Biometrical Journal () , zzz–zzz / DOI: 10.1002/

Identifying critical shortcomings of estimators of discriminative performance in time-to-event analyses: a comparison study

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

[1] Department of Biostatistics and Informatics, Colorado School of Public Health, Fitzsimons building, 4th floor, 13100 E. 17th Place, Aurora, CO 80045

Predicting timing and occurrence of events of interest has been a major topic in biostatistical research. Many estimators have been proposed to assess discriminative performance of time-to-event models, including time-dependent AUC and concordance. Based on their formulation, these estimators can be categorized into semi-parametric estimators, which weighs subjects by the exponential of estimated risk, and non-parametric estimators which are built on plug-in principal. Although theoretical properties of many of these estimators are well established, empirical behavoirs are less well understood. In this paper we identify a previously unknown flaw of the class of semi-parametric estimators that can result in vastly over-optimistic out-of-sample estimation of discriminative performance in common applied tasks. Although these semi-parametric estimators are popular in practice, the phenomena we identify here suggests this class of estimators is inappropriate for use in model assessment and selection based on out-of-sample evaluation criteria (e.g. cross-validated discrimination) in general. Fully non-parametric estimators, on the other hand, do not suffer from this problem. However, non-parametric estimates of time-varying AUC are highly variable. We propose to address this problem through smoothing using penalized splines. Using a simulation study, we illustrate the occurrence of the identified phenomena under two different mechanisms (model overfitting and data contamination), and show the superiority of non-parametric estimators under all scenarios considered. The estimators are further compared via a case study using data from the National Health and Nutrition Examination Survey (NHANES) 2011-2014.

Key words: Concordance; C-index; Proportional hazard model; Survival prediction; Time-dependent AUC

The data that supports the findings of this study are available in the supplementary material of this article

1 Introduction

Modelling time-to-event outcomes, often referred to as survival analysis, is a major area of methodologic development in biostatistical research. Broadly, performance of time-to-event models is evaluated using discrimination or calibration criteria, with the former criteria being similar to Area under the Receiver Operating Characteristic Curve (AUC) for binary outcomes. While several estimators have been proposed to assess discriminative performance of time-to-event models, including time-dependent AUC and concordance, there exists a previously unidentified flaw of a specific class of estimators which renders them inappropriate for use in many contexts. Specifically, the semi-parametric estimators, proposed by [10], have the potential to substantially overestimate out-of-sample discriminative performance, even when the model is correctly specified. It often fails to appropriately reflect the model performance and generalizability, and as a result, misleads the process of model assesment or selection.

This poor behavior is most easily seen in the context of: 1) model overfit, and 2) contaminated data. Both are common issues when developing models based on existing real-world data. Overfit can happen

*Corresponding author: e-mail: ying.jin@cuanschutz.edu

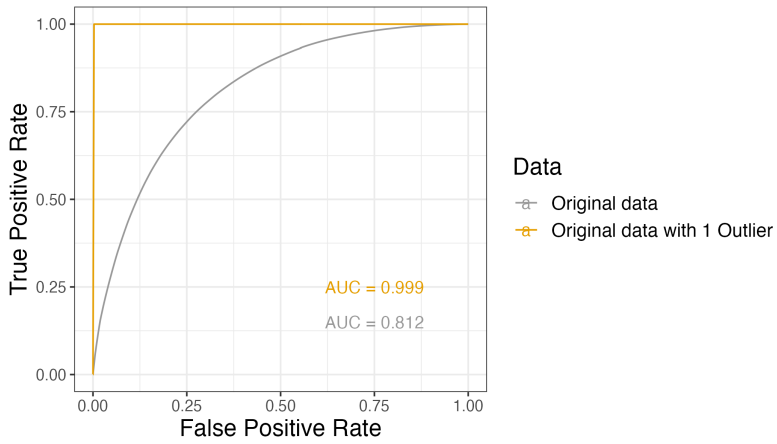


Figure 1: Change of out-of-sample semi-parametric estimation of discriminative performance after introducing one outlier to the training sample. The datasets represented by the yellow and grey lines are identical, except for one outlier with outlying values of covariates.

when the model is too complicated with too many free parameters or noise signals, for which out-of-sample evaluation is often used to evaluate model generalizability. However, as we will see in the following sections, semi-parametric estimators can severly overestimate out-of-sample discriminative performance, even close to perfect discrimination. In this case the out-of-sample performance appears to be much better than in-sample performance, which clearly is not realistic. Data contamination, on the other hand, happens when the dataset is collected with outlying or noisy observations who do not follow the sample distribution as population of interest. These observations can also drive the out-of-sample semi-parametric estimates higher simply because of abnormal values of covariates, regardless of whether their outcome is accurately predicted. In both cases the semi-parametric estimators can show over-optimistic discriminative performance when the underlying model in fact fits poorly.

Figure 1 is an example of such flaw, where we fit the same cox regression model on two datasets and evaluate their out-of-sample discriminative performance using time-dependent AUC [10] at a specific time point. The two training datasets are identical except for one subject: the dataset represented by the yellow line introduced one outlier with abnormally large values of covariates. As the figure revealed, this one single observation has driven out-of-sample AUC to 99.8% at this time point, while it in fact had little contribution to the discriminative ability of the underlying model. In this case, the evaluation metric failed to properly reflect model performance or generalizabiliy. The poor behavior of this estimator of time-dependent AUC is caused by its disconnection to the actually event status of subjects at this time. Specifically, the formulation of such estimators is only supposed to tell us whether our model *ranks* subjects' risks correctly, but says nothing about the *accuracy* of risk estimation or prediction (e.g. probability of an event).

It has been previously noted that semi-parametric estimators of discrimination designed for data generated according to a Cox proportional hazards model are inaccurate and potentially over-optimistic when the proportional hazards assumption is violated [20]. Here, the behavior we identify is different and potentially more worrisome, since it can occur even under a correctly specified proportional hazards model. Although compelling arguments have been made that discrimination is an inappropriate criteria for model selection in the context of time-to-event models [2], such measures provide one piece of useful information about the predictions made by a particular model, thus have been and continue to be used frequently for model assessment and selection. Due to their utilization by practitioners, understanding the properties and

shortcomings of various estimators is critical. Thus, we add to the literature by identifying both the phenomena of inflated out-of-sample estimation of discriminative performance and the mechanism by which it occurs. In addition, we provide recommendations for alternative non-parametric estimators that are able to appropriately reflect discriminative performance.

In the following section, we define the evaluation metrics of discriminative performance for time-to-event models, such as Incident/Dynamic AUC and concordance. Section 3 then introduces their estimators, identifies the source of inflated out-of-sample estimation, and proposes alternative non-parametric estimators to be used in practice. A simulation study in Section 4 compares the behavior of different classes of estimators in the context of model overfit and data contamination. Section 5 further illustrates their practical utility with an application to the 2011-2014 National Health and Nutrition Examination Survey (NHANES) data.

2 Measures of Discrimination for Time-to-event Models

Accuracy of discrimination of risk in time-to-event models can be assessed locally (i.e. for a fixed time point) or globally (summarized over a set of time points). Local discrimination involves defining a set of "cases" (incident events) and "controls" (individuals without events) specific to time t, and then calculating the corresponding Receiver Operating Characteristic (ROC) curve at this time point. The local Area-under-the-ROC-Curve (AUC) are broadly referred to as time-dependent AUC, with two popular estimands being Incident/Dynamic AUC and Cumulative/Dynamic AUC [10]. The former compares discrimination of risk for incident events and the latter focuses on historical events. Global summaries of discrimination over the follow-up period are generally referred to as measures of Concordance [7]. Both local and global measures assess how well estimated risk compares to observed event times, accounting for censoring. In this paper we restrict our focus to the local measure of Incident/Dynamic AUC ($AUC^{I/D}(t)$) and the global measure of Concordance (C), which is a weighted average of $AUC^{I/D}(t)$ over time. The time-to-event outcomes are subject to right censoring independent from covariates.

