Published in final edited form as:

Biometrics. 2017 September ; 73(3): 725–734. doi:10.1111/biom.12628.

A risk-based measure of time-varying prognostic discrimination for survival models

C. Jason Liang* and Patrick J. Heagerty**

Department of Biostatistics, University of Washington, F-600 Health Sciences Building, Campus Mail Stop 357232, Seattle, Washington 98195-7232, U.S.A

Summary

Prognostic survival models are commonly evaluated in terms of both their calibration and their discrimination. Comparing observed and predicted survival curves can assess calibration, while discrimination is typically summarized through comparison of the properties of cases or subjects who experience an event, and the properties of controls represented by event-free individuals. For binary data, discrimination is characterized either by using the relative ranks of cases and controls and a receiver operating characteristic (ROC) curve, or by summarizing the magnitude of risk placed on cases and controls through calculation of the discrimination slope (DS). In this article we propose a risk-based measure of time-varying discrimination that generalizes the discrimination slope to allow use with incident events and hazard models. We refer to the new measure as the hazard discrimination summary (HDS) since it compares the relative risk among incident cases to their associated dynamic risk set controls. We introduce both a model-based estimation procedure that adopts the Cox model, and an alternative approach that locally relaxes the proportional hazards assumption. We illustrate the proposed methods using both a benchmark survival data set, and an oncology study where primary interest is in the time-varying performance of candidate biomarkers.

Keywords

Biomarkers; Cox regression; Discrimination; Survival; Longitudinal analysis; Predictive accuracy

1. Introduction

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 measure, though notably as the change in discrimination slope (DS) which summarizes the mean risk placed on cases and controls respectively; (Yates, 1982; Pepe et al., 2008); 3) under certain

* liangcj@uw.edu. ** heagerty@uw.edu.

6. Supplementary Materials

Web Appendices and Tables referenced in Sections 2.1, 3.1, 3.5, 3.6, and 4, and R code are available with this paper at the *Biometrics* website on Wiley Online Library. An R package implementing the methods in this paper is available at <https://github.com/liangcj/hds>.

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 used in conjunction with other summary measures such as the area under the receiver operating characteristic curve (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 receiver operating characteristic (ROC) curves and C-index or AUC 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.

In applications we typically seek to summarize the discrimination performance of a specific marker, and then to evaluate the incremental value that the marker provides beyond existing standard markers or models. For example, with binary outcomes the AUC or C-index is rarely calculated once, since interest often focuses on the change in AUC for two models, because the difference characterizes improvement in performance associated with a new marker. Similarly, the discrimination slope characterizes the performance of one marker while IDI is the change in the discrimination slope between two models. In order to generalize IDI to models for incident risk we first propose a generalization of the discrimination slope appropriate for incident events and we use the hazard function and ratios of expected risk in order to provide an interpretable time-specific survival model performance summary. In this paper we propose an incident generalization of DS, which we term the hazard discrimination summary (HDS).

We illustrate the new methods by first focusing on a well-known survival prognostic model where IDI can be compared to other performance measures published for these data. The Mayo Primary Biliary Cirrhosis study has been used to construct a mortality prediction model using a small number of clinical measurements (Dickson et al., 1989). We use these data to demonstrate the scale on which incident risk can be summarized, and how performance changes over time. Our second example comes from a multiple myeloma study (Barlogie et al., 2011) where scientific interest is in the performance of individual markers over time, and in the net potential of a multi-marker risk prediction model. We show that HDS can easily summarize the performance of individual markers or models, and then summarize the resulting improvement associated with the use of a multivariate marker model.

1.1 Notation

Let $T_1, \dots, T_n$ be i.i.d. event times and $C_1, \dots, C_n$ be i.i.d. censoring times. Let Y_i be defined as $\min(T_i, C_i)$. In practice, for each i we will only observe Y_i rather than T_i and C_i . Let M_i be the vector of p predictors for observation i . We assume that conditional on M , T and C are independent.

Let $\lambda(t)$ be the hazard function for T and $\lambda(t|M)$ be the conditional hazard function. Two hazard models which we will be referring to in proceeding 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 an array of parameters in $\mathbb{R}^p$. The local-in-time version 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 Background

As mentioned in the introduction, an attractive feature of DS is that it measures discrimination using the native risk scale. For binary outcomes, DS is defined as

$$DS = E\{P(D = 1|M) | D = 1\} - E\{P(D = 1|M) | D = 0\}.$$

Above, D is the binary outcome, with $D = 1$ denoting cases and $D = 0$ controls. DS uses the risk scale to measure how well cases and controls are separated. Here, the natural measure of risk is the probability of experiencing an outcome. For binary data there is a single unambiguous definition of cases and controls, and prognostic risk.

