Journal of Data Science 23 (3), 470–490
July 2025

DOI: 10.6339/25-JDS1163
Statistical Data Science

Comparing Estimators of Discriminative Performance of Time-to-Event Models

Ying Jin[1],* and Andrew Leroux[1]

[1] *Department of Biostatistics & Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, U.S.*

Abstract

Predicting the timing and occurrence of events is a major focus of data science applications, especially in the context of biomedical research. Performance for models estimating these outcomes, often referred to as time-to-event or survival outcomes, is frequently summarized using measures of discrimination, in particular time-dependent AUC and concordance. Many estimators for these quantities have been proposed which can be broadly categorized as either semi-parametric estimators or non-parametric estimators. In this paper, we review the mathematical construction of the two classes of estimators and compare their behavior. Importantly, we identify a previously unknown feature of the class of semi-parametric estimators that can result in vastly overoptimistic out-of-sample estimation of discriminative performance in common applied tasks. Although these semi-parametric estimators are popular in practice, the phenomenon we identify here suggests that this class of estimators may be inappropriate for use in model assessment and selection based on out-of-sample evaluation criteria. This is due to the semi-parametric estimators' bias *in favor* of models that are overfit when using out-of-sample prediction criteria (e.g. cross-validation). Non-parametric estimators, which do not exhibit this behavior, are highly variable for local discrimination. We propose to address the high variability problem through penalized regression splines smoothing. The behavior of various estimators of time-dependent AUC and concordance are illustrated via a simulation study using two different mechanisms that produce overoptimistic out-of-sample estimates using semi-parametric estimators. Estimators are further compared using a case study using data from the National Health and Nutrition Examination Survey (NHANES) 2011–2014.

Keywords *C-index; concordance; proportional hazard model; survival prediction; time-dependent AUC*

1 Introduction

Modeling time-to-event outcomes, also known as survival analysis, is a major area of methodological development in statistics and machine learning and is relevant to many applied data science tasks. Broadly, the prediction accuracy of time-to-event outcomes can be assessed locally (i.e., for a fixed time point) or globally (summarized over a set of time points). Local performance is often assessed by the time-dependent Receiver Operating Characteristics (ROC) curve and the area under the ROC curve (AUC), while global performance is often assessed by concordance (C-index). In this paper, we review the formulation of these estimators, compare their

*Corresponding author. Email: ying.jin@cuanschutz.edu.

Received June 5, 2024; Accepted January 1, 2025

behavior in realistic data science scenarios, and identify the cause of undesirable out-of-sample behavior in semi-parametric estimators of discrimination. We restrict our focus to the Cox proportional hazards model framework (Cox, 1972), one of the fundamental statistical approaches to modeling time-to-event outcomes (Harrell et al., 1996; Abd ElHafeez et al., 2021). The behavior of discrimination estimators have been studied in previous literature, although the focus has been on the violations of proportional hazard or independent censoring assumptions. The behavior of estimators discussed in this paper is wholly different from the effect of such model misspecification and applies to correctly specified models in which the assumptions above holds.

Some semi-parametric estimators (Heagerty and Zheng, 2005; Song and Zhou, 2008) are consistent under the assumption of proportional hazard and independent censoring. However, a previously unidentified flaw exists for a specific class of estimators, making them inappropriate for use in many data science contexts where out-of-sample accuracy is used to perform model selection (Yates et al., 2023; Burman, 1989). Specifically, the semi-parametric estimators considered in the current work have the potential to substantially overestimate out-of-sample discriminative performance, even when the model is well calibrated to the training sample and the majority of the testing sample. This poor behavior is most easily seen in the context of 1) model overfit and 2) covariate misalignment in combination with poorly calibrated risk prediction for a subset. The latter we define to mean situations where the *test* sample is contaminated with subjects not from the population of interest. The model fit on the training data, even when correctly specified for the training sample and most of the testing sample, will poorly predict this subset of subjects. However, this decreased accuracy of the overall risk prediction in the test data is not reflected in semi-parametric estimators of discriminative performance.

The use of out-of-sample prediction accuracy as the gold standard for model selection and assessment is done in large part to avoid the tendency of in-sample estimates to be overly optimistic due to model overfit to the training data (Yates et al., 2023; Arlot and Celisse, 2010), allowing for a more accurate estimate of a model's generalizability in terms of prediction. As a result, it is important for out-of-sample estimators to accurately identify when a model provides poorly calibrated risk predictions, such as in the case where the model is overfit to the data, or where the test sample is contaminated by a different population. To understand the potential magnitude of the phenomena we've identified, we provide an illustration using a case of covariate misalignment via the presence of an outlier in the testing data. Figure 1 provides an illustration of out-of-sample overestimation under covariate misalignment caused by the presence of one outlier, where we evaluate the out-of-sample performance of the same Cox regression model on two different datasets using the Incident/Dynamic AUC (Heagerty and Zheng, 2005) at a particular follow-up time ($t = 0.27$). Specifically, Figure 1 presents estimated Incident/Dynamic ROC curves, the integral of which is Incident/Dynamic AUC. Values of Incident/Dynamic AUC close to 1 indicate near-perfect discrimination (or, equivalently, ROC curves which tend toward the upper left quadrant of the plot), while a value close to 0.5 (ROC curves near the identity line) indicates the predictor is no more prognostic than a coin flip. The Cox model was correctly specified and fit to a training dataset with 300 subjects, with 228 subjects at risk at $t = 0.27$. Out-of-sample performance is evaluated on two testing datasets that are identical except for one subject: the dataset represented by the yellow line introduced one outlier with abnormally large values of covariates. Both testing sets have a sample size of 500. At the $t = 0.27$, the number of subjects at risk is 364 in the dataset without the outlier and 365 in the dataset with the outlier. As Figure 1 shows, this one single observation has driven out-of-sample semi-parametric AUC from 0.812 to 0.999 at this time point. Intuitively, a single observation out of 365, whether their risk was *accurately* predicted by the model or not, should not shift AUC by such a large margin.

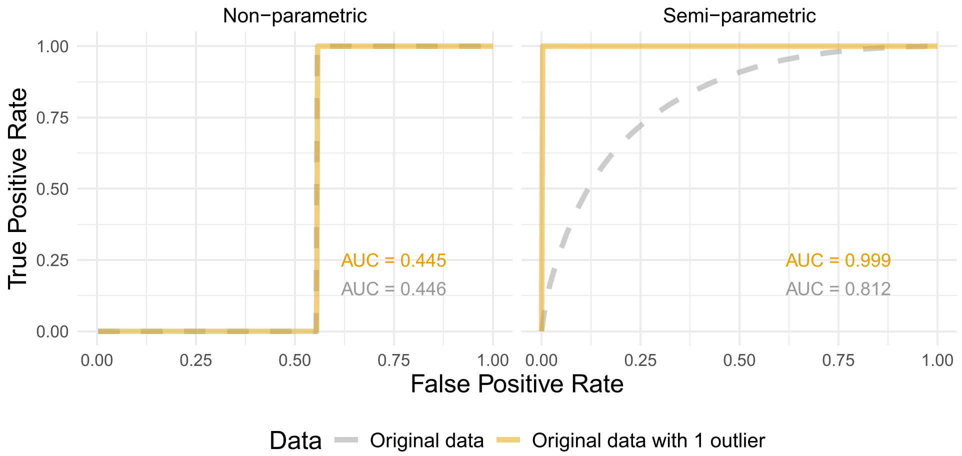


Figure 1: Change of out-of-sample semi-parametric estimation of discriminative performance after introducing one outlier to the testing sample (N = 500). The datasets represented by the yellow and grey lines are identical, except for one outlier with outlying values of covariates.

Moreover, this individual's risk was poorly predicted by the model, suggesting discrimination should *decrease* rather than substantially increase to near *perfect* discrimination. That is, the estimated log hazard of this subject is 3.6 times higher than the second largest log hazard in the same sample at this time point, with a predicted probability of survival beyond time $t = 0.27$ of essentially 0. However, the subject experienced the event at time 0.85, indicating the Cox model has poorly estimated the observation's true risk. In this example, the semi-parametric estimator of Incident/Dynamic AUC failed to properly reflect model performance or generalizability. The poor behavior of this estimator of time-dependent AUC at some time t, AUC(t), is caused by the disconnection between the estimator (which uses the magnitude of model estimated risk in it's construction) and the actual event status of samples in the risk set who experience an event beyond the current time (i.e., for some $t^* > t$). We provide additional details in Section 2.3.

Non-parametric estimators, on the other hand, do not exhibit this behavior as the estimators do not include the *magnitude* of the estimated risk and instead consider only the *relative ranking* of risk. However, non-parametric estimators of time-dependent AUC exhibit higher instability, frequently with large jumps between neighboring time points. This is because, at each event time point, the number of events is usually low. Only one or a few events would be observed, causing the sensitivity estimates to be highly unstable, even fluctuating between extreme values such as 0 and 1. As shown in the left panel of Figure 1, the non-parametric incident/dynamic ROC curve at this time point is a step function, jumping from 0 to 1 when the cutoff value of the risk score reaches a certain point. The resulting AUC at this point is low (0.44), but remains unchanged after introducing an extreme outlier. Therefore, these estimators often require additional smoothing. For example, Shen et al. (2015) explored fractional polynomial regression for smoothing. Song et al. (2012) use kernel functions for smoothing. In this paper, we address this instability issue by smoothing the estimates over time using penalized regression splines (Wood, 2004).

