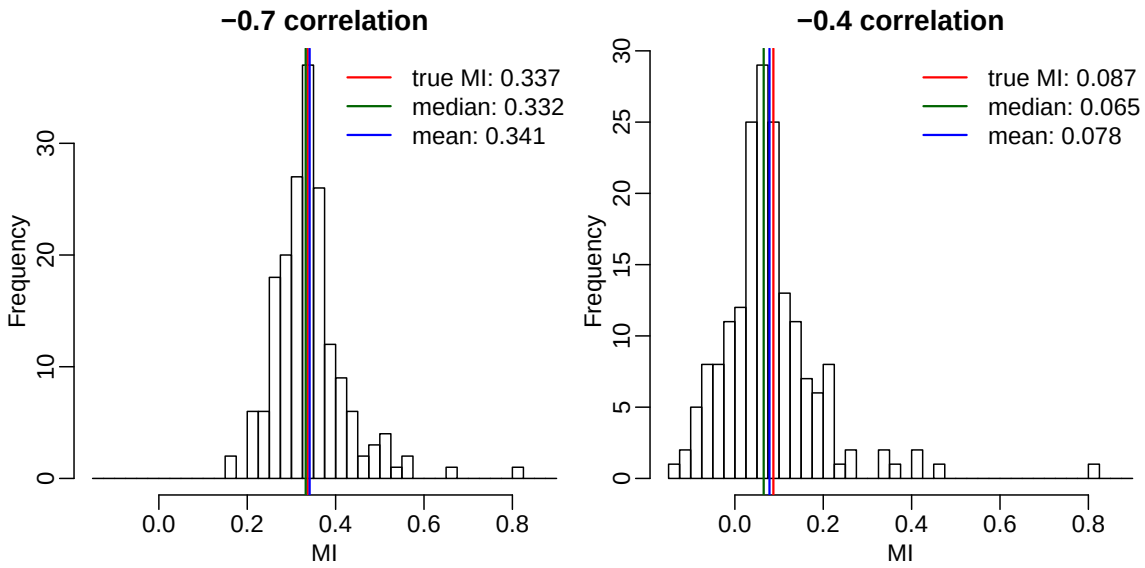

Figure 4.7: Mutual information estimates for markers of two different predictive strengths. Data for **left plot** generated from a model where survival time and M are bivariate normal with correlation -0.7, and roughly 33% censoring. The **right plot** is the same, but with correlation -0.4. For each plot, MI estimates from 200 replications (each $n = 500$) are shown. The bandwidths for each replication were chosen automatically using the selection procedure in Section 4.4.

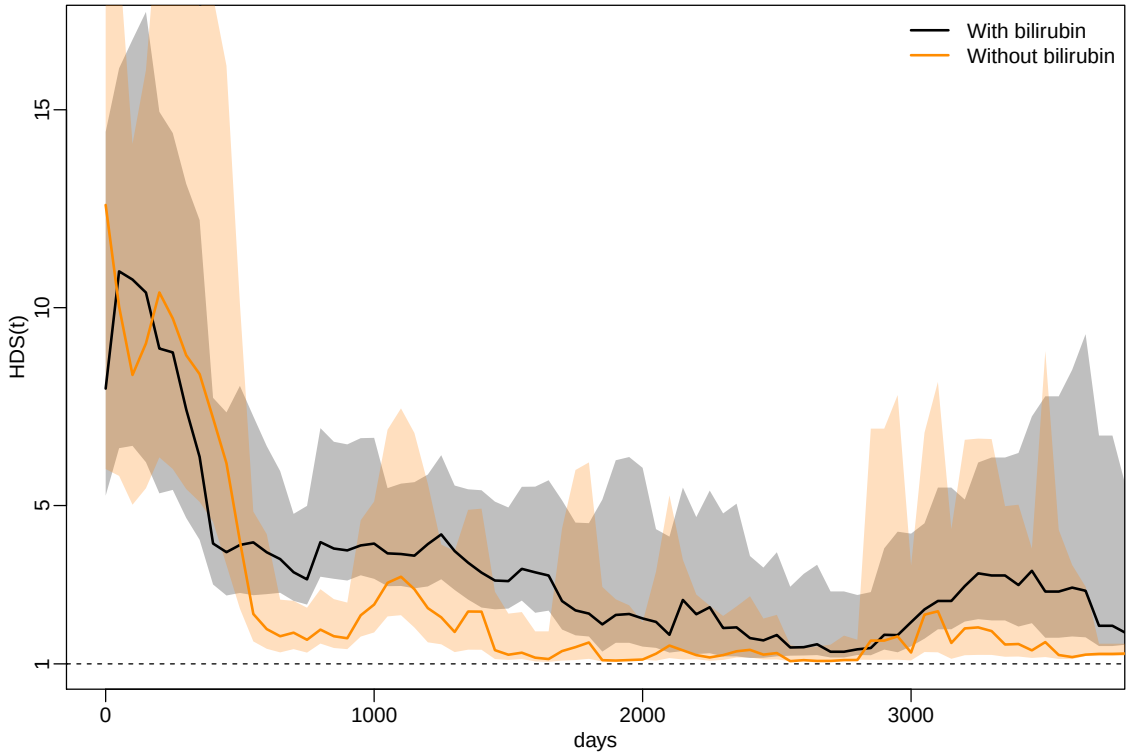

Figure 4.8: $HDS(t)$ estimates using $\widehat{HDS}^{NP}(t)$. For the black line, M is the linear predictor from a Cox regression using log(bilirubin), log(prothrombin time), edema, albumin, and age. For the orange line, M is the same except with log(bilirubin) excluded. The black line calculated using time bandwidth of 400 days and marker bandwidth of 0.5; orange line calculated using 250 days and marker bandwidth of 0.75. Bandwidths were selected using method described in Section 4.4.