With survival data there are more choices in definitions. Two natural measures of risk in survival analysis are cumulative risk, or the probability of having an event by a certain time; and incident risk, or the hazard rate at a specific time. Analogously, we can also define cases and controls in a cumulative (Heagerty et al., 2000) or incident manner (Heagerty and Zheng, 2005).

In this section we review both cumulative and incident definitions of cases and controls in the setting of survival data. We also summarize the current definition for DS that allows extension to survival outcomes, noting that it uses the cumulative risk scale to measure how well cumulative cases and controls are discriminated.

1.2.1 Discrimination slope and time-varying cases/controls—Chambless et al. (2011) partially extended DS to survival outcomes by proposing a model-based estimator for a time-specific discrimination slope, $DS(t)$, which is the difference in average predicted failure probabilities comparing those with an outcome by time t and those still outcome-free at time t :

$$DS(t) = E_{M|T \leq t}\{P(T \leq t|M) | T \leq t\} - E_{M|T > t}\{P(T \leq t|M) | T > t\}.$$

In addition, $DS(t)$ can be used as a way to compare two models by calculating the difference in $DS(t)$ between the two models. This difference is known as the integrated discrimination improvement:

$$IDI(t) = DS_{new}(t) - DS_{old}(t).$$

Note that we have added subscripts “new” and “old” to indicate $DS(t)$ values for the two different models. The difference in $DS(t)$ is analogous to the commonly used rank-based

method of calculating the difference in AUC associated with two markers. Estimators of $DS(t)$ and $IDI(t)$ that allow for censoring while assuming the Cox model were proposed by Chambless et al. (2011) and Uno et al. (2012).

With survival outcomes, we can refer to those with and without an outcome by t as the *cumulative* cases and dynamic controls, respectively. Specifically, cumulative cases are those with an event before or at t , $\{i: T_i \leq t\}$, and dynamic controls are those still outcome-free at time t , $\{i: T_i > t\}$. Therefore, $DS(t)$ is the difference in expected *cumulative* risk between cases and controls using a time-dependent categorization of cumulative cases.

However, our motivation is to obtain a time-varying measure of discrimination that is appropriate for *incident* risk, and therefore we define cases and controls accordingly. Throughout this article we will refer to those who experience an incident outcome at time t as “cases at time t ”, $\{i: T_i = t\}$, and those still outcome-free at time t as “controls at time t ”, $\{i: T_i > t\}$. This is the same as the definition of *incident* cases and dynamic controls in Heagerty and Zheng (2005).

2. Proposed incident time-dependent discrimination measure

In the previous section 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 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:

$$\frac{\text{mean case hazard}}{\text{mean control hazard}} = \frac{E_{M|T=t} \{\lambda(t|M) | T = t\}}{E_{M|T>t} \{\lambda(t|M) | T > t\}} > 1.$$

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

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

As a ratio measure $HDS(t)$ is not bounded, and can be interpreted relative to a value of one that indicates no discriminatory performance, while increasingly better performance is indicated as $HDS(t)$ approaches infinity. $HDS(t)$ can be shown to be bounded below at 1. The derivation is in Web Appendix A.

2.2 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 (details provided in the appendix):

$$E_{M|T>t} \{ \lambda(t|M) | T > t \} = \lambda(t), \quad (1)$$

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

$$HDS(t) = E_{M|T=t} \left[\frac{\lambda(t|M)}{\lambda(t)} \mid T = t \right],$$

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

To illustrate $HDS(t)$ under different known scenarios, we consider situations where $\{M, \log(T)\}$ are 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 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 comparison, analogous plots of $AUC^{TD}(t)$ (Heagerty and Zheng, 2005) are also shown. We see that $HDS(t)$ and $AUC^{TD}(t)$ 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 as controls, or equivalently that knowing the marker increases the risk assigned to cases by a factor of 2.5 relative to not having marker information. For the same time and correlation, $AUC^{TD}(t)$ 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 simply expresses the relationship between the marker and the outcome using different fundamental scales.

