Concordance Measures for Variable Screening and Model Evaluation with Competing Risks Data

by

Yang Qu

B.S. in Statistics, Shandong University, China, 2016

M.S. in Statistics, University of Wisconsin-Madison, 2017

Submitted to the Graduate Faculty of

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Yang Qu

It was defended on

September 19th 2022

and approved by

Yu Cheng, Department of Statistics, University of Pittsburgh

Satish Iyengar, Department of Statistics, University of Pittsburgh

Kehui Chen, Department of Statistics, University of Pittsburgh

Ying Ding, Department of Biostatistics, University of Pittsburgh

Concordance Measures for Variable Screening and Model Evaluation with Competing Risks Data

Yang Qu, PhD

University of Pittsburgh, 2022

We focus on analysis of time-to-event data with competing risks. In the first project, we make additional assumption of natural ordered event status, and propose a time-dependent model-free variable screening method for high-dimensional data that evaluate the discrimination ability of a biomarker to distinguish multiple event status simultaneously. The proposed method utilizes the Volume under the ROC surface (VUS), which measures the concordance between values of biomarkers and event status at certain time points. We show that the VUS possesses the sure screening property, i.e., true important covariates can be retained with probability tending to one. Simulations and data analysis show that VUS appears to be a viable screening metric, and is robust to data contamination.

In the second project, we provide a systematic examination of model evaluation metrics that evaluate the discrimination ability of prognostic models. Most of the existing metrics focus on how a particular cause of event can be discriminated from the healthy control by the prognostic models when competing events exist, and one metric, the polytomous discrimination index (PDI), additionally provides an overall evaluation of diagnostic accuracy of a group of models for predicting all competing events. A systematic comparison of PDI with other existing methods is missing. We thus fill this gap and illustrate the performance of different model evaluation metrics under various scenarios via simulation studies and data analyses. Several natural extensions of concordance index are also considered, and their performance of model evaluation is also assessed. An R package is developed to provide model evaluation and model comparisons based on existing methods and extended concordance indices.

Table of Contents

List of Tables

List of Figures

Preface

This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681.

1.0 Introduction

1.1 Background

Variable selection and model evaluation are important tasks in statistical analysis. The area under the receiver operating characteristic (ROC) curve (AUC) has been widely used in summarizing sensitivity and specificity, and evaluating the discrimination ability of a marker with binary outcomes. The marker can either be a single covariate or a regression model that estimates the probability of belonging to a class. The AUC measures the probability of concordance between the marker values and the underlying class membership (Pepe, 2003).

In survival analysis, there exists an event of interest, and time-to-event data are observed. At each time point t, subjects either have experienced the event of interest or are still event-free. Subject's status can thus be considered as a time-dependent dichotomous outcome. One challenge with time-to-event data is that event time is subject to censoring, which is a special type of missingness. There are different types of censoring, and this dissertation only focuses on right-censored time-to-event data, where subjects may be lost to follow-up before the end of study, leaving their event status being unobserved. AUC has been extended to analyze survival outcomes with adjustment being made to handle censoring and to evaluate the ability of a marker to discriminate subjects who are at a higher risk of developing an event. Heagerty et al. (2000) and Heagerty and Zheng (2005) proposed time-dependent ROC curves for evaluating the discrimination ability of a diagnostic marker or a regression model. Heagerty and Zheng (2005) also showed that the proposed time-dependent AUC is directly related to the global concordance summary for survival data, which is the probability that the subject experienced the event of interest has a larger value of the marker than the healthy control, assuming that larger values of the marker indicate more severe conditions.

The existence of multiple competing risks brings new challenges. In the real world, subjects can experience events of different causes, which are called competing risks. The occurrence of one cause of event may hinder or alter the probability of the occurrence of other causes of events, therefore may prevent other causes of events from being observed.

Ignoring competing risks or simply treating competing risks as censoring would lead to biased analysis (Gooley et al., 1999). Therefore, it is desirable to investigate methods for variable selection and model evaluation under the competing risks setting. Efforts have been made to extend AUC methods to competing risks data. Saha and Heagerty (2010) extended the concept of the ROC curve by estimating sensitivity based on cumulative or incident cases of one specific cause of event and estimating specificity based on controls that are free of any event. Zheng et al. (2012) further allowed additional covariates that may affect the accuracy of the marker of interest in the estimation and provided inference procedure. Blanche et al. (2013) considered two alternative definitions of the control group in defining specificity. For one definition, the control group was defined as subjects who were free of any event by the pre-specified time t. For the other, subjects who didn't experience the event of interest were defined as controls, which included subjects who were event free and subjects who developed competing events before t. Estimators of sensitivity and each definition of specificity were proposed with the use of inverse probability of censoring weighting (IPCW) to handle independent censoring. Wolbers et al. (2014) proposed an IPCW estimate of a cause-specific concordance index for prognostic scores, which was shown to be a weighted average of time-dependent AUC over time, with the control group defined as subjects who experienced the event of interest later or not at all. Wu and Li (2018) provided cause-specific AUC estimates based on two different definitions of time-dependent specificity by weighting censored subjects with the conditional probability of experiencing event of interest, where again AUCs were defined as probabilities of concordance of prognostic scores.

We see that in all above works, AUC can only evaluate the predictive accuracy of a marker or a prognostic model for one specific cause of event at a time. For a single marker, AUC would be deficient if the marker is related to a sequence of competing events, for example, the progression of cancer or cognitive impairment, as it cannot evaluate the relationship between the marker and the whole disease progress. When evaluating a prognostic model, AUC can only assess how one specific cause of event is discriminated from the control group or all other events at a time, thus lacking an overall evaluation of effectiveness of a set of models fitted for all causes of event. Besides, the definition of different types of control groups may be hard to interpret. As we see above, a control group can either include only event-free

subjects or all subjects who haven't experienced the event of interest. The latter definition combines subjects who developed all other events before the t and event-free subjects.

When there are two competing events, at any time t, subjects can be divided into three groups: subjects who developed cause-1 event before t, those who developed cause-2 event before t, and those who are still event-free by time t. We have even more groups when there are more than two competing events. Therefore one natural way to extend AUC to competing risks data is to find its multi-dimensional analog. Metrics have been developed to evaluate decision making procedures or diagnostic tests with traditional multi-category outcomes. Scurfield (1996) and Mossman (1999) independently introduced the three-way ROC to handle diagnostic tests with three possible outcomes. Mossman (1999) showed that the volume under the three-way ROC surface is equal to the probability that three subjects each from three distinct categories are correctly sorted. Mossman (1999) also proposed three decision rules that yield the ROC surface. Dreiseitl et al. (2000) provided the standard deviation of the VUS estimate proposed by Mossman (1999) based on the Mann-Whitney U-statistic. Nakas and Yiannoutsos (2004) provided a non-parametric estimator of VUS assuming that a natural order exists for the three possible outcomes, and modifications were given to handle ties in the marker. The idea can be further extended to multi-category outcomes with more than three categories, which was done by Li and Fine (2008), where they rigorously defined the hypervolume under the ROC manifold (HUM). Li and Zhou (2009) again focused on ordinal outcomes and proposed non-parametric and semi-parametric estimates of the three-dimensional ROC surface itself. For survival outcomes with competing risks, recent work by Zhang et al. (2021) utilized VUS to evaluate the time-varying prognostic accuracy of a biomarker to competing risks outcomes with a natural order, and proposed VUS estimator based on different interpretation: the volume under the ROC surface and the concordance index. We've seen that AUC has been generalized to various conditions, yet the usage of VUS under the competing risks setting has not been fully discovered.

1.2 Overview of the Dissertation

My dissertation consists of two parts: risk screening and model evaluation for right-censored survival outcome with competing risks. We will focus on conditions with two competing events. Ideas can be generalized to multi-category outcomes, though inferences may become more complicated. The aim of risk screening is to reduce the number of covariates to a moderate size while keeping all important covariates, so that existing variable selection methods can be applied.

In Chapter 2, we adopt the VUS proposed by Zhang et al. (2021) as a model-free ultra-high dimensional variable screening method for competing risks outcomes. The VUS is shown to possess the sure screening property by locating markers that are associated with event times from all causes simultaneously under certain conditions. We run simulation studies under different settings to test the performance of the VUS as a screening metric, and conduct a real data analysis of gene-expression data from a breast cancer study (van de Vijver et al., 2002).

Model evaluation follows naturally after we have a candidate sets of covariates and prognostic models. In Chapter 3. we provide a comparison of model evaluation metrics that are commonly used for competing risks data. Meanwhile, we examine the ability of two natural extensions of concordance index to perform as model evaluation metrics. These extensions are supposed to measure the probability that randomly selected subjects from distinct classes are correctly assigned by the model. We run extensive simulation studies to examine the performance of model evaluation metrics under different circumstances. The behaviors of commonly used prognostic models are also evaluated through multiple real data analyses. The extensions of concordance index are shown to have nice performance in simulations. We thus develop an R package to provide model evaluation and model comparisons based on existing methods and extended concordance indices.

2.0 VUS as a Screening Metric

2.1 Introduction

Traditional variable selection methods have been fully investigated with moderately high dimensional covariates. However, the performance of these methods is not guaranteed when the dimension of covariates p is ultra-high, i.e., $\log(p)$ is of the order of sample size. Advanced technology facilitates the availability of ultra-high dimensional data, for example, gene expressions, and the development of variable screening methods becomes increasingly necessary.

The goal of a variable screening method is to reduce the number of covariates to a moderate size, which would then allow the use of traditional or high-dimensional variable selection methods (Fan and Lv, 2008). Fan and Lv (2008) introduced the concept of sure screening and proposed the sure independent screening (SIS) method simply by keeping covariates that are independently highly correlated with the outcome of interest. They showed that their SIS method possesses the sure screening property, which means that the screening method would keep all important covariates with probability tending to one.

The use of the sure screening method was extended to survival outcomes. Fan et al. (2010) extended the SIS to variable screening for Cox's proportional hazard model via a penalization method, without rigorous proof of the sure screening property. Zhao and Li (2012) proposed the principle sure independence screening (PSIS) method using a Wald-type statistic from the marginal Cox regression model for each single covariate. A cutoff point for variable selection was provided by controlling the false positive rate. Gorst-Rasmussen and Scheike (2013) defined the 'feature aberration at survival times' (FAST) statistic to measure marginal association between each covariate and survival. The usage of the FAST statistic was justified for single-index hazard rate models, where hazard function $\lambda(t)$ is of the form $\lambda(t) = \lambda(t, Z^T\alpha^0)$ with α^0 being a vector of regression coefficients. These methods all made assumptions that covariates are associated to the hazard rate via some functional form and thus may not be adequate when the model assumption is violated. Various model-free

methods were developed to overcome the problem. Song et al. (2014) proposed a censored rank independence screening method using an IPCW Kendall's τ statistic that measures marginal rank correlation. Li et al. (2016) proposed a survival impact index as a weighted average distance between the covariate-stratified survival function and the marginal survival function, using the Kaplan-Meier estimator of survival function to avoid strong model assumptions. Hong et al. (2018) proposed an integrated powered density (IPOD) criterion to screen variables, where for each covariate the maximum absolute difference in IPODs at all discretized covariate values is used as the screening criterion. Pan et al. (2018) proposed a non-parametric independence feature screening procedure that utilizes the correlation between each covariate and the indicator function of whether an event occurred before a specific time. All of these methods were shown to possess the sure screening property.

We see various extensions to survival outcomes, but the investigation of sure screening method under competing risks setting is still limited. A naive way to deal with competing risks problems is to treat failures from other causes as censoring. However, event times from different causes are usually correlated, which leads to dependent/informative censoring, and thus the assumption of independent censoring in most screening methods for survival outcomes is violated. Besides, the naive method can only focus on one cause at a time, hence it is deficient in evaluating how a marker can predict different competitive failures simultaneously. Peng (2019) proposed a joint correlation rank screening method for semi-competing risks data, which utilized the squared correlation between each covariate and the joint survival function of two competing events as a marginal utility measure. Lu et al. (2020) adopted the distance correlation between each covariate's survival function and the joint survival function of non-terminal event and terminal events as the screening metric. The latter seems to perform better by using a covariate's survival function to avoid the subexponential tail probablity assumption for the covariate and by using the distance correlation to enjoy the independent property when the distance is zero. However, both methods are developed under the semi-competing risks setting, and require that the minimum of non-terminal and terminal events and censoring time and the minimum of the terminal event time and censoring time are both observed, which can be restrictive under the competing risks setting when only time to the first event is observed.

We aim to develop a model-free variable screening metric for ultra-high dimensional data under the competing risks setting. Zhang et al. (2021) proposed to use VUS as an assessment of the overall discrimination capacity of a continuous marker for multi-level categorical outcomes with a natural ordinal disease status. It measures the time-dependent probability of concordance between the marker and disease status, with which we can observe how the discrimination ability changes over time for each marker of interest. Besides, the estimator of VUS utilizes a U-statistic and does not depend on any model assumption. In this chapter, we adopt VUS as the screening metric and show that it possesses the sure screening property.

2.2 Volume under the ROC Surface

In this section, we provide a brief review of VUS and focus on the case with two ordered competing events assuming that the cause-1 event is more severe than the cause-2 event. Cases with more than two competing events can be generalized. Let T be the time to the first event and $\epsilon = 1, 2$ be the cause indicator, with $\epsilon = i$ indicating subject experienced event i. Let $D(t)$ denote the disease status of a subject at time t. At a fixed time point t_0, subjects can be divided into three classes:

$$\begin{cases} D(t_0) = 1, \text{if } T \leq t_0, \epsilon = 1, \\ D(t_0) = 2, \text{if } T \leq t_0, \epsilon = 2, \\ D(t_0) = 3, \text{if } T > t_0. \end{cases}$$

In semi-competing risks setting, it's possible that only the most severe event is observed because patients are not followed up closely. VUS can still work for semi-competing risks data, and we define subject's status at time t_0 as the last event observed before t_0.

Let $\mathbf{Z}$ be a p-dimensional vector of covariates. For each covariate, we assume that smaller values indicate more severe medical conditions. For a specific single covariate Z_l, let $c_1, c_2 \in \mathcal{R}$ be two cutpoints with $c_1 \leq c_2$. We assign a subject into class 1 if its $Z_l \leq c_1$, class 2 if $c_1 < Z_l \leq c_2$ and class 3 if $Z_l > c_2$. Then the correct classification probabilities are given by

$$\begin{cases} CCP_1 = P(Z_l \leq c_1 | T \leq t_0, \epsilon = 1), \\ CCP_2 = P(c_1 < Z_l \leq c_2 | T \leq t_0, \epsilon = 2), \\ CCP_3 = P(Z_l > c_2 | T > t_0). \end{cases}$$

The plot of (CCP_1, CCP_2, CCP_3) for all (c_1, c_2) forms the ROC surface, and the VUS is defined as the volume under the ROC surface. Mossman (1999) showed that the VUS also has an interpretation as a concordance index. Suppose we randomly select three subjects i, j, k, such that $T_i \leq t_0, \epsilon_i = 1$, $T_j \leq t_0, \epsilon_j = 2$ and $T_k > t_0$, then for a single covariate Z_l,

$$VUS_l(t_0) = P(Z_{il} < Z_{jl} < Z_{kl} | T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0),$$

where Z_{il} is the lth covariate for subject i.