as the linear predictor from a Cox regression with the latter four predictors; and M_{new}, defined as the linear predictor from a Cox regression with log(bilirubin) added. Note that for convenience we used Cox regression to reduce the dimension of our predictor sets, but in practice would not be restricted to doing so. Figure 4.8 shows a plot comparing two $\widehat{HDS}^{NP}(t)$ estimates: using M_{old} versus M_{new}.

We can see that the "new" prognostic score M_{new} that incorporates information from log(bilirubin) tends to outperform M_{old} over time.

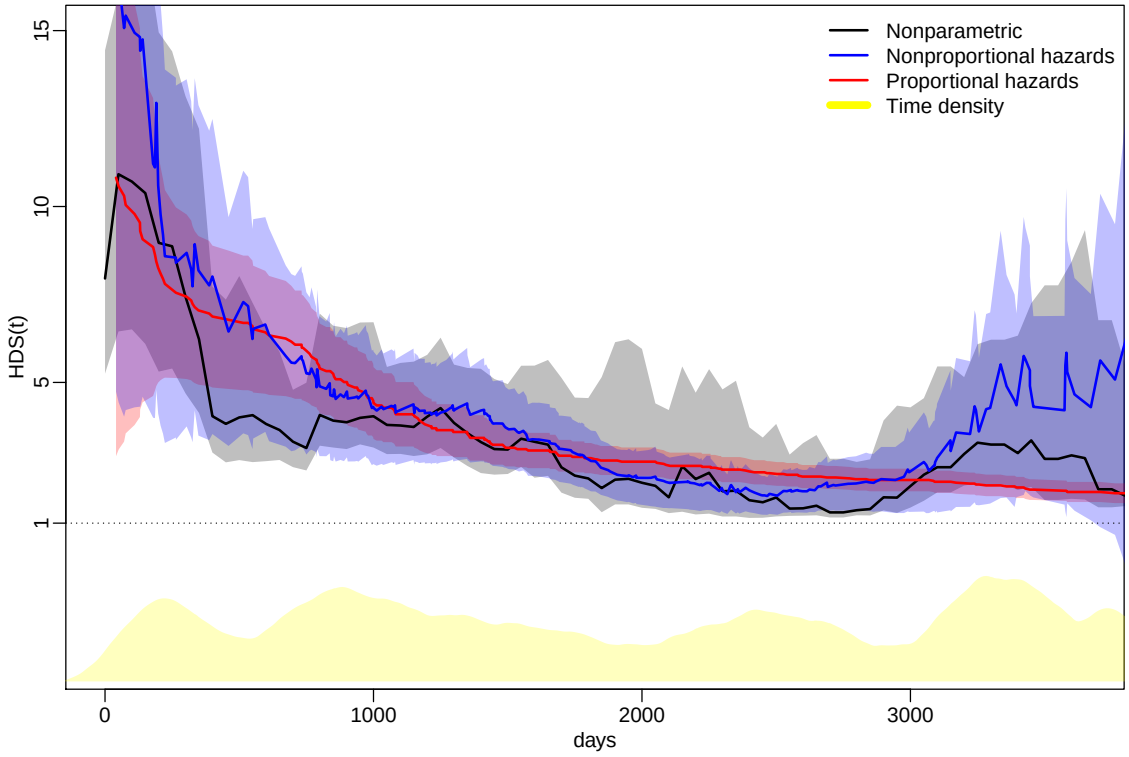

Figure 4.9: $HDS(t)$ estimates using $\widehat{HDS}^{NP}(t)$ (black), $\widehat{HDS}^{LC}(t)$ (blue), and $\widehat{HDS}(t)$ (red). Truncated marginal time density (kernel-smoothed Kaplan-Meier estimator) shown in yellow at the bottom of the graph. This is an updated version of Figure 2.4, where the nonparametric estimator $\widehat{HDS}^{NP}(t)$ has been added.

4.7.2 Comparing estimators

In this dissertation we have offered three estimators of $HDS(t)$: $\widehat{HDS}(t)$, $\widehat{HDS}^{LC})t)$, and $\widehat{HDS}^{NP}(t)$. We provide a comparison of the three predictors using the Mayo PBC data. Specifically, we use M_{new} as defined above and estimate $HDS(t)$ using different predictors. The results, along with pointwise 95% confidence intervals are shown in Figure 4.9.