Let $i = 1, \ldots, N$ denotes individual, and T_i^* denotes a non-negative random variable (e.g. time-to-event) subject to right censoring at C_i. For each individual i 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^*). $\delta_i = 1(T_i^* \leq C_i)$ is the event indicator, and $\mathbf{X}_i \in R^p$ is a vector of time-fixed covariates. Censoring time C_i is independent of event time T_i^* conditional on $\mathbf{X}_i$. Additionally, the data is assumed to be generated from a proportional hazards model [23], where the log hazard of the event time takes on the additive form

$$\log \lambda(t|\mathbf{X}_i) = \log \lambda_0(t) + \mathbf{X}_i^t \boldsymbol{\beta} = \log \lambda_0(t) + \eta_i; \qquad t > 0 \tag{1}$$

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 constant across population. $\boldsymbol{\beta}$ is a vector of unknown parameters, corresponding to the linear contribution of each element of $\mathbf{X}_i$ to the log hazard, and η_i is the overall contribution of covariates to log hazard, indicating the subject-specific deviation of log hazard from the population level ($\log \lambda_0(t)$). Hereafter we refer to η_i as the risk score of subject i.

2.1 Incident/Dynamic AUC

Incident/Dynamic AUC, or $AUC^{I/D}(t)$ [10], generalizes the notion of AUC for binary data, allowing for time-dependent discrimination in time-to-event models. At a specific time t, $AUC^{I/D}(t)$ is achieved by calculating the incident sensitivity ($sensitivity^I(c,t)$) and dynamic specificity ($specificity^D(c,t)$) at a series of unique thresholds c for risk score η_i, deriving a time-specific ROC curve and estimating the area under it. As with AUC for binary data, $AUC^{I/D}(t) \in [0,1]$, with values closer to 1 indicating better discrimination and values near 0.5 indicating the risk score η_i is no 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

$$sensitivity^I(c,t) = TP_t^I(c) = Pr(\eta_i > c | T_i^* = t) \tag{2}$$

$$specificity^D(c,t) = 1 - FP_t^D(c) = Pr(\eta_i \leq c | T_i^* > t) \tag{3}$$

where $TP_t^I(c)$ and $FP_t^D(c)$ are abbreviations for time-specific incident true-positive and dynamic false-positive rate respectively, and c is a threshold of 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 $FP_t^D(c)$, then

$$ROC_t^{I/D}(p) = TP_t^I\{[FP_t^D]^{-1}(p)\}$$

From which it follows that $AUC^{I/D}(t)$:

$$AUC^{I/D}(t) = \int_0^1 ROC_t^{I/D}(p)dp$$

In practice $AUC_t^{I/D}(t)$ is generally approximated by numeric integration, evaluating $[TP_t^I(c), FP_t^D(c)]$ for $c \in \{\eta_i : 1 \leq i \leq N\} \cup -\infty$.

2.2 Concordance

Concordance, defined as $C = Pr(\eta_i < \eta_j | T_i^* > T_j^*)$, represents the overall agreement between true event times and risk scores. As with $AUC^{I/D}(t)$, $C \in [0,1]$, 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, resulting in a need to administratively censor participants at some follow-up time τ (e.g. the end of the study). In the context of administrative censoring, the estimand becomes $C^\tau = Pr(\eta_i < \eta_j | T_i^* > T_j^*, T_j^* < \tau)$. It has been shown that this truncated concordance is a weighted-average of Incident/Dynamic AUC [10]:

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

where $S(t)$ is the marginal survival function of event times (not conditional on covariates) and $f(t)$ is the marginal probability density function of time to event.

3 Estimators of Discrimination for Time-to-event Models

In this section we discuss methods of estimating $AUC^{I/D}(t)$ and C^τ defined in Section 2 above. We distinguish between semi- and non-parametric estimators as the different classes of estimators with regard to both their formulation and out-of-sample behavior.

3.1 Incident/Dynamic AUC

As in Section 2, estimation of $AUC^{I/D}(t)$ can be achieved by numeric approximation of the integral of the Incident/Dynamic ROC curve at time t. Procedurally, this is done by obtaining estimates of incident sensitivity and dynamic specificity at all thresholds $c \in \{\eta_i : T_i \geq t\} \cup -\infty$, which is the set of unique values of individual risk score of subjects at risk at time t. Evaluating at $c = -\infty$ ensures the estimated $ROC_t^{I/D}$ passes through the point $(1,1)$. Thus, different estimators of $AUC^{I/D}(t)$ arise from the use of different estimators of dynamic specificity and/or incident sensitivity. Here, all estimators considered use the same non-parametric estimator of dynamic specificity, but differ in their approach to estimate incident sensitivity.

Specifically, suppose we have obtained 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 then can be estimated as follows:

$$1 - \widehat{specificity}^{D}(c,t) = \hat{FP}_t(c) = \frac{\sum_k I(\hat{\eta}_k > c) I(T_k > t)}{\sum_j I(T_j > t)} \tag{5}$$

This estimator of dynamic false-positive rate is built based on the plug-in principal, counting up the proportion of individuals who have an estimated risk greater than a particular threshold among those individuals 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 principal:

$$\widehat{sensitivity}^{I}(c,t) = \hat{TP_t}^{NP}(c) = \frac{\sum_k I(\hat{\eta}_k > c) I(T_k = t) I(\delta_k = 1)}{\sum_j I(T_j = t) I(\delta_j = 1)} \tag{6}$$

This non-parametric estimator is not continuous over time because it needs to be evaluated at time points with more than one event ($\{t : \sum_j I(T_j = t) I(\delta_j = 1) > 0\}$). It is inherently more variable than the non-parametric estimator of dynamic specificity in Equation (5). To see this, note that Equation (6) 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. In practice there is often only one event at a single time point, and the estimated value of sensitivity would fluctuate between 0 and 1. The resulting time-specific ROC curve would be a step function. The estimator of $AUC^{I/D}(t)$ obtained by non-parametric specificity (5) and sensitivity (6) is therefore a non-parametric estimator.

Then consider the semi-parametric estimator of $AUC^{I/D}(t)$ proposed by [10]. It uses the same non-parametric estimator of $\hat{FP}_t(c)$ in Equation (5), but differs in their estimator of incident sensitivity:

$$\hat{TP_t}^{SP}(c) = \frac{\sum_k I(\hat{\eta}_k > c) I(T_k \geq t) exp(\hat{\eta}_k)}{\sum_j I(T_j \geq t) exp(\hat{\eta}_j)} \tag{7}$$

Instead of counting the proportion of true-positive subjects, this estimator conditions on all subjects at risk at time t and weigh the subjects by their exponential estimates of risk score.