Let C denote the time of censoring, then for subjects who are censored before t_0, their disease statuses are not observable. Zhang et al. (2021) proposed a VUS estimator using the inverse probability of censoring weighting (IPCW) method to accommodate censoring. The idea of inverse probability method is to inversely weight subjects by the probability of being observed. We assume that C is independent of T and covariates Z. Define $X = \min(T, C)$, and $\eta = I(T \leq C)\epsilon$, and the observed data consists of independent and identically distributed triplets $\{(X_i, \eta_i, \mathbf{Z}_i)\}, i = 1 \ldots n$. Let G be the survival function of C and denote its Kaplan-Meier estimator as $\hat{G}$. Then from Zhang et al. (2021), an estimator of the VUS for the l-th covariate based on U-statistics and IPCW is given by

$$\widehat{VUS}_l(t_0) = \frac{\sum_{i \neq j \neq k} \frac{I(X_i \leq t_0, \eta_i = 1, X_j \leq t_0, \eta_j = 2, X_k > t_0, Z_{il} < Z_{jl} < Z_{kl})}{\hat{G}(X_i -)\hat{G}(X_j -)\hat{G}(t_0)}}{\sum_{i \neq j \neq k} \frac{I(X_i \leq t_0, \eta_i = 1, X_j \leq t_0, \eta_j = 2, X_k > t_0)}{\hat{G}(X_i -)\hat{G}(X_j -)\hat{G}(t_0)}}.$$

In the presence of tied biomarkers, a modification is made by substituting $I(Z_{il} < Z_{jl} < Z_{kl})$ with $I(Z_{il} < Z_{jl} < Z_{kl}) + \frac{1}{2} I(Z_{il} < Z_{jl} = Z_{kl}) + \frac{1}{2} I(Z_{il} = Z_{jl} < Z_{kl}) + \frac{1}{6} I(Z_{il} = Z_{jl} = Z_{kl})$. Zhang et al. (2021) showed the consistency and weak convergence of $\widehat{VUS}(t_0)$.

2.3 VUS for screening

We now use $VUS(t)$ as a metric for variable screening. Define $\mathcal{M}_*$ as the set of true active variables such that

$$\mathcal{M}_* = \{l : P(T \leq t_0, \epsilon = 1 \text{ or } 2|\mathbf{Z}) \text{ depends on } Z_l\}.$$

We select a set of important covariates by comparing their VUS estimates, $\widehat{VUS}(t)$ with 1/6, which is the value of VUS when there's no association between the covariates and event time. Suppose γ_n is a pre-specified threshold. The selected set is denoted by

$$\hat{\mathcal{M}} = \{l : |\widehat{VUS}_l(t_0) - 1/6| \geq \gamma_n\}.$$

Note that there are several cases where the VUS can be below 1/6. In constructing the ROC surface, we assume that a smaller covariate value indicates a more severe condition. This assumption may be violated in practice. Under a two-dimensional case, the area under the ROC curve (AUC) can be smaller than 1/2 if the assumed order of the biomarker is incorrect and we can simply flip the AUC value as 1-AUC. With three categories, things become more complicated. For example, the important covariates can still be related to the outcome in an ordinal manner, but with larger values indicating more severe conditions. These covariates are less likely to be selected than variables satisfying the assumption, because the distance between their VUSs and 1/6 are bounded by 1/6. For these covariates, we would reverse their signs to satisfy the assumption, and VUS would still work. However, if the relationship between the covariate and the outcomes is not ordinal, we cannot tell where the value of VUS would locate and our metric can fail to pick such covariates.

2.4 Sure Screening Property

In this section, we establish the sure screening property of the VUS method. The following conditions are required.

Condition 1 There exists a $\nu > 0$ such that $P(C = \nu) > 0$ and $P(C > \nu) = 0$.

Condition 2 $min_{l \in \mathcal{M}_*} |VUS_l(t_0) - 1/6| \geq c_0 n^{-\kappa}$ for some $0 < \kappa < 1/2$ and $c_0 > 0$.

Condition 3 There exists $\delta > 0$, such that $P(T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0) > \delta$.

Condition 1 is used to show asymptotic properties in Song et al. (2014) for survival outcomes. Condition 2 indicates that true active covariates can be distinguished from pure noise. Condition 3 is used to show asymptotic properties and can be easily satisfied with a properly selected t_0.

Theorem 1. *Under Conditions 1-3, for any positive constant c_6, there exist positive constants c_3, c_4 and c_5 such that for any single covariate Z,*

$$\begin{aligned} P(|\widehat{VUS}(t_0) - VUS(t_0)| \geq c_6 n^{-\kappa}) \leq\ & 10n^3 \exp\{-\frac{1}{36}\epsilon^8 n\} + 4\exp\{-\frac{2}{3}c_3^2\epsilon^6 n\} \\ & + 4\exp\left\{-\frac{2}{27}c_4^2\epsilon^6\delta^2(1 + c_4 n^{-\kappa})^{-2} n^{1-2\kappa}\right\} \\ & + 2.5n^3 \exp\left\{-\frac{1}{9}c_4^2\delta^2\epsilon^8(3 + 3c_4 n^{-\kappa} + c_4 n^{-\kappa}\delta)^{-2} n^{1-2\kappa}\right\} \\ & + 4\exp\{-\frac{2}{27}c_5^2\epsilon^6 n^{1-2\kappa}\} \\ & + 2.5n^3 \exp\{-\frac{1}{9}c_5^2\epsilon^8(3 + c_5 n^{-\kappa})^{-2} n^{1-2\kappa}\}. \end{aligned} \tag{2.4.1}$$

Taking $\gamma_n = cn^{-\kappa}$ with $c = c_0 - c_6$, we have

$$P(\mathcal{M}_* \subset \hat{\mathcal{M}}) \geq 1 - sP(|\widehat{VUS}(t_0) - VUS(t_0)| \geq (c_0 - c)n^{-\kappa}),$$

where $s = |\mathcal{M}_*|$ is the cardinality of $\mathcal{M}_*$.

Theorem 2. *Under the conditions of Theorem 1, with* $p = o(\exp(n^{1-2\kappa}))$ *and assuming* $\sum_{l=1}^{p} |VUS_l(t_0) - 1/6| = O(n^{\xi})$ *for some* $\xi > 0$*, we have*

$$P(|\hat{\mathcal{M}}| \leq O(n^{\xi+\kappa})) \geq P(\max_{1\leq l\leq p} |\widehat{VUS}_l(t_0) - VUS_l(t_0)| \leq \frac{1}{2}c_6 n^{-\kappa})$$
$$\geq 1 - pP(|\widehat{VUS}_l(t_0) - VUS_l(t_0)| \geq \frac{1}{2}c_6 n^{-\kappa}).$$

The detailed proofs of Theorem 1 and Theorem 2 can be found in the Appendix A. We assume the sparsity of true active covariates, thus Theorem 1 shows the sure screening property of the VUS method. Theorem 2 shows that the size of the selected set $\hat{\mathcal{M}}$ can be controlled when $p = o(\exp(n^{1-2\kappa}))$ and $\sum_{l=1}^{p} |VUS_l(t_0) - 1/6| = O(n^{\xi})$.

2.5 Simulation

In this section, we evaluate finite sample performance of the VUS-based screening under different scenarios. The VUS is compared with two existing methods, PSIS (Zhao and Li, 2012) and Kendall's τ (Song et al., 2014). For the PSIS method, we fit a cause-specific hazard model for each cause of event by treating events from the other cause as if they were independent censoring, and look for important biomarkers associated with each cause-specific hazard. The important set will contain the union of important biomarkers from the two events. For Kendall's τ method, it can only handle typical survival outcomes with independent censoring, and cause-1 and cause-2 events are typically not independent given the covariates. Thus, subjects who have experienced either event 1 or event 2 are combined together as the overall event group in implementing this method.

We considered the following three scenarios. Under each scenario, we simulated 200 datasets with number of subjects $n = 200$ and number of covariates $p = 5000$. Covariates $Z = (Z_1, \ldots, Z_p)'$ under scenario 1 are generated from a multivariate normal distribution with mean 0 and correlation $0.5^{|i-j|}$ between Z_i and Z_j. For each scenario, there are four true covariates Z_1, Z_2, Z_3, Z_4.

Scenario 1: Latent event times were generated from log-logistic models

$$\log(T_j) = \beta_j' \mathbf{Z} + \sigma e, \quad j = 1, 2,$$

with $\sigma = 0.2$, $\beta_1' = (1, 0.9, 0.8, 0.5, 0, \cdots, 0)$ for the cause 1 event and $\beta_2' = (0.5, 0.3, 0.2, 0.1, 0, \cdots, 0)$ for the cause 2 event. If $T_1 < T_2$, the time to first event T was set as T_1 and the event indicator ϵ was set as 1; otherwise, the first event time was set as T_2 with ϵ being 2. As only the time to first event is recorded, T_1 and T_2 cannot be observed simultaneously and they are thus referred to as latent event times. Censoring time was generated from a mixture of uniform distributions with censoring rates of 20% and 40%. The observed event time is the minimum of time to first event and time of censoring, and the corresponding indicator is the product of the censoring indicator and the cause indicator. We estimated VUS at $t_0 = 1$ for which there is about 20% censoring and $t_0 = 1.7$ for 40% censoring.

Scenario 2: Event times were generated from a Fine-Gray model (Fine and Gray, 1999) with the cumulative incidence function for cause 1 being

$$F_1(t|\mathbf{Y}) = 1 - \left[1 - 0.8\left\{1 - \exp\left(-(t/20)^5\right)\right\}\right]^{\exp(\beta_1'\mathbf{Y})}.$$

As noted in the next Chapter, F_1 is an improper distribution and may not be invertible. Let V be a uniform random variable. If V is smaller than $F_1(\infty|\mathbf{Y})$ which is the asymptote of F_1, we invert $F_1^{-1}(V)$ to simulate the event time and set the cause indicator as 1. If V exceeds $F_1(\infty|\mathbf{Y})$, F_1 is not invertible and the corresponding subject is assumed to have experienced cause 2 event first, i.e., $T_2 < T_1$ under our previous latent framework. With this the event time is simulated based on the conditional distribution for T given that the cause 2 event has occurred first. That is,

$$P(T \leq t|\epsilon = 2, \mathbf{Y}) = 1 - \exp\Big(-\exp(\beta_2'\mathbf{Y})(t/20)^5\Big).$$

$\mathbf{Y}$ is a long vector containing all discretized biomarkers. Each observed continuous biomarker Z_k is discretized into three categories $Z_k < -0.5, -0.5 \leq Z_k \leq 0.5$ and $Z_k \geq 0.5$, and Z_k indicates which category Z_k falls into. The associated coefficients for each category are $\beta_{1k}' = (\log 3, \log 1/3, \log 1/6)$ and $\beta_{2k}' = (\log 9, 0, \log 1/2)$ for Z_1, Z_2, Z_3, Z_4 and zero otherwise. Censoring rates were set to be 20% and 40%. The VUS was estimated at $t_0 = 17$ in all cases in this scenario.

Scenario 3: Event times were generated from Gerds' multinomial logistic regression model (Gerds et al., 2012) with cause-1 CIF

$$F_1(t|\mathbf{Y}) = \frac{\exp(a_1 t + b_1 + \beta_1'\mathbf{Y})}{\exp(a_1 t + b_1 + \beta_1'\mathbf{Y}) + \exp(a_2 t + b_2 + \beta_2'\mathbf{Y}) + 1}$$

and cause-2 CIF

$$F_2(t|\mathbf{Y}) = \frac{\exp(a_2 t + b_2 + \beta_2'\mathbf{Y})}{\exp(a_1 t + b_1 + \beta_1'\mathbf{Y}) + \exp(a_2 t + b_2 + \beta_2'\mathbf{Y}) + 1},$$

where $a_1 = a_2 = 2$ and $b_1 = b_2 = -15$. $\mathbf{Y}$ is the same as scenario 2 and the associated coefficients for each category are $\beta_{1k}' = (\log 0.9, \log 0.1, \log 0.05)$ and $\beta_{2k}' = (\log 0.1, \log 0.9, \log 0.45)$ for Z_1, Z_2, Z_3, Z_4 and zero otherwise. The cause indicator was generated from a Bernoulli distribution with probability $F_1/(F_1+F_2)$, where F_1 and F_2 were calculated at the simulated event time for each subject. We again considered censoring rates of 20% and 40%. The VUS was estimated at $t_0 = 10$ and $t_0 = 9$, respectively.

For VUS and Kendall's τ, we summarize how many true variables can be captured when we select 8, 20, 40, 60, 80 variables. For the PSIS method, we selected 4, 10, 20, 30, 40 variables for each cause of event, and finally the selected covariates were the union of important biomarkers from each type of event. Following the investigation in Song et al. (2014), under each setting, we also examined the performance of three metrics when observed covariates $\mathbf{Z}$ were contaminated; with a probability of 0.1 each covariate could be contaminated by a t distribution with mean 0 and 1 degree of freedom.

Results are summarized in Tables 1, 2 and 3. We first noted that these methods would not necessarily have a better performance with a lower rate of censoring. It is particurly clear in Table 3, where data were generated from Gerds' multinomial regression model. After a careful investigation, we found that what really matters is the number of subjects falling in each disease category at the time of prediction, especially for the VUS method, which requires enough samples in each category. Therefore, when the rate of censoring is low, there may be few subjects in the 'survivor' category at the time of prediction, resulting in a less satisfactory performance of VUS.

Comparing VUS, Kendall's τ and PSIS, we see that VUS cannot beat the other two methods under the log-logistic model with no contamination in covariates. When data

Table 1: Number of true variables captured under latent log-logistic models

	Non-contaminated						Contaminated					
	20%			40%			20%			40%		
size	VUS	τ	PSIS	VUS	τ	PSIS	VUS	τ	PSIS	VUS	τ	PSIS
8	3.89	4	4	3.74	3.99	4	3.815	3.98	2.42	3.535	3.965	2.295
20	3.94	4	4	3.805	3.995	4	3.905	3.99	2.615	3.67	3.995	2.51
40	3.95	4	4	3.845	4	4	3.925	3.99	2.69	3.765	4	2.61
60	3.955	4	4	3.865	4	4	3.93	3.99	2.77	3.82	4	2.685
80	3.955	4	4	3.88	4	4	3.935	3.99	2.815	3.84	4	2.78

Table 2: Number of true variables captured under Fine-Gray model

	Non-contaminated						Contaminated					
	20%			40%			20%			40%		
size	VUS	τ	PSIS	VUS	τ	PSIS	VUS	τ	PSIS	VUS	τ	PSIS
8	3.99	3.98	3.995	3.96	3.955	3.99	3.905	3.94	2.155	3.825	3.765	2.07
20	3.995	3.995	4	3.97	3.985	4	3.905	3.98	2.375	3.915	3.88	2.32
40	4	4	4	3.995	3.99	4	3.985	3.98	2.5	3.935	3.935	2.455
60	4	4	4	4	3.99	4	3.985	3.98	2.555	3.965	3.955	2.525
80	4	4	4	4	3.99	4	3.99	3.99	2.62	3.975	3.97	2.575

Table 3: Number of true variables captured under Gerds model

	Non-contaminated						Contaminated					
	20%			40%			20%			40%		
size	VUS	τ	PSIS	VUS	τ	PSIS	VUS	τ	PSIS	VUS	τ	PSIS
8	1.505	0.005	3.630	2.56	0.005	3.67	1.010	0	1.240	1.79	0.005	1.125
20	1.920	0.025	3.870	3.09	0.01	3.89	1.375	0.005	1.460	2.45	0.015	1.345
40	2.155	0.045	3.940	3.335	0.025	3.955	1.685	0.020	1.645	2.8	0.025	1.515
60	2.315	0.075	3.955	3.55	0.05	3.965	1.910	0.030	1.780	3.005	0.045	1.67
80	2.470	0.085	3.960	3.565	0.075	3.975	2.080	0.055	1.885	3.12	0.08	1.745

are contaminated, PSIS is affected the most and has worse performance than the other two methods. When data were generated from Fine-Gray model, again PSIS fails when covariates are contaminated. Both VUS and Kendall's τ perform well. Under the Gerds multinomial regression model, Kendall's τ method completely fails, and VUS has a better performance than PSIS when covariates are contaminated. In this model, after collapsing cause-1 and cause-2 events for Kendall's τ method, the overall risk is the same for the two categories $Z < -0.5$ and $-0.5 < Z < 0.5$, although the overall risk for subjects with large covariates is small. Due to the lack of ability to distinguish these two categories, Kendall's τ isn't able to capture the relationship between the covariate and the outcome, while VUS still works. Overall, we observe that when covariates are contaminated, performances of all three metrics are negatively affected, but compared to the PSIS method, VUS and Kendall's τ are more roust to contamination. It can be seen that with higher rate of censoring, VUS outperformed the other two metrics under both Fine-Gray and Gerds models.