Overall, regardless of which estimator we had chosen, we would have drawn the same conclusion: the panel of five markers has extremely good discriminatory performance early on, with a steep drop-off (in relative terms) over time to "just" good performance over the rest of the follow-up period.

The similarity among the curves suggests that the proportional hazards assumptions are

not wildly violated. Specifically, the similarity between $\widehat{HDS}(t)$ and $\widehat{HDS}^{LC}(t)$ suggests that the assumption of time-invariant coefficients is not grossly violated. Similarities between $\widehat{HDS}(t)$ and $\widehat{HDS}^{NP}(t)$ suggest that if we view the linear combination of the five markers as a univariate score, the score appears to follow a monotone relationship with the hazard rate (since $\widehat{HDS}^{NP}(t)$ is able to detect non-monotone relationships).

4.7.3 Illustration of $HDS(t)$ as the time-varying prognostic value of a marker

A visual aide for how $HDS(t)$ is interpreted as *the time-varying prognostic value of a marker* M (as defined in Section 4.2) is shown in Figure 4.10. The figure contains a series of three plots, using the Mayo PBC data, and where M is defined as the linear predictor from a Cox regression using log(bilirubin), log(prothrombin time), edema, albumin, and age.

The visualization shows that we can approximate how much each *individual's* risk prediction is improved by incorporating information from our marker M. We can then see how each individual's risk improvement (or degredation) is quantified (as the ratio) and finally incorporated into the $HDS(t)$ estimate. While $HDS(t)$ quantifies time-varying prognostic value in aggregate, such a visualization can provide additional information, such as which individuals actually had worse risk predictions, or whether the $HDS(t)$ estimate is driven by extreme values at any point.

4.7.4 Modifying alternate $HDS(t)$ estimator to estimate $MI(T, M)$

Continuing with the same setup as the previous section, the series of three plots in Figure 4.11 illustrate how the alternate nonparametric $HDS(t)$ estimator can be adapted for $MI(T, M)$. The top plot is identical to the top plot in Figure 4.10. From there, instead of taking the ratio of hazard estimates $\frac{\hat{\lambda}(t|m)}{\hat{\lambda}(t)}$, we take the log ratio $\log\frac{\hat{\lambda}(t|m)}{\hat{\lambda}(t)}$. Mutual information is then estimated by taking a weighted sum of the log ratios, where individual weights are equal to the jump in the Kaplan-Meier estimator corresponding to the event time of each log ratio. This weighted sum is illustrated by the bottom plot, where the x-axis has been transformed to the $S(t)$ scale, and is 0.49.

Note that as will often be the case in practice, the time support only goes from 1 to