Other work looking at properties of discrimination estimators has focused on the assumptions of proportional hazard and independent censoring. As examples, van Geloven et al. (2021) showed that the incident/dynamic AUC estimator may be biased when the proportional hazard assumption is violated. Schmid and Potapov (2012) compared different concordance estimators through an extensive simulation study with violation of either assumption and found that the estimator proposed by Heagerty and Zheng (2005) showed bias in both cases. However, the behavior of semi-parametric estimators we describe in our work here is seen even when the conditions above are satisfied and represents a different, and heretofore unidentified, feature. Although compelling arguments have been made that discrimination is an inappropriate criterion for model selection in the context of time-to-event analyses, such measures provide one piece of useful information about the predictions made by a particular model and thus have been and continue to be used frequently for model assessment and selection. For example, Cornec-Le Gall et al. (2016) used the semi-parametric AUC estimator proposed by Heagerty and Zheng (2005) to predict renal survival in Autosomal Dominant Polycystic Kidney Disease. The non-parametric Harrell's C index was used in Stephenson et al. (2005) to predict the 10-year risk of recurrence of prostate cancer after radical prostatectomy. Due to their utilization by practitioners, understanding the properties and shortcomings of various estimators is critical. Thus, we add to the literature by identifying the pros and cons of different estimators. We hope to provide appropriate references for investigators who need these measures for tasks such as model selection, model assessment, etc.

In the following section, we define the evaluation metrics of discriminative performance, introduce their estimators, and identify the source of undesirable behaviors. A simulation study in Section 3 compares the behavior of different classes of estimators in the context of model overfit and covariate misalignment. Section 4 further illustrates their practical utility with an application to the 2011–2014 National Health and Nutrition Examination Survey (NHANES, 2011–2014) data.

2 Methods

For risk prediction in time-to-event models, an estimate of risk (or, more precisely, a ranking of risk) is required for all units/subjects at a set of times for the range of the time-to-event random variable. Given these risk estimates, local discrimination involves defining a set of "cases" and "controls" specific to time t and then calculating the corresponding ROC curve at this time point using the time-specific risk estimates. This "local" AUC is broadly referred to as time-dependent AUC, with two popular estimands being Incident/Dynamic AUC and Cumulative/Dynamic AUC (Heagerty and Zheng, 2005). The former compares the discrimination of risk for incident cases (events observed at a specific time point) to dynamic controls (subjects at risk at the same time point). It is often used when researchers are interested in the discriminative ability of makers with temporal updates, such as dynamic prediction/forecasting (Cornec-Le Gall et al., 2016; van Geloven et al., 2021). The use of incident cases prevents redundant information over time (Blanche et al., 2013). The latter focuses on historical cases (events observed up to a specific time point) and dynamic controls, which are often used for prognosis from baseline (Mortensen et al., 2017). Blanche et al. (2013) also suggested its use for clinical decision making, such as enrollment in clinical trials. Global summaries of discrimination look at how well risk predictions discriminate across a range of (or all) time points, t, with the Concordance or C-index (Harrell et al., 1996; Uno et al., 2011) being common measures. Both local and global measures assess how well risk rankings compare to observed event times. The focus of this paper is restricted to the local measure of Incident/Dynamic AUC ($\text{AUC}^{I/D}(t)$) and the global measure of Concordance ($\mathcal{C}$), a weighted average of $\text{AUC}^{I/D}(t)$ over time. We discuss the definition and different estimators of $\mathrm{AUC}^{I/D}(t)$ and $\mathcal{C}$, and distinguish between semi- and non-parametric estimators as the different classes of estimators with regard to both their formulation and out-of-sample behavior.

Let $i = 1, \ldots, N$ denote an individual for whom we observe a time-to-event outcome, T_i^*, subject to right censoring, denoted as C_i. For each individual, we observe $[T_i, \delta_i, \mathbf{X}_i^t]$ where $T_i = \min(T_i^*, C_i)$ is the observed time (minimum of censoring time C_i and true event time T_i^*). $\delta_i = 1(T_i^* \leqslant C_i)$ is the event indicator and $\mathbf{X}_i \in R^p$ is a vector of time-fixed covariates. The censoring time C_i is assumed to be independent of the event time T_i^* conditional on $\mathbf{X}_i$. Furthermore, the data is assumed to be generated from a proportional hazards model (Cox, 1972), where the log hazard of the event time takes on the additive form:

$$\log \lambda(t|\mathbf{X}_i) = \log \lambda_0(t) + \mathbf{X}_i^t \boldsymbol{\beta} \ = \log \lambda_0(t) + \eta_i; \qquad t > 0.$$

In the proportional hazards model, $\log \lambda(t|\mathbf{X}_i)$ is the conditional log-hazard for subject i given their covariate vector $\mathbf{X}_i$, and $\log \lambda_0(t)$ is the log baseline hazard which is shared across the population and left unspecified. The vector of regression coefficients, $\boldsymbol{\beta}$, is unknown and corresponds to the linear contribution of each element of $\mathbf{X}_i$ to the log hazard. Finally, η_i is the overall contribution of covariates to log hazard, indicating the subject-specific deviation of log hazard from the population level. Hereafter we refer to η_i as the risk score of the subject i. Note that the results and findings below readily extend to extensions of the classical Cox model where risk depends on time ($\eta_i(t)$) through a time-specific coefficient ($\beta(t)$), a time-specific covariate ($\mathbf{X}_i(t)$), or both.

2.1 Incident/Dynamic AUC

Definition Incident/Dynamic AUC, or $\mathrm{AUC}^{I/D}(t)$ (Heagerty and Zheng, 2005), generalizes the notion of AUC for binary data, allowing time-dependent discrimination for time-to-event models. At a specific time t, $\mathrm{AUC}^{I/D}(t)$ is achieved by calculating the incident sensitivity (sensitivity$^I(c, t)$) and the dynamic specificity (specificity$^D(c, t)$) at a series of unique thresholds c for the risk score η_i, deriving a time-specific ROC curve and estimating the area under it. As in the AUC for binary data, $\mathrm{AUC}^{I/D}(t) \in [0, 1]$, with values closer to 1 indicating better discrimination and values near 0.5 indicating that the risk score η_i is not better at discriminating events at time t than a flip of a coin. We describe these estimands in more detail below.

Incident sensitivity and dynamic specificity are defined as:

$$\text{sensitivity}^I(c, t) = \mathrm{TP}_t^I(c) = \Pr(\eta_i > c | T_i^* = t);$$
$$\text{specificity}^D(c, t) = 1 - \mathrm{FP}_t^D(c) = \Pr(\eta_i \leqslant c | T_i^* > t).$$

where $\mathrm{TP}_t^I(c)$ and $\mathrm{FP}_t^D(c)$ are abbreviations for time-specific incident true-positive and dynamic false-positive rate, respectively, and c is a threshold of the risk score η_i. Using the above definitions for incident sensitivity and dynamic specificity, we can then define the Incident/Dynamic ROC curve. Let p denote the value of $\mathrm{FP}_t^D(c)$, then $\mathrm{ROC}_t^{I/D}(p) = \mathrm{TP}_t^I\{[\mathrm{FP}_t^D]^{-1}(p)\}$, from which it follows that $\mathrm{AUC}^{I/D}(t) = \int_0^1 \mathrm{ROC}_t^{I/D}(p)dp$.

Estimation In practice, $\mathrm{AUC}_t^{I/D}(t)$ is generally approximated by numeric integration: $\sum_p \delta_p \widehat{\mathrm{ROC}}^{I/D}(p)$, where p is the estimated dynamic false positive rate at every unique value of the risk score at a given time. Evaluating at $c = -\infty$ ensures that the estimated $\mathrm{ROC}_t^{I/D}$ passes through the point (1, 1). δ_p here is the quadrature weight that is typically determined using the trapezoid rule. Thus, different estimators of $\mathrm{AUC}^{I/D}(t)$ arise from the use of different estimators of dynamic specificity and/or incident sensitivity. In this section, all estimators considered use the same non-parametric estimator of dynamic specificity but differ in their approach to estimating incident sensitivity.

Specifically, suppose that we have obtained the estimated coefficient $\hat{\boldsymbol{\beta}}$ and used it to estimate individual risk scores $\hat{\eta}_i = \boldsymbol{X}_i^t \hat{\boldsymbol{\beta}}$. Dynamic specificity can then be estimated as follows:

$$1 - \widehat{\text{specificity}}^D(c, t) = \widehat{\mathrm{FP}}_t(c) = \frac{\sum_k I(\hat{\eta}_k > c) I(T_k > t)}{\sum_j I(T_j > t)}. \tag{1}$$

This estimator of dynamic false-positive rate is built based on the plug-in principle, counting up the proportion of individuals with an estimated risk greater than a particular threshold among those who have not experienced the event by time t.