2.6 Data Analysis

We applied our screening method to a gene-expression dataset from a breast cancer study (van de Vijver et al., 2002). This dataset is obtained from the R package "cancerdata" (Budczies and Kosztyla, 2021) and contains 295 women with breast cancer and expression values of 24881 genes in tumor samples of each woman. Two events of interest are distant metastasis and death. Among 295 patients, five (1.7%) patients died without metastasis, 101 (34.2%) experienced metastasis, and 74 (25.1%) died after metastasis. The overall censoring rate, i.e, patients survived without metastasis and death, is 64.1%. The objective of our analysis is to capture the genes that are associated with the progression from breast cancer to distant metastasis and/or death.

We looked at patients' survival at $t_0 = 5$ and considered the most severe event each patient experienced before t_0. At five years, 48 patients died either with or without metastasis, 32 patients were alive but with metastasis, 207 patients were alive and metastasis-free and eight patients censored before t_0.

Different from simulation studies, in real data analysis, we are not sure how covariates are associated with the outcomes. As mentioned in Section 2.3, for covariates whose larger values are associated with more severe conditions, we will reverse the relationship. In this dataset, it is not clear which genes violated our assumption. Therefore, for each gene, we calculated two VUS estimates, one assuming that lower values are associated with more severe states, and the other assuming that the relationship is in the opposite direction. For the latter, we used two minus the observed value of gene expression as values of the covariates to estimate VUS, where two is the maximum value of all covariates. We kept the VUS estimate that was further away from 1/6.

Following Fan and Lv (2008), we selected $[n/\log(n)] = 51$ important variables for VUS and Kendall's τ. For the PSIS method, 26 variables were selected for each type of event, and the final important set contained the union of important variables of two events.

We show the top 51 genes selected by VUS in Table 4. Genes that are selected by both VUS and Kendall's τ are denoted by '*', and those selected by both VUS and PSIS are denoted by '**'. Bold-faced genes are selected by all three methods, which include seven

Table 4: Top 51 genes selected by VUS

U96131**	NM_005480*	Contig29555_RC
Contig31288_RC	M96577	NM_003494
NM_003295	NM_000987	NM_004219
NM_005733	NM_004701*	Contig57173_RC
NM_001605**	NM_019059	NM_004217
NM_002466*	**Contig57584_RC**	NM_007019*
NM_006607	NM_006579*	NM_001809
NM_006845	Contig6498	D38553*
Contig38288_RC	NM_005804	NM_006623**
D43950**	Contig35629_RC	NM_018188*
NM_001673	**D14678**	Contig41828_RC
NM_018410*	NM_002624	NM_001255
NM_020313	NM_014454	NM_018688
Contig56390_RC*	Contig38901_RC	NM_003504
NM_001333	NM_007195	NM_019597**
NM_017761*	NM_018834	NM_003600*
Contig41977_RC	Contig43747_RC	NM_001168*

Bold-faced genes are selected by all three metrics; '*' are selected by VUS and Kendall's τ; '**' are selected by VUS and PSIS.

genes. The same dataset was also analyzed in Song et al. (2014) and Lu et al. (2020). In Song et al. (2014) the event of interest was the overall survival time, and only the top 20 selected genes were shown. We compared the results and found that 13 genes were both selected by our VUS method and by Song et al. (2014). Lu et al. (2020) treated the data as semi-competing risks outcomes. While explicitly handling semi-competing risks is beyond the scope of this work, the proposed VUS is still applicable by counting the most severe event that occurred. The authors selected 51 genes using the proposed method in Lu et al. (2020) and improved the results based on an adaptive threshold rule, for which 25 genes were selected. They showed that the adaptive threshold rule would perform better than the proposed method itself, so we compared the selected genes with their 25 selected genes. Among the 25 selected genes, 13 genes were selected by both our VUS method and Lu et al. (2020), and these 13 genes were partially different from those selected by VUS and Kendall's τ in Song et al. (2014). These comparisons imply that there is no one-size-fits-all metric, and our VUS method is a viable variable screening metric for competing risk outcomes as it focuses on a different aspect of the association between the covariates and the outcomes than existing methods.

2.7 Discussion

In this project we have shown that VUS possesses the sure screening property. The VUS can provide an overall assessment of diagnostic accuracy of covariates in predicting ordinal outcomes and has a straightforward interpretation as the concordance probability between the value of covariates and the disease status, and simulation studies and data analysis have shown that it can serve as an alternative for variable screening, especially with data contamination in covariates.

A limitation of VUS is that it is designed to pick covariates with an ordinal relationship with the outcomes. One possible solution is that instead of measuring the concordance between the value of a biomarker itself and the disease status, we may look at the concordance between some functions based on each single biomarker and the true event status, for exam-

ple, the estimated CIF based on each single Z. Similarly, we may evaluate how a group of biomarkers can be associated with the competing risks outcomes by modeling CIF based on this group of biomarkers to handle correlated biomarkers and categorical biomarkers. However, this solution will rely on additional model assumptions on CIF. Besides, when modeling CIF, we typically assume that the transformed CIF is a linear function of the covariates, which indicates that a monotone relationship exists between the biomarker and the CIF, and that is similar to the assumption we've made for VUS. We may consider including higher order terms in CIF estimation, but further investigation is beyond the scope of this work.

We surprisingly found that the PSIS method, which utilizes the Wald-type statistic from Cox's proportional hazard model, is quite robust to model misspecification. In all three scenarios we considered, the PSIS method performed pretty well when there's no covariate contamination. There is one limitation with the PSIS method though. When applying it to competing risks settings, we don't know how many variables should be selected for cause-1 event and cause-2 event respectively. What we have observed in our simulation is that when we select equal number of variables for two types of events, variables selected for cause 1 event contain most important variables, yet for cause 2 event, its performance is deficient. However, we often have no prior knowledge when we are analysing real data and thus cannot adjust the number of important variables selected for each type of event in advance to further improve the performance of the PSIS method.

Besides, in both simulation studies and real data analysis, we decided how many variables k to be selected in an empirical way and then kept those variables whose absolute distances from 1/6 are among the top k. Variables can also be selected by picking a threshold via controlling the false positive rate, which was done by Zhao and Li (2012). Zhang et al. (2021) showed the weak convergence of VUS estimate and provided an estimate of standard error. Thus, Wald statistics can be computed for each VUS estimate based on some transformation (e.g., arcsine square root transformation) to improve practial performance. Following Zhao and Li (2012), we could fix a false positive rate, and the threshold of p-value would be fixed accordingly. After an inverse transformation, we would get the threshold γ_n for the distance between VUS estimate and the null value 1/6.

3.0 Model Evaluation with Competing Risks Data

3.1 Introduction

In biomedical studies, treatments can be effective for one type of disease, but not for the other. Also, high-risk individuals may get more benefits from the treatment than those low-risk ones. Therefore, when analyzing competing risks data, one important goal is to predict the disease status of a patient at certain time points based on baseline measurements. The classification of subjects may depend on values of biomarkers, or the absolute risk of certain events predicted from a prognostic model.

Cumulative incidence function (CIF), or the absolute risk, is used to describe the probability of the occurrence of an event in existence of competing events, and is defined as

$$F_l(t|\mathbf{Z}) = P(T \leq t, \epsilon = l|\mathbf{Z})$$

for cause l event at time t given covariates $\mathbf{Z}$. It is not a proper distribution, as $F_l(\infty|\mathbf{Z}) = P(\epsilon = l|\mathbf{Z}) < 1$. When there are L ($L \geq 2$) competing events, we have $\sum_{l=1}^{L} F_l(t|\mathbf{Z}) + S(t|\mathbf{Z}) = 1$, where $S(t|\mathbf{Z})$ is the survival function of event time T.

Various models have been developed to analyze competing risks data, and many of them provide estimates of CIF and survival functions. The most common model for analyzing competing risks data is the proportional hazards model on cause-specific hazard function (Kalbfleisch and Prentice, 2011), which is defined as

$$\lambda_l(t|\mathbf{Z}) = \lim_{h \to 0} \frac{P(t \leq T < t + h, \epsilon = l|T \geq t, \mathbf{Z})}{h}.$$

It is the instantaneous rate of occurrence of event l among event-free subjects given covariates $\mathbf{Z}$. To calculate CIF for event l, cause-specific hazards for all events need to be modeled, and

$$F_l(t|\mathbf{Z}) = \int_0^t \lambda_l(s|\mathbf{Z}) S(s|\mathbf{Z}) ds,$$

where $S(t|\mathbf{Z}) = \exp\left(-\sum_{l=1}^{L}\int_0^t \lambda_l(s|\mathbf{Z})ds\right)$. Another model that is frequently considered is the Fine-Gray subdistribution hazard (Fine and Gray, 1999), where the subdistributional hazard is defined as

$$\lambda_l(t|\mathbf{Z}) = \lim_{h\to 0}\frac{P(t \le T < t+h, \epsilon = l | T \ge t \cup (T \le t \cap \epsilon \ne l, \mathbf{Z})}{h}$$

and $F_l(t|\mathbf{Z}) = \exp(-\int_0^t \lambda_l(s|\mathbf{Z})ds)$. A proportional hazard model of the form $\lambda_l(t|\mathbf{Z}) = \lambda_{l0}(t)\exp(\mathbf{Z}^T\boldsymbol{\beta}_l)$ was adopted for modeling the subdistribution hazard (Fine and Gray, 1999). Different from the cause-specific hazard, now the interpretation of covariate effects on CIF is straightforward. Scheike et al. (2008) extended the model proposed by Fine and Gray (1999) via binomial regression and allowed for time-varying coefficients. In the Fine-Gray and Scheike's models, it is not guaranteed that $\sum_{l=1}^{L}\hat{F}_l(t|\mathbf{Z}) \le 1$, where $\hat{F}_l$ are the estimated CIFs. To fix the issue, Gerds et al. (2012) proposed to use the multinomial logistic regression to model the probabilities of the occurrence of competing risks and survival as the following:

$$F_l(t|\mathbf{Z}) = \frac{\exp(A_l(t) + \mathbf{Z}^T\boldsymbol{\beta}_l)}{1 + \sum_{k=1}^{L}\exp(A_k(t) + \mathbf{Z}^T\boldsymbol{\beta}_k)}$$

and

$$S(t|\mathbf{Z}) = \frac{1}{1 + \sum_{k=1}^{L}\exp(A_k(t) + \mathbf{Z}^T\boldsymbol{\beta}_k)}.$$

There are other models that can provide CIF estimates such as parametric models proposed by Jeong and Fine (2006), Shi et al. (2013) and Haile et al. (2016), among many others. With the availability of different models, one natural problem to consider is that given the observed data, which model can provide better prognostic accuracy, in other words, which model can better predict the probability of occurrence of certain events and discriminate subjects from different causes of event group.

There are metrics developed to assess the discrimination power of a prognostic model with survival outcomes. Harrell et al. (1982) introduced the C-index that measures the probability that the patient with a higher prognostic score would live longer. It is a commonly-used concordance index to evaluate the goodness-of-fit of the models that estimate prognostic scores. As mentioned in Chapter 1, Heagerty and Zheng (2005) used AUC to evaluate the accuracy of a regression model. Besides, Gerds and Schumacher (2006) provided an

estimation of mean squared prediction error based on IPCW, which is a weighted average of squared distance between the survivor indicator and the estimated survival function, an extension of Brier score (Brier, 1950).

Extensions have also been made to competing risks outcomes. Schoop et al. (2011) extended the Brier score based on IPCW to competing risks data, which estimates the prediction error of prognostic model for the event of interest. Blanche et al. (2013) extended AUC to competing risks model evaluation focusing on the specific event of interest and used the IPCW method to deal with censoring. Two definitions of specificity were provided depending on how the control group was defined. Gerds et al. (2014) proposed the calibration plot of predicted risk versus expected risk. Wolbers et al. (2014) proposed a concordance probability for cause-specific model evaluation and showed how it was related to AUC. Wu and Li (2018) also extended AUC and Brier scores, with a different weighting scheme to deal with censoring. The weight was given by the conditional probability of the event of interest given observed data instead of the probability of censoring.

All above methods focus on evaluating diagnostic accuracy of models for predicting one particular type of event, given the existence of competing risks. They cannot provide an overall assessment of diagnostic accuracy of a group of models for predicting all competing events. To the best of our knowledge, only one work provides an overall evaluation of competing risks prediction models. Ding et al. (2021) extended the polytomous discrimination index (PDI) (Van Calster et al., 2012) to competing risks data. The PDI of a specific cause j measures the probability that, of $L+1$ randomly selected subjects from $L+1$ distinct groups, the subject from the j-th group has the largest estimate of cause-j CIF compared to subjects from other groups. The overall evaluation of diagnostic accuracy is the average probability over all causes, including the survivor, and a larger value of PDI indicates better discrimination power of the model.

PDI is the first metric that provides both overall and cause-specific evaluations of diagnostic accuracy, therefore Ding et al. (2021) only investigated the statistical properties of PDI itself, and a comparison between PDI and other existing metrics for cause-specific evaluation is missing. However, PDI may cause confusion in what is considered a correct classification. Consider two subjects i and j with two competing events. Subject i experi-

enced cause-1 event with estimated CIFs and survival functions being $\hat{F}_{1i} = 0.35$, $\hat{F}_{2i} = 0.33$, and $\hat{S}_i = 0.32$ at the time of assessment. Subject j experienced cause-2 event with estimated functions being $\hat{F}_{1j} = 0.4$, $\hat{F}_{2j} = 0.5$, and $\hat{S}_j = 0.1$. Intuitively, given the estimated functions, we would assign subject i into cause-1 group and subject j into cause-2 group given the estimated probabilities, while based on the definition of PDI, although subject i belongs to cause-1 group, the estimated CIF $\hat{F}_{1i}$ is smaller than $\hat{F}_{1j}$ of subject j, and assigning subject i to cause-1 would not be counted as a correct classification in a triplet of subjects containing subjects i and j. Thus, the naive way of assigning the subject to the class with the largest estimated CIF or survival as compared to other classes may not be considered a correct classification by PDI.

In this project, we aim to provide a systematic evaluation of the performance of commonly used model evaluation metrics including AUC, Brier score (Schoop et al., 2011) and PDI (Ding et al., 2021) under the competing risks setting, and provide a practical guide of which metrics to use under different circumstances. For AUC, the more recent estimator proposed by Wu and Li (2018) was shown to perform similarly as the IPCW estimator proposed by Blanche et al. (2013) with independent censoring, and the latter was fully investigated and has a complete testing procedure. Therefore we used the IPCW based AUC estimator (Blanche et al., 2013) in this project. We would also like to test ideas of natural extensions of concordance index based on the naive way of classification and the idea of PDI. Details of the extension are provided in Section 3.2, and we call the extensions Extended Concordance Indices.

3.2 Methods

3.2.1 Existing Methods

We provide a brief review of AUC, Brier score and PDI methods in this subsection.