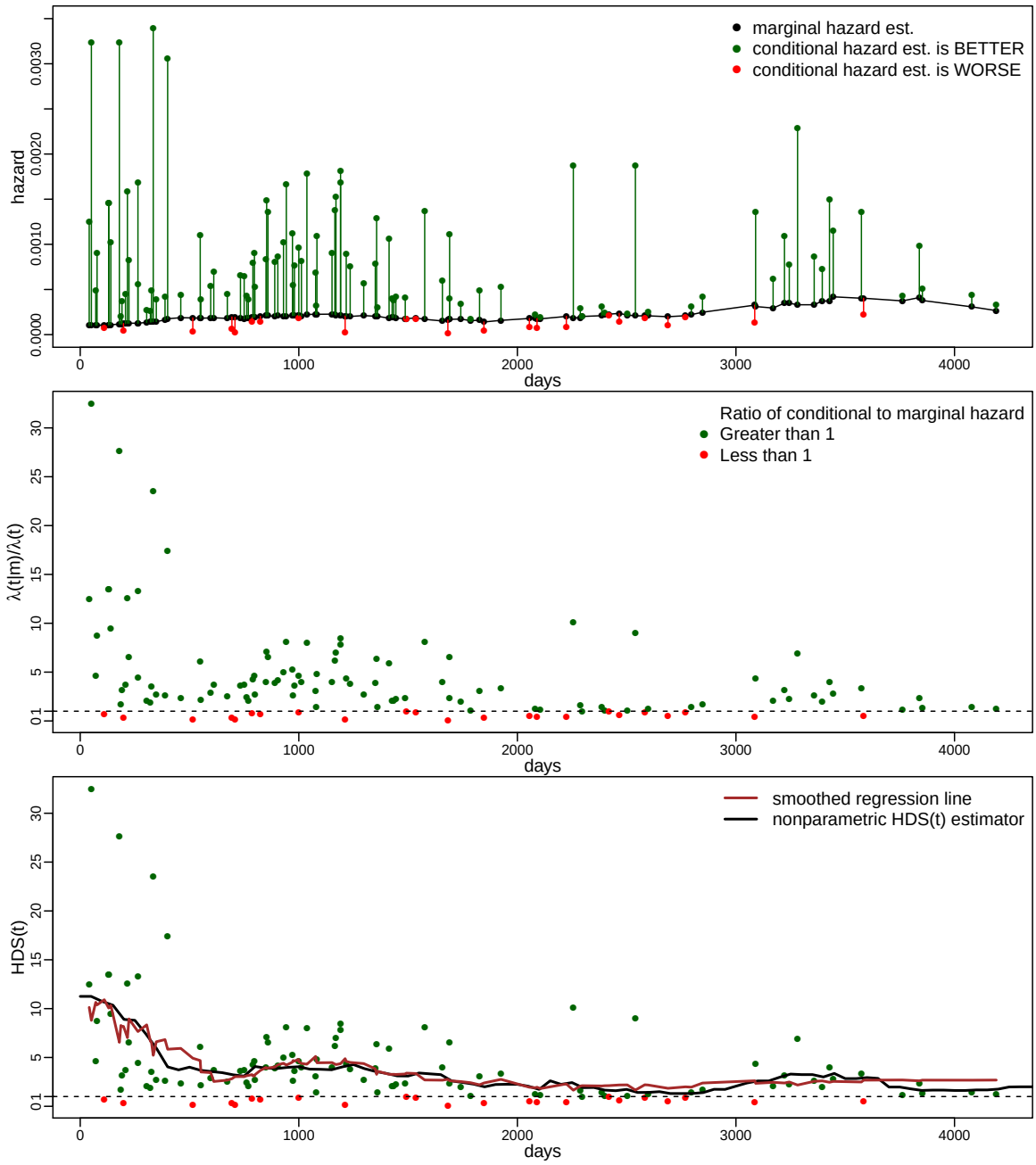

Figure 4.10: A visual aide for how $HDS(t)$ is interpreted as the **"time-varying prognostic value of a marker** M**"**. We use the Mayo PBC data, and M is the linear predictor from a Cox regression of the usual 5 predictors. **The first graph** shows marginal hazard $\lambda(t)$ estimates (black dots) and conditional hazard $\lambda(t|M)$ estimates (green and red dots) at the event times for the 125 observations that experienced events. For each green/red dot M is chosen to be the linear predictor for that observation. M is clearly "predictive", as most hazard predictions are improved (green) when using M. **The second graph** plots the ratios of the conditional (green and red dots) to marginal (black dots) hazard estimates. **The third graph** adds a brown kernel smoother through the scatterplot. The black line is $\widehat{HDS}^{NP}(t)$ for comparison.

0.34, and thus our weighted sum is not strictly an estimate of $MI(T, M)$. To place the weighted sum on the mutual information scale, we could divide it by the length of the support ($0.49/0.66 = 0.74$), effectively "extrapolating" the value out through the entire survival time period.

4.8 Discussion

In this chapter we introduced two nonparametric estimators for $HDS(t)$ and outlined some novel connections that risk-based discrimination measures $HDS(t)$ and $DS(t)$ have with long-standing information theory concepts such as mutual information and f-divergences.

Just as the definition of $AUC^{I/D}(t) = \mathrm{P}\left(M_i > M_j | T_i = t, T_j > t\right)$ implies M is univariate, we consider the scenario where M is restricted to be univariate for $HDS(t)$. Such a restriction achieves two goals: 1) when evaluating multiple predictors, one must first combine them into a univariate risk score, forcing a decoupling of model building and model evaluation; 2) estimating $HDS(t)$ nonparametrically becomes tractable.

We introduced two nonparametric $HDS(t)$ estimators: $\widehat{HDS}^{NP}(t)$ and an alternative graphical smoothing estimator $\widehat{HDS}^{G}(t)$. Both $\widehat{HDS}^{NP}(t)$ and $\widehat{HDS}^{G}(t)$ are dependent on reasonable estimates of $\Lambda(t|M)$ and $\lambda(t|M)$, the latter of which requires two bandwidths: one for time and one for marker. We proposed an automated 2-D bandwidth selection method based on a nonparametric extension of Schoenfeld residuals that is analogous to residual-based procedures for typical kernel regression methods.

Returning to our two estimators, while $\widehat{HDS}^{NP}(t)$ is expected to be a more efficient estimator, the $\widehat{HDS}^{G}(t)$ also offers advantages. Specifically, it is pedagogically valuable in that it provides a graphical connection between the estimator and an empirical formulation of the underlying data. Also, it can be easily modified into an estimator for $MI(T, M)$. The connection to $MI(T, M)$ is valuable, as to our knowledge, little if any has been published on estimating mutual information with survival data. In fact, even the problem of estimating mutual information between uncensored variables is still considered an open problem (Kinney and Atwal, 2014).