Moving on to estimators of incident sensitivity, first consider a non-parametric estimator based similarly on the plug-in principle:

$$\widehat{\text{sensitivity}}^I(c, t) = \widehat{\mathrm{TP}}_t^{NP}(c) = \frac{\sum_k I(\hat{\eta}_k > c) I(T_k = t) I(\delta_k = 1)}{\sum_j I(T_j = t) I(\delta_j = 1)}. \tag{2}$$

This non-parametric estimator is inherently more variable than the non-parametric estimator of dynamic specificity in Equation (1). To see this, note that Equation (2) is based on counting the proportion of individuals who have a risk score above a particular threshold c among those with an observed event at t. As such, the resulting time-specific ROC curve would be a step function with a single step (See Figure 1 left panel). The estimator of $\mathrm{AUC}^{I/D}(t)$ obtained by non-parametric specificity (Equation (1)) and sensitivity (Equation (2)) is therefore a non-parametric estimator.

Next, consider the semi-parametric estimator of $\mathrm{AUC}^{I/D}(t)$ proposed by Heagerty and Zheng (2005). It uses the same non-parametric estimator of $\widehat{\mathrm{FP}}_t(c)$ in Equation (1), but differs in their estimator of incident sensitivity:

$$\widehat{\mathrm{TP}}_t^{SP}(c) = \frac{\sum_k I(\hat{\eta}_k > c) I(T_k \geqslant t) \exp(\hat{\eta}_k)}{\sum_j I(T_j \geqslant t) \exp(\hat{\eta}_j)}. \tag{3}$$

Instead of counting the proportion of true-positive subjects, this estimator considers all subjects at risk at time t and weighs the subjects' contribution to the estimated true positive rate as their exponentiated estimated risk score. The subject-specific weight $\frac{\exp(\hat{\eta}_k)}{\sum_j I(T_j \geqslant t) \exp(\hat{\eta}_j)}$ depends on the *values* coefficient estimates $\hat{\boldsymbol{\beta}}$, thus *parametric* in nature. In addition, we note the lack of dependence on actual observed events at t (i.e. δ_i appears nowhere in this formula). These two points are the key marks to distinguish the formulation of semi-parametric estimators from the non-parametric ones. They are also the reason why semiparametric estimators, although consistent in our proposed framework (Xu and O'Quigley, 2000) and relatively smooth in practice, suffer from overoptimistic inflation of out-of-sample discrimination. We return to this point in Section 2.3.

2.2 Concordance

Definition Concordance, defined as $\mathcal{C} = \Pr(\eta_i < \eta_j | T_i^* > T_j^*)$ for a randomly selected pair (i, j), represents the overall agreement between true event times and risk scores. As in $\mathrm{AUC}^{I/D}(t)$,

$C \in [0,1]$, with values closer to 1 denoting better global discrimination of the risk score. In practice, T_i^* may have support beyond the duration of a study, which results in the need to administratively censor participants at some follow-up time τ (for example, the end of the study). In the context of administrative censoring, the estimand becomes $\mathcal{C}^\tau = \Pr(\eta_i < \eta_j | T_i^* > T_j^*, T_j^* < \tau)$. It has been shown that this truncated concordance is a weighted average of Incident/Dynamic AUC (Heagerty and Zheng, 2005):

$$\mathcal{C}^\tau = \int_0^\tau \mathrm{AUC}^{I/D}(t) w^\tau(t) dt; \qquad w^\tau(t) = \frac{2f(t)S(t)}{1 - S^2(\tau)}, \tag{4}$$

where $S(t)$ is the marginal survival function (not conditional on covariates) and $f(t)$ is the marginal probability density function of time to event. Using this result, $\mathcal{C}^\tau$ (Equation (4)) may be estimated by numerically integrating semi- and non-parametric estimators of Incident/Dynamic AUC.

Estimation First, consider estimating $\mathcal{C}^\tau$ as the weighted integral of $\mathrm{AUC}^{I/D}(t)$. This can be done using any estimator of $\mathrm{AUC}^{I/D}(t)$, with weights derived from estimated marginal survival function: $\hat{w^\tau}(t) = \frac{2\hat{f}(t)\hat{S}(t)}{1-\hat{S}^2(\tau)}$.

However, as mentioned previously, the non-parametric estimator of $\mathrm{AUC}^{I/D}(t)$ derived from non-parametric specificity (Equation (1)) and sensitivity (Equation (2)) is highly variable, which presents a challenge for numeric integration. Some previous literature has approached this issue by smoothing the $\mathrm{AUC}^{I/D}(t)$ estimates, using Lowess, kernel (Song et al., 2012), or fractional polynomial (Shen et al., 2015). We therefore propose a complementary approach. Specifically, we propose to smooth the non-parametric $\widehat{\mathrm{AUC}}^{I/D}(t)$ using penalized regression splines via the *gam* function in the *mgcv* package (Wood, 2003, 2011, 2017) in *R* (R Core Team, 2021). That is, we estimate the additive model $\widehat{\mathrm{AUC}}^{I/D}(t) = \widetilde{\mathrm{AUC}}^{I/D}(t) + \epsilon(t) = \sum_{k=1}^K \xi_k B_k(t) + \epsilon(t)$, where $\widetilde{\mathrm{AUC}}^{I/D}(t)$ is the smoothed non-parametric Incident/Dynamic AUC estimates modeled as the linear combination of a set of cubic spline basis functions $B_1(t) \dots B_K(t)$ subject to penalty on second derivative (L2 penalty). $\epsilon(t)$ denotes a random Gaussian noise that is independent and identically distributed. Unknown parameters are estimated using the restricted maximum likelihood (REML) method.

Please note that the weight estimator $\hat{w^\tau}(t)$ requires estimating both the marginal survival function $\hat{S}(t)$ and the survival time density $\hat{f}(t)$. While the Kaplan-Meier curve is commonly used to estimate $S(t)$, it is unrealistic to estimate $f(t)$ by taking the derivative of $\hat{S}(t)$ since $\hat{S}(t)$ would be a step function. Therefore, it has also been proposed to use a smoothed version of the Kaplan-Meier curve. Many smoothing methods have been suggested for either the counting process (Ramlau-Hansen, 1983) or hazard function (Wang, 2014). In this paper, we smooth the estimated Kaplan-Meier curve using a Constrained Additive Model (Pya, 2021): $\hat{S}(t) = \tilde{S}(t) + \epsilon(t) = \sum_{k=1}^K \zeta_k M_k(t) + \epsilon(t)$, where $\hat{S}(t)$ is the Kaplan-Meier estimators of marginal survival function, and smoothed survival function $\tilde{S}(t)$ is modeled as a linear combination of P-spline basis functions $M_1(t) \dots M_K(t)$ that are subject to a monotonicity constraint: $\tilde{S}(t_1) > \tilde{S}(t_2)$ for $t_1 < t_2$. While other options with the same constraint are available, the estimated survival functions stay robust against the choice of basis $M_1(t) \dots M_K(t)$. (See Figure S.7 in Supplement S.6.)

We hereafter refer to the estimator of concordance estimated using unsmoothed non-parametric $\widehat{\mathrm{AUC}}^{I/D}(t)$ as non-parametric concordance $\hat{\mathcal{C}}_{NP}$. The estimator by integrating $\widetilde{\mathrm{AUC}}^{I/D}(t)$, the smoothed non-parametric estimator of Incident/Dynamic AUC, will be referred to as smoothed non-parametric concordance $\hat{\mathcal{C}}_{SNP}$. The estimator from the semi-parametric $\widehat{\mathrm{AUC}}^{I/D}(t)$, since it was introduced by Heagerty and Zheng (2005), will be referred to as the Heagerty-Zheng semi-parametric concordance $\hat{\mathcal{C}}_{HZ}$.

In addition to estimators of Concordance based on integrating estimates of $\mathrm{AUC}^{I/D}(t)$, we consider one additional semi-parametric estimator proposed by Gonen and Heller (2005):

$$\hat{\mathcal{C}}_{GH} = \frac{2}{n(n-1)} \sum_{i<j} \frac{I(\hat{\eta}_j - \hat{\eta}_i < 0)}{1 + \exp(\hat{\eta}_j - \hat{\eta}_i)} + \frac{I(\hat{\eta}_i - \hat{\eta}_j < 0)}{1 + \exp(\hat{\eta}_i - \hat{\eta}_j)}, \tag{5}$$

and one additional non-parametric estimator of concordance by Harrell et al. (1996):

$$\hat{\mathcal{C}}_{Harrell} = \frac{\sum_{i<j} I(T_i < T_j) I(\hat{\eta}_i > \hat{\eta}_j) I(\delta_i = 1) + I(T_i > T_j) I(\hat{\eta}_i < \hat{\eta}_j) I(\delta_j = 1)}{\sum_{i<j} I(T_i < T_j) I(\delta_i = 1) + I(T_i > T_j) I(\delta_j = 1)}. \tag{6}$$

Similar to the semi-parametric estimator of incident sensitivity, the semi-parametric estimator of Concordance proposed by Gonen and Heller (2005) includes terms of the form e^{η} and excludes event status δ, while Harrell's C only compares relative ranking of risk estimation to event time. We further note that Harrell's C can be inconsistent for the estimand of interest in the presence of censoring. Alternative consistent non-parametric estimators have been proposed (Uno et al., 2011), using the inverse probability weighting technique to modify Harrell's C-index. In this paper, we focus on the original Harrell's C for simplicity of presentation and due to the fact that in our simulations and applications, the estimator proposed by Uno et al. (2011), which is consistent for truncated Concordance shows the same in- versus out-of-sample predictive performance effects with a slightly shifted distribution.