AUC (Blanche et al., 2013) provides a cause-specific evaluation of prognostic models with competing risks data. It measures the probability that the subject that experienced the event

of interest has a higher risk score than the subject from the control group. Two definitions of the control group are provided. The standard control group contains subjects who never experienced any event, and the augmented control group includes subjects who were event-free, or experienced other events by the time of evaluation. With different definitions of control groups, AUCs can be defined as:

$$AUC_l(t_0) = P(F_{li} > F_{lk} | T_i \leq t_0, \epsilon_i = l, T_k > t_0),$$

$$AUCA_l(t_0) = P(F_{li} > F_{lk} | T_i \leq t_0, \epsilon_i = l, \{T_k > t_0\} \cup \{T_k \leq t_0, \epsilon_k \neq l\}),$$

where F_{li} is the risk score of subject i for cause-l event. Here AUC uses the standard control group and AUCA is defined based on the augmented control group with A standing for augmented. Blanche et al. (2013) proposed the IPCW estimator of AUC and AUCA, and showed their large sample properties.

The adaption of Brier score to competing risks data was proposed by Schoop et al. (2011), which is a modification of Brier score proposed by Gerds et al. (2014) for standard survival outcomes. For each cause of event, the prediction error is defined as the mean squared distance between the true event status and the CIF:

$$\text{PE}_l(t_0) = E[I(T \leq t_0, \epsilon = l) - F_l(t_0|\mathbf{Z})]^2.$$

It measures both discrimination and calibration accuracy of the prognostic model. An estimator based on IPCW was introduced and shown to be uniformly consistent. Different from other metrics, smaller values of Brier score is preferred since it estimates the prediction error. We denote Brier score as BS in all tables.

PDI proposed by Ding et al. (2021) provides both cause-specific and overall evaluations of the discrimination power of prognostic models for competing risks data. Suppose there are L competing events. Subjects $i_1, \cdots, i_L, i_{L+1}$ are randomly selected from distinct event group $1, \cdots, L, L+1$, such that subject i_l experienced event l or no events for $L+1$, and we denote event $\mathcal{C}$ as $\{T_{i_1} \leq t_0, \epsilon_{i_1} = 1, \cdots, T_{i_L} \leq t_0, \epsilon_{i_L} = L, T_{i_{L+1}} > t_0\}$. Then for cause-$l$,

PDI_l is the probability that subject i_l has the largest cause-l CIF estimation among subjects $i_1, \cdots, i_L, i_{L+1}$:

$$\text{PDI}_l(t_0) = E\left[\prod_{j=1, j\neq l}^{j=L+1} I\{F_{li_l}(t_0|\mathbf{Z}_{i_l}) \geq F_{li_j}(t_0|\mathbf{Z}_{i_j})\}|\mathcal{C}\right].$$

When $L = 1$ and there's no competing events, PDI_l reduce to AUC with the standard control group. The overall PDI is the average of cause-specific PDIs over all event groups, including the survivor group. PDI can be estimated using the IPCW method. Ding et al. (2021) showed the consistency and asymptotic normality of the estimator.

3.2.2 Extended Concordance Indices

In this subsection we consider natural extensions of concordance index for evaluating discrimination ability. Ideally, we would like the predicted event status to agree with the true event status for all subjects, and the extent of agreement indicates the discrimination ability of the prognostic model. Therefore, a natural measure for evaluation would be the probability that subjects are correctly classified by the prognostic model. The two indices under consideration differ in how a successful classification is defined. The construction of these evaluation metrics were inspired by the concordance probability and VUS in Chapter 2. The first index comes directly from the idea of PDI, but instead of taking the average over all causes, we consider the probability that subjects are correctly assigned simultaneously.

At a specific time point of interest t_0, we consider three randomly selected subjects i, j, k such that $T_i \leq t_0, \epsilon_i = 1$, $T_j \leq t_0, \epsilon_j = 2$ and $T_k > t_0$. We are interested in the probability that all three subjects are correctly sorted simultaneously such that subject from cause-j has the largest cause-j CIF compared to the other two subjects. We define the first extended concordance index (ExC*) as

$$\begin{aligned}\text{ExC}^*(t_0) = P(\max \mathbf{F}_1 = F_{1i}(t_0), \max \mathbf{F}_2 = F_{2j}(t_0), \max \mathbf{S} = \\ S_k(t_0)|T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0),\end{aligned} \tag{3.2.1}$$

where $\mathbf{F}_l(t_0) = (F_{li}(t_0|\mathbf{Z}_i), F_{lj}(t_0|\mathbf{Z}_j), F_{lk}(t_0|\mathbf{Z}_k)), l = 1, 2$, $F_{li}(t_0), l = 1, 2$, is the cause-l CIF for subject i given covariates $\mathbf{Z}_i$, and $\mathbf{S} = (S_i(t_0|\mathbf{Z}_i), S_j(t_0|\mathbf{Z}_j), S_k(t_0|\mathbf{Z}_k))$. We skip $\mathbf{Z}$ in

(3.2.1) for simplicity. Supposedly, ExC* would provide an overall evaluation of the ability of prognostic models to discriminate subjects from all causes, just like PDI.

For cause-specific evaluation, we consider randomly picking a subject i from the cause of interest and a subject k from the control group, and again we measure the probability that these two subjects are correctly assigned simultaneously. Based on different definitions of the control group as shown in Blanche et al. (2013), we come up with two related metrics:

$$\text{ExC}^*_l(t_0) = P(F_{li} > F_{lk}, S_k > S_i)|T_i \leq t_0, \epsilon_i = l, T_k > t_0),$$

and

$$\text{ExC}^*_{l_2}(t_0) = P(F_{li} > F_{lk}|T_i \leq t_0, \epsilon_i = l, \{T_k > t_0\} \cup \{T_k \leq t_0, \epsilon_k \neq l\}),$$

for cause-l event. Note that $\text{ExC}^*_{l_2}(t_0)$ is exactly the augmented AUC in Blanche et al. (2013), so we will just use the notation AUCA_l instead, and l denotes the cause of event.

Now we introduce the second extended concordance index, which we denoted as ExC. For three randomly selected subjects i, j, k such that $T_i \leq t_0, \epsilon_i = 1$, $T_j \leq t_0, \epsilon_j = 2$ and $T_k > t_0$, this time we assign a subject to class l if for this subject the probability of belonging to class l is the largest compared to the probabilities of belonging to other classes. Then ExC is the probability that three subjects from different groups are all correctly classified based on our rule of assignment, and is defined as:

$$\begin{aligned}\text{ExC}(t_0) = P(&\max \mathbf{p}_i = F_{1i}(t_0), \max \mathbf{p}_j = F_{2j}(t_0), \max \mathbf{p}_k = \\ &S_k(t_0)|T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0),\end{aligned} \tag{3.2.2}$$

where

$$\mathbf{p} = p(t_0|\mathbf{Z}) = (F_1(t_0|\mathbf{Z}), F_2(t_0|\mathbf{Z}), S(t_0|\mathbf{Z}))^T.$$

We will simply refer to it as $\mathbf{p}$ when there is no confusion. Cause-specific indices can be defined similarly:

$$\text{ExC}_l(t_0) = P(\max \mathbf{p}_i = F_{li}(t_0), \max \mathbf{p}_k = S_k(t_0)|T_i \leq t_0, \epsilon_i = l, T_k > t_0),$$

Table 5: Notation for Evaluation Metrics

	Cause-specific evaluations
AUC_l	subject i_l has a larger cause-l CIF than subject i_{L+1}
AUCA_l	subject i_l has a larger cause-l CIF than a subject from the augmented control
BS_l	mean squared distance between true event status and cause-l CIF
PDI_l	subject i_l has the largest cause-l CIF compared to all other subjects
ExC_l^*	subject i_l has a larger cause-l CIF and subject i_{L+1} has a larger survival
ExC_l	$\max \mathbf{p_{i_l}} = F_{i_l}$, $\max \mathbf{p_{i_{L+1}}} = S_{i_{L+1}}$
	Overall evaluations
PDI	average of PDI_ls, $l = 1, \ldots, L$
ExC^*	subject i_l has the largest cause-l CIF or survival ($l = L+1$) among all selected subjects
ExC	subject i_l' cause-l CIF or survival ($l = L+1$) is the largest among all probabilities

and

$$\text{ExCA}_l(t_0) = P(\max \mathbf{p}_i = F_{li}(t_0), \max \mathbf{p}_k \neq F_{lk}(t_0)| \\ T_i \leq t_0, \epsilon_i = l, \{T_k > t_0\} \cup \{T_k \leq t_0, \epsilon_k \neq l\}).$$

We summarize each aforementioned metric in Table 5 for easy reference, where we assume subjects $i_1, \cdots, i_L, i_{L+1}$ are randomly selected from $L+1$ distinct groups $1, \cdots, L, L+1$ with group $L+1$ being survivors, as we described before.

Although the idea behind ExC seems natural, we have found conditions where this idea doesn't work as expected. Consider two competing events, where cause-1 CIF is larger than cause-2 CIF all the time for each subject. If the prognostic model is well fitted for each cause of event, then the estimated CIF will be very close to the true CIF, and we will see that the estimated cause-1 CIF is always larger than the estimated cause-2 CIF for each subject, even for a subject who actually experienced a cause-2 event. In this case, if we assign each subject to the group that has the largest estimated CIF compared to other groups, then

subjects who had cause-2 event are much likely to be assigned to the cause-1 group, causing ExC estimates to be zero or extremely small, as illustrated in simulation studies.

3.2.3 Estimation of ExC's and Inference

The estimator of ExC* can be modified from Zhang et al. (2021) to accommodate independent censoring:

$$\widehat{\mathrm{ExC}^*}(t_0) = \frac{\sum_{i\neq j\neq k} \frac{I(X_i\leq t_0,\eta_i=1,X_j\leq t_0,\eta_j=2,X_k>t_0,\max\hat{\mathbf{F}}_1=\hat{F}_{1i}(t_0),\max\hat{\mathbf{F}}_2=\hat{F}_{2j}(t_0),\max\hat{\mathbf{S}}=\hat{S}_k(t_0))}{\hat{G}(X_i-)\hat{G}(X_j-)\hat{G}(t_0)}}{\sum_{i\neq j\neq k} \frac{I(X_i\leq t_0,\eta_i=1,X_j\leq t_0,\eta_j=2,X_k>t_0)}{\hat{G}(X_i-)\hat{G}(X_j-)\hat{G}(t_0)}},$$

where $\hat{F}_1(t_0|\mathbf{Z}), \hat{F}_2(t_0|\mathbf{Z}), \hat{S}(t_0|\mathbf{Z})$ can be estimated from any competing risks model under evaluation and are allowed to depend on different sets of covariates. $\hat{G}$ is the estimated survival function of censoring. Unlike the screening metric in Chapter 2, where censoring time is required to be independent of T and covariates $\mathbf{Z}$, and $\hat{G}$ is estimated using the Kaplan-Meier estimator, in this project, censoring time C is allowed to be independent of T given covariates $\mathbf{Z}$, and $\hat{G}$ may be provided by a regression model such as Cox's proportional hazards model. Weak convergence of $\widehat{\mathrm{ExC}^*}(t_0)$ to ExC^* can be shown similarly to Ding et al. (2021), and the asymptotic normality of $\widehat{\mathrm{ExC}^*}(t_0)$ is established in Appendix B.

The cause-specific metric $\mathrm{ExC}^*_l(t_0)$ can be estimated straightforwardly by

$$\widetilde{\mathrm{ExC}^*}_l(t_0) = \frac{\sum_{i\neq k} \frac{I(X_i\leq t_0,\eta_i=l,X_k>t_0)I(\hat{F}_{li}>\hat{F}_{lk},\hat{S}_k>\hat{S}_i)}{\hat{G}(X_i-)\hat{G}(t_0)}}{\sum_{i\neq k} \frac{I(X_i\leq t_0,\eta_i=l,X_k>t_0)}{\hat{G}(X_i-)\hat{G}(t_0)}}.$$

The estimator of ExC and its cause-specific relatives are similarly defined as ExC* and are thus not explicitly given here.

Variance estimation relies on the particular prognostic models that provide the CIF estimates, and varies across different models. Therefore, an explicit form of the variance is unavailable, and the bootstrap method may be used to get an estimate of the variance.

In practice, when we have categorical covariates, we may have two subjects from two distinct groups have same CIF estimates for some causes, which results in ties in CIF estimates. Modifications can further be made in the same way as in Ding et al. (2021) to handle

ties in CIF estimates. Take ExC* for example, if there are ties in cause-l CIF estimates for $l = 1, 2$ or the survival function, then $I(\max \hat{\mathbf{F}}_l = \hat{F}_{li_l}(t_0))$ can be substituted by

$$\frac{I(\max \hat{\mathbf{F}}_l = \hat{F}_{li_l}(t_0))}{1 + \sum_{j=1, j\neq l}^{L+1} \{\hat{F}_{li_l}(t_0) = \hat{F}_{li_j}(t_0)\}},$$

where $\hat{F}_{li_j}$ is cause-l CIF estimate for subject i_j.

3.2.4 Model Comparison

In practice we often encounter problems of model comparison. For example, we would like to investigate whether the inclusion of a covariate would significantly improve the discrimination ability of the prognostic model. Therefore, it is desirable for a model evaluation metric to have a testing procedure for whether two models have a significant difference in diagnostic accuracy. That is, for two different prognostic models m_1 and m_2 and for any model evaluation metric R, we want to test $H_0 : R(m_1) - R(m_2) = 0$.

Due to the difficulty in deriving the test statistic, especially of the variance, we can use the Bootstrap method to build a confidence interval for $R(m_1) - R(m_2)$ and see whether the confidence interval covers zero. More specifically, we use the bias-corrected and accelerated (BCa) bootstrap procedure (Efron, 1987), which was implemented in function "bcanon" in R package "bootstrap".

3.3 Simulation

We compare the performance of model evaluation metrics under various settings. Metrics examined include AUC (Blanche et al., 2013), Brier score (Schoop et al., 2011), PDI (Ding et al., 2021) and the two sets of overall and cause-specific metrics we proposed in Section 3.2.

We generated data from three different models: Cox's cause-specific proportional hazard (Kalbfleisch and Prentice, 2011), Gerds' multinomial logistic regression (Gerds et al., 2012) and Fine-Gray subdistribution hazard model (Fine and Gray, 1999). For Cox's and Gerds'

models, we consider two scenarios: cause-1 and cause-2 CIFs are equal everywhere, and cause-1 and cause-2 CIFs are different. For the Fine-Gray model, we mainly examine the scenario where cause-1 and cause-2 are different. Under each scenario, we consider two model evaluation problems that generally arise in practice: which model to fit among Cox's, Fine-Gray, and Gerds' models, and which combination of covariates can provide better prediction. For both problems, we would like to see whether the model evaluation metrics are able to select the true model under which the data have been generated. For each problem under different scenarios and models, we generated data with sample size $n = 500$. Five-folds cross-validation was implemented, with training sets used for model fitting and test sets for CIF prediction and model evaluation. We considered 30% and 15% overall rates of censoring, and conducted 1000 simulations in each case. The proportions of each model selected and the average of estimated values over 1000 simulations are summarized for each metric.