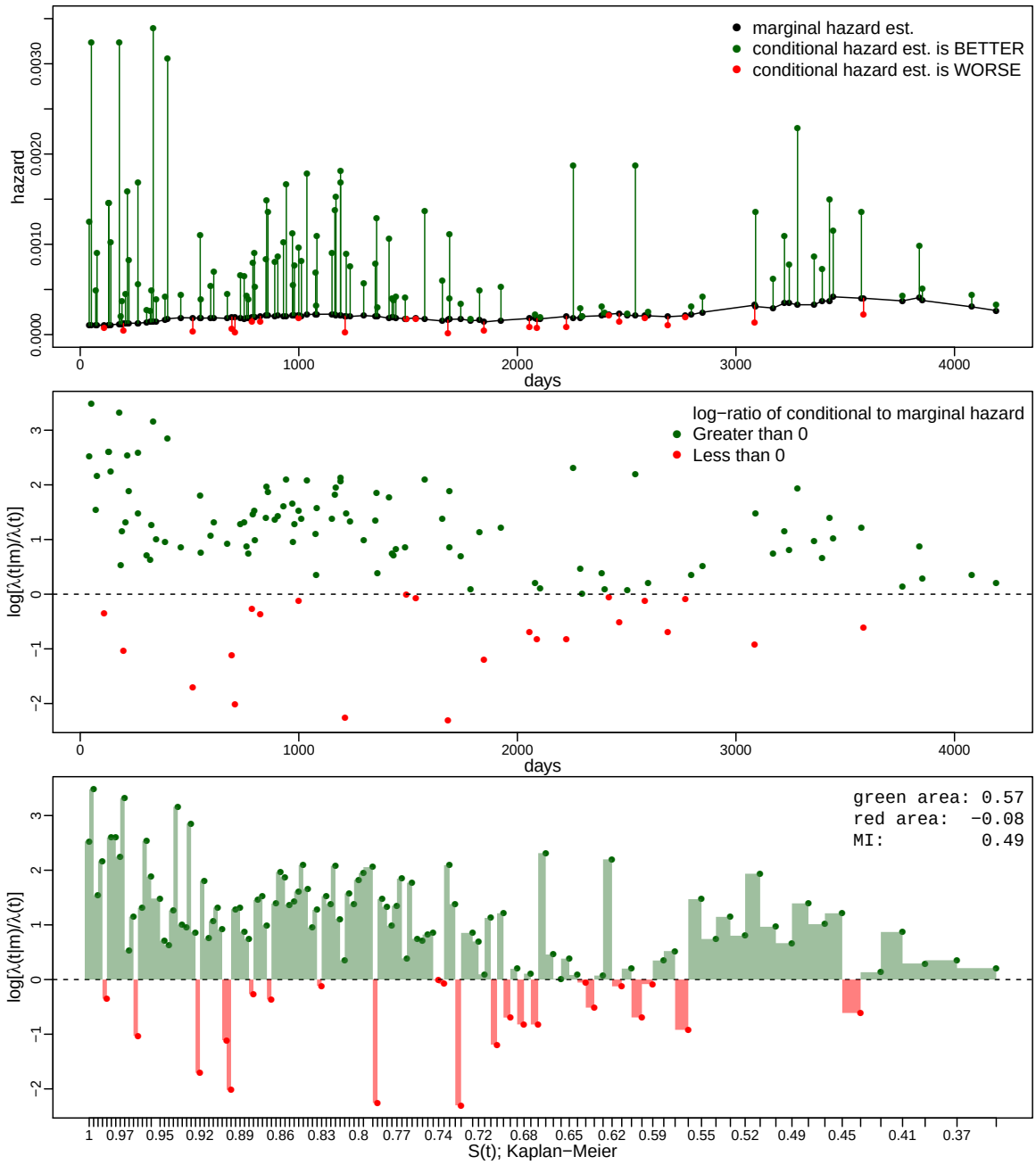

Figure 4.11: A visual aide for how to modify the alternate nonparametric $HDS(t)$ estimator for $MI(T, M)$. We use the Mayo PBC data, and M is the linear predictor from a Cox regression of the usual 5 predictors. **The first graph** shows marginal hazard $\lambda(t)$ estimates (black dots) and conditional hazard $\lambda(t|M)$ estimates (green and red dots) at the event times for the 125 observations that experienced events. **The second graph** plots the *log* ratios of the conditional (green/red dots) to marginal (black dots) hazard estimates. **The third graph** transforms the x-axis to the $S(t)$ scale using the Kaplan-Meier estimator. Mutual information is then estimated by taking the "area under the curve" (green area minus red area) using a piecewise constant interpolation between points.

Future work Our bandwidth selection method is fairly computationally intensive. Assuming no censoring, the function that is minimized over all candidate bandwidths requires calculating $\frac{n(n-1)}{2}$ estimates of $\lambda(t|m)$ at different values of t and m. If we also assume a fixed number of bandwidths being searched regardless of n, the computational complexity is on the order of $O(n^2)$. Strictly, if the data is sorted by marker values, the complexity is actually $O(n^2 \cdot nb)$, since estimating $\lambda(t|m)$ has complexity $O(nb)$, where b is the marker bandwidth. However, since in practice nb is small, we have found that the main source of complexity is in the bandwidth selection procedure.

The computational complexity of the bandwidth selection method could be limiting if many estimates of $HDS(t)$ or $MI(T, M)$ are needed. A recent example where $HDS(t)$ or $MI(T, M)$ could be useful is the development of a 70 gene profile for predicting distant metasteses among sporadic lymph-node-negative breast cancer patients (Van't Veer et al., 2002). Among the steps necessary to reduce 25000 candidate genes to the 70 gene profile, one step requires ranking all genes according to each gene's pairwise correlation coefficient with the outcome. The correlation coefficient is known to only detect linear associations. Using a measure such as $MI(T, M)$ or integrated $HDS(t)$ could detect a wider class of associations, including U-shaped or J-shaped associations. Furtherfore, with survival data, one would not need to choose an arbitrary cut-off (such as five years in the Van't Veer et al. (2002) study).