2.3 Mechanism of Inflated Out-of-Sample Estimation

As described in Section 2.1, the cause of the observed out-of-sample inflation of $\widehat{\mathrm{AUC}}^{I/D}(t)$ lies in the semi-parametric estimator of incident sensitivity. Remember the estimator in Equation (3) weighs observations at risk at time t by the exponential of their estimated risk scores $\frac{\exp(\hat{\eta}_k)}{\sum_j I(T_j \geqslant t) \exp(\hat{\eta}_j)}$. When an observation has a high estimated risk score $\hat{\eta}$, its corresponding weight would be very large. As a result, the semi-parametric estimator would be dominated by observations with high estimated risk, regardless of their actual event status at time t or the accuracy of risk estimation. Take Figure 1 as an example. In the original dataset, the largest value of the sensitivity weight above is 2.8%. Considering the sample size of 500, the contribution to the semi-parametric sensitivity estimation is similar across subjects. However, after introducing one outlier, the weight of this outlier is 99.99%, and the second largest weight in the sample is 0.00018%. That is, the semi-parametric estimator of incident sensitivity depends almost entirely on this outlier when the observation itself is not even well predicted.

As a result of the inflated estimates of incident sensitivity, the Incident/Dynamic AUC will also be overestimated at the corresponding time points. Figure 1 provides a straightforward illustration. As the right panel shows, the semi-parametric ROC curve was shifted markedly to the upper left corner through the single introduced outlier, causing the area under the curve to increase. The semi-parametric concordance, as a weighted integral of a series of overestimated AUC, would also be overestimated as a result.

The Gonen-Heller estimator also has a semi-parametric composition, as Equation (5) has included the term $\exp(\hat{\eta}_j - \hat{\eta}_i)$ in the denominator. Similar to the Heagerty-Zheng estimator of incident sensitivity, the Gonen-Heller estimator can be perceived as the weighted sum of all random pairs (i, j). Specifically, Equation (5) can be rewritten as $\frac{2}{n(n-1)} \sum_{i<j} \frac{I(\hat{\eta}_j-\hat{\eta}_i<0)+I(\hat{\eta}_i-\hat{\eta}_j<0)}{1+\exp(-|\hat{\eta}_i-\hat{\eta}_j|)}$. Here, the contribution to the estimation from each random pair of subject (i, j) $(i < j)$ is the term $\frac{1}{1+\exp(-|\hat{\eta}_i-\hat{\eta}_j|)}$. It is straightforward to see that this term increases with the difference between $\hat{\eta}_i$ and $\hat{\eta}_j$. That is, pairs that have larger difference between estimated risk will drive the concordance estimate higher. However, note that the exponentiated risk difference appears only in the denominator, stabilizing the maximum contribution of a single observation on the overall estimate. As a result, the Gonen-Heller estimator tends to inflate less than the Heagerty-Zheng estimator.

Take Figure 1 as an example. Compared to the original dataset, the introduction of one outlier with a very large risk score would lead to many pairs of observations with large difference between their risk scores. The contribution to the Gonen-Heller estimator, $\frac{1}{1+\exp(-|\hat{\eta}_i-\hat{\eta}_j|)}$, will be larger for these pairs, leading to an inflated estimate of concordance. However, the degree of the inflation will be limited. In this case, this one outlier has driven the Gonen-Heller concordance estimator from 80.05% to 80.14%. For details about the distribution of difference of risk scores between pairs of observations, please see Figure S.4 in the Supplement.

The observations above are illustrated empirically in the following sections through a simulation study and a real-world data application.

3 Simulation Study

3.1 Data Generating Mechanism

We designed a simulation study to illustrate the in- and out-of-sample behavior of semi- and non-parametric estimators introduced in Section 2 in finite samples. We simulate data under a Cox proportional hazards model framework with independent censoring.

The specific model for data generation is as follows:

$$\log \lambda(t|\boldsymbol{X}) = \log \lambda_0(t) + \boldsymbol{X}^t \boldsymbol{\beta} = \log(p\theta t^{p-1}) + \eta, \quad t > 0. \tag{7}$$

$\boldsymbol{X} = [X_1, X_2, X_3]^t \in \mathbf{R}^3$ are three covariates simulated as independent $N(0, 1)$ random variables, and $\boldsymbol{\beta} = (1, -1, 0.25)$ are the true values of coefficients. $\lambda_0(t) = p\theta t^{p-1}$ is the Weibull baseline hazard with $\theta = 2$ and $p = 2$, resulting in $E[T|\boldsymbol{X} = \boldsymbol{0}] = 0.63$. Censoring times are simulated uniformly from $(0, 1)$ independent of event times, with administrative censoring for all individuals at $\tau = 1$. Across simulated datasets this resulted in an average of 58.6% of participants being censored, and an observed median event time of 0.29. To examine the effect of censoring rate, we have also explored a few different censoring mechanisms and found that the behavior of estimators remained similar regardless. Therefore, we present the uniform censoring alone in the manuscript for conciseness and leave some selected output from other censoring rates to the Supplement S.5. Figure S.2 in the supplement presents the distribution of event and censoring times under this data generating mechanism.

We simulate 1000 datasets, each containing $N = 250$ individuals generated under Equation (7) to be used for model fitting (the *training set*) and estimation of in-sample discriminative performance. Each simulated dataset contains an additional 250 individuals simulated whose data are not used in model fitting. These individuals represent the *testing set* and are used to evaluate out-of-sample discrimination. In- and out-of-sample discrimination estimates are compared to the true quantities whose values were estimated using methods described in supplement S.3. The behavior of semi- and non-parametric estimators is compared under the two scenarios mentioned in Section 1: model overfit and covariate misalignment with poorly calibrated risk, each of which are described in more details below. Note that for both of these scenarios, the data generating mechanism of the training samples remains as described above. The difference is in the model fit to the data (model overfit) and the distribution of a small subset of the testing sample, as well as how well the risk of this small subset is calibrated (covariate misalignment).

3.1.1 Model Overfit

Model overfit in our simulation study is induced by fitting a Cox model presented in Equation (7) along with a set of covariates simulated independently from the outcome (T), censoring time (C), and covariates which define the true model ($\boldsymbol{X}$). That is, we simulate a new random vector $\boldsymbol{Z} \in \mathbf{R}^m$ for $m \in \{0, 20, 100\}$. When $m = 0$, no additional covariates are used as predictors, and the model is correctly specified for both training and testing sets. $m = 20$ corresponds to 20 additional covariates used as predictors, and $m = 100$ 100 additional covariates. Variables in $\boldsymbol{Z}$ follow the standard normal distribution $N(0, 1)$ and are mutually independent. We then fit the model

$$\log \lambda(t|\boldsymbol{X}) = \log \lambda_0(t) + \boldsymbol{X}^t \boldsymbol{\beta} + \boldsymbol{Z}^t \boldsymbol{\gamma} \tag{8}$$

to the training data, using the results to obtain estimates of in- and out-of-sample discrimination. Given the sample size for the training sample and the expected number of observed events in the training sample, $m = 20$ and $m = 100$ represent moderate and severe overfitting issues. We note a few points here. First, the behavior we observe when fitting Model (8) to data generated by Model (7) would still occur even if the new covariates were associated with the outcome (i.e., Model (8) with $\boldsymbol{\gamma} \neq \mathbf{0}$ in the data generating mechanism). We simply chose no association ($\boldsymbol{\gamma} = \mathbf{0}$) as it is cleaner in that it keeps the true values of discrimination the same across all scenarios. Second, maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (and thus risk scores) are consistent in this scenario ($[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}]^t \xrightarrow{p} [\boldsymbol{\beta}, \mathbf{0}]^t$).

3.1.2 Covariate Misalignment in Combination with Poorly Calibrated Risk Prediction

In the second scenario, the training sets are draw from the same data generation mechanism as Model (7). However, the testing sets includes a subset of subjects whose covariates are *scrambled* by a different distribution, after the event and censor times are generated the same way as the training sets. We introduce a small proportion (10%) of observations generated from a different distribution from the training set to mimic a sample contaminated by subjects not from the population of interest. The testing sample, though very similar to the training sample, includes a small subset of subjects whose estimated risk is inaccurate and uncorrelated with their true risk. Specifically, after generating a sample under the same mechanism as Model (7), the covariates of a small portion of these testing samples are regenerated and replaced from a different distribution. As a result, model predicted risk will be well calibrated for 90% of the test data while poorly calibrated for 10% (the “misaligned” subjects), as the observed risk is uncorrelated with true risk due to the *replaced* covariate values. Thus, discrimination will be, on average, lower in the test data than in the training data by construction.