Setting 1: Event times were generated from Cox's cause-specific proportional hazard model. Cause-1 and cause-2 hazards are given by

$$\lambda_1(t|\mathbf{Z}) = t^{1/2}\exp(\boldsymbol{\beta}_1^T\mathbf{Z}), \text{ and } \lambda_2(t|\mathbf{Z}) = t^{1/2}\exp(\boldsymbol{\beta}_2^T\mathbf{Z}).$$

When comparing Cox's, Gerds' and Fine-Gray models, we generated $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4)$, with $Z_1, Z_2 \overset{iid}{\sim} N(0,1), Z_3 \sim \text{Binom}(1, 0.6), Z_4 \sim \text{Binom}(1, 0.3)$. Under the scenario where cause-1 and cause-2 CIFs are equal, we set $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (0.5, 0.5, 0.5, 0.5)$, with censoring time being generated from $U(0, 2)$ for 30% censoring and from $U(0, 4.2)$ for 15% censoring. Models were evaluated at $t_0 = 0.8$. Under the scenario where cause-1 and cause-2 CIFs are different, we used $\boldsymbol{\beta}_1 = (0.3, 0.3, 0.3, 0.3)$ and $\boldsymbol{\beta}_2 = (0.5, 0.5, 0.5, 0.5)$, $U(0, 2.1)$ for 30% censoring, and $U(0, 4.3)$ for 15% censoring, and predicted time at $t_0 = 0.8$. We fitted Cox's, Gerds' and the Fine-Gray models on training datasets with cause-1 and cause-2 events both relying on Z_1 to Z_4. We examine whether these model evaluation methods can successfully pick the true underlying Cox model.

When comparing models depending on different combinations of covariates, we generated $(Z_1, Z_4) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.04\\0.04 & 1\end{pmatrix}\right)$, $Z_2, Z_3 \overset{iid}{\sim} N(0,1)$, and $\mathbf{Z} = (Z_1, Z_2, Z_3)$. Z_4 is only used in model fitting, and the true model doesn't rely on Z_4. True coefficients were set as

$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (0.5, 0.5, 0.5)$ for equal cause-1 and cause-2 CIFs, with censoring time generated from $U(0, 2.8)$ for 30% censoring and from $U(0, 5.9)$ for 15% censoring. The prediction time was set to be $t_0 = 1.1$. For different cause-1 and cause-2 CIFs, we used $\boldsymbol{\beta}_1 = (0.3, 0.3, 0.3)$ and $\boldsymbol{\beta}_2 = (0.5, 0.5, 0.5)$, with $U(0, 2.6)$ for 30% censoring, $U(0, 5.4)$ for 15% censoring and $t_0 = 1$. We fitted two sets of Cox's cause-specific model; one model, denoted as "Cox", was fitted with both cause-1 and cause-2 hazards depending on (Z_1, Z_2, Z_3), and the other model with hazards of both causes depending on (Z_2, Z_3, Z_4) is denoted as "CoxR" with "R" standing for the reduced model. We examine the probability that the model depending on (Z_1, Z_2, Z_3) can be selected. Results on proportions of models being selected and mean values of each evaluation metric are summarized in Table 6 and Table 7, respectively.

Table 6: Proportion of Model Selected under Cox's Model

Metrics	Different Models												Combination of Covariates							
	Same Hazard						Diff Hazard						Same Hazard				Diff Hazard			
	30%			15%			30%			15%			30%		15%		30%		15%	
	Cox	FG	Gerds	Cox	FG	Gerds	Cox	FG	Gerds	Cox	FG	Gerds	Cox	CoxR	Cox	CoxR	Cox	CoxR	Cox	CoxR
ExC	0.401	0.215	0.384	0.458	0.130	0.412	0.427	0.088	0.485	0.361	0.199	0.440	0.744	0.256	0.742	0.258	0.768	0.232	0.789	0.211
ExC^*	0.145	0.002	0.853	0.166	0.483	0.351	0.106	0.004	0.890	0.121	0.111	0.768	0.738	0.262	0.744	0.256	0.823	0.177	0.826	0.174
PDI	0.420	0.468	0.112	0.637	0.074	0.289	0.643	0.134	0.224	0.692	0.006	0.302	0.994	0.006	0.996	0.004	0.982	0.018	0.991	0.009
ExC_1	0.519	0.267	0.214	0.610	0.036	0.354	0.476	0.068	0.456	0.403	0.143	0.454	0.839	0.161	0.851	0.149	0.724	0.276	0.751	0.249
$ExCA_1$	0.414	0.268	0.318	0.417	0.252	0.331	0.341	0.113	0.546	0.190	0.503	0.307	0.640	0.360	0.656	0.344	0.576	0.424	0.593	0.407
ExC_1^*	0.904	0.045	0.051	0.879	0.000	0.121	0.976	0.005	0.019	0.915	0.000	0.085	0.986	0.014	0.994	0.006	0.853	0.147	0.887	0.113
$AUCA_1$	0.544	0.347	0.108	0.697	0.053	0.250	0.657	0.100	0.243	0.666	0.023	0.311	0.900	0.100	0.922	0.078	0.722	0.278	0.740	0.260
PDI_1	0.380	0.487	0.133	0.576	0.155	0.269	0.617	0.133	0.250	0.648	0.007	0.345	0.771	0.229	0.786	0.214	0.644	0.356	0.650	0.350
AUC_1	0.686	0.209	0.105	0.798	0.015	0.187	0.929	0.028	0.043	0.887	0.000	0.113	0.992	0.008	0.996	0.004	0.814	0.186	0.847	0.153
BS_1	0.465	0.171	0.365	0.490	0.044	0.465	0.400	0.204	0.396	0.439	0.040	0.521	0.913	0.087	0.924	0.076	0.717	0.283	0.735	0.265
ExC_2	0.283	0.414	0.303	0.414	0.084	0.502	0.480	0.400	0.120	0.640	0.033	0.327	0.784	0.216	0.786	0.214	0.940	0.060	0.957	0.043
$ExCA_2$	0.255	0.352	0.394	0.255	0.333	0.412	0.525	0.245	0.230	0.589	0.060	0.350	0.603	0.397	0.609	0.391	0.891	0.109	0.912	0.088
ExC_2^*	0.868	0.053	0.080	0.817	0.001	0.182	0.809	0.056	0.136	0.823	0.007	0.169	0.992	0.008	0.996	0.004	0.993	0.007	0.997	0.003
$AUCA_2$	0.516	0.341	0.142	0.631	0.054	0.315	0.545	0.273	0.182	0.668	0.097	0.235	0.913	0.087	0.935	0.065	0.981	0.019	0.992	0.008
PDI_2	0.373	0.491	0.136	0.533	0.162	0.305	0.481	0.344	0.175	0.620	0.135	0.244	0.799	0.201	0.818	0.182	0.963	0.037	0.971	0.029
AUC_2	0.629	0.245	0.126	0.714	0.014	0.272	0.558	0.254	0.188	0.654	0.101	0.245	0.993	0.007	0.995	0.005	0.995	0.005	1.000	0.000
BS_2	0.475	0.136	0.389	0.477	0.047	0.477	0.477	0.212	0.312	0.554	0.081	0.364	0.932	0.068	0.947	0.053	0.983	0.017	0.988	0.012

Table 7: Averaged Estimated Values under Cox's Model

Metrics	Different Models												Combination of Covariates							
	Same Hazard						Diff Hazard						Same Hazard				Diff Hazard			
	30%			15%			30%			15%			30%		15%		30%		15%	
	Cox	FG	Gerds	Cox	FG	Gerds	Cox	FG	Gerds	Cox	FG	Gerds	Cox	CoxR	Cox	CoxR	Cox	CoxR	Cox	CoxR
ExC	0.079	0.074	0.078	0.077	0.067	0.077	0.039	0.032	0.041	0.036	0.030	0.037	0.075	0.061	0.074	0.061	0.065	0.051	0.064	0.050
ExC^*	0.130	0.102	0.141	0.123	0.129	0.128	0.157	0.125	0.170	0.156	0.117	0.167	0.122	0.102	0.114	0.096	0.155	0.126	0.152	0.124
PDI	0.527	0.527	0.525	0.529	0.521	0.527	0.501	0.491	0.498	0.503	0.477	0.501	0.543	0.499	0.544	0.500	0.516	0.475	0.517	0.477
ExC_1	0.254	0.241	0.236	0.253	0.197	0.241	0.062	0.048	0.062	0.055	0.041	0.056	0.273	0.222	0.275	0.224	0.119	0.101	0.116	0.098
$ExCA_1$	0.256	0.246	0.251	0.252	0.241	0.247	0.090	0.078	0.098	0.081	0.096	0.086	0.253	0.236	0.252	0.235	0.146	0.140	0.143	0.137
ExC_1^*	0.751	0.731	0.737	0.758	0.694	0.750	0.563	0.439	0.519	0.572	0.274	0.540	0.787	0.709	0.791	0.713	0.612	0.546	0.620	0.554
$AUCA_1$	0.622	0.619	0.618	0.624	0.608	0.622	0.539	0.511	0.533	0.541	0.486	0.537	0.634	0.605	0.635	0.606	0.552	0.535	0.554	0.537
PDI_1	0.450	0.451	0.446	0.453	0.441	0.450	0.370	0.342	0.364	0.373	0.303	0.368	0.460	0.437	0.461	0.438	0.376	0.363	0.378	0.365
AUC_1	0.779	0.771	0.770	0.783	0.744	0.778	0.628	0.550	0.596	0.634	0.431	0.612	0.807	0.740	0.809	0.743	0.660	0.618	0.665	0.623
BS_1	0.225	0.227	0.225	0.224	0.227	0.224	0.218	0.219	0.218	0.217	0.220	0.217	0.223	0.228	0.222	0.227	0.225	0.226	0.224	0.225
ExC_2	0.242	0.247	0.238	0.248	0.208	0.251	0.409	0.402	0.385	0.419	0.317	0.407	0.254	0.213	0.252	0.213	0.385	0.321	0.391	0.328
$ExCA_2$	0.249	0.250	0.253	0.248	0.247	0.253	0.367	0.359	0.358	0.368	0.335	0.364	0.240	0.229	0.237	0.228	0.380	0.339	0.383	0.342
ExC_2^*	0.752	0.733	0.740	0.757	0.698	0.751	0.758	0.748	0.752	0.762	0.742	0.758	0.786	0.707	0.790	0.712	0.785	0.716	0.788	0.721
$AUCA_2$	0.623	0.620	0.619	0.624	0.608	0.622	0.677	0.675	0.675	0.679	0.674	0.677	0.634	0.605	0.635	0.606	0.690	0.651	0.691	0.652
PDI_2	0.451	0.452	0.448	0.452	0.441	0.450	0.516	0.514	0.514	0.518	0.513	0.516	0.460	0.437	0.461	0.438	0.533	0.491	0.533	0.492
AUC_2	0.779	0.772	0.771	0.783	0.746	0.779	0.783	0.780	0.780	0.785	0.777	0.783	0.806	0.739	0.808	0.742	0.805	0.743	0.805	0.745
BS_2	0.225	0.227	0.225	0.224	0.227	0.224	0.222	0.223	0.223	0.221	0.223	0.222	0.223	0.228	0.222	0.227	0.213	0.221	0.212	0.221

Setting 2: Event times were generated from Gerds' multinomial logistic regression model (Gerds et al., 2012). Cause-1 and cause-2 CIFs are given by

$$F_1(t|\mathbf{Z}) = \frac{\exp(a_1 t + b_1 + \boldsymbol{\beta}_1^T \mathbf{Z})}{\exp(a_1 t + b_1 + \boldsymbol{\beta}_1^T \mathbf{Z}) + \exp(a_2 t + b_2 + \boldsymbol{\beta}_2^T \mathbf{Z}) + 1},$$

and

$$F_2(t|\mathbf{Z}) = \frac{\exp(a_2 t + b_2 + \boldsymbol{\beta}_2^T \mathbf{Z})}{\exp(a_1 t + b_1 + \boldsymbol{\beta}_1^T \mathbf{Z}) + \exp(a_2 t + b_2 + \boldsymbol{\beta}_2^T \mathbf{Z}) + 1}.$$

Again we consider two problems: comparison among different models and comparison among different combinations of covariates. For each problem under each scenario, the same distributions were used to generate covariates as in Setting 1. We used $a_1 = a_2 = 1$ and $b_1 = b_2 = -15$ acroass all scenarios in Setting 2. $\boldsymbol{\beta}$s were the same as those in Setting 1. When comparing which model fits the data better, whether cause-1 and cause-2 CIFs are equal or not, censoring times were generate from $U(0, 46)$ for 30% censoring and from $U(0, 93)$ for 15% censoring, with $t_0 = 15$. When comparing models depending on different combination of covariates, we used $U(0, 47)$ for 30% censoring and $U(0, 94)$ for 15% censoring, with $t_0 = 15$. Results are shown in Table 8 and Table 9.

Table 8: Proportion of Model Selected under Gerds Model

Metrics	Different Models												Combination of Covariates							
	Same CIF						Diff CIF						Same CIF				Diff CIF			
	30%			15%			30%			15%			30%		15%		30%		15%	
	Cox	FG	Gerds	Cox	FG	Gerds	Cox	FG	Gerds	Cox	FG	Gerds	Gerds	GerdsR	Gerds	GerdsR	Gerds	GerdsR	Gerds	GerdsR
ExC	0.379	0.027	0.595	0.404	0.013	0.583	0.315	0.114	0.571	0.318	0.085	0.597	0.637	0.363	0.646	0.354	0.668	0.332	0.661	0.339
ExC*	0.069	0.722	0.209	0.042	0.836	0.123	0.157	0.058	0.785	0.133	0.042	0.825	0.543	0.457	0.495	0.505	0.757	0.243	0.782	0.218
PDI	0.398	0.028	0.575	0.402	0.007	0.591	0.308	0.088	0.604	0.298	0.063	0.639	0.948	0.052	0.968	0.032	0.909	0.091	0.930	0.070
ExC_1	0.451	0.002	0.547	0.449	0.002	0.549	0.329	0.101	0.570	0.329	0.078	0.593	0.642	0.358	0.645	0.355	0.597	0.403	0.573	0.427
$ExCA_1$	0.307	0.266	0.427	0.316	0.265	0.419	0.061	0.797	0.142	0.033	0.881	0.086	0.564	0.436	0.591	0.409	0.548	0.452	0.537	0.463
ExC_1^*	0.401	0.000	0.599	0.355	0.000	0.645	0.231	0.004	0.765	0.173	0.003	0.824	0.932	0.068	0.955	0.045	0.692	0.308	0.712	0.288
$AUCA_1$	0.431	0.042	0.527	0.445	0.018	0.537	0.241	0.325	0.435	0.201	0.324	0.474	0.824	0.176	0.852	0.148	0.613	0.387	0.607	0.393
PDI_1	0.447	0.068	0.486	0.497	0.024	0.479	0.295	0.175	0.530	0.265	0.148	0.587	0.690	0.310	0.726	0.274	0.569	0.431	0.578	0.422
AUC_1	0.397	0.010	0.593	0.414	0.003	0.583	0.229	0.039	0.731	0.189	0.023	0.788	0.951	0.049	0.967	0.033	0.657	0.343	0.664	0.336
BS_1	0.281	0.094	0.626	0.314	0.046	0.640	0.319	0.123	0.558	0.359	0.061	0.580	0.830	0.170	0.846	0.154	0.585	0.415	0.596	0.404
ExC_2	0.408	0.012	0.580	0.412	0.004	0.584	0.401	0.001	0.598	0.401	0.001	0.598	0.648	0.352	0.667	0.333	0.827	0.173	0.838	0.162
$ExCA_2$	0.295	0.308	0.396	0.302	0.272	0.426	0.378	0.062	0.560	0.373	0.057	0.571	0.578	0.422	0.598	0.402	0.804	0.196	0.834	0.166
ExC_2^*	0.353	0.000	0.647	0.330	0.000	0.670	0.359	0.017	0.624	0.342	0.004	0.654	0.932	0.068	0.956	0.044	0.957	0.043	0.978	0.022
$AUCA_2$	0.390	0.032	0.578	0.391	0.012	0.598	0.390	0.124	0.486	0.409	0.084	0.507	0.846	0.154	0.859	0.141	0.949	0.051	0.956	0.044
PDI_2	0.429	0.060	0.511	0.429	0.030	0.540	0.406	0.129	0.466	0.434	0.099	0.467	0.711	0.289	0.728	0.272	0.898	0.102	0.916	0.084
AUC_2	0.368	0.017	0.615	0.359	0.003	0.638	0.378	0.129	0.493	0.378	0.102	0.521	0.952	0.048	0.970	0.030	0.966	0.034	0.980	0.020
BS_2	0.276	0.082	0.641	0.293	0.044	0.663	0.277	0.142	0.581	0.294	0.093	0.613	0.824	0.176	0.865	0.135	0.952	0.048	0.961	0.039