We considered residual methods and extensions of $\lambda(t)$ bandwidth methods to $\lambda(t|M)$, but there are other avenues that may be worth exploring. Some candidates are: using other measures of model fit such as the Brier score or R^2 extensions, methods based on maximizing some type of likelihood (nonparametric, partial, conditional, etc.) for the data.

In this chapter we also described some preliminary connections between $HDS(t)$ and central information theoretic concepts. This is another potentially fruitful avenue for exploration. For example, a deeper study of our $MI(T, M)$ estimator might be useful – both to compare it with existing mutual information estimators and to evaluate its viability for use with predictive survival models.

Finally, the methods in this chapter are only suitable for time-invariant predictors. With the increasing ubiquity of longitudinal markers, further extensions based on $HDS(t)$ or

$MI(T, M)$ that incorporate such markers could be of great interest.

Chapter 5

DISCUSSION AND FUTURE WORK

5.1 Summary

In this dissertation we developed a new method for evaluating the time-varying predictive accuracy of survival models, called the hazard discrimination summary, $HDS(t)$. We motivated $HDS(t)$ as an incident extension of the discrimination slope $DS(t)$, and as a risk-based complement to the rank-based $AUC^{I/D}(t)$.

We proposed two semiparametric estimators: $\widehat{HDS}(t)$ and $\widehat{HDS}^{LC}(t)$. The former is is based on the Cox model, and the latter is a less parametric version that relaxes the proportional hazards assumption to be only local in time. For both estimators, we showed asymptotic normality, as well as accompanying analytic standard error estimators that appear to have reasonable small-sample coverage rates.

Practical applications often involve longitudinal markers, but the literature for corresponding predictive accuracy measures is still somewhat small. An advantage of incident measures of predictive accuracy is that they are amenable to evaluating models with longitudinal markers, and $HDS(t)$ is no exception. We proposed another semiparametric estimator, $\widehat{HDS}^{R}(t)$ that is suitable for evaluating survival models with longitudinal markers. The estimator is based on the Cox model, but also requires the slightly more restrictive assumption of random censoring.

Just as the definition of $AUC^{I/D}(t) = \mathrm{P}\left(M_i > M_j | T_i = t, T_j > t\right)$ implies M is univariate, for $HDS(t)$ we consider the scenario where M is restricted to be univariate. Such a restriction achieves two goals: 1) when evaluating multiple predictors, one must first combine them into a univariate risk score, forcing a decoupling of model building and model evaluation; 2) estimating $HDS(t)$ nonparametrically becomes tractable.

Taking advantage of the univariate restriction, we introduced two nonparametric $HDS(t)$ estimators: $\widehat{HDS}^{NP}(t)$ and an alternative graphical smoothing estimator $\widehat{HDS}^{G}(t)$. Both

$\widehat{HDS}^{NP}(t)$ and $\widehat{HDS}^{G}(t)$ are dependent on reasonable estimates of $\Lambda(t|M)$ and $\lambda(t|M)$, the latter of which requires two bandwidths: one for time and one for marker. We proposed an automated 2-D bandwidth selection method based off a nonparametric extension of Schoenfeld residuals that is analogous to residual-based procedures for typical kernel regression methods.

Alternative $HDS(t)$ interpretation While we motivated $HDS(t)$ as measure of time-varying discrimination and extension of $DS(t)$, we showed that $HDS(t)$ can be equivalently interpreted as the *value of a prognostic marker*, or $HDS(t) = \mathrm{E}_{M|T=t}\left\{ \frac{\lambda(t|M)}{\lambda(t)} \middle| T = t \right\}$. The alternative interpretation also facilitates several useful connections. For the semiparametric setting (including longitudinal markers), we were able to derive an $HDS(t)$ estimator $\widehat{HDS}^{PL}(t)$ based on smoothing scaled partial likelihood contributions against time. This serves as an interesting pedagogical connection to the well-known partial likelihood, and also graphically connects $HDS(t)$ to an empirical representation of the underlying data. For the nonparametric setting, this continues to apply, by replacing scaled partial likelihood contributions with a nonparametric extension. Furthermore, the nonparametric graphical estimator $\widehat{HDS}^{G}(t)$ can be easily modified to serve as a reasonable estimator for $MI(T, M)$. The connection to $MI(T, M)$ is valuable, as to our knowledge, little if any work has been published on estimating mutual information with survival data.

5.2 Future work

Information theory connections The alternative interpretation of $HDS(t)$ illuminated some preliminary connections between $HDS(t)$ and important information theoretic concepts. We are interested in further exploring this connection. For example, a deeper study of our $MI(T, M)$ estimator would be useful – both to compare it with existing mutual information estimators and to evaluate its viability for use with predictive survival models. With the increasing ubiquity of longitudinal markers, further nonparametric extensions based on $HDS(t)$ or $MI(T, M)$ that incorporate such markers could be of great interest.