We study the effect on out-of-sample behavior from two different types of change of covariate space: 1) a mean shift, where misaligned covariates are generated with a larger mean ($N(5, 1)$);

and 2) a variation change, where misaligned covariates are generated with are larger variance ($N(0, 5)$).

3.2 Results

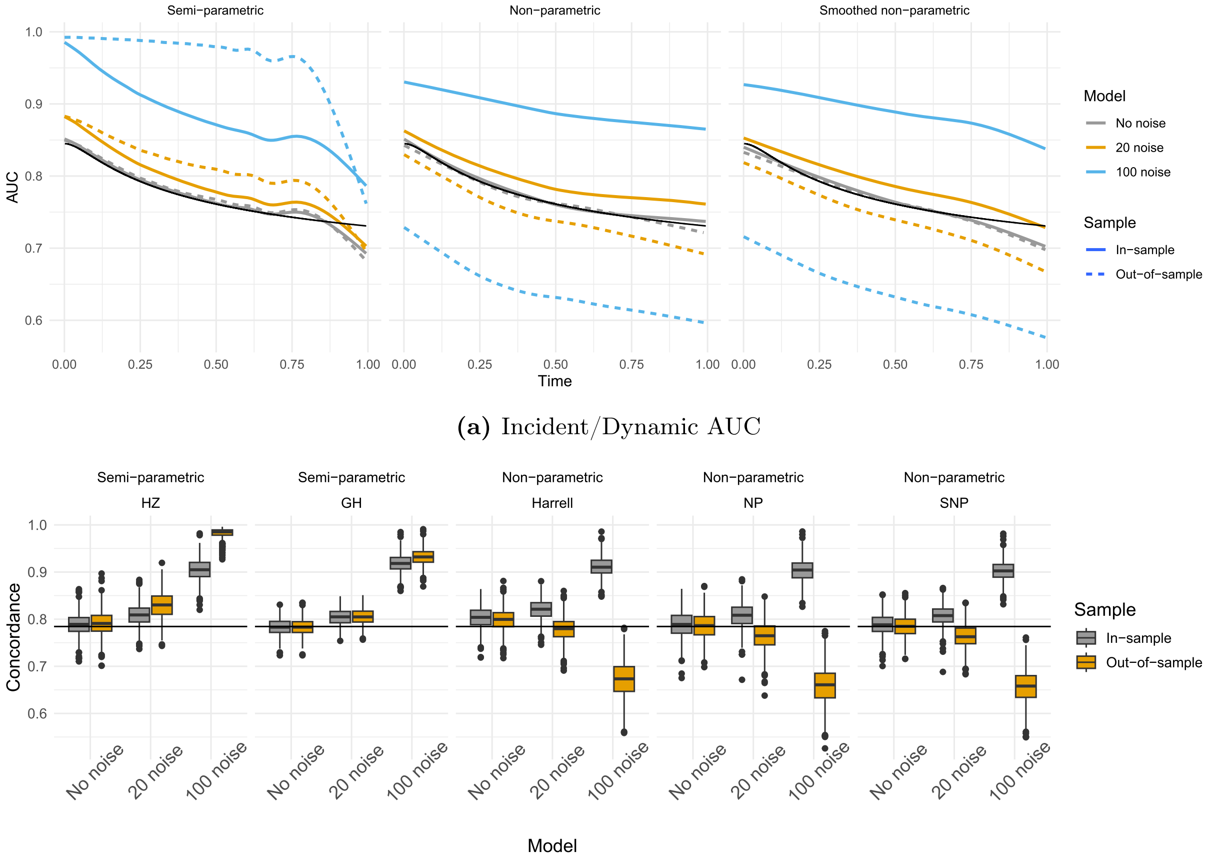
Results are presented separately by scenario (model overfit vs covariate misalignment) below. Within each scenario, we first discuss in- and out-of-sample local discrimination estimates ($\text{AUC}^{I/D}(t)$), followed by global discrimination (concordance).

3.2.1 Model Overfit

Local discrimination $\hat{\text{AUC}}^{I/D}(t)$ is presented in Figure 2a, showing trends in average discrimination. Estimated discrimination under the various models fit to the data are indicated by color, with grey, yellow, and blue indicating results from models fit with no unrelated variables ($\hat{\eta}_i = \boldsymbol{X}_i^t\boldsymbol{\beta}$), 20 unrelated variables ($\hat{\eta}_i = \boldsymbol{X}_i^t\boldsymbol{\beta} + \boldsymbol{Z}_i^t\hat{\gamma}$, $\boldsymbol{Z}_i \in \mathbf{R}^{20}$), and 100 unrelated variables ($\hat{\eta}_i = \boldsymbol{X}_i^t\boldsymbol{\beta} + \boldsymbol{Z}_i^t\hat{\gamma}$, $\boldsymbol{Z}_i \in \mathbf{R}^{100}$), respectively. The black solid line represents the true values of $\text{AUC}^{I/D}(t)$. Each panel corresponds to a different estimator, with the semi-parametric estimator of Heagerty and Zheng (2005), a non-parametric estimator, and the smoothed non-parametric estimator in the left, middle, and right panels, respectively. Solid and dashed lines indicate in-sample versus out-of-sample estimated discrimination. Each line represents the average values of $\text{AUC}^{I/D}(t)$ *across* simulated datasets smoothed by generalized additive model.

In the scenario where a model is overfit to the data, we would expect the model to perform better on the training set than the testing set due to poor generalization, and the discrepancy between in-sample and out-of-sample performance should increase with the severity of model overfit. Visually, in Figure 2a this would correspond to solid lines being higher than the dashed lines. However, the semi-parametric estimator behaved in an opposite way. The estimates are substantially higher on the testing samples than training samples, even near perfect out-of-sample discrimination for the highly overfit model. It reveals that this estimator can overestimate the discriminative performance significantly in the presence of model overfit. However, under the correctly specified model without noise signals (grey lines), the semi-parametric estimators appear unbiased and also much smoother than the other estimators. The non-parametric estimator and smoothed non-parametric estimator, on the other hand, behaved consistently with the expectations of overfit models, where in-sample estimates have higher values than out-of-sample estimates. When the model is not overfit, the fully non-parametric estimator appears unbiased, while the smoothed non-parametric estimator showed a slight downward bias at both ends of the follow-up period. This is likely due to oversmoothing of individual time-dependent AUC curves.

Figure 2b visualizes the estimates of global discriminative performance across all simulations. Here, concordance estimates are summarized using boxplots, with grey boxes representing in-sample and yellow out-of-sample estimates. Each penal shows a different estimator. The two left panels, Heagerty-Zheng and Gonen-Heller estimators, are both semi-parametric. All Heagerty-Zheng estimators here and forth are derived using survival functions smoothed by shape-constrained P-spline basis. Although other options are available, the choice of spline basis does not affect the behavior of estimators (see Figure S.8 in Supplement S.6), thus not included for conciseness. The three right panels, including Harrell's C-index, (unsmoothed) non-parametric and smoothed non-parametric estimators, falls into the category of non-parametric. Similar to $\text{AUC}^{I/D}$, we expect model overfit would cause in-sample estimates to be higher than out-of-sample estimates. But the two semi-parametric estimators both have higher out-of-sample



(a) Incident/Dynamic AUC

(b) Concordance

Figure 2: Behavior of estimators of model discrimination under the effect of model overfit. Estimates of Incident/Dynamic AUC are presented in (a) are smoothed across all simulations for better visualization. Solid lines represent in-sample estimates and dashed lines represent out-of-sample estimates. Line color indicates the underlying model, where grey corresponds to the correctly specified model, yellow a moderately overfit model with 20 additional signals, blue a highly overfit model with 100 additional signals. The solid black line represents true value of AUC. Estimates of concordance are presented in (b) with grey indicating in-sample and yellow out-of-sample estimates. The black horizontal line is the true value of concordance.

estimates than in-sample. The Heagerty-Zheng estimator exhibits more severe out-of-sample inflation than the Gonen-Heller estimator. On the other hand, the behavior of non-parametric estimators is more reasonable. Harrell's C-index showed upward bias, while the others appear unbiased under the non-overfit model.

However, the non-parametric estimators of AUC$^{I/D}$ with proper out-of-sample behavior are numerically more unstable. Its value frequently fluctuates between two extremes 0 and 1. For example, under the correct model with no additional covariates, the standard deviation of non-parametric estimates along the first 20% of the follow-up period (0.203) is about 7 times of semi-parametric estimates (0.0283). Smoothed non-parametric estimators is not as variable, but also not as smooth as the semi-parametric estimator. Along the same interval, its standard deviation (0.0325) is about 15% greater than the semi-parametric ones. All three estimators showed increasing variation over time. For more details on the variation of estimators, please refer to S.1 in the supplement.