Table 9: Averaged Estimated Values under Gerds Model

Metrics	Different Models												Combination of Covariates							
	Same CIF						Diff CIF						Same CIF				Diff CIF			
	30%			15%			30%			15%			30%		15%		30%		15%	
	Cox	FG	Gerds	Cox	FG	Gerds	Cox	FG	Gerds	Cox	FG	Gerds	Gerds	GerdsR	Gerds	GerdsR	Gerds	GerdsR	Gerds	GerdsR
ExC	0.052	0.027	0.054	0.052	0.021	0.055	0.032	0.015	0.035	0.031	0.011	0.034	0.044	0.037	0.044	0.038	0.037	0.032	0.036	0.031
ExC*	0.113	0.128	0.118	0.107	0.129	0.111	0.123	0.089	0.133	0.123	0.083	0.135	0.087	0.085	0.078	0.079	0.133	0.115	0.133	0.115
PDI	0.453	0.437	0.454	0.454	0.436	0.455	0.434	0.423	0.437	0.435	0.423	0.438	0.471	0.443	0.471	0.444	0.452	0.425	0.452	0.426
ExC_1	0.169	0.074	0.173	0.169	0.057	0.173	0.050	0.021	0.055	0.046	0.015	0.052	0.210	0.190	0.207	0.187	0.078	0.075	0.073	0.068
$ExCA_1$	0.231	0.219	0.232	0.228	0.215	0.230	0.120	0.166	0.122	0.113	0.171	0.116	0.217	0.205	0.212	0.202	0.106	0.105	0.100	0.098
ExC_1^*	0.601	0.498	0.606	0.611	0.494	0.616	0.326	0.172	0.354	0.322	0.152	0.356	0.664	0.610	0.670	0.616	0.441	0.402	0.442	0.404
$AUCA_1$	0.572	0.546	0.573	0.572	0.544	0.573	0.506	0.499	0.509	0.507	0.499	0.511	0.608	0.584	0.607	0.584	0.520	0.511	0.519	0.511
PDI_1	0.406	0.382	0.406	0.408	0.380	0.408	0.334	0.313	0.339	0.335	0.310	0.341	0.421	0.404	0.420	0.404	0.345	0.339	0.345	0.339
AUC_1	0.662	0.602	0.664	0.667	0.599	0.669	0.498	0.425	0.514	0.496	0.416	0.515	0.700	0.656	0.702	0.659	0.554	0.535	0.557	0.537
BS_1	0.231	0.232	0.231	0.230	0.232	0.230	0.221	0.223	0.221	0.221	0.222	0.220	0.213	0.216	0.213	0.215	0.217	0.218	0.217	0.217
ExC_2	0.165	0.077	0.172	0.168	0.059	0.175	0.250	0.094	0.261	0.257	0.067	0.267	0.212	0.192	0.214	0.194	0.336	0.298	0.344	0.306
$ExCA_2$	0.228	0.220	0.231	0.228	0.216	0.232	0.311	0.281	0.315	0.308	0.274	0.313	0.217	0.206	0.218	0.208	0.351	0.319	0.355	0.324
ExC_2^*	0.604	0.504	0.611	0.612	0.498	0.618	0.641	0.614	0.645	0.647	0.619	0.650	0.664	0.609	0.670	0.617	0.683	0.631	0.687	0.636
$AUCA_2$	0.571	0.546	0.573	0.571	0.543	0.573	0.630	0.624	0.630	0.630	0.625	0.630	0.608	0.584	0.608	0.586	0.655	0.623	0.655	0.624
PDI_2	0.405	0.382	0.406	0.406	0.379	0.407	0.470	0.465	0.471	0.471	0.466	0.472	0.420	0.404	0.422	0.406	0.490	0.456	0.491	0.458
AUC_2	0.662	0.604	0.664	0.666	0.600	0.670	0.687	0.679	0.688	0.689	0.682	0.690	0.700	0.656	0.702	0.659	0.713	0.669	0.713	0.671
BS_2	0.231	0.233	0.231	0.230	0.232	0.230	0.231	0.232	0.231	0.231	0.231	0.230	0.214	0.216	0.213	0.216	0.211	0.217	0.211	0.216

Setting 3: We generated data from the Fine-Gray subdistribution hazard model (Fine and Gray, 1999). Following their simulation strategy, we generated cause-1 event times from subdistribution hazard and cause-2 event times from the exponential distribution given that subject experienced a cause-2 event. Note that cause-2 event times generated in this way don't follow subdistribution hazard, and we cannot let cause-1 and cause-2 CIFs be equal. The cause-1 CIF was given by

$$F_1(t|\mathbf{Z}) = 1 - [1 - \gamma\{1 - \exp(-t)\}]^{\exp(\boldsymbol{\beta}_1^T \mathbf{Z})}$$

and the conditional distribution of cause-2 event was

$$P(t|\mathbf{Z}, \epsilon = 2) \sim \text{Exp}(\exp(\boldsymbol{\beta}_2^T \mathbf{Z})).$$

Covariates were generated the same way as in Setting 1 for each scenario. For comparison among Cox's, Gerds' and Fine-Gray models, parameter values were $\boldsymbol{\beta}_1 = (0.3, 0.3, 0.3, 0.3)$, $\boldsymbol{\beta}_2 = (-0.3, -0.3, -0.3, -0.3)$ and $\gamma = 0.48$. $C = U(0, 3.2)$ was used for 30% censoring and $C = U(0, 6.8)$ was for 15% censoring with prediction time $t_0 = 1.2$. When comparing two models depending on (Z_1, Z_2, Z_3) and (Z_2, Z_3, Z_4), we had $\boldsymbol{\beta}_1 = (0.3, 0.3, 0.3)$, $\boldsymbol{\beta}_2 = (-0.3, -0.3, -0.3)$ for (Z_1, Z_2, Z_3) and $\gamma = 0.48$. Censoring times were generated from $U(0, 2.9)$ for 30% censoring and from $U(0, 6.2)$ for 15% censoring, with $t_0 = 1.2$. Results are shown in Table 10 and Table 11.

First, for the overall evaluation, when we look at the proportion of correct selection, we see that PDI has a higher probability of selecting the true model than the two extended overall evaluation methods ExC and ExC* most of the time under Cox's and Gerds' models. ExC can select the true model sometimes, but with a lower proportion. Besides, when we look at the averaged estimated values of each metric, we can see that ExC tends to have a lower estimated value when cause-1 and cause-2 CIFs are different. This may be because many estimated values are zero, as discussed earlier. ExC* is more likely to prefer Gerds' model when data are generated from a Cox model, and fails again when data are generated from Gerds' model with equal cause-1 and cause-2 CIFs. Besides, ExC* tends to have worse performance when cause-1 and cause-2 CIFs are equal, as we can see when choosing different combinations, the probability of selecting true model is lower in this case. The estimated

Table 10: Proportion of Model Selected under Fine-Gray Model

Metrics	Different Models						Combination of Covariates			
	Diff CIF						Diff CIF			
	30%			15%			30%		15%	
	Cox	FG	Gerds	Cox	FG	Gerds	FG	FGR	FG	FGR
ExC	0.211	0.759	0.030	0.219	0.632	0.149	0.832	0.168	0.878	0.122
ExC^*	0.146	0.821	0.033	0.174	0.802	0.025	0.888	0.112	0.932	0.068
PDI	0.311	0.438	0.251	0.379	0.458	0.163	0.957	0.043	0.981	0.019
ExC_1	0.197	0.770	0.033	0.225	0.610	0.165	0.820	0.180	0.867	0.133
$ExCA_1$	0.161	0.667	0.172	0.089	0.715	0.196	0.827	0.173	0.845	0.155
ExC_1^*	0.030	0.786	0.184	0.003	0.878	0.118	0.734	0.266	0.781	0.219
$AUCA_1$	0.179	0.479	0.342	0.137	0.649	0.215	0.959	0.041	0.972	0.028
PDI_1	0.184	0.500	0.316	0.149	0.641	0.209	0.878	0.122	0.917	0.083
AUC_1	0.241	0.419	0.340	0.223	0.536	0.241	0.645	0.355	0.660	0.340
BS_1	0.134	0.598	0.268	0.087	0.676	0.237	0.961	0.039	0.978	0.022
ExC_2	0.185	0.787	0.028	0.207	0.646	0.147	0.827	0.173	0.865	0.135
$ExCA_2$	0.191	0.228	0.581	0.203	0.411	0.386	0.931	0.069	0.941	0.059
ExC_2^*	0.625	0.135	0.240	0.811	0.041	0.148	0.803	0.197	0.814	0.186
$AUCA_2$	0.361	0.320	0.319	0.409	0.340	0.251	0.995	0.005	0.998	0.002
PDI_2	0.344	0.326	0.330	0.395	0.364	0.241	0.989	0.011	0.996	0.004
AUC_2	0.336	0.327	0.337	0.385	0.366	0.249	0.958	0.042	0.981	0.019
BS_2	0.244	0.213	0.543	0.341	0.300	0.359	0.998	0.002	1.000	0.000

Table 11: Averaged Estimated Values under Fine-Gray Model

Metrics	Different Models						Combination of Covariates			
	Diff CIF						Diff CIF			
	30%			15%			30%		15%	
	Cox	FG	Gerds	Cox	FG	Gerds	FG	FGR	FG	FGR
ExC	0.027	0.036	0.016	0.020	0.028	0.018	0.030	0.018	0.020	0.010
ExC^*	0.178	0.188	0.167	0.173	0.188	0.163	0.174	0.141	0.180	0.146
PDI	0.482	0.483	0.480	0.483	0.482	0.478	0.490	0.455	0.491	0.456
ExC_1	0.059	0.080	0.036	0.045	0.063	0.039	0.045	0.027	0.029	0.015
$ExCA_1$	0.317	0.328	0.312	0.303	0.321	0.308	0.368	0.340	0.373	0.345
ExC_1^*	0.327	0.361	0.326	0.309	0.381	0.324	0.274	0.240	0.289	0.251
$AUCA_1$	0.654	0.655	0.654	0.653	0.656	0.654	0.663	0.632	0.665	0.633
PDI_1	0.500	0.502	0.501	0.500	0.503	0.500	0.489	0.459	0.490	0.460
AUC_1	0.577	0.577	0.577	0.576	0.577	0.576	0.559	0.547	0.559	0.547
BS_1	0.229	0.229	0.229	0.228	0.228	0.228	0.213	0.219	0.211	0.218
ExC_2	0.039	0.053	0.023	0.029	0.041	0.024	0.059	0.037	0.037	0.020
$ExCA_2$	0.376	0.374	0.384	0.372	0.376	0.377	0.436	0.391	0.435	0.389
ExC_2^*	0.474	0.449	0.450	0.490	0.421	0.444	0.553	0.512	0.538	0.499
$AUCA_2$	0.721	0.721	0.721	0.722	0.721	0.721	0.729	0.685	0.729	0.684
PDI_2	0.573	0.573	0.573	0.574	0.573	0.573	0.588	0.534	0.588	0.534
AUC_2	0.687	0.687	0.687	0.688	0.687	0.687	0.708	0.665	0.708	0.664
BS_2	0.177	0.177	0.177	0.177	0.177	0.177	0.202	0.214	0.201	0.213

values, which are the conditional probabilities of subjects being correctly sorted, are also smaller. One possible explanation is that, when models are well fitted and cause-1 and cause-2 CIFs are equal, for any two subjects i and j with i from cause-1 group and j from cause-2 group, if i had larger cause-1 CIF estimate than j, then it is much likely that i also had larger cause-2 CIF estimate than j, since for each subject the true cause-1 and cause-2 CIF are set to be equal and their estimates are likely to be close to each other. Then these two subjects could not be correctly sorted by the definition of ExC*. PDI is not affected because it is the average of correct classification probabilities across all groups, while ExC* requires all groups are well distinguished simultaneously.

For the cause-specific evaluation, the cause-specific ExC* outperforms the standard AUC, the augmented AUC, PDI, and the Brier score for both cause-1 and cause-2 events when data are generated from Cox's or Gerds' model and the goal is to compare models from three different families. ExC* and AUC are comparable in distinguishing models relying on different combinations of covariates, with AUC working better when cause-1 and cause-2 CIFs are equal and ExC* working better otherwise, and both of them perform better than other metrics in cause-specific evaluations. When data are generated from the Fine-Gray model and the goal is to pick the right underlying model from three fitted prognostic models, it is not surprising that almost all metrics fail for the cause-2 event, because the cause-2 event time was actually generated from a conditional distribution that does not follow the subdistribution hazard.

We observe that all metrics perform better when the goal is to select the combination of covariates that can provide better prediction. This can be seen from both the proportion of correct selection and the averaged estimated values. When choosing whether to fit Cox's, Gerds' or the Fine-Gray model, the discrepancies among estimated values for the three models are not obvious, especially between the Cox and Gerds models, while for selecting the combination of covariates, models depending on different sets of covariates can be well distinguished. This shows that even the model is misspecified, the Cox model and the Gerds model may still provide reliable prediction of event status. Gerds' model is often preferred when the true model is not Gerds'. The use of b-splines method in estimating the non-parametric part in the CIF makes Gerds' model quite flexible. Yet failing to include

important covariates may cause poorer prediction and should be avoided.

Overall, we think PDI provides a better overall evaluation of prognostic models. For cause-specific evaluation, both ExC* and AUC perform well. Besides, Brier score evaluates not only the diagnostic accuracy, but also the calibration accuracy. We observe that Brier score also works quite reliable for cause-specific evaluation, although sometimes it is not sensitive to the difference between models, as can be seen in the estimated values.

3.4 Data Analysis

We apply AUC (Blanche et al., 2013), Brier score (Schoop et al., 2011), PDI (Ding et al., 2021) and ExCs to the following datasets to see how they perform in real data.

3.4.1 Malignant Melanoma

Malignant melanoma data are available from the R package "riskRegression," where 205 patients with malignant melanoma, a skin cancer, were followed from 1962 to 1977. By the end of 1977, 134 patients were still alive, 57 died from the cancer and 14 died from other causes. We show the evaluation of Cox's, the Fine-Gray and Gerds' models in predicting event status at different time points. Models are evaluated at $t_0 = 1000, 2000, 3000$ to avoid extrapolation issues.

It is known that age, sex, tumor thickness and ulceration are risk factors on survival. Here we used the tumor thickness on log-scale instead of the original measurement. Models were fitted with both cause-1 and cause-2 events relying on these four covariates, and the cause-1 event, death from the malignant melanoma, was of primary interest. We summarize estimated values of each model evaluation metric in Table 12.