3.2 Concordance

Truncated concordance, C^τ, may be estimated using the result linking C^τ to $AUC^{I/D}(t)$ (weighted integral) or using other semi- and non-parametric estimators. First consider estimating C^τ as the weighted integral of $AUC^{I/D}(t)$. This can be done using either non- or semi-parametric estimates of $AUC^{I/D}(t)$, 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 was mentioned previously, the non-parametric estimator of $AUC^{I/D}(t)$ derived from non-parametric specificity (5) and sensitivity (6) is highly variable, which presents a challenge for numeric integration. We therefore propose to smooth the non-parametric $\hat{AUC}^{I/D}(t)$ using penalized regression splines via the *mgcv* package [26, 28, 29] in *R* [19].

$$\hat{AUC}^{I/D}(t) = \tilde{AUC}^{I/D}(t) + \epsilon(t) = \sum_{k=1}^{K} \xi_k B_k(t) + \epsilon(t)$$

Here $\tilde{AUC}^{I/D}(t)$ is the smoothed non-parametric Incident/Dynamic AUC estimates, modelled as the linear combination of a set of spline basis functions $B_1(t)...B_K(t)$ subject to penalty on second derivative.

$\epsilon(t)$ denotes a random noise that follows a zero-mean Gaussian process across time. Other options, such as kernel smoothing [25] are possible.

Please note that the weight estimator $\hat{w^{\tau}}(t)$ requires estimating both marginal survival function $\hat{S}(t)$ and density of survival time $\hat{f}(t)$. While 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 Kaplan-Meier curve. In this paper, we use a Constrained Additive Model [18]:

$$\hat{S}(t) = \tilde{S}(t) + \epsilon(t) = \sum_{k=1}^{K} \zeta_k M_k(t) + \epsilon(t)$$

where $\hat{S}(t)$ is the Kaplan-Meier estimators of marginal survival function, and smoothed survival function $\tilde{S}(t)$ is modelled as a linear combination of P-spline basis functions $M_1(t)...M_K(t)$ that are subject to the following constrains:

1. Monotonicity: $\tilde{S}(t_1) > \tilde{S}(t_2)$ for $t_1 < t_2$
2. $\tilde{S}(0) = 1$
3. Positivity: $\tilde{S}(t) > 0$

We hereafter refer to the estimator of concordance derived from non-parametric $\hat{AUC}^{I/D}(t)$ as non-parametric concordance $\hat{C}_{NP}$. As for the estimator from the semi-parametric $\hat{AUC}^{I/D}(t)$, since it was introduced by Heagerty and Zheng in 2005 [10], we refer the this estimator as the Heagerty-Zheng semi-parametric concordance $\hat{C}_{HZ}$. The estimator by integrating $\tilde{AUC}^{I/D}(t)$, the smoothed non-parametric estimator of Incident/Dynamic AUC, will be referred to as smoothed non-parametric concordance $\hat{C}_{SNP}$.

In addition to estimators of Concordance based on integrating estimates of $AUC^{I/D}(t)$, we consider one additional semi-parametric estimator proposed by Gonen and Heller [6]

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

and one additional non-parametric estimator of concordance [7]

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

Similar to the semi-parametric estimator of incident sensitivity, the semi-parametric estimator of Concordance proposed by [6] 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. In addition, we note that Harrell's C is biased for the estimand of interest in the presence of censoring. Alternative unbiased estimators have been proposed by [24], but we use Harrell's C here for simplicity of presentation.

3.3 Mechanism for Inflated Estimation of Out-of-Sample Discrimination

As mentioned in the introduction, semi-parametric estimators are prone to overestimate model performance on new samples, even when the model is correctly specified. To understand the mechanism behind this critical flaw, we need to revisit the formula of semi-parametric estimator of incident sensitivity, or equivalently, true positive rate:

$$\hat{TP}_t^{SP}(c) = \frac{\sum_k I(\hat{\eta}_k > c) I(T_k \geq t) exp(\hat{\eta}_k)}{\sum_j I(T_j \geq t) exp(\hat{\eta}_j)} = \sum_k I(\hat{\eta}_k > c) I(T_k \geq t) \frac{exp(\hat{\eta}_k)}{\sum_j I(T_j \geq t) exp(\hat{\eta}_j)} \tag{8}$$

While counting the number of observations at risk at time t with risk score over a threshold c, this estimator also weighs such observations by the exponential of their estimated risk scores. The subject-specific weight $\frac{exp(\hat{\eta}_k)}{\sum_j I(T_j \geq t) exp(\hat{\eta}_j)}$ depends on 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 reasons that semi-parametric estimators, though consistent under our proposed framework [30] and relatively smooth in practice, suffer from over-optimistic inflation.

Specifically, the true positive rate estimates depend heavily on the value of $\hat{\eta}$ of subjects at risk, regardless of whether the risk is correctly estimated or whether an event actually happened to the subjects at the specific time point. It causes the estimate to be dominated by a small subset of observations with large estimated risk, even when they may not be well predicted by the model. For example in Figure 1, the single outlying observation has a very large estimate of risk score, as a result of the outlying values of covariates. Its weight is thus close to 1, meaning the estimates AUC almost entirely depend on this one single subject, regardless of the event status of this subject at the time of interest.

4 Simulation Study

We designed a simulation study to illustrate the in- and out-of-sample behavior of the three classes of estimators introduced in Section 3 in finite samples. As such, we generate data under a single data generating mechanism according to a Cox proportional hazards model with independent censoring, a framework under which each semi- and non-parametric estimator of $AUC^{I/D}(t)$ considered here are unbiased and consistent.

The specific model for data generation is as follows:

$$\log \lambda(t|\boldsymbol{X}) = \log \lambda_0(t) + \boldsymbol{X}^t \boldsymbol{\beta} = \log(p\theta t^{p-1}) + \eta, \quad t > 0 \tag{9}$$

$\boldsymbol{X} = (X_1, X_2, X_3)$ 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 $[\theta, p]^t = [2, 2]^t$. Censoring times are simulated uniformly from two discrete values $(0.5, 1)$ independently of event times, with administrative censoring for all individuals at $\tau = 1$.

We simulate 1000 datasets containing $N = 250$ individuals as the **training set** used for model fitting and estimation of in-sample discriminative performance. An additional 250 individuals are simulated under the same data generating mechanism in (9) as the **testing set** to estimate out-of-sample discrimination. The behavior of semi- and non-parametric estimators is compared under the two scenarios mentioned in Section 1: model overfit and data contamination.

4.1 Model overfit

In this scenario, we set up overfitted models by adding noise signals. These noise signals are generated independently from both covariates and the time-to-event outcomes, thus should not contribute to the discriminative ability of fitted models. To introduce different severity of overfit, we construct three different models: 1) a not overfitted model with no noise signals; 2) a moderately overfitted model with 20 noise signals; and 3) a much overfitted model with 100 noise signals.

Figure 2 presents the in-sample and out-of-sample estimates of model discrimination ability across all simulated datasets of models with different degrees of overfit. Estimated $AUC^{I/D}(t)$ is visualized in Figure 2a, smoothed across all simulations for better visual presentation, since these estimated values tend to be highly variable especially for the non-parametric and smoothed non-parametric estimators. 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 to 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 (left panel) behaved in an opposite way. The estimates are substantially higher on the testing samples (dashed line) than training samples (solid line), even near perfect out-of-sample discrimination for the highly overfit model (blue lines). The non-parametric estimator (middle panel) and smoothed non-parametric estimator (right panel) behaved consistently with the expectaions of overfitted models.

The same trend is observed in the concordance estimators in Figure 2b. We expect model overfit would cause in-sample estimates (grey boxes) to be higher than out-of-sample estimates (yellow boxes). But the two semi-parametric estimators, Heagerty-Zheng (first panel) and Gonen-Heller concordance (second panel) both have higher out-of-sample estimates than in-sample. Heagerty-Zheng estimator has a more severe out-of-sample inflation than the Gonen-Heller. On the other hand, the behavior of non-parametric and smoothed non-parametric estimators is reasonable regarding overfitted models.