3.2.2 Covariate Misalignment in Combination with Poorly Calibrated Risk Prediction

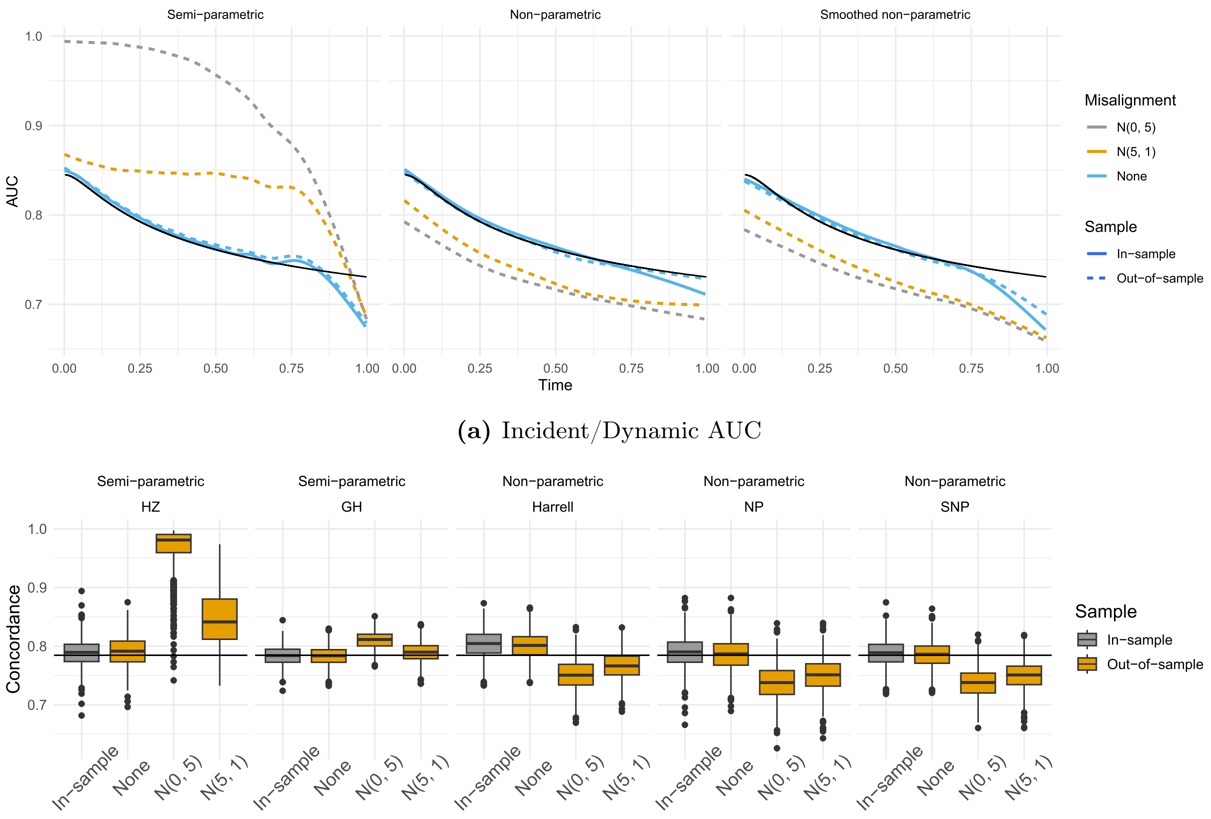
Local and global discrimination for the covariate misalignment scenario is presented in Figures 3a and 3b, respectively. These figures present results in a similar format as the model overfit scenario (Section 3.2.1). Looking at Figure 3, the model with three covariates is fitted on training samples and then tested on different *misaligned* testing samples. The color here represents the type of covariate misalignment, with grey, yellow, and blue corresponding to variation change, mean shift, and no misalignment, respectively.

With the misaligned covariate space between the training and testing sets, the model, though correct for the training set and the majority (90%) of the testing sample, would not be able to predict the risk of the misaligned subjects well. Therefore we expect out-of-sample $\text{AUC}^{I/D}$ estimates to be lower than the in-sample ones. While the behavior of non-parametric estimators is consistent with such expectations, semi-parametric estimators show the opposite, anti-intuitive trend. In the left panel of 3a the out-of-sample semi-parametric estimates on misaligned test samples are clearly higher than both their corresponding in-sample estimates and the true values. This inflation is more prominent when covariates have larger spread compared to mean shift. The non-parametric estimators did not suffer from overoptimistic estimation on the misaligned testing sets. On the testing sets without misalignment, all three $\widehat{\text{AUC}}^{I/D}$ showed slight bias downwards at the end of the follow up period. In terms of stability, non-parametric estimator showed much more drastic fluctuation than the semi-parametric one, while smoothed non-parametric has a moderate variability in between the two (Supplement S.1).

The concordance estimators in Figure 3b also showed similar behavior. The semi-parametric estimators, including Heagerty-Zheng and Gonen-Heller showed higher out-of-sample than in-sample estimates on misaligned testing sets (yellow boxes). When the covariates are misaligned with larger variance, the Heagerty-Zheng estimator can be inflated to nearly 1, which can be interpreted as perfect discrimination. The Gonen-Heller estimator seems more robust against covariate misalignment especially when the source of misalignment is a mean shift. However, without covariate misalignment, estimators appear unbiased and smooth, except for Harrell's $\mathcal{C}$ with a small upward bias.

4 Data Application

We illustrate the potential impact of the choice of estimator on model selection using a mortality prediction model. Recent work has used cross-validated Concordance as a criterion for evaluating the relative predictive power of data derived from wearable accelerometers versus demographic, lifestyle, and health variables known to be associated with mortality (Smirnova et al., 2020; Leroux et al., 2021) and to evaluate the added value gained by considering diurnal activity patterns (Cui et al., 2021) using techniques from functional regression (Crainiceanu et al., 2024; Ramsay and Silverman, 2005). These complex models are highly parameterized in which overfitting is a concern. In these contexts overinflated cross-validated estimates of Concordance could incorrectly lead to choosing a more complex model and result in substantially reduced accuracy of mortality predictions.



(a) Incident/Dynamic AUC

(b) Concordance

Figure 3: Behavior of estimators of model discrimination under the effect of covariate misalignment. Estimates of Incident/Dynamic AUC are presented in (a), smoothed across all simulations for better visualization. Solid lines represent in-sample estimates and dashed lines out-of-sample estimates. Line color indicates the distribution from which misaligned covariates are generated, where grey corresponds to a greater variation, yellow a mean shift, and blue the same distribution as the training sample (no misalignment). The solid black line represents true value of AUC. Estimates of concordance are presented in (b) with grey indicating in-sample and yellow out-of-sample estimates. The black horizontal line is the true value of concordance.

Using this framework as a motivation, here we use data from the National Health and Nutrition Examination Survey (NHANES) 2011–2014 to predict all-cause mortality using physical activity features derived from wearable accelerometers (Leroux et al., 2019) and other participant characteristics associated with mortality as predictors. Note that the NHANES study is a multi-stage probabilistic sample from the non-institutionalized US population. Results are thus generalizable with appropriate uncertainty estimates when survey sampling methods are accounted for in regression modeling (i.e. survey weights, cluster sampling. etc). We do not account for survey design features in our application for simplicity of presentation. Specifically, eliminating the question of the impact of the distribution of survey weights on our results.

4.1 NHANES 2011–2014 Data

The analytic sample includes 3556 participants aged 50–80, with at least three days of accelerometry data with 95% estimated wear time and complete data in covariates of interest. The total number of observed all-cause mortality events was 424, with a total of 23587.17 person-years of follow-up. Accelerometry data was processed in this spirit of the pipeline used by Leroux et al. (2021, 2019) for the NHANES 2003–2006 and the UK Biobank data, respectively. The predictor vector $\boldsymbol{X_i}$ included five variables: age, Body Mass Index (BMI), Active-to-Sedentary Transition Probability (ASTP), Relative Amplitude (RA), and Total Monitor-Independent Movement Summary units (TMIMS). The latter three variables (ASTP, RA, and TMIMS) are derived from participants' accelerometry data.