From Table 12 it can be seen that different models may be preferred by the same metric at different time of evaluation. For example, the overall PDI prefers the Fine-Gray model at $t_0 = 1000$, cannot distinguish the Fine-Gray and Gerds models at $t_0 = 2000$, and selects Gerds' model at $t_0 = 3000$. Besides, metrics we listed here fail to reach an agreement of

Table 12: Estimated Values of Model Evaluation for Malignant Melanoma Data

	$t_0 = 1000$			$t_0 = 2000$			3000		
Estimates	Cox	FG	Gerds	Cox	FG	Gerds	Cox	FG	Gerds
ExC	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ExC^*	0.195	0.195	0.205	0.222	0.212	0.245	0.248	0.224	0.286
PDI	0.589	0.596	0.593	0.602	0.604	0.604	0.619	0.616	0.623
ExC_1	0.038	0.000	0.038	0.295	0.318	0.385	0.404	0.407	0.443
ExCA_1	0.038	0.000	0.038	0.288	0.311	0.376	0.392	0.397	0.431
ExC_1^*	0.774	0.778	0.781	0.756	0.759	0.761	0.745	0.747	0.736
AUCA_1	0.799	0.803	0.800	0.765	0.767	0.765	0.746	0.747	0.740
PDI_1	0.485	0.491	0.520	0.454	0.458	0.480	0.471	0.465	0.493
AUC_1	0.806	0.814	0.805	0.779	0.784	0.773	0.763	0.764	0.753
BS_1	0.094	0.094	0.093	0.150	0.149	0.151	0.185	0.184	0.187
ExC_2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ExCA_2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ExC_2^*	0.694	0.670	0.696	0.732	0.715	0.716	0.737	0.725	0.730
AUCA_2	0.825	0.825	0.817	0.860	0.858	0.850	0.894	0.896	0.888
PDI_2	0.647	0.656	0.619	0.693	0.694	0.678	0.741	0.743	0.732
AUC_2	0.838	0.832	0.823	0.889	0.885	0.878	0.951	0.955	0.945
BS_2	0.032	0.032	0.032	0.045	0.045	0.045	0.051	0.051	0.050

what model to use. The zeros in ExC estimates are not surprising since cause-1 and cause-2 CIFs are well separated, as can be seen in Figure 1. From the estimated values we see that Gerds' model and the Fine-Gray model are selected by more metrics, but the discrepancies between models are really small regarding the discrimination of subjects with different event status at the time of prediction. Both the Fine-Gray and Gerds' models seem to provide reliable prediction of event status.

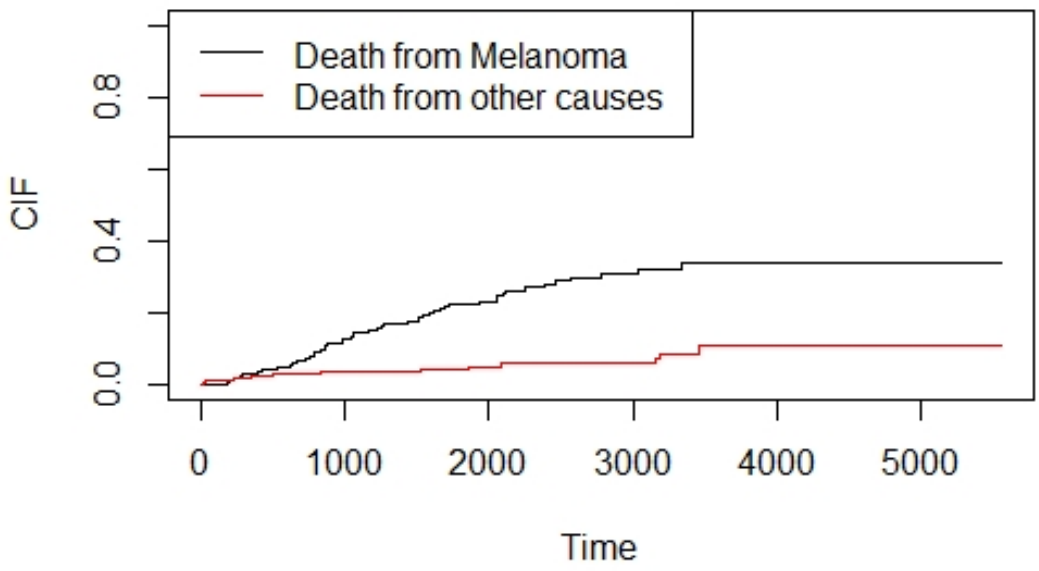

Figure 1: Raw CIF Estimates for Melanoma Data

3.4.2 Primary Biliary Cholangitis

We use Mayo clinic primary biliary cholangitis (PBC) data to show that failure to involve important risk factors in prognostic models can result in worse prediction of event status, which can be detected by the model evaluation metrics. The PBC data are available in R package "survival."

The PBC data contain 418 subjects in total who were followed between 1974 and 1984.

From the analysis conducted by Mayo clinic (Dickson et al., 1989), patients' age, total serum bilirubin, serum albumin concentrations, prothrombin time and severity of edema are important risk factors for PBC. A Cox's proportional hazard model was fitted for survival time, with transplant being treated as independent censoring (Dickson et al., 1989). Here we re-analyze the data by treating transplant as competing risks, as transplant (cause-1 event) can alter the probability of the event of primary interest, which is death (cause-2 event). Two subjects had missing values in prothrombin time and were removed for model evaluation. After removal, there were 231 subjects alive at the end of study, 25 had transplant and 160 were dead. We fitted two models, both assuming cause-specific proportional hazards, with one model relying on all important risk factors captured by Dickson et al. (1989) for both causes of event, and the other model relying on risk factors other than bilirubin for both causes. Following Dickson et al. (1989), natural logarithm was taken for bilirubin, albumin and prothrombin time, and we evaluate the model performance at predict time $t_0 = 3000$.

We summarize estimated values of each metric for the two fitted models. Besides, to see whether evaluation metrics can capture the significant reduction of the discrimination ability of prognostic model caused by the exclusion of important covariates, we constructed 95% BCa bootstrap confidence intervals based on 1000 bootstrap samples. We didn't provide the confidence intervals in simulation studies, since we know the true model in simulation settings, and we have instead focused on the probability that the true model can be selected. For the PBC data, results are summarized in Table 13. The 95% BCa confidence intervals of ExC_2^*, AUCA_2, AUC_2 and BS_2 indicate that the exclusion of total serum bilirubin does significantly affect model's ability to predict the probability of death. Regarding the cause-1 event, which is the transplant, there is some disagreement among metrics. ExC_1 and ExCA_1 completely fail in this example. The small sample size of subjects who had transplant and the bootstrap method caused unstable estimates. Besides, from Figure 2 we can see that the cause-2 CIF is constantly larger than the cause-1 CIF, and subjects who actually had transplant, which is the cause-1 event, would have a large estimated probability of experiencing cause-2 event, causing zero estimates. The same problem also happened in the Melanoma example with the cause-2 event. For the overall evaluation that measures how different causes of events were fitted simultaneously, we see that PDI shows a significant

difference between two models. ExC fails for the same reason as ExC_1 and ExCA_1. Both ExC_1^* and ExC_2^* suggest significant improvement in predicting transplant and death with bilirubin in the model. PDI_1 and PDI_2 suggest improvement when bilirubin is added in the model, though the improvement has not reached statistical significance. All these indicate the important role of bilirubin in predicting transplant and death.

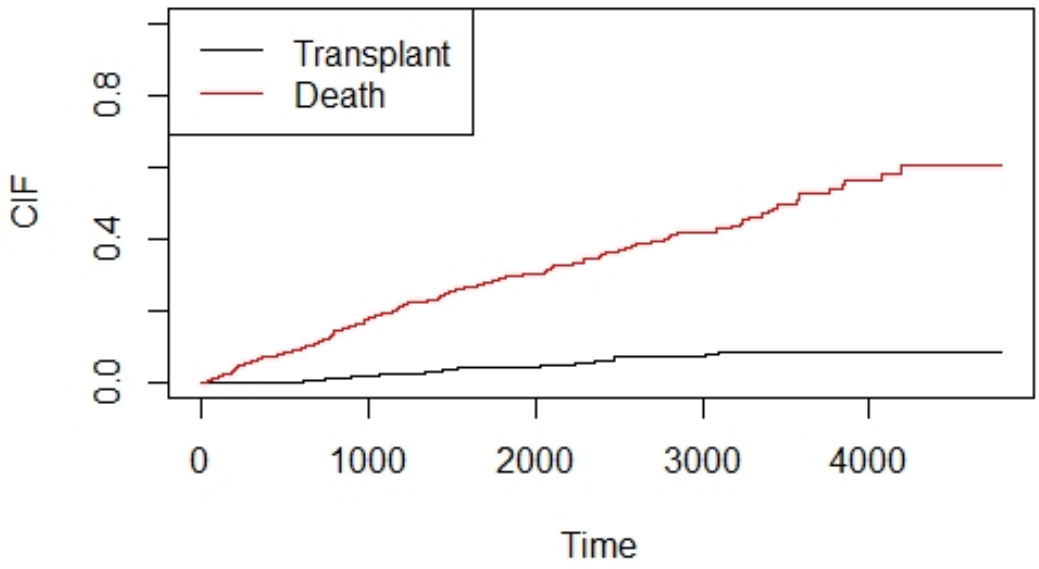

Figure 2: Raw CIF Estimates for PBC Data

3.5 Discussion

In this project, we performed a systematic examination of various model evaluation metrics with competing risks data. These model evaluation metrics were designed to measure the ability of prognostic models to estimate CIF and to predict event status, where AUC (Blanche et al., 2013) and Brier score (Schoop et al., 2011) focus on cause-specific evaluation

Table 13: Model Evaluation for PBC Data

Estimates	W/ Bilirubin	W/out Bilirubin	95% BCa CI
ExC	0	0	(-0.074, 0)
ExC^*	0.459	0.355	(-0.011, 0.194)
PDI	0.723	0.644	(0.015, 0.127)
ExC_1	0	0	(-0.147, 0.003)
$ExCA_1$	0	0	(-0.252, 0)
ExC_1^*	0.716	0.547	(0.031, 0.341)
$AUCA_1^*$	0.892	0.876	(-0.010, 0.061)
PDI_1	0.809	0.785	(-0.015, 0.088)
AUC_1	0.892	0.876	(-0.017, 0.049)
BS_1	0.056	0.057	(-0.005,0.001)
ExC_2	0.525	0.447	(-0.021, 0.202)
$ExCA_2$	0.52	0.444	(-0.022, 0.19)
ExC_2^*	0.798	0.684	(0.055, 0.184)
$AUCA_2$	0.806	0.727	(0.025, 0.132)
PDI_2	0.675	0.634	(-0.044, 0.093)
AUC_2	0.806	0.727	(0.029, 0.128)
BS_2	0.173	0.210	(-0.057, -0.017)

and PDI (Ding et al., 2021) can additionally provide overall assessment of diagnostic accuracy of models for predicting all competing events. Besides, we proposed and considered several overall and cause-specific extensions of concordance index that measure the probability of all subjects randomly selected from distinct groups are correctly sorted. The extended concordance indices and the bootstrap-based model comparisons are implemented in the R package "crModelEval", which is available for download at `https://github.com/YAQgh/crModelEval`.

In practice, model evaluation are performed for different purposes, such as comparing models from different families, and comparing models from the same family that depend on different combinations of covariates. We investigated these two problems in simulation studies under different scenarios and provided two data examples for further illustration. Overall, these model evaluation metrics are quite sensitive to the exclusion of important covariates in fitting prognostic models. While for comparison among models from different classes, although some metrics were able to select the true model with a relatively high proportion, we saw from simulation studies and data analysis that estimates didn't differ too much among different models, and metrics can even select models from different families at different times of evaluation. This indicates that regarding the prediction of event status, a misspecified model may still have reliable prediction, but the failure to include important covariates would more likely impair the ability to discriminate different classes.

Appendix A Supplement to Chapter 2

A.1 Proof of Theorem 1

A.1.1 Notation

We consider a single covariate Z here instead of a set of covariates $\mathbf{Z} = \{Z_l\}_{l=1}^{p}$. Subscripts i, j and k in the following notations simply denote subjects i, j and k.

$$\widehat{VUS} = \frac{\sum_{i\neq j\neq k} \frac{I(X_i\leq t_0,\eta_i=1,X_j\leq t_0,\eta_j=2,X_k>t_0,Z_i<Z_j<Z_k)}{\hat{G}(X_i-)\hat{G}(X_j-)\hat{G}(t_0)}}{\sum_{i\neq j\neq k} \frac{I(X_i\leq t_0,\eta_i=1,X_j\leq t_0,\eta_j=2,X_k>t_0)}{\hat{G}(X_i-)\hat{G}(X_j-)\hat{G}(t_0)}}$$

$$VUS = P(Z_i < Z_j < Z_k | T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0)$$

$$\hat{A} = \frac{1}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \frac{I(X_i \leq t_0, \eta_i = 1, X_j \leq t_0, \eta_j = 2, X_k > t_0, Z_i < Z_j < Z_k)}{\hat{G}(X_i-)\hat{G}(X_j-)\hat{G}(t_0)}$$

$$\tilde{A} = \frac{1}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \frac{I(X_i \leq t_0, \eta_i = 1, X_j \leq t_0, \eta_j = 2, X_k > t_0, Z_i < Z_j < Z_k)}{G(X_i-)G(X_j-)G(t_0)}$$

$$A = Pr(T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0, Z_i < Z_j < Z_k)$$

$$\hat{B} = \frac{1}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \frac{I(X_i \leq t_0, \eta_i = 1, X_j \leq t_0, \eta_j = 2, X_k > t_0)}{\hat{G}(X_i-)\hat{G}(X_j-)\hat{G}(t_0)}$$

$$\tilde{B} = \frac{1}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \frac{I(X_i \leq t_0, \eta_i = 1, X_j \leq t_0, \eta_j = 2, X_k > t_0)}{G(X_i-)G(X_j-)G(t_0)}$$

$$B = Pr(T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0)$$

A.1.2 Conditions

Condition 1 There exists a $\nu > 0$ such that $P(C = \nu) > 0$ and $P(C > \nu) = 0$.

Condition 2 $min_{l\in\mathcal{M}_*}|VUS_l(t_0) - 1/6| \geq c_0 n^{-\kappa}$ for some $0 < \kappa < 1/2$ and $c_0 > 0$.

Condition 3 There exists $\delta > 0$, such that $P(T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0) > \delta$.

A.1.3 Lemmas

Lemma 1 Bitouzé et al. (1999)

Let $\{T_i\}_{i=1}^n$ and $\{C_i\}_{i=1}^n$ be independent sequences of independently identically distributed non-negative random variables with distribution functions F and G, respectively. Let $\hat{F}_n$ be the Kaplan-Meier estimator of the distribution function F. There exists a positive constant, D, such that for any positive constant λ,

$$P(\sqrt{n}\|(1-G)(\hat{F}_n - F)_\infty > \lambda\|) \leq 2.5\exp\{-2\lambda^2 + D\lambda\}.$$

Lemma 2 Hoeffding (1963)

Let $g = g(x_1, \ldots, x_m)$ be a kernel of the U-statistic, U, with

$$a \leq g(x_1, \ldots, x_m) \leq b.$$

For any $t > 0$ and $m \leq n$, we have

$$P(|U - EU| > t) \leq 2\exp\left\{\frac{-2[n/m]t^2}{(b-a)^2}\right\}$$

Lemma 3 Pan et al. (2018)

Under Condition 1, for any $c_1 > 0$, when $n \geq D^2 c_2^{-1}$ with D being the constant from Lemma 1,

$$P\left(\max_{i,j,k}|\frac{G(X_i)G(X_j)G(X_k)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(X_k)} - 1| \geq c_1\right) \leq 2.5n^2\exp\{-c_2 n\},$$

where $c_2 = \frac{1}{9}(\frac{c_1}{1+c_1})^2\epsilon^8$, and $\epsilon > 0$ such that $\epsilon < S(x) < 1$ and $\epsilon < G(x) < 1$ under Condition 1.