2.3 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 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 proceeding presentation, we will define the following notation for the numerator and denominator of $HDS(t)$

$$\begin{aligned}\bar{H}_1(t) &= E_{M|T=t} \{ \lambda(t|M) | T = t \}, \\ \bar{H}_0(t) &= E_{M|T>t} \{ \lambda(t|M) | T > t \} = \lambda(t).\end{aligned}$$

In section (2.2) 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:

$$\begin{aligned}HDS(t;M1) / HDS(t;M2) &= \frac{\bar{H}_1(t;M1) / \bar{H}_0(t;M1)}{\bar{H}_1(t;M2) / \bar{H}_0(t;M2)} = \frac{\bar{H}_1(t;M1) / \lambda(t)}{\bar{H}_1(t;M2) / \lambda(t)} \\ &= \frac{\bar{H}_1(t;M1)}{\bar{H}_1(t;M2)} = \frac{E_{M1|T=t} \{ \lambda(t|M1) | T = t \}}{E_{M2|T=t} \{ \lambda(t|M2) | T = t \}}\end{aligned}$$

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 $X2$.

3. Estimation

In this section we detail estimation and inference for $HDS(t)$, and for ratios of these discrimination summaries. We first provide alternative mathematical representations that allow us to provide estimates directly based on standard Cox regression estimators and the marginal distribution of the marker.

3.1 HDS(t) in terms of marginal expectations

As defined in Section 2, $HDS(t)$ 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 (details are given in the Web Appendix B):

$$\begin{aligned}\bar{H}_1(t) &= E_{M|T=t} \{ \lambda(t|M) | T = t \} \\ &= \frac{E_M \{ \lambda^2(t|M) \exp \{ -\Lambda(t|M) \} \}}{E_M \{ \lambda(t|M) \exp \{ -\Lambda(t|M) \} \}} \\ \bar{H}_0(t) &= E_{M|T>t} [\lambda(t|M) | T > t] \\ &= \frac{E_M \{ \lambda(t|M) \exp \{ -\Lambda(t|M) \} \}}{E_M \{ \exp \{ -\Lambda(t|M) \} \}}\end{aligned}$$

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).

3.2 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 $HDS(t)$:

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

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

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

We can relax the proportional hazards assumption with a local 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:

$$\widehat{HDS}^{LC}(t) = \frac{\sum_{i=1}^n \exp\left\{2\hat{\beta}_h'(t) m_i - e^{\hat{\beta}_h'(t) m_i} \hat{\Lambda}_h(t)\right\} \sum_{i=1}^n \exp\left\{-e^{\hat{\beta}_h'(t) m_i} \hat{\Lambda}_h(t)\right\}}{\left[\sum_{i=1}^n \exp\left\{\hat{\beta}_h'(t) m_i - e^{\hat{\beta}_h'(t) m_i} \hat{\Lambda}_{a_n}(t)\right\}\right]^2} \quad 3$$

$\hat{\beta}_h(t)$ 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.

3.4 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 (1), $\bar{H}_0(t)$ is equivalent to the marginal hazard $\lambda(t)$. We can again exploit this relation to rewrite $HDS(t)$, under the proportional hazards assumption:

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

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

$$\left\{ \left(y_i, \frac{\exp(\hat{\beta}' m_i)}{(1/n_i) \sum_{j \in R_i} \exp(\hat{\beta}' m_j)} \right) : t_i < c_i \right\}, \text{ where } R_i \text{ represents the risk set at time}$$

y_i and $\sum_{j=1}^n I(y_j \geq y_i)$. A simple graphical estimator of $HDS(t)$ would then be 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 4 we use a myeloma data set to illustrate the graphical connection between $HDS(t)$ and the partial likelihood.