Studying additional bandwidth selection methods Our proposed bandwidth selection method for the nonparametric estimators is fairly computationally intensive, which could be limiting if many estimates of $HDS(t)$ or $MI(T, M)$ are needed (e.g. screening a large number of pairwise associations). There are other avenues that may be worth exploring. Some candidates are: using other measures of model fit such as the Brier score or R^2 extensions, and methods based on maximizing some type of likelihood (nonparametric, partial, conditional, etc.) for the data.

Longitudinal markers and time-dependent coefficients For survival models with longitudinal markers, we proposed an $HDS(t)$ estimator based on the proportional hazards model. We believe a further extension that relaxes the proportional hazards model to be local in time (i.e. time-dependent coefficients) would require additional work but would also be feasible and useful.

Competing risks, endpoint severity All of the methods in this dissertation have been developed for models with "standard" time-to-event endpoints. However, it is not uncommon to have data with competing risks, all of which are of interest (e.g. death and cancer recurrence). It is also not uncommon to have survival data with additional endpoint severity measurements (e.g. viral load measurements on those who experience the event of becoming HIV positive). Investigating where extensions of risk-based predictive measures such as $HDS(t)$ can be successfully applied to competing risk and endpoint severity models could also yield useful results.

Extensions to other definitions of cases and controls As outlined in Section 1.4, there are several ways to define cases and controls for survival data. We have mostly reviewed methods for cumulative cases and dynamic controls; and incident cases and dynamic controls, while our contributions in this dissertation have been for the latter. It would be interesting to explore $HDS(t)$ extensions for interval cases and/or static controls.

BIBLIOGRAPHY

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

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, Technical Report, Univ. California, Berkeley.

Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Roux, A. V. D., Folsom, A. R., Greenland, P., JacobsJr, D. R., Kronmal, R., Liu, K., et al. (2002). Multi-ethnic study of atherosclerosis: objectives and design. *American Journal of Epidemiology*, 156(9):871–881.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*, 30(1):93–111.

Chambless, L. E., Cummiskey, C. P., and Cui, G. (2011). Several methods to assess improvement in risk prediction models: extension to survival analysis. *Statistics in medicine*, 30(1):22–38.

Collins, F. S. and Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795.

Cook, N. R. and Ridker, P. M. (2009). Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Annals of internal medicine*, 150(11):795–802.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

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

Dabrowska, D. M. (1987). Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics*, pages 181–197.

D'Agostino, R. and Nam, B.-H. (2004). Evaluation of the performance of survival analysis models: discrimination and calibration measures. *Handbook of statistics*, 23:1–25.

Fisher, L. D. and Lin, D. Y. (1999). Time-dependent covariates in the cox proportional-hazards regression model. *Annual review of public health*, 20(1):145–157.

Ganna, A. and Ingelsson, E. (2015). 5 year mortality predictors in 498 103 uk biobank participants: a prospective population-based study. *The Lancet*.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goff, D. C., Lloyd-Jones, D. M., Bennett, G., O'Donnell, C., Coady, S., and Robinson, J. (2014). 2013 acc/aha guideline on the assessment of cardiovascular risk. *J Am Coll Cardiol*.

González-Manteiga, W., Cao, R., and Marron, J. S. (1996). Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *Journal of the American Statistical Association*, 91(435):1130–1140.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons.

Greenland, P., Alpert, J. S., Beller, G. A., Benjamin, E. J., Budoff, M. J., Fayad, Z. A., Foster, E., Hlatky, M. A., Hodgson, J. M., Kushner, F. G., et al. (2010). 2010 accf/aha guideline for assessment of cardiovascular risk in asymptomatic adults: A report of the american college of cardiology foundation/american heart association task force on practice guidelines developed in collaboration with the american society of echocardiography, american society of nuclear cardiology, society of atherosclerosis imaging and prevention, society for cardiovascular angiography and interventions, society of cardiovascular computed tomography, and society for cardiovascular magnetic resonance. *Journal of the American College of Cardiology*, 56(25):e50–e103.

Grønnesby, J. K. and Borgan, Ø. (1996). A method for checking regression models in survival analysis based on the risk score. *Lifetime data analysis*, 2(4):315–328.

Guffey, D. (2012). *Hosmer-Lemeshow goodness-of-fit test: Translations to the Cox Proportional Hazards Model*. PhD thesis, University of Washington.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):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(2):337–344.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105.

Henderson, R., Diggle, P., and Dobson, A. (2002). Identification and efficacy of longitudinal markers for survival. *Biostatistics*, 3(1):33–50.

Hess, K. R., Serachitopol, D. M., and Brown, B. W. (1999). Hazard function estimators: a simulation study. *Statistics in medicine*, 18(22):3075–3088.

Hilden, J. and Gerds, T. A. (2013). A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in medicine*.

Hjort, N. L. (1985). Bootstrapping cox's regression model. Technical report, DTIC Document.