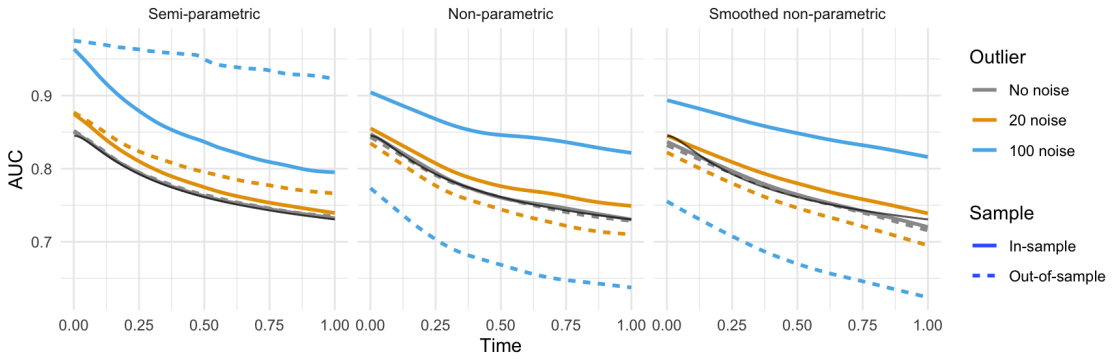
Under the correctly specified model with no noise signals (gray lines), the semi-parametric and non-parametric estimators of $\hat{AUC}^{I/D}$ in Figure 2a are unbiased. The smoothed non-parametric estimator (right panel) showed slight downward bias at both ends of the follow-up period. In terms of variability, the semi-parametric estimator is smooth across the follow-up period. The non-parametric estimator is very unstable, which can be mitigated by smoothing. Figure 4a provides a more detailed comparison of variation between different $AUC^{I/D}$ estimators, where the entire follow-up period is divided into five equal-length interval, and the distribution of estimates within each interval is summarized by boxplots. As is revealed, all three estimators showed increasing variation over time. Within the same time interval, the semi-parametric estimates is centered, while non-parametric estimates spread out to a large range. The smooth non-parametric estimators is not as centered as the semi-parametric ones, but much more stable than the non-parametric ones.

4.2 Data contamination

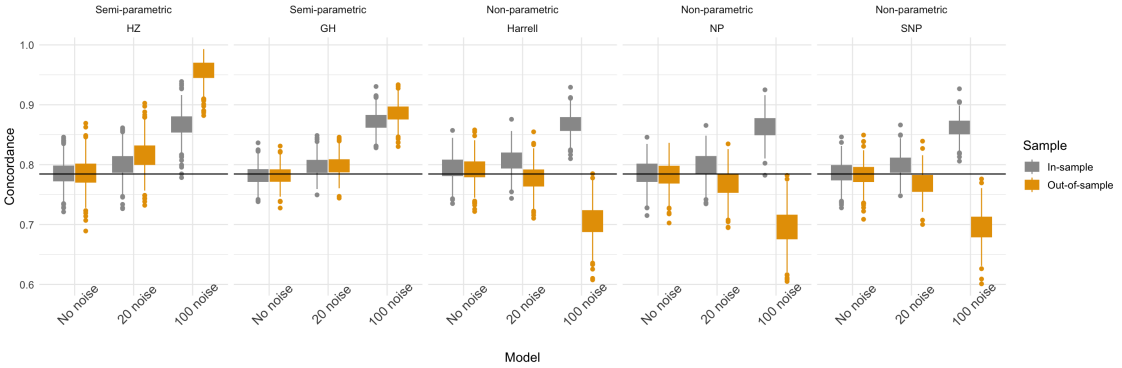
In the second scenario, we follow the same data generation mechanism as model (9). However in the testing sets, we introduce 10% of contaminated observations whose covariates are generated from a different distribution. Specifically, we study the effect on out-of-sample behaviour of two different types of data contamination: 1) a mean shift, where the covariates are generated from $N(5, 1)$; and 2) a spread change, where covariates are generated from $N(0, 5)$. Since these contaminated observations are not used for model fitting, they should not contributed to the discrimitive ability of models fitted on the training sets. We would expect the models to predict these observations poorly, causing low out-of-sample estimates of Incident/Dynamic AUC and concordance.

Figure 3 compares the behavior of different estimators of model discriminative ability when testing samples are contaminated. In this scenario, the same, correctly specified model with three covariates is fitted on training samples but tested on different contaminated testing samples. Similar to the previous scenario, estimated $AUC^{I/D}$ here is visualized in Figure 3a, also smoothed across simulations for presentation purposes. Estimates of concordance are summaried in Figure 3b.

Since the contaminated observations are generated by different distributions from the sample used for model fitting, we expect the model to fit poorly to these observations, causing worse out-of-sample discriminative performance. This would be reflected as dashed lines being lower than the solid lines in Figure 3a , and Yellow boxes being lower than grey boxes in Figure 3b. While the behavior of non-parametric estimators is consistent with such expectations, semi-parametric estimators show the opposite trend. In the



(a) Incident/Dynamic AUC

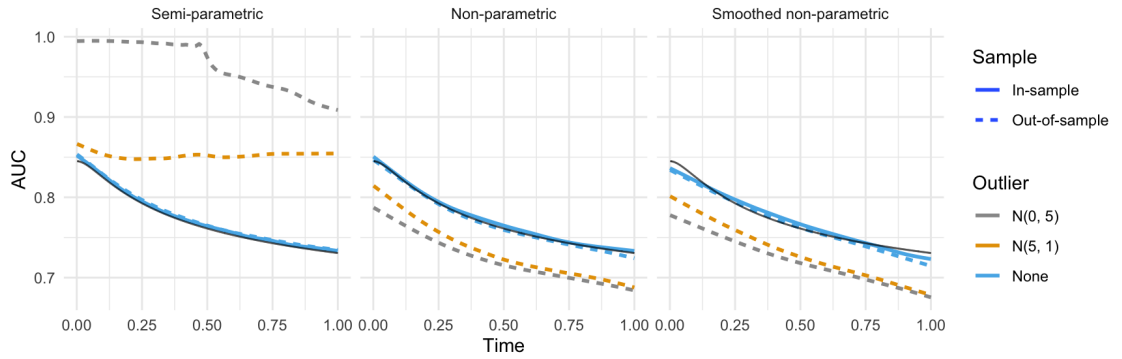


(b) Concordance

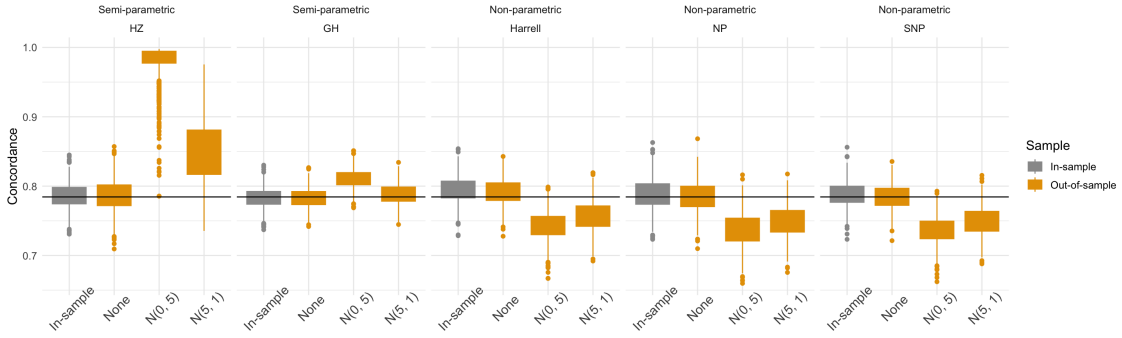
Figure 2: Behavior of estimators of model discrimination under the effect of model overfit. Estimates of Incident/Dynamic AUC are presented in (a), where solid lines represent in-sample estimates and dashed lines represent out-of-sample estimates. Color of lines indicates the underlying model, where grey corresponds to the correctly specified model, yellow a moderately overfit model with 20 noise signals, blue a highly overfit model with 100 noise 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.