3.5 Inference for $HDS(t)$

Under the Cox model we show that $\sqrt{n} \{ \widehat{HDS}(t) - HDS(t) \}$ is asymptotically normal, pointwise for each t . While the details are in the Web Appendix C, we will outline the main steps in this section. Since $\lambda(t|M) = \lambda_0(t) \exp(\beta' M)$ we can write $HDS(t)$ and $\widehat{HDS}(t)$ as follows:

$$\begin{aligned} HDS(t) &= \frac{E_M \left[\exp \{ 2\beta' M - e^{\beta' M} \Lambda_0(t) \} \right] E_M \left[\exp \{ -e^{\beta' M} \Lambda_0(t) \} \right]}{E_M \left[\exp \{ \beta' M - e^{\beta' M} \Lambda_0(t) \} \right]^2} \\ \widehat{HDS}(t) &= \frac{\sum_{i=1}^n \exp \{ 2\hat{\beta}' m_i - e^{\hat{\beta}' m_i} \hat{\Lambda}_0(t) \} \sum_{i=1}^n \exp \{ -e^{\hat{\beta}' m_i} \hat{\Lambda}_0(t) \}}{\left[\sum_{i=1}^n \exp \{ \hat{\beta}' m_i - e^{\hat{\beta}' m_i} \hat{\Lambda}_0(t) \} \right]^2} \end{aligned}$$

If we let $a_\theta = E_M[\exp \{ \theta \cdot \beta' M - e^{\beta' M} \Lambda_0(t) \}]$ and

$$\hat{a}_\theta = \sum_{i=1}^n \exp \left\{ \theta \cdot \hat{\beta}' m_i - e^{\hat{\beta}' m_i} \hat{\Lambda}_0(t) \right\} \text{ for } \theta \in 0, 1, 2, \text{ we have}$$

$$\widehat{HDS}(t) = \frac{a_2 a_0}{a_1^2} \text{ and } \widehat{HDS}^{LC}(t) = \frac{\hat{a}_2 \hat{a}_0}{\hat{a}_1^2}.$$

Tsiatis (1981) showed consistency of $\hat{\beta}$ and $\hat{\Lambda}_0(t)$, and combined with results from van der Vaart and Wellner (2007, Theorem 2.1) it is straightforward to show joint asymptotic normality of the three $\hat{a}_\theta$ estimators. Once joint normality is established, the delta method can be applied to finish the proof. Furthermore, the resulting asymptotic variance is of a form that lends itself to an analytic estimator.

Standard errors for $\widehat{HDS}^{LC}(t)$ —The estimator $\widehat{HDS}^{LC}(t)$ is similar to $\widehat{HDS}(t)$, with the only difference being that $\hat{\beta}$ and $\hat{\Lambda}_0(t)$ are replaced with local versions $\hat{\beta}_h(t)$ and $\hat{\Lambda}_h(t)$. The subscript h corresponds to a smoothing parameter that varies with n . $\hat{\beta}_h(t)$ and $\hat{\Lambda}_h(t)$ are consistent estimators of $\beta(t)$ and $\Lambda_0(t)$, and asymptotically normal at a $\sqrt{nh}$ rate when $h = O(n^{-\nu})$ with $1/4 < \nu < 1/2$ (Cai and Sun, 2003; Tian et al., 2005). Since $\hat{\beta}_h(t)$ and $\hat{\Lambda}_h(t)$ are consistent estimators, inferential procedures analogous to those for $\widehat{HDS}(t)$ can still be adapted, though the rate of convergence will no longer be $\sqrt{n}$. The details are in Web Appendix C.

Standard errors for ratio of two $\widehat{HDS}(t)$ functions—Assuming $\widehat{HDS}_{new}(t)$ and $\widehat{HDS}_{old}(t)$, 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 $\widehat{HDS}_{new}(t) / \widehat{HDS}_{old}(t)$.

3.6 Simulations to evaluate $\widehat{HDS}(t)$ and $\widehat{HDS}^{LC}(t)$

We did two simulations studies to evaluate the small sample properties of $\widehat{HDS}(t)$ and $\widehat{HDS}^{LC}(t)$.

For the first simulation study, we assumed a Cox model with approximately 25% uninformative censoring. The data generating mechanism was specified as

$$\lambda(t|M) = 0.5 \cdot \exp(0.5 \cdot M_1 + 1.5 \cdot M_2), \\ M_1 \text{Unif}(0, 2), \quad M_2 \text{Unif}(0, 2).$$

We simulated 1000 datasets, each $n = 500$; for each dataset we calculated $\widehat{HDS}(t)$ and $\widehat{HDS}^{LC}(t)$, along with analytic standard error estimates from Section 3.5. An