4.2 Models

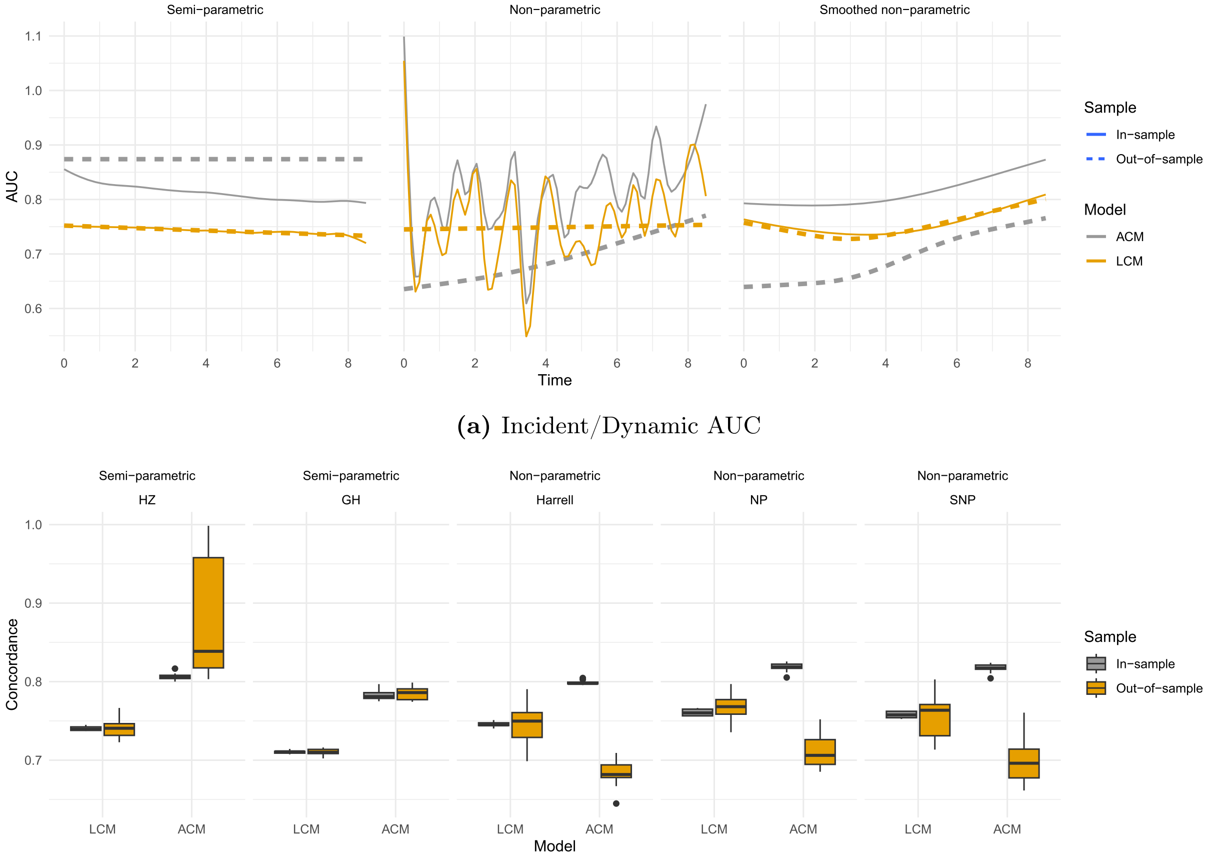
We apply models to the data in the spirit of the "model overfit" framework discussed in our simulation study in Section 3.2.1. Specifically, we consider a highly parameterized model that allows for non-linear associations and interactions between the predictor vector and risk of mortality as compared to a linear model with no interactions. The first model, an additive Cox model, is parameterized as $\log \lambda(t|\mathbf{X}_i) = \log \lambda_0(t) + f(\mathbf{X}_i)$. The risk score in this model, $f(\mathbf{X}_i) : \mathbf{R}^5 \to \mathbf{R}^1$, is a smooth function of the predictors (in this case a 5-dimensional vector) estimated using rank 200 unpenalized thin plate regression splines (Wood, 2003) via the *mgcv* package (Wood, 2017) in *R* (R Core Team, 2021). This model, which contains a five-dimensional smooth function, estimates 200 coefficients *without* regularization. With a ratio of a number of parameters to a number of events of 2.12, this will tend to result in serious overfitting to the data (Harrell et al., 1996). We will refer to this Model as the "additive Cox model (ACM)". Note that the use of unpenalized regression splines is not the standard recommended in applied work. Regularization in the form of a penalty on the curvature of the estimated surface is used to control the tendency of this class of models to overfit the observed data. Here, we specifically do not use regularization/penalization to illustrate the behavior of various estimators in a clear context of model overfitting, as in other contexts, the tendency of the model to overfit to a particular dataset may not be as clear.

The second model has a simpler linear form for the risk score: $\log \lambda(t|\boldsymbol{X}_i) = \log \lambda_0(t) + \boldsymbol{X}_i^t \boldsymbol{\beta}$. Given the size of the data and the number of observed events, a linear model of this form will be substantially less prone to overfitting, particularly in comparison to the ACM. We will refer to this Model as the "linear Cox model (LCM)".

The discriminative performance of the ACM and LCM models are evaluated using 10-fold cross-validation, using both semi- and non-parametric estimators. The former includes the Heagerty-Zheng estimators of Incident/Dynamic AUC and concordance (Equation (3)), as well as the Gonen-Heller estimator of concordance (Equation (5)). The latter includes non-parametric and smoothed non-parametric estimators of Incident/Dynamic AUC, their corresponding concordance (Equation (2)) (weighted by smoothed survival function), and Harrell's C-index (Equation (6)).

4.3 Results

Figure 4 compares the in- and out-of-sample behavior of discrimination estimators from the simpler LCM and the more complex ACM models. Results are presented in the same format as Figure 2a and Figure 2b for $\widehat{\mathrm{AUC}}^{I/D}(t)$ and Concordance, respectively.



(a) Incident/Dynamic AUC

(b) Concordance

Figure 4: Incident/Dynamic AUC and concordance estimates on NHANES data through 10-fold cross validation. Incident/Dynamic AUC estimates are presented in (a) and smoothed over all 10 testing folds. The grey color indicates the complicated additive Cox model and yellow the simple linear Cox model; solid lines indicate in-sample and dashed lines indicate out-of-sample estimates. Concordance estimates are presented in (b), where grey indicates in-sample while yellow out-of-sample estimates.

First, consider local discrimination presented in Figure 4a, where grey and yellow lines indicate the additive and linear model, respectively, and solid/dashed lines indicate in- versus out-of-sample estimates, respectively. We note a few key findings. First, the in-sample behavior of the three estimators is similar, with estimates from ACM being higher than estimates from LCM (compare solid lines across panels). That the ACM has higher in-sample discrimination is to be expected due to overfitting to the training folds. However, the out-of-sample semi-parametric estimates of $\widehat{\mathrm{AUC}}^{I/D}(t)$ from the ACM are not only much higher than the estimates obtained from the LCM, but also from its own in-sample counterpart, with a striking out-of-sample increase as high as 0.15. If an analyst were to make a model selection decision based on cross-validated discrimination using the semi-parametric estimator, they would incorrectly choose the more complicated ACM, which fits poorly to the data, over the simpler but better fitting LCM. In contrast, the non-parametric and smoothed non-parametric estimators showed the opposite trend when comparing in-sample (training folds) versus out-of-sample (testing folds) discrimination when evaluating the ACM. That is, the dashed grey lines presenting below the solid grey lines in the right two panels of Figure 4a are consistent with the expectation that the ACM is overfitting to the data. The more parsimonious LCM shows essentially equivalent in- and out-of-sample discrimination across estimators considered here. We note that the fully non-parametric estimator of in-sample discrimination (Figure 4a, middle panel) appears highly non-linear for both the LCM and ACM due to its high instability and the inconsistency of estimates across training splits. This point is expounded upon in the Appendix.

Figure 4b summarizes the in-sample (grey boxplots) and cross-validated/out-of-sample (yellow boxplots) estimates of global discriminative performance separately by estimator (panels). Note that from the perspective of model selection, for a given estimator, an analyst would choose the ACM over the LCM if the center of the yellow boxplot (average across testing splits) is higher for the ACM than the LCM. From this perspective, it is clear that both semi-parametric estimators would result in the analyst incorrectly choosing the ACM over the LCM. In contrast, using any of the non-parametric estimators would result in choosing the LCM over the ACM. Analyzing the findings in a bit more detail, we see that comparing the grey to the yellow boxplots for the LCM and ACM for a fixed estimator, the same out-of-sample inflation of estimated local discrimination ($\widehat{\text{AUC}}^{I/D}(t)$) in semi-parametric estimators is also observed in Concordance. Because the Heagerty-Zheng estimator of Concordance is just a weighted integral of local discrimination, this finding for this particular estimator is unsurprising. The Gonen-Heller estimator showed higher out-of-sample estimates from ACM than LCM, though the discrepancy is much smaller than was observed in the Haegerty-Zheng estimator. The difference between mean estimates over all ten testing folds between ACM and LCM is 0.136 for Heagerty-Zheng, and 0.075 for Gonen-Heller.

5 Discussion

In this paper, we compared the behavior of several semi-parametric and non-parametric estimators of discrimination performance, including the local measure Incident/Dynamic AUC ($\widehat{\text{AUC}}^{I/D}(t)$) and the global measure concordance through a simulation study in Section 3 and a case study in Section 4. The global measure concordance can be estimated either by integrating the local measure $\widehat{\text{AUC}}^{I/D}(t)$ over the follow-up period, or directly using non- or semi-parametric estimators proposed in previous literature. In the context of out-of-sample evaluation, the class of semi-parametric estimators, including the Heagerty-Zheng estimator of Incident/Dynamic AUC, its corresponding concordance, and the Gonen-Heller concordance, show the tendency to overestimate out-of-sample estimates of model performance, especially when the model is overfit, or the test sample used is contaminated with participants from a different population. We have also identified the source of this phenomenon in Section 2.3 by pointing out the lack of dependence on the accuracy of risk estimation and actual event status of the incident sensitivity estimator. These estimators can thus lead analysts to incorrectly believe that a model performs much better than it actually does, resulting in their choice of an overfit, complex model over a more appropriate, simple model. Therefore, we caution their use for model assessment, comparison, and selection purposes, particularly when out-of-sample evaluation is used. On the other hand, when the model is correctly specified and the sample is not misaligned, these estimators have desirable properties, such as consistency and smoothness.