left panel of 3a the out-of-sample semi-parametric estimates (dashed lines) on contaminated test samples are clearly higher than both their corresponding in-sample estimates (solid lines) and the true values (solid black lines). This inflation is more prominent when covariates have larger spread (grey lines) compared to mean shift (yellow lines). The concordance estimators in 3b also showed similar behaviors. The semi-parametric estimators, including Heagerty-Zheng (first panel) and Gonen-Heller (second panel) showed higher out-of-sample than in-sample estimates on contaminated testing sets (yellow boxes). When the data is contaminated with a larger variation, Heagerty-Zheng estimates can be inflated close to 1, indicating perfect discrimination. Gonen-Heller seems more robust against data contamination especially when the source of contamination is a mean shift. The Harrel's C, non-parametric and smoother non-parametric estimators all behaved consistently with expectation.

Since the $\hat{AUC}^{I/D}$ showed in Figure 3a without data contamination (blue lines) is essentially the same as the correctly-specific model in the first simulation scenario in Section 4.1. We therefore see the same



(a) Incident/Dynamic AUC



(b) Concordance

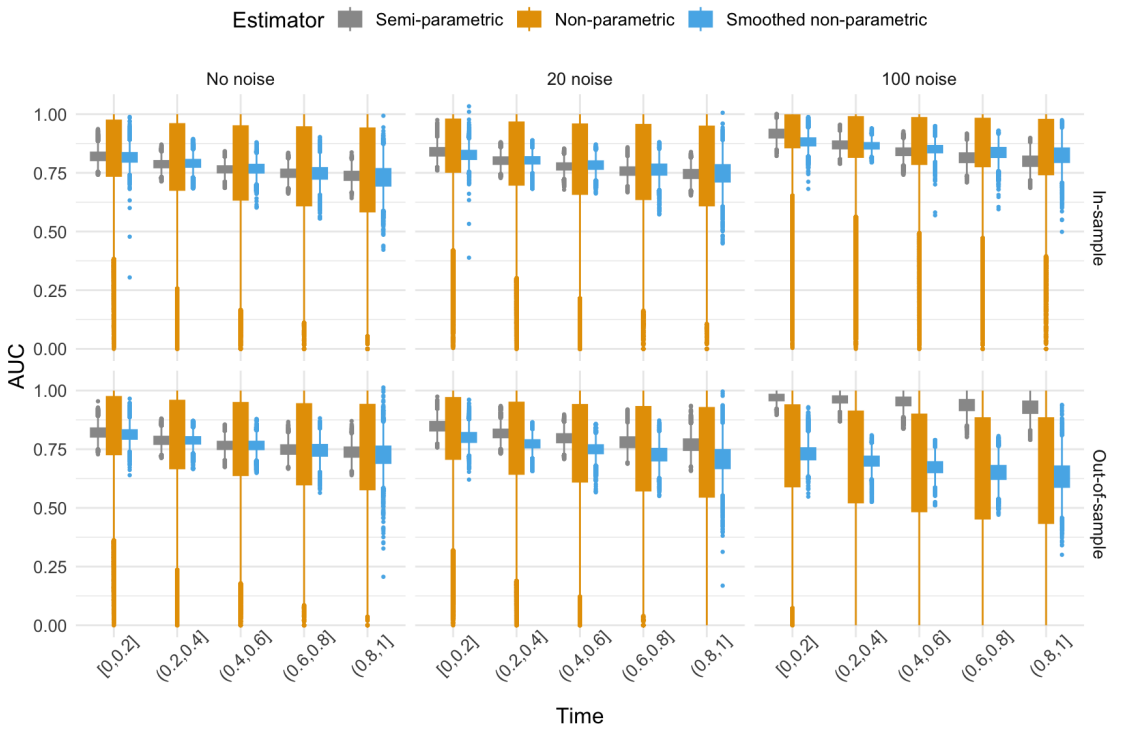
Figure 3: Behavior of estimators of model discrimination under the effect of data contamination. Estimates of Incident/Dynamic AUC are presented in (a), where solid lines represent in-sample estimates and dashed lines out-of-sample estimates. Color of lines indicates the distribution from which outlying observations are generated, where grey corresponds to a greater variation, yellow a mean shift, and blue no contamination . 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.

unbiased semi-parametric estimates, unbiased non-parametric estimates and slightly biased smoothed non-parametric estimates. For concordance, all the estimators are unbiased without data contamination, except for Harrel's C-index (middel panel) which biased upward. Like the scenario of model overfit, non-parametric estimates of $AUC^{I/D}$ are much more unstable than semi-parametric estimates and smoothed non-parametric estiamtes, as presented below in Figure 4b.

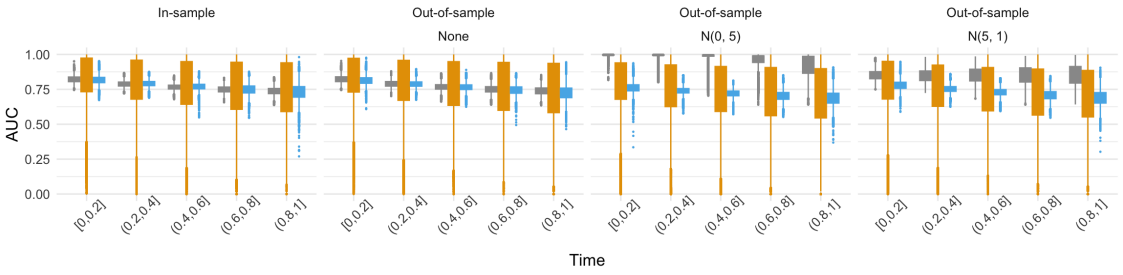
4.3 Mechanism for Inflated Estimation of Out-of-Sample Discrimination

As described in Section 3.3, the cause of the observed out-of-sample inflation lies in the semi-parametric estimator of incident sensitivity. Remember the estimator in Equation (7) weighs observations at risk at time t by the exponential of their estimated risk scores:

$$\frac{exp(\hat{\eta}_k)}{\sum_j I(T_j \geq t) exp(\hat{\eta}_j)}$$



(a) Model overfit



(b) Data contamination

Figure 4: Comparing variability of in-sample and out-of-sample Incident/Dynamic AUC estimates. The top and bottom panels respectively reflect the effect of model overfit or data contamination. The entire follow-up period is divided into five equal-length intervals, and each box represents AUC estimates in the corresponding time interval. Color of boxes represents class of estimator, with gray for semi-parametric, yellow for non-parametric and blue for smoothed non-parametric estimator.