Epanechnikov kernel $K(u) = \frac{3}{4}(1 - u^2)1_{|u| \leq 1}$ scaled by a bandwidth of $h = 1/5$,

or $5K(5u)$, was used for all $\widehat{HDS}^{LC}(t)$ calculations. The bandwidth was chosen based on guidelines by Tian et al. (2005). Cai and Sun (2003) suggest a bandwidth of $h = O(n^{-1/5})$, which optimizes the mean squared error for $\hat{\beta}_h(t)$ but results in asymptotically biased estimates of $\beta(t)$. Tian et al. (2005) recommend choosing h that converges to zero at a faster rate; specifically, they show that when $h = O(n^{-\nu})$ with $1/5 < \nu < 1$, $\hat{\beta}_h(t)$ is no longer

asymptotically biased. Further, when $1/4 < \nu < 1/2$, $\sqrt{n} \{ \hat{\Lambda}_h(t) - \Lambda_0(t) \}$ converges to a mean-0 Gaussian process. Thus, as outlined in Section 3.5, when $1/4 < \nu < 1/2$,

$\sqrt{nh} \{ \widehat{HDS}^{LC}(t) - HDS(t) \}$ is pointwise asymptotically normal. For our simulations we chose $h = 1/5$, which is slightly less than $n^{-1/4}$ for $n = 500$. $\widehat{HDS}(t)$ shows very little bias and coverage rates close to the nominal 95% at all times. $\widehat{HDS}^{LC}(t)$ shows very little bias at earlier times, with moderate bias at later times when the data is sparser. Approximately 10% of the data lies after $t = 0.5$ in each dataset. The coverage rates fluctuate between 0.87 and 0.99 at different times.

For the second simulation study, we assumed a local-in-time Cox model with approximately 25% uninformative censoring. The data generating mechanism was specified as a non-proportional hazards model:

$$\begin{aligned} \lambda(t|M) &= \exp(t \cdot M_1 + 0.5 \cdot M_2), \\ M_1 &\text{Unif}(1, 3), \quad M_2 \sim N(0, 1). \end{aligned}$$

We simulated 1000 datasets, each $n = 500$; for each dataset we calculated $\widehat{HDS}(t)$ and $\widehat{HDS}^{LC}(t)$, along with analytic standard error estimates from Section 3.5. As in the first simulation study, an Epanechnikov kernel with fixed bandwidth of $h = 1/5$ was used for $\widehat{HDS}^{LC}(t)$ calculations. $\widehat{HDS}(t)$, which is misspecified, becomes increasingly biased for greater times, and coverage rates quickly deteriorate as the bias increases. $\widehat{HDS}^{LC}(t)$ shows little bias, with coverage rates between 0.91 and 0.93 at different times.

Detailed results are shown in Table 1. We also note that if percentile bootstrap confidence intervals are used and the data is generated from a Cox model, then the coverage rates for $\widehat{HDS}(t)$ are within 0.01 of 0.95 at all evaluated times. Likewise, if percentile bootstrap confidence intervals are used and the data is generated from a local-in-time Cox model, then the coverage rates for $\widehat{HDS}^{LC}(t)$ are within 0.01 of 0.95 at nearly all evaluated times. Details of the coverage rates for the percentile bootstrap confidence intervals are shown in Web Appendix D. Based on the results from our simulation study, if computation time is not an issue, we recommend percentile bootstrap confidence intervals. However, if computation

time is important, the analytic standard errors can still provide approximate nominal coverage.

4. Examples

Mayo PBC data

We first use the Mayo PBC data to illustrate $HDS(t)$ on a benchmark dataset. The data contains 312 subjects, for whom 125 had events. We use $\log(\text{bilirubin})$, $\log(\text{prothrombin time})$, edema, albumin, and age as the predictors. The hazard ratios from a multivariate Cox regression using the five predictors is in Web Appendix E.

Using all five predictors and assuming proportional hazards, we can estimate $HDS(t)$ using methods detailed in Section 3.2. 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. 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.

We provide two additional examples using the Mayo PBC data. First, we compare $HDS(t)$, $DS(t)$, and $AUC^{LD}(t)$. For each of the three measures, using the five predictors listed above, we calculate both Cox model estimators and local Cox model estimators. Comparing $HDS(t)$ to $DS(t)$ illustrates differences between an incident and cumulative measure; comparing $HDS(t)$ to $AUC^{LD}(t)$ illustrates differences between a risk-based and rank-based measure. Detailed results are in Web Appendix F. Guidance for how to choose among the three measures is in Section 5. Second, to demonstrate how one can evaluate a single marker, we estimate the ratio $HDS_{new}(t)/HDS_{old}(t)$. $HDS_{new}(t)$ is $HDS(t)$ for all five markers; $HDS_{old}(t)$ is $HDS(t)$ for the same markers but with $\log(\text{bilirubin})$ omitted. When $\log(\text{bilirubin})$ is added, average risk estimates are improved by a factor of roughly 1.5 over all times. Detailed results are in Web Appendix F.