Hlatky, M. A., Greenland, P., Arnett, D. K., Ballantyne, C. M., Criqui, M. H., Elkind, M. S., Go, A. S., Harrell, F. E., Hong, Y., Howard, B. V., et al. (2009). Criteria for evaluation of novel markers of cardiovascular risk a scientific statement from the american heart association. *Circulation*, 119(17):2408–2416.

Hosmer, D., Lemeshow, S., and May, S. (2008). *Applied survival analysis: Regression modeling of time to event data.* Wiley-Interscience.

Hosmer, D. W. and Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, 9(10):1043–1069.

Hosmer, D. W. and Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

Kerr, K. F., McClelland, R. L., Brown, E. R., and Lumley, T. (2011). Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *American journal of epidemiology*, 174(3):364–374.

Kinney, J. B. and Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359.

Kumar, S., Nilsen, W. J., Abernethy, A., Atienza, A., Patrick, K., Pavel, M., Riley, W. T., Shar, A., Spring, B., Spruijt-Metz, D., et al. (2013). Mobile health technology evaluation: the mhealth evidence workshop. *American journal of preventive medicine*, 45(2):228–236.

McKeague, I. W. and Utikal, K. J. (1990). Inference for a nonlinear counting process regression model. *The Annals of Statistics*, pages 1172–1187.

Metz, C. E. (1986). Roc methodology in radiologic imaging. *Investigative radiology*, 21(9):720–733.

Morimoto, T. (1963). Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331.

Moskowitz, C. S. and Pepe, M. S. (2004). Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics*, 5(1):113–127.

Muller, H.-G. and Wang, J.-L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, pages 61–76.

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., Cronin, M., Baehner, F. L., Watson, D., Bryant, J., et al. (2006). Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor–positive breast cancer. *Journal of clinical oncology*, 24(23):3726–3734.

Patil, P. (1993). Bandwidth choice for nonparametric hazard rate estimation. *Journal of statistical planning and inference*, 35(1):15–30.

Paul, P., Pennell, M. L., and Lemeshow, S. (2013). Standardizing the power of the hosmer–lemeshow goodness of fit test in large data sets. *Statistics in medicine*, 32(1):67–80.

Pencina, M. J., D'Agostino, R. B., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172.

Pepe, M., Feng, Z., and Gu, J. (2008). Comments on evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyondby mj pencina et al., statistics in medicine (doi: 10.1002/sim. 2929). *Statistics in medicine*, 27(2):173–181.

Pepe, M. S., Kerr, K. F., Longton, G., and Wang, Z. (2013). Testing for improvement in prediction model performance. *Statistics in medicine*, 32(9):1467–1482.

Peterson, W. W., Birdsall, T. G., and Fox, W. (1954). The theory of signal detectability. *Information Theory, Transactions of the IRE Professional Group on*, 4(4):171–212.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, pages 453–466.

Saha-Chaudhuri, P. and Heagerty, P. (2013). Non-parametric estimation of a time-dependent predictive accuracy curve. *Biostatistics*, 14(1):42–59.

Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in cox regression. *Biometrics*, 56(1):249–255.

Schoop, R., Graf, E., and Schumacher, M. (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*, 64(2):603–610.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., and Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21(1):128.

Tanner, M. A. and Wong, W. H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *The Annals of Statistics*, pages 989–993.

Tian, L., Zucker, D., and Wei, L. (2005). On the cox model with time-varying regression coefficients. *Journal of the American statistical Association*, 100(469):172–183.

Tsiatis, A. A. (1981). A large sample study of cox's regression model. *The Annals of Statistics*, 9(1):93–108.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14(3):809–834.

Uno, H., Tian, L., Cai, T., Kohane, I. S., and Wei, L. (2012). A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Statistics in Medicine*.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence*. Springer.

van der Vaart, A. W. and Wellner, J. A. (2007). Empirical processes indexed by estimated functions. *Lecture Notes-Monograph Series*, pages 234–252.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536.

Wang, J.-L. (2005). Smoothing hazard rates. *Encyclopedia of biostatistics*.

Watson, G. and Leadbetter, M. (1964a). Hazard analysis. i. *Biometrika*, 51(1/2):175–184.

Watson, G. and Leadbetter, M. (1964b). Hazard analysis ii. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 101–116.

Yandell, B. S. (1983). Nonparametric inference for rates with censored survival data. *The Annals of Statistics*, pages 1119–1135.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30(1):132–156.

Zheng, Y., Cai, T., Pepe, M. S., and Levy, W. C. (2008). Time-dependent predictive values of prognostic biomarkers with failure time outcome. *Journal of the American Statistical Association*, 103(481):362–368.

Zheng, Y. and Heagerty, P. J. (2004). Semiparametric estimation of time-dependent roc curves for longitudinal marker data. *Biostatistics*, 5(4):615–632.

Zheng, Y. and Heagerty, P. J. (2005). Partly conditional survival models for longitudinal data. *Biometrics*, 61(2):379–391.

Zheng, Y. and Heagerty, P. J. (2007). Prospective accuracy for longitudinal markers. *Biometrics*, 63(2):332–341.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577.