When an observation has a large estimated risk score $\hat{\eta}$, its corresponding weight would be very large. As a result, the semi-parametric estimator would be dominated by the observations with large estimated risk score, regardless of their actual event status at time t or the accuracy of risk estimation. In the simulation study above, both noise signals and data contamination cause the covariates to be more variable, leading to large estimated risk scores for some subjects. Thus, this small proportion of subjects would contribute more to the estimation of sensitivity. For example, when the model is overfitted, one single observation

can weight as high as 70% at a time point. When the test samples include outliers with a larger variation, one outlier can take on 99% of the weight at each time point, while subjects from the population of interest have weights less than 1%. For more details about the distribution of subject weights, please refer to the Section A.1. in the Appendix .

5 Data Application

In this section, we present a real-data example of the out-of-sample overestimation behavior from semi-parametric estimators. Specifically, we used data from the National Health and Nutrition Examination Survey (NHANES) from the year 2011-2014 to predict all-cause mortality, using physical activity features and demographics as predictors. Note that the NHANES study is a multi-stage probabilistic sample from the non-institutionalized US population. Results are thus generalizable when survey sampling methods are accounted for in regression modelling (i.e. survey weights, cluster sampling. etc). As our data application is for illustrative purposes only, we do not account for survey design. We refer interested readers to [17], [13], and [22] for overviews of survey methodology for regression analyses.

5.1 NHANES 2011-2014 Data

The analytic sample includes 3556 participants with age 50-80, 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 [15] for the NHANES 2003-2006 and the UK Biobank data, respectively. The predictor vector $\boldsymbol{X_i}$ included five variables: age, BMI, active-to-sedentary transition probability (ASTP), relative amplitude (RA), and total MIMS units (TMIMS). The latter three variables (ASTP, RA, and TMIMS) are variables derived from participants' wearable accelerometry data. These five variables exhibit moderate pairwise correlation.

5.2 Models

As the simulation study, we would like to compare the in-sample and out-of-sample behavior of semi- and non-parametric estimators on this real dataset. Specifically, we would like to examine their practical utility, e.g. if these estimators assess discrimintive performance properly, and if they are able to identify a overfitted model, etc. According to these purposes, we set up two different Cox models to predict time to mortality. The first model is an additive Cox model as follows:

$$\log \lambda(t|\mathbf{X}_i) = \log \lambda_0(t) + f(\mathbf{X}_i) \tag{10}$$

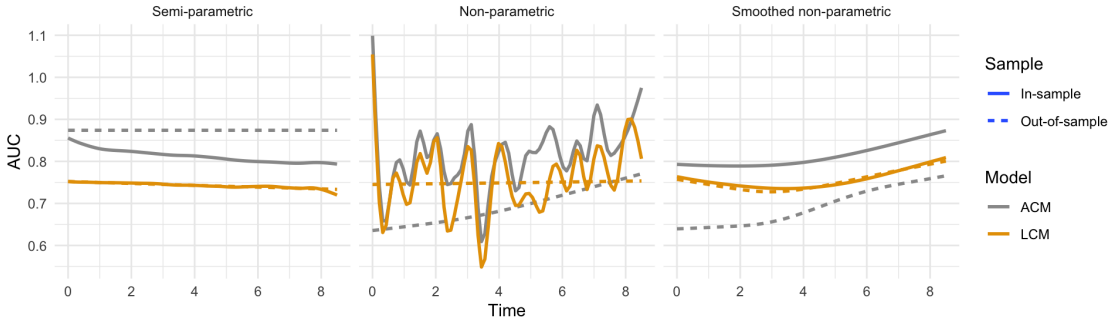
The risk score in this model, $f(\mathbf{X}_i)$, is a smooth function of predictors modelled as a linear combination of 200 unpenalized thin plate regression splines [26] via the *mgcv* package [29] in *R* [19]. This model, which contains a five-dimensional smooth function, estimates 200 coefficients without regularization, thus tends to overfit to the data. We will refer to Model (10) as the "additive Cox model (ACM)".

The second model has a simpler linear form for the risk score as follows:

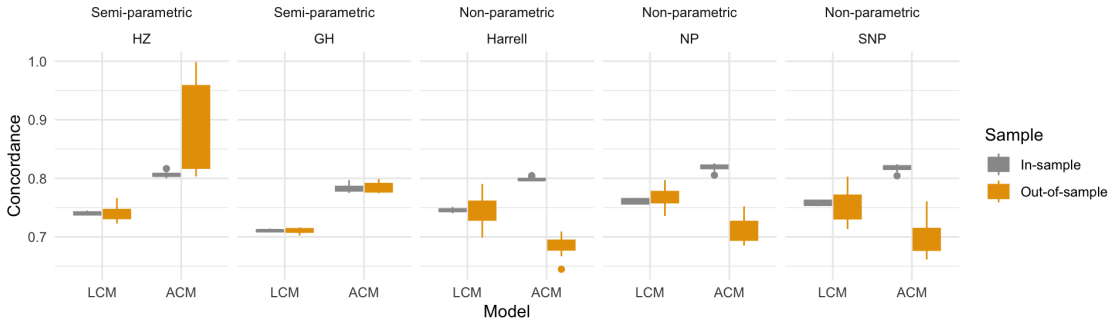
$$\log \lambda(t|\boldsymbol{X}_i) = \log \lambda_0(t) + \boldsymbol{X}_i\boldsymbol{\beta} \tag{11}$$

Given the size of the data and the number of observed events, a linear model of this form will tend not to overfit, particularly in comparison to the ACM in (10). We will refer to Model (11) 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 Heagerty-Zheng estimators of Incident/Dynamic AUC and concordance, as well as Gonen-Heller concordance. The latter includes non-parametric and smoothed non-parametric estimators of Incident/Dynamic AUC, their corresponding concordance (weighted by smoothed survival function) and Harrell's C-index.



(a) Incident/Dynamic AUC



(b) Concordance

Figure 5: Incident/Dynamic AUC and concordance estimates on NHANES data using 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.

5.3 Results

Figure 5 compares the in- and out-of-sample behavior of $\hat{AUC}^{I/D}(t)$ and concordance estimators from the simpler LCM and the more complex ACM models. In Figure 5a, the estimates of $AUC^{I/D}$ are smoothed across all ten testing folds throughout the cross validation process. We note a few key findings. First, the in-sample behavior of three estimators are similar, with estimates from ACM (gray solid lines) higher than estimates from LCM (yellow solide lines). This is to be expected since ACM is more likely to overfit to the training folds. However, the out-of-sample semi-parametric estimates of ACM (left panel, gray dashed line) go much higher not only from the LCM (yellow dashed line), but also from its own in-sample counterpart, with a striking difference as high as 0.15. Indeed, this difference would lead an analyst to choose the more complicated ACM which fits poorly over the simpler LCM if they were to use the semi-parametric estimator with cross-validation for model selection. On the other hand, the non-parametric (middle panel) and smoothed non-parametric estimators (right panel) showed the opposite trend on testing folds consistent with the expectation of overfitted models, favoring the less complex LCM model. As an aside, the highly non-linear shape of the in-sample non-parametric estimator (yellow solid line) of $\hat{AUC}^{I/D}(t)$ for the LCM and ACM appears to be driven by a high degree of inconsistency in the estimator at each time point. This finding may be an artifact of the data and the highly variable nature of the non-parametric estimator.