Multiple myeloma trial S9321 data

To illustrate $HDS(t)$ as a tool for comparing time-varying performance, we use a multiple myeloma dataset from the Southwest Oncology Group (SWOG) trial S9321. The dataset has 775 patients, with a median survival of 48 months. Of the 775 patients, 505 experienced events. We also have data on eight baseline markers: platelet count; $\log(\text{SB2M})$; $\log(\text{creatinine})$; age; albumin; calcium; lactate dehydrogenase; and hemoglobin. SB2M is short for “beta-2-microglobulin”

The hazard ratios from a Cox regression with the eight predictors are shown in Web Appendix G. Most of the confidence intervals include or are close to one, suggesting that the predictors are individually only weakly informative. 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. Both the time-invariant and time-varying hazard ratios are shown in Figure 3.

To more formally assess the time-varying predictive performance of each marker, we estimate $HDS(t)$ for each of the eight individual markers. We also assess the combined time-varying predictive performance of all eight markers by estimating $HDS(t)$ using all eight markers together. 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 4.

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.

We use the myeloma data set to illustrate the graphical connection between $HDS(t)$ and the partial likelihood outlined in Section 3.4. Figure 5 shows a scatterplot of scaled partial

$$\text{likelihood contributions versus time, } \left\{ \left(y_i, \frac{\exp(\hat{\beta} m_i)}{(1/n_i) \sum_{j \in R_i} \exp(\hat{\beta} m_j)} \right) : t_i < c_i \right\},$$

where m is the vector of all eight markers. The smoothed scatterplot line and $\widehat{HDS}(t)$ estimated using formula (2) are both shown as well, and the two summaries curves are in close agreement.

Finally, we provide two additional examples using the myeloma data. First, we compare $HDS(t)$, $DS(t)$, and $AUC^{LD}(t)$. For each of the three measures, using only log(SB2M), we calculate both Cox model estimators and local Cox model estimators. Comparing $HDS(t)$ to $DS(t)$ illustrates differences between an incident and cumulative measure; comparing $HDS(t)$ to $AUC^{LD}(t)$ illustrates differences between a risk-based and rank-based measure. Detailed results are in Web Appendix H. Guidance for how to choose among the three measures is in Section 5. Second, to illustrate how one can evaluate the addition of multiple markers to a panel, we estimate the ratio $HDS_{new}(t)/HDS_{old}(t)$. $HDS_{new}(t)$ is $HDS(t)$ for all eight markers; $HDS_{old}(t)$ is $HDS(t)$ using only log(SB2M). Adding the seven predictors to

$\log(\text{SB2M})$ improves the hazard predictions by a modest factor of between 1.1 and 1.2 over all times. Detailed results are in Web Appendix H.

5. Discussion

We have proposed a time-varying measure of discrimination for survival models, $HDS(t)$, that naturally generalizes $DS(t)$ 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. While cumulative measures can be seen as an extension of predictive accuracy measures for binary outcome models, incident measures are more naturally coupled to hazard models. Incident cases also do not nest over time as cumulative cases do, which means incident measures at different times can be compared directly to summarize time-varying performance. Another consequence of the non-nesting property is that incident measures, when integrated over time into a single number summary, tend to be more meaningful global measures of performance.

While $HDS(t)$ is a risk-based incident measure, $AUC^{ID}(t)$ is a rank-based incident measure. Rank-based measures are invariant to monotone transformations of the marker; however, they do not provide an assessment of the magnitude of the difference in risk between cases and controls. Some may find risk-based measures more intuitive for incorporating into subsequent decision making. Another consequence the rank-based scale is that $AUC^{ID}(t)$ only will detect monotone dependencies between risk and marker at each time. In contrast, it is possible for $HDS(t)$ to detect non-monotone dependencies by adding splines to the Cox model or discretizing the markers. Finally, note that when markers are continuous, $AUC^{ID}(t) = P(M_1 > M_2 | T_1 = t, T_2 > t) = P(M_1 \geq M_2 | T_1 = t, T_2 > t)$. However, when markers are discrete, making ties in marker values between cases and controls possible, the latter two values can differ. No analogous ambiguity exists for $HDS(t)$.