While the fully non-parametric estimators do not suffer from overoptimistic out-of-sample estimation, they are highly unstable due to the relatively small number of events at any given time point, motivating the need for a smoothed estimator. As such, we proposed a method for smoothing these non-parametric estimators using penalized regression splines, though other smoothers (e.g. kernel smoothing) may be used. In our simulation study, this smoothing approach works well for reducing instability but could result in slight bias for time-dependent Incident/Dynamic AUC. In addition to the oversmoothing for visualization purposes, we believe the bias is also likely a result of heteroskedasticity of the residual process and correlation of estimates for $\text{AUC}^{I/D}(t)$ across time, which are not accounted for in classical additive models. Further methodological work for identifying a better smoother of non-parametric estimators and establishing accurate inferential procedures is needed. Although defining an optimal smoothing approach is beyond the scope of this paper, the GAM smoother used here provides a good illustration of expected in- and out-of-sample behavior for an estimator of discrimination.

In this paper, our aim is to focus on the illustration and explanation of the behavior of different estimators of measures of discrimination. Measures of calibration, on the other hand, may not exhibit the same behavior, since the inaccuracy of risk prediction should be reflected in their construction. Though beyond the scope of this work, we provide an example of the behavior of Brier score in the Supplement S.4 as an illustration. In addition, the results presented in this work are derived from one single model framework, the Cox proportional hazards framework. However, we note that the semi-parametric estimators discussed in the paper (i.e. the semi-parametric AUC estimator proposed by Heagerty and Zheng) can be potentially extended to more general hazard models (Heagerty and Zheng, 2005), such as the Accelerated Failure Time models (AFT) model, though we are not aware of specific work on that point. In contrast, non-parametric estimators depend only on relative rankings for risk predictions and do not depend on the model formulation or data generation mechanism. Thus, they are directly applicable under alternative modeling frameworks. While the exploration of more flexible model structures is not within the scope of this work, it is certainly an area we intend to explore in future research.

In summary, this work represents an important step forward in identifying the conditions under which various estimators of discrimination in time-to-event models are appropriate. Critically, we identified a previously ignored intrinsic flaw in a class of popular evaluation criterion for the discriminative performance of time-to-event models. Evidence are provided through simulation and case study to illustrate how such a flaw can mislead the process of model assessment and selection, and how alternative non-parametric estimators are better options for such purposes. Finally, we propose one smoothing method to mitigate the high stability of non-parametric estimators, at the cost of introducing slight bias.

Conflict of Interest

The authors have declared no conflict of interest.

Supplementary Material

The supplementary material includes additional information that is relevant but not included in the manuscript, including figures, mathematical derivation and data file used for the data application section. It also includes a zipped file containing code scripts to reproduce the results presented above. Here is a brief summary of is content:

- outlier_exp.R: to generate data and produce Figure 1 in the Introduction.
- Simulation: code scripts used to implement the simulation study.
  - Sim_overfit.R: for the first scenario of model overfit in Section 3.2.1.
  - Sim_contamination.R: for the second scenario of covariate misalignment in Section 3.2.2.
  - helpers.R: functions to calculate discussed estimators.
  - trueAUC.R: calculate the true values of incident/dynamic AUC.
  - SimFigs.R: produce Figures 2 and 3.
- DataAppl: scripts to reproduce the data application section.
  - data_appl.R: scripts to reproduce the data application results.
  - helpers_appl.R: functions to calculate discussed estimators.
  - DataApplFigs.R: produce Figure 4.
- SuppFigs.R: to produce figures included in the supplement.

References

Abd ElHafeez S, D'Arrigo G, Leonardis D, Fusaro M, Tripepi G, Roumeliotis S (2021). Methods to analyze time-to-event data: The Cox regression analysis. *Oxidative Medicine and Cellular Longevity*.

Arlot S, Celisse A (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4: 40–79. https://doi.org/10.1214/09-SS054

Blanche P, Latouche A, Viallon V (2013). Time-dependent auc with right-censored data: A survey. In: *Risk Assessment and Evaluation of Predictions* (MLT Lee, M Gail, R Pfeiffer, G Satten, T Cai, A Gandy, eds.), 239–251. Springer New York, New York, NY.

Burman P (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3): 503–514. https://doi.org/10.1093/biomet/76.3.503

Cornec-Le Gall E, Audrézet MP, Rousseau A, Hourmant M, Renaudineau E, Charasse C, et al. (2016). The propkd score: A new algorithm to predict renal survival in autosomal dominant polycystic kidney disease. *Journal of the American Society of Nephrology*, 27(3): 942–951. https://doi.org/10.1681/ASN.2015010016

Cox D (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B, Methodological*, 34(2): 187–220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Crainiceanu C, Goldsmith J, Leroux A, Cui E (2024). *Functional Data Analysis with R*, 1st ed. Chapman and Hall/CRC.

Cui E, Crainiceanu C, Leroux A (2021). Additive functional Cox model. *Journal of Computational and Graphical Statistics*, 30(3): 780–793. https://doi.org/10.1080/10618600.2020.1853550

Gonen M, Heller G (2005). Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4): 965–970. https://doi.org/10.1093/biomet/92.4.965

Harrell FE, Lee KL, Mark DB (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4): 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Heagerty PJ, Zheng Y (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61(1): 92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x

Leroux A, Di J, Smirnova E, McGuffey EJ, Cao Q, Bayatmokhtari E, et al. (2019). Organizing and analyzing the activity data in NHANES. *Statistics in Biosciences*, 11(2): 262–287. https://doi.org/10.1007/s12561-018-09229-9

Leroux A, Xu S, Kundu P, Muschelli J, Smirnova E, Chatterjee N, et al. (2021). Quantifying the predictive performance of objectively measured physical activity on mortality in the UK Biobank. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, 76(8): 1486–1494. https://doi.org/10.1093/gerona/glaa250

Mortensen RN, Gerds TA, Jeppesen JL, Torp-Pedersen C (2017). Office blood pressure or ambulatory blood pressure for the prediction of cardiovascular events. *European Heart Journal*, 38(44): 3296–3304. https://doi.org/10.1093/eurheartj/ehx464

Pya N (2021). *scam: Shape Constrained Additive Models*. R package version 1.2-12.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramlau-Hansen H (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, 11(2): 453–466.

Ramsay JO, Silverman BW (2005). *Functional Data Analysis*. Springer New York, NY.

Schmid M, Potapov S (2012). A comparison of estimators to evaluate the discriminatory power of time-to-event models. *Statistics in Medicine*, 31(23): 2588–2609. https://doi.org/10.1002/sim.5464

Shen W, Ning J, Yuan Y (2015). A direct method to evaluate the time-dependent predictive accuracy for biomarkers. *Biometrics*, 71(2): 439–449. https://doi.org/10.1111/biom.12293

Smirnova E, Leroux A, Cao Q, Tabacu L, Zipunnikov V, Crainiceanu C, et al. (2020). The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: National health and nutritional examination survey 2003–2006. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, 75(9): 1779–1785. https://doi.org/10.1093/gerona/glz193

Song X, Zhou XH (2008). A semiparametric approach for the covariate specific roc curve with survival outcome. *Statistica Sinica*, 18(3): 947–965.

Song X, Zhou XH, Ma S (2012). Nonparametric receiver operating characteristic-based evaluation for survival outcomes. *Statistics in Medicine*, 31(23): 2660–2675. https://doi.org/10.1002/sim.5386

Stephenson AJ, Scardino PT, Eastham JA, Bianco FJ, Dotan ZA, DiBlasio CJ, et al. (2005). Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *Journal of Clinical Oncology*, 23(28): 7005–7012. https://doi.org/10.1200/JCO.2005.01.867

Uno H, Cai T, Pencinac MJ, D'Agostinod RB, Weib LJ (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10): 1105–1117. https://doi.org/10.1002/sim.4154

van Geloven N, He Y, Zwinderman A, Putter H (2021). Estimation of incident dynamic auc in practice. *Computational Statistics & Data Analysis*, 154: 107095. https://doi.org/10.1016/j.csda.2020.107095

Wang JL (2014). *Smoothing Hazard Rates*. John Wiley & Sons, Ltd.

Wood S (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 65(1): 95–114. https://doi.org/10.1111/1467-9868.00374

Wood S (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467): 673–686. https://doi.org/10.1198/016214504000000980

Wood S (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 73(1): 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood S (2017). *Generalized Additive Models: An Introduction with R*, 2 edition. Chapman and Hall/CRC.

Xu R, O'Quigley J (2000). Proportional hazards estimate of the conditional survival function. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 62(4): 667–680. https://doi.org/10.1111/1467-9868.00256

Yates LA, Aandahl Z, Richards SA, Brook BW (2023). Cross validation for model selection: A review with examples from ecology. *Ecological Monographs*, 93(1): e1557. https://doi.org/10.1002/ecm.1557