Estimator	Bias	Variavility	Out-of-sample behavior
Time-dependent AUC			
Semi-parametric	Unbiased	Low	Over-optimistic
Non-parametric	Unbiased	High	Appropriate
Smoothed non-parametric	Slightly biased	Low	Appropriate
Concordance (semi-parametric)			
Heagerty-Zheng	Unbiased	Low	Over-optimistic
Gonen-Heller	Unbiased	Low	Over-optimistic
Concordance (non-parametric)			
Harrell	Biased upwards	Low	Appropriate
Non-parametric	Unbiased	High	Appropriate
Smoothed non-parametric	Unbiased	Low	Appropriate/slightly over-optimistic

Table 1: Summary of behavior of estimators for discriminative performance of time-to-event models

The same out-of-sample inflation is also observed in Concordance estimators, as in the first and second panels of Figure 5b. Both Heagerty-Zheng (first panel) and Gonen-Heller (second panel) showed higher out-of-sample estimates from ACM than LCM, though the discrepancy is smaller for the latter. 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. If we use these estimators as criterion for model selection, both would mislead us to chose the complex ACM model over LCM when in fact it does not perform better. Again, these results would lead an analyst to choose the more complex ACM, despite overfitting to the data, over the LCM when using cross-validated semi-parametric concordance as a model selection criteria. The Harrell's C-index (third panel), non-parametric concordance (fourth panel) and smoothed non-parametric concordance (fifth panel) have lower out-of-sample estimates from ACM than LCM. All three estimator have led us to the concise linear model, instead of the overfitted additive model.

6 Discussion

The simulation study in Section 4 and case study in Section 5 revealed a troublesome behavior of the class of semi-parametric estimator for discriminative performance of time-to-event models. These estimators, including the Heagerty-Zheng estimator of Incident/Dynamic AUC, its corresponding concordance and Gonen-Heller concordance, suffer from an intrinsic tendency to overestimate out-of-sample discrimination even when the model does not fit well to the data. We have also identified the source of this phenomena in Sections 3.3 and 4.3 by pointing out the lack of dependence on the accuracy of risk estimation and actual event status of the incident sensitivity estimator. This property calls into serious question the appropriateness of this class of semi-parametric estimators for model assessment, comparison and selection purposes, particularly when comparing complicated models to simple models or when data is contaminated. It is likely to lead analysts to choose an overfitted, complex model over an appropriate, simple model. If the data is contaminated, these estimators can show misleading result that a model performs much better than it actually does. It is worrisome that regular measures against model overfit, such as out-of-sampl evaluation and cross validation, could not make up for this critical flaw.

The behavior we identified suggests that non-parametric estimators should generally be preferred over semi-parametric estimators of Incident/Dynamic discrimination. Such estimators are unbiased when model is correctly specified, but 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 variability, but

resulted in slight bias for time-dependent Incident/Dynamic AUC. We believe the bias is likely a result of heteroskedasticity of the residual process and correlation of estimates for $AUC^{I/D}(t)$ 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.

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 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. The behaviors and properties of all estimators are summarized in Table 1 above.

Conflict of Interest

The authors have declared no conflict of interest. (or please state any conflicts of interest)

Appendix

A.1. The distribution of subject weights in the semi-parametric incident sensitivity estimator

Below the distribution of subject-specific weights in the estimator of incident sensitivity from one dataset in the simulation study in Section 4 is visualized. From (a) it is easy to see how the additional noise signals has caused the weight estimates to spread a lot higher for some observations in the test samples. When the model is moderately overfit with 20 noise signals (middle panel), the highest weight is 0.3. However with the severely overfitted model with 100 noise signals (lower penal), the weights can end up as high as 0.7. This explains why the out-of-sample inflation from semi-parametric estimators of $AUC^{I/D}$ and concordance would increase with more noise signals in the model.

Similarly, Figure (b) shows the values of weight when 10% of the testing samples is contaminated. A mean shift of covariate distribution (lower-left panel) induced larger weights for these outlying observations. However, the increase of weights is much more significant when the outlying covariates have a larger variance (lower-right penal). For some observations here, the weights go up to close to 1, meaning the sensitivity estimation depends almost entirely on these observations at some time points.

A.2. Evaluating True Incident/Dynamic AUC

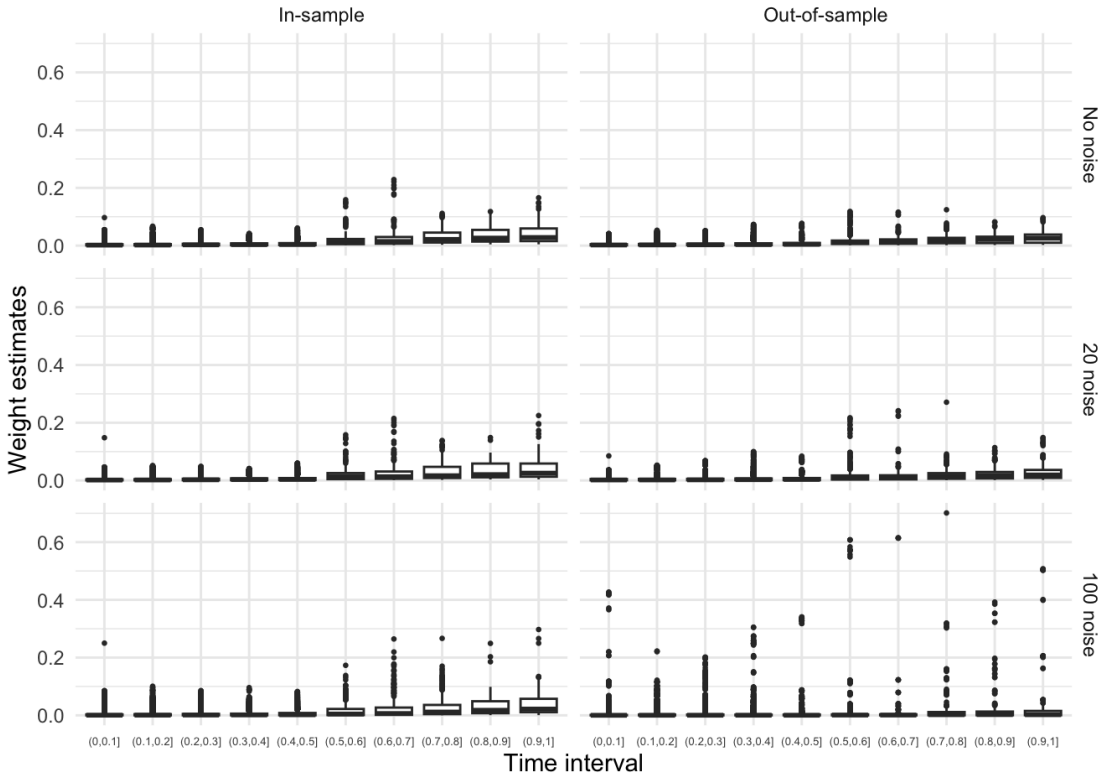
We can obtain true incident sensitivity and dynamic specificity under our data generating mechanism by monte-carlo integration. Specifically, consider incident sensitivity

$$\begin{aligned}\Pr(\eta > c|T = t) &= E[1(\eta > c)|T = t] \\ &= \int 1(\eta > c) f(\eta|t) d\eta \\ &= \int 1(\eta > c) \frac{f(t|\eta) f(\eta)}{\int f(t|\eta) f(\eta) d\eta} d\eta\end{aligned}$$