We have presented $HDS(t)$ as a risk-based alternative to rank-based methods such as $AUC^{ID}(t)$. However, 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. In other words, since the AUC is only well-defined when M is univariate, a model must first be used to combine the predictors into a univariate prognostic score. Only then is the prognostic value of the score evaluated using AUC. In contrast, $HDS(t)$ is well-defined even when M is multidimensional, so one can forego the task of using a separate model to first generate a prognostic score from the multiple predictors. This requires either the Cox model (or the local Cox model) 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 $HDS(t)$.

We mention three potential directions for further related research. First, just as the C-index is recovered by taking a weighted average over time of time-dependent incident AUC, it may be useful to consider whether some type of time-averaging of $HDS(t)$ also results in a meaningful quantity. Second, we have focused on the scenario where M is only a baseline marker. However, given how amenable incident measures are to time-varying markers, it would be of interest to extend $HDS(t)$ so as to accomodate time-varying markers. Third, it would be interesting to further explore an extension of $HDS(t)$ that, like AUC curves, requires separately generating a prognostic score, after which $HDS(t)$ is estimated nonparametrically.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful for NIH grants that have partially supported this work, including NIH R01 HL072966, and UL1 TR000423.

Appendix

$$\text{Equality of } E_{M|T>t} \{ \lambda(t|M) | T > t \} \text{ and } \lambda(t)$$

This equality is used as an intermediate step for both demonstrating the graphical connection of $HDS(t)$ to the partial likelihood (Section 3.4); and re-expressing $HDS(t)$ in terms of marginal expectations to facilitate estimation (Section 3.1).

$$\begin{aligned} E_{M|T>t} \{ \lambda(t|M) | T > t \} &= \int \lambda(t|m) f(m|T > t) dm = \int \frac{f(t|m)}{S(t|m)} \cdot \frac{S(t|m)f(m)}{S(t)} dm \\ &= \int \frac{f(t|m)f(m)}{S(t)} dm = \frac{1}{S(t)} \int f(t, m) dm \\ &= \frac{1}{S(t)} f(t) = \lambda(t) \end{aligned}$$

References

- Barlogie B, Bolejack V, Schell M, Crowley J. Prognostic factor analyses of myeloma survival with intergroup trial s9321 (int 0141): examining whether different variables govern different time segments of survival. *Annals of Hematology*. 2011; 90:423–428. [PubMed: 21153898]
- Cai Z, Sun Y. Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*. 2003; 30:93–111.
- Chambless LE, Cummiskey CP, Cui G. Several methods to assess improvement in risk prediction models: extension to survival analysis. *Statistics in Medicine*. 2011; 30:22–38. [PubMed: 20827726]
- Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*. 1972; 34:187–220.
- Dickson ER, Grambsch PM, Fleming TR, Fisher LD, Langworthy A. Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology*. 1989; 10:1–7. [PubMed: 2737595]
- Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, Greenland P, Lackland DT, Levy D, O'Donnell CJ, et al. 2013 acc/aha guideline on the assessment of cardiovascular risk: a

report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*. 2014; 63

Greenland P, Alpert JS, Beller GA, Benjamin EJ, Budoff MJ, Fayad ZA, Foster E, Hlatky MA, Hodgson JM, Kushner FG, et al. 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*. 2010; 56:e50–e103. [PubMed: 21144964]

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

Heagerty PJ, Lumley T, Pepe MS. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*. 2000; 56:337–344. [PubMed: 10877287]

Heagerty PJ, Zheng Y. Survival model predictive accuracy and roc curves. *Biometrics*. 2005; 61:92–105. [PubMed: 15737082]

Hjort NL. Bootstrapping cox's regression model. 1985 Technical report, DTIC Document.

Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, Go AS, Harrell FE, Hong Y, Howard BV, et al. Criteria for evaluation of novel markers of cardiovascular risk a scientific statement from the american heart association. *Circulation*. 2009; 119:2408–2416. [PubMed: 19364974]

Kerr KF, McClelland RL, Brown ER, Lumley T. Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *American Journal of Epidemiology*. 2011; 174:364–374. [PubMed: 21673124]

Pencina MJ, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in medicine*. 2008; 27:157–172. [PubMed: 17569110]

Pepe M, Feng Z, Gu J. Commentary on 'evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond'. *Statistics in medicine*. 2008; 27:173–181. [PubMed: 17671958]

Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010; 21:128. [PubMed: 20010215]

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

Tsiatis AA. A large sample study of cox's regression model. *The Annals of Statistics*. 1981; 9:93–108.