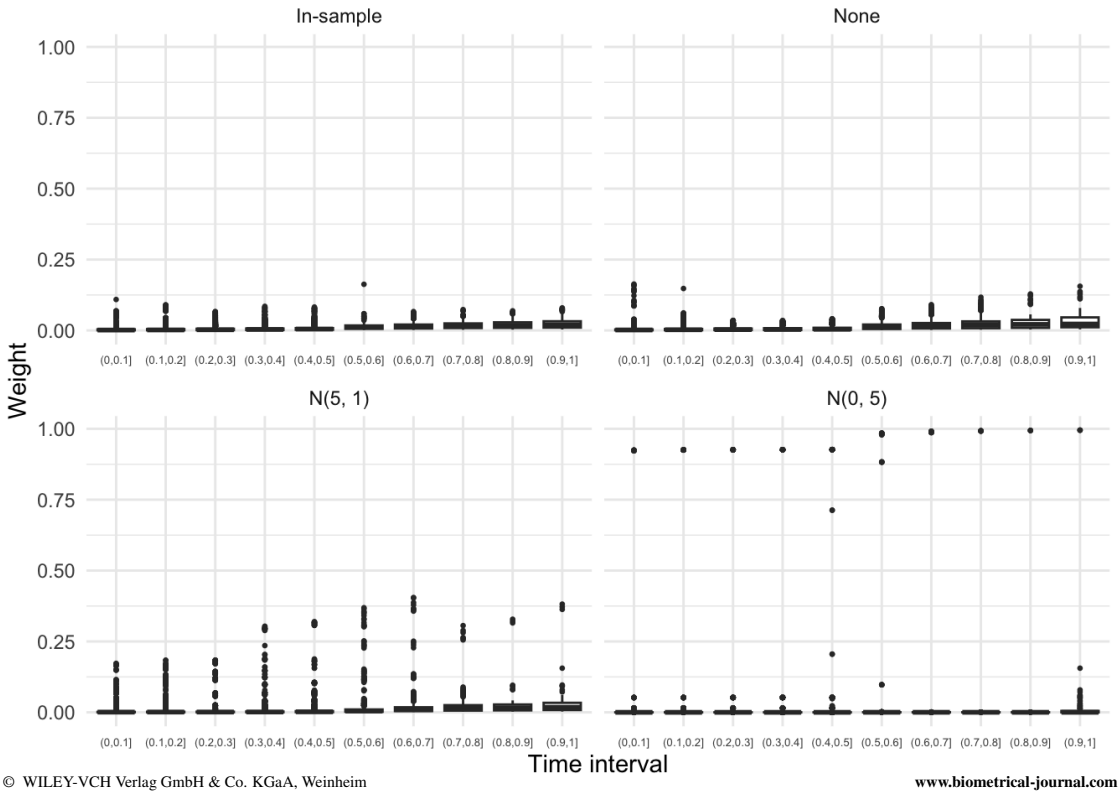
Since $\eta = x^t\beta$ is a linear combination of normal random variables, η is normally distributed. In addition, under the assumption of a Weibull baseline hazard, we can obtain

$$\begin{aligned}f(t|\eta) &= \lambda(t|\eta) S(t|\eta) \\ &= (\theta e^{\eta}) p t^{p-1} e^{-(\theta e^{\eta}) t^p}\end{aligned}$$

We can then estimate incident sensitivity using numeric integration via, e.g., the *integrate()* function in *R*.



(a) Model overfit



(b) Data contamination

Next, consider dynamic specificity

$$\begin{aligned}\Pr(\eta \leq c|T > t) &= \frac{\Pr(\eta \leq c \cap T > t)}{\Pr(T > t)} \\ &= \frac{\int_t^\infty \int_{-\infty}^c f(t,\eta)d\eta dt}{\int_t^\infty [\int f(t|\eta)f(\eta)d\eta]dt} \\ &= \frac{\int_t^\infty \int_{-\infty}^c f(t|\eta)f(\eta)d\eta dt}{\int_t^\infty [\int f(t|\eta)f(\eta)d\eta]dt}\end{aligned}$$

The double integrals involved can again be evaluated using numeric integration via, e.g., the *cubature::adaptIntegrate()* function in *R*.

References

[1] Paul Blanche, Jean François Dartigues, and Hélène Jacqmin-Gadda. Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5):687–704, 2013.

[2] Paul Blanche, Michael W Kattan, and Thomas A Gerds. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics*, 20(2):347–357, 02 2018.

[3] N. E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review / Revue Internationale de Statistique*, 43(1):45–57, 1975.

[4] Centers for Disease Control and Prevention (CDC) National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsvile, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. 2011-2014.

[5] Ruth Etzioni, Margaret Pepe, and Gary Longton. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Med Decis Making*, 19(3):242–251, 1999.

[6] Mithat Gonen and Glenn Heller. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

[7] Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–87, 1996.

[8] Patrick J. Heagerty, Thomas Lumley, and Margaret S. Pepe. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

[9] Patrick J. Heagerty and packaging by Paramita Saha-Chaudhuri. *risksetROC: Riskset ROC curve estimation from censored survival data*, 2012. R package version 1.0.4.

[10] Patrick J. Heagerty and Yingye Zheng. Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105, 2005.

[11] Adina Najwa Kamarudin, Trevor Cox, and Ruwanthi Kolamunnage-Dona. Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, 17(53), 2017.

[12] Marta Karas, John Muschelli, Andrew Leroux, Jacek K Urbanek, Amal A Wanigatunga, Jiawei Bai, Ciprian M Crainiceanu, and Jennifer A Schrack. Comparison of accelerometry-based measures of physical activity: Retrospective observational data analysis study. *JMIR Mhealth Uhealth*, 10(7):e38077, Jul 2022.

[13] Edward L Korn and Barry I Graubard. *Analysis of health surveys*, volume 323. John Wiley & Sons, 2011.

[14] Andrew Leroux, Junrui Di, Ekaterina Smirnova, Elizabeth J. McGuffey, Quy Cao, Elham Bayatmokhtari, Lucia Tabacu, Vadim Zipunnikov, Jacek K. Urbanek, and Ciprian Crainiceanu. Organizing and analyzing the activity data in nhanes. *Statistics in Biosciences*, 11(2):262–287, 2019.

[15] Andrew Leroux, Shiyao Xu, Prosenjit Kundu, John Muschelli, Ekaterina Smirnova, Nilanjan Chatterjee, and Ciprian Crainiceanu. Quantifying the Predictive Performance of Objectively Measured Physical Activity on Mortality in the UK Biobank. *The Journals of Gerontology: Series A*, 76(8):1486–1494, 09 2020.

[16] Jorge M. Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145–151, 2008.

[17] Thomas Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19, 2004.

[18] Natalya Pya. *scam: Shape Constrained Additive Models*, 2021. R package version 1.2-12.

[19] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[20] Matthias Schmid, Hans A Kestler, and Sergej Potapov. On the validity of time-dependent AUC estimators. *Briefings in Bioinformatics*, 16(1):153–168, 09 2013.

[21] Venkatraman E. Seshan and Karissa Whiting. *clinfun: Clinical Trial Design and Data Analysis Functions*, 2022. R package version 1.1.0.

[22] Chris Skinner, Jon Wakefield, et al. Introduction to the design and analysis of complex survey data. *Statistical Science*, 32(2):165–175, 2017.

[23] Royal Statistical Society. Regression models and life-tables. *Journal of the Royal Statistical Society*, 34(2):187–202, 1972.

[24] Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011.

[25] N. Van Geloven, Y. He, A.H. Zwinderman, and H. Putter. Estimation of incident dynamic auc in practice. *Computational Statistics & Data Analysis*, 154:107095, 2021.

[26] Simon Wood. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)*, 65(1):95–114, 2003.

[27] Simon Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686, 2004.

[28] Simon Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36, 2011.

[29] Simon Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017.

[30] Ronghui Xu and John O'Quigley. Proportional hazards estimate of the conditional survival function. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(4):667–680, 2000.