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

van der Vaart, AW., Wellner, JA. Weak Convergence. Springer; 1996.

van der Vaart AW, Wellner JA. Empirical processes indexed by estimated functions. *Lecture Notes-Monograph Series*. 2007; 55:234–252.

Yates JF. External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*. 1982; 30:132–156.

Figure 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: $AUC^{TD}(t)$ for varying correlations of $(M, \log(T))$.

Figure 2.

$HDS(t)$ estimates using the Mayo PBC data. Shown are estimates under the proportional hazards assumption, $\widehat{HDS}(t)$, and the relaxed local-in-time (bandwidth = 730 days) proportional hazards assumption, $\widehat{HDS}^{LC}$, with pointwise 95% confidence intervals. Truncated marginal time density (estimated by using a kernel-smoothed Kaplan-Meier estimator) shown at the bottom. This figure appears in color in the electronic version of this article.

Figure 3.

Adjusted time-varying hazard ratio estimates for each predictor, using the multiple myeloma data. $\beta(t)$ values were estimated with the Cai and Sun (2003) local constant method. Straight lines extending to the far left edge 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. This figure appears in color in the electronic version of this article.

Figure 4.

Using the multiple myeloma data, $HDS(t)$ estimates 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. For each plot, the eight thinner lines correspond to $HDS(t)$ estimates for each of the eight univariate markers; the thick line with a confidence band is the $HDS(t)$ estimate using all eight markers together. This figure appears in color in the electronic version of this article.

Figure 5.

Multiple myeloma data. 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 i th partial likelihood contribution (evaluated at $\hat{\beta}$) scaled by the size of the risk set. The dotted line is $\widehat{HDS}(t)$, calculated using (2). The solid line is a smoother through the scaled partial likelihood contributions. This figure appears in color in the electronic version of this article.

Table 1

Pointwise coverage results for $\widehat{HDS}^{LC}(t)$ and $\widehat{HDS}(t)$ 95% confidence intervals using analytic estimators.

Top: data simulated from Cox model. Bottom: data simulated from Cox model with time-varying coefficients as in (Cai and Sun, 2003).

time	$HDS(t)$ – Cox model	Cox model		Local Cox model	
		$\widehat{HDS}(t)$	Coverage	$\widehat{HDS}^{LC}(t)$	Coverage
0.025	1.83	1.83	0.95	1.83	0.87
0.050	1.84	1.84	0.95	1.84	0.88
0.100	1.84	1.84	0.94	1.84	0.90
0.200	1.76	1.76	0.92	1.76	0.97
0.300	1.66	1.66	0.93	1.66	0.99
0.400	1.57	1.57	0.93	1.57	0.99
0.500	1.49	1.49	0.95	1.51	0.97
0.600	1.43	1.43	0.96	1.47	0.95
0.700	1.39	1.39	0.97	1.45	0.95
0.800	1.35	1.35	0.97	1.43	0.94
0.900	1.32	1.32	0.97	1.43	0.95
1.000	1.29	1.29	0.98	1.43	0.92

time	$HDS(t)$ – Local Cox model	Cox model		Local Cox model	
		$\widehat{HDS}(t)$	Coverage	$\widehat{HDS}^{LC}(t)$	Coverage
0.025	1.28	1.34	0.97	1.31	0.91
0.050	1.28	1.33	0.97	1.31	0.91
0.100	1.28	1.32	0.96	1.30	0.92
0.200	1.28	1.30	0.96	1.29	0.92
0.300	1.28	1.29	0.96	1.28	0.93
0.400	1.29	1.28	0.93	1.28	0.93
0.500	1.30	1.26	0.81	1.29	0.93
0.600	1.32	1.24	0.57	1.30	0.93
0.700	1.34	1.23	0.27	1.31	0.92
0.800	1.36	1.22	0.27	1.32	0.92
0.900	1.37	1.20	0.27	1.34	0.92
1.000	1.38	1.19	0.01	1.35	0.93