A.1.4 Proof

We first show that $|\widehat{VUS} - VUS| > cn^{-\kappa}$ with low probability. For any single covariate Z in general, we have

$$\begin{aligned}
|\widehat{VUS} - VUS| &= \left|\frac{\hat{A}}{\hat{B}} - \frac{A}{B}\right| \\
&= \left|\frac{\hat{A}B - A\hat{B}}{\hat{B}B}\right| \\
&= \left|\frac{\hat{A}B - \hat{A}\hat{B} + \hat{A}\hat{B} - A\hat{B}}{\hat{B}B}\right| \\
&= \left|\hat{A}(\frac{1}{\hat{B}} - \frac{1}{B}) + \frac{1}{B}(\hat{A} - A)\right| \\
&\leq |\hat{A}|\left|\frac{1}{B}\right|\left|\frac{B}{\hat{B}} - 1\right| + \left|\frac{1}{B}\right||\hat{A} - A| \\
&= p_1 + p_2.
\end{aligned} \tag{A.1.1}$$

We start with bounding $\hat{A}$ in p_1.

$$\begin{aligned}
|\hat{A}| &= |\hat{A} - \tilde{A} + \tilde{A} - A + A| \\
&\leq |\hat{A} - \tilde{A}| + |\tilde{A} - A| + |A|.
\end{aligned} \tag{A.1.2}$$

Denote $I(X_i \leq t_0, \eta_i = 1, X_j \leq t_0, \eta_j = 2, X_k > t_0, Z_i < Z_j < Z_k)$ by KN.

$$\begin{aligned}
|\hat{A} - \tilde{A}| &= \frac{1}{n(n-1)(n-2)}\left|\sum_{i\neq j\neq k}\frac{KN}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - \sum_{i\neq j\neq k}\frac{KN}{G(X_i)G(X_j)G(t_0)}\right| \\
&= \frac{1}{n(n-1)(n-2)}\left|\sum_{i\neq j\neq k} KN\left(\frac{1}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - \frac{1}{G(X_i)G(X_j)G(t_0)}\right)\right| \\
&= \frac{1}{n(n-1)(n-2)}\left|\sum_{i\neq j\neq k}\frac{KN}{G(X_i)G(X_j)G(t_0)}\left(\frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1\right)\right| \\
&\leq \max_{i\neq j\neq k}\left|\frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1\right||\tilde{A}| \\
&\leq \max_{i\neq j\neq k}\left|\frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1\right||\tilde{A} - A| + \max_{i\neq j\neq k}\left|\frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1\right||A|.
\end{aligned} \tag{A.1.3}$$

By Lemma 3,

$$P\left(\max_{i\neq j\neq k}\left|\frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)}-1\right|\geq 1\right)\leq 2.5n^3\exp\{-\frac{1}{36}\epsilon^8 n\}. \tag{A.1.4}$$

By Lemma 2 and Condition 1, for any $c_3>0$,

$$P(|\tilde{A}-A|\geq c_3)\leq 2\exp\{-\frac{2}{3}c_3^2\epsilon^6 n\}. \tag{A.1.5}$$

Since A is a probability, we always have $|A|\leq 1$. By A.1.3, A.1.4, and A.1.5,

$$\begin{aligned}
P(|\hat{A}-\tilde{A}|\geq c_3+1) &\leq P(\max_{i\neq j\neq k}\left|\frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)}-1\right||\tilde{A}-A|\geq c_3)\\
&+P(\max_{i\neq j\neq k}\left|\frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)}-1\right||A|\geq 1)\\
&\leq 2\times P\left(\max_{i\neq j\neq k}\left|\frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)}-1\right|\geq 1\right)\\
&+P(|\tilde{A}-A|\geq c_3)\\
&=5n^3\exp\{-\frac{1}{36}\epsilon^8 n\}+2\exp\{-\frac{2}{3}c_3^2\epsilon^6 n\}.
\end{aligned} \tag{A.1.6}$$

From A.1.2, A.1.5 and A.1.6,

$$\begin{aligned}
P(|\hat{A}|\geq 2c_3+2) &\leq P(|\hat{A}-\tilde{A}|\geq c_3+1)+P(|\tilde{A}-A|\geq c_3)+P(|A|\geq 1)\\
&\leq 5n^3\exp\{-\frac{1}{36}\epsilon^8 n\}+2\exp\{-\frac{2}{3}c_3^2\epsilon^6 n\}+2\exp\{-\frac{2}{3}c_3^2\epsilon^6 n\}\\
&=5n^3\exp\{-\frac{1}{36}\epsilon^8 n\}+4\exp\{-\frac{2}{3}c_3^2\epsilon^6 n\}
\end{aligned} \tag{A.1.7}$$

We now show that $|\frac{B}{\hat{B}}-1|$ can be bounded with high probability.

We start with showing that for any $c_4>0$,

$$\left|\frac{B}{\hat{B}}-1\right|\geq c_4 n^{-\kappa}\implies|\hat{B}-B|\geq\frac{c_4 n^{-\kappa}}{1+c_4 n^{-\kappa}}B$$

1. $B>\hat{B}$

   From $B-\hat{B}\geq c_4 n^{-\kappa}\hat{B}$, we have $\hat{B}\leq\frac{1}{c_4 n^{-\kappa}+1}B$, and it follows that $\frac{c_4 n^{-\kappa}}{1+c_4 n^{-\kappa}}B=(1-\frac{1}{c_4 n^{-\kappa}+1})B\leq B-\hat{B}$.

2. $B < \hat{B}$

   We have $B - \hat{B} \leq -c_4 n^{-\kappa}\hat{B}$. Add $2c_4 n^{-\kappa}\hat{B}$ and minus $c_4 n^{-\kappa} B$ on both side, we get $(1 + c_4 n^{-\kappa})(\hat{B} - B) \geq 2c_4 n^{-\kappa}\hat{B} - c_4 n^{-\kappa} B$. Since $B < \hat{B}$, it follows that $2c_4 n^{-\kappa}\hat{B} - c_4 n^{-\kappa} B > c_4 n^{-\kappa} B$, which leads to $\hat{B} - B \geq \frac{c_4 n^{-\kappa}}{1+c_4 n^{-\kappa}} B$

By Condition 3, it follows that

$$
\begin{aligned}
& P\left(\left|\frac{B}{\hat{B}} - 1\right| \geq c_4 n^{-\kappa}\right) \\
& \leq P\left(|\hat{B} - B| \geq \frac{c_4 n^{-\kappa}}{1 + c_4 n^{-\kappa}} B\right) \\
& \leq P\left(|\hat{B} - B| \geq \frac{c_4 n^{-\kappa}}{1 + c_4 n^{-\kappa}} \delta\right) \\
& \leq P\left(|\hat{B} - \tilde{B}| + |\tilde{B} - B| \geq \frac{c_4 n^{-\kappa}}{1 + c_4 n^{-\kappa}} \delta\right) \\
& \leq P\left(|\hat{B} - \tilde{B}| \geq \frac{2}{3}\left(\frac{c_4 n^{-\kappa}}{1 + c_4 n^{-\kappa}}\right)\delta\right) + P\left(|\tilde{B} - B| \geq \frac{1}{3}\left(\frac{c_4 n^{-\kappa}}{1 + c_4 n^{-\kappa}}\right)\delta\right)
\end{aligned}
\tag{A.1.8}
$$

Similar to Equation A.1.3, $|\hat{B} - \tilde{B}|$ can be decomposed as

$$
|\hat{B} - \tilde{B}| \leq \max_{i \neq j \neq k} \left| \frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1 \right| |\tilde{B} - B| + \max_{i \neq j \neq k} \left| \frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1 \right| |B|
$$

By Lemma 2,

$$
P\left(|\tilde{B} - B| \geq \frac{1}{3}\left(\frac{c_4 n^{-\kappa}}{1 + c_4 n^{-\kappa}}\right)\delta\right) \leq 2\exp\left\{-\frac{2}{27} c_4^2 \epsilon^6 \delta^2 (1 + c_4 n^{-\kappa})^{-2} n^{1-2\kappa}\right\}.
$$

Again by Lemma 3,

$$
\begin{aligned}
& P\left(\max_{i \neq j \neq k} \left| \frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1 \right| \geq \frac{1}{3}\left(\frac{c_4 n^{-\kappa}}{1 + c_4 n^{-\kappa}}\right)\delta\right) \\
& \leq 2.5 n^3 \exp\left\{-\frac{1}{9} c_4^2 \delta^2 \epsilon^8 (3 + 3c_4 n^{-\kappa} + c_4 n^{-\kappa}\delta)^{-2} n^{1-2\kappa}\right\}.
\end{aligned}
$$

By Equation A.1.8,

$$
\begin{aligned}
P\left(\left|\frac{B}{\hat{B}}-1\right| \geq c_4 n^{-\kappa}\right) &\leq P\left(\max_{i\neq j\neq k}\left|\frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)}-1\right| |\tilde{B}-B| \geq \frac{1}{3}\left(\frac{c_4 n^{-\kappa}}{1+c_4 n^{-\kappa}}\right)\delta\right) \\
&+ P\left(\max_{i\neq j\neq k}\left|\frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)}-1\right| |B| \geq \frac{1}{3}\left(\frac{c_4 n^{-\kappa}}{1+c_4 n^{-\kappa}}\right)\delta\right) \\
&+ P\left(|\tilde{B}-B| \geq \frac{1}{3}\left(\frac{c_4 n^{-\kappa}}{1+c_4 n^{-\kappa}}\right)\delta\right) \\
&\leq 2.5n^3 \exp\{-\frac{1}{36}\epsilon^8 n\} + 2\exp\left\{-\frac{2}{27}c_4^2\epsilon^6\delta^2(1+c_4 n^{-\kappa})^{-2}n^{1-2\kappa}\right\} \\
&+ 2.5n^3 \exp\left\{-\frac{1}{9}c_4^2\delta^2\epsilon^8(3+3c_4 n^{-\kappa}+c_4 n^{-\kappa}\delta)^{-2}n^{1-2\kappa}\right\} \\
&+ 2\exp\left\{-\frac{2}{27}c_4^2\epsilon^6\delta^2(1+c_4 n^{-\kappa})^{-2}n^{1-2\kappa}\right\} \\
&= 2.5n^3 \exp\{-\frac{1}{36}\epsilon^8 n\} + 4\exp\left\{-\frac{2}{27}c_4^2\epsilon^6\delta^2(1+c_4 n^{-\kappa})^{-2}n^{1-2\kappa}\right\} \\
&+ 2.5n^3 \exp\left\{-\frac{1}{9}c_4^2\delta^2\epsilon^8(3+3c_4 n^{-\kappa}+c_4 n^{-\kappa}\delta)^{-2}n^{1-2\kappa}\right\}
\end{aligned}
\tag{A.1.9}
$$

Equation A.1.7 and Equation A.1.9 together showed p_1 can be bounded with high probability:

$$
\begin{aligned}
&P\left(|\hat{A}|\left|\frac{1}{B}\right|\left|\frac{B}{\hat{B}}-1\right| \geq \frac{1}{\delta}(2c_3+2)c_4 n^{-\kappa}\right) \\
&\leq P(|\hat{A}| \geq (2c_3+2)) + P\left(\left|\frac{B}{\hat{B}}-1\right| \geq c_4 n^{-\kappa}\right) \\
&\leq 7.5n^3 \exp\{-\frac{1}{36}\epsilon^8 n\} + 4\exp\{-\frac{2}{3}c_3^2\epsilon^6 n\} \\
&+ 4\exp\left\{-\frac{2}{27}c_4^2\epsilon^6\delta^2(1+c_4 n^{-\kappa})^{-2}n^{1-2\kappa}\right\} \\
&+ 2.5n^3 \exp\left\{-\frac{1}{9}c_4^2\delta^2\epsilon^8(3+3c_4 n^{-\kappa}+c_4 n^{-\kappa}\delta)^{-2}n^{1-2\kappa}\right\}.
\end{aligned}
\tag{A.1.10}
$$

Now we show p_2 can be bounded with high probability, where $p_2 = \left|\frac{1}{B}\right| |\hat{A}-A|$. For any $c_5 > 0$,

$$
\begin{aligned}
&P(|\hat{A} - A| \geq c_5 n^{-\kappa}) \\
&\leq P(|\hat{A} - \tilde{A}| \geq \frac{2}{3} c_5 n^{-\kappa}) + P(|\tilde{A} - A| \geq \frac{1}{3} c_5 n^{-\kappa}) \\
&\leq P\left(\max_{i \neq j \neq k} \left| \frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1 \right| |\tilde{A} - A| \geq \frac{1}{3} c_5 n^{-\kappa} \right) \\
&+ P\left(\max_{i \neq j \neq k} \left| \frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1 \right| |A| \geq \frac{1}{3} c_5 n^{-\kappa} \right) + P(|\tilde{A} - A| \geq \frac{1}{3} c_5 n^{-\kappa}),
\end{aligned}
\tag{A.1.11}
$$

where we used Equation A.1.3 again for the second inequality, and by Lemma 2 and Lemma 3, we can bound each component as the following

$$
P\left(\max_{i \neq j \neq k} \left| \frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1 \right| \geq 1 \right) \leq 2.5 n^3 \exp\{-\frac{1}{36}\epsilon^8 n\},
$$

$$
P(|\tilde{A} - A| \geq \frac{1}{3} c_5^{-\kappa}) \leq 2 \exp\{-\frac{2}{27} c_5^2 \epsilon^6 n^{1-2\kappa}\},
$$

$$
P\left(\max_{i \neq j \neq k} \left| \frac{G(X_i)G(X_j)G(t_0)}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - 1 \right| \geq \frac{1}{3} c_5 n^{-\kappa} \right) \leq 2.5 n^3 \exp\{-\frac{1}{9} c_5^2 \epsilon^8 (3 + c_5 n^{-\kappa})^{-2} n^{1-2\kappa}\}.
$$

Plug into Equation A.1.11, we have

$$
\begin{aligned}
P(|\hat{A} - A| \geq c_5 n^{-\kappa}) &\leq 2.5 n^3 \exp\{-\frac{1}{36}\epsilon^8 n\} + 2 \exp\{-\frac{2}{27} c_5^2 \epsilon^6 n^{1-2\kappa}\} \\
&+ 2.5 n^3 \exp\{-\frac{1}{9} c_5^2 \epsilon^8 (3 + c_5 n^{-\kappa})^{-2} n^{1-2\kappa}\} + 2 \exp\{-\frac{2}{27} c_5^2 \epsilon^6 n^{1-2\kappa}\} \\
&= 2.5 n^3 \exp\{-\frac{1}{36}\epsilon^8 n\} + 4 \exp\{-\frac{2}{27} c_5^2 \epsilon^6 n^{1-2\kappa}\} \\
&+ 2.5 n^3 \exp\{-\frac{1}{9} c_5^2 \epsilon^8 (3 + c_5 n^{-\kappa})^{-2} n^{1-2\kappa}\}.
\end{aligned}
\tag{A.1.12}
$$

Now for any positive constant c_6, there exist positive constants c_3, c_4 and c_5 such that $c_6 = \frac{1}{\delta}(2c_3+2)c_4 + \frac{1}{\delta}c_5$. By Equation A.1.10 and Equation A.1.12

$$\begin{aligned}
&P(|\widehat{VUS} - VUS| \geq c_6 n^{-\kappa}) \leq P\left(|\hat{A}|\left|\frac{1}{B}\right|\left|\frac{B}{\hat{B}}-1\right| \geq \frac{1}{\delta}(2c_3+2)c_4 n^{-\kappa}\right) \\
&+ P\left(\left|\frac{1}{B}\right||\hat{A}-A| \geq \frac{1}{\delta}c_5 n^{-\kappa}\right) \\
&\leq P\left(|\hat{A}|\left|\frac{B}{\hat{B}}-1\right| \geq (2c_3+2)c_4 n^{-\kappa}\right) + P\left(|\hat{A}-A| \geq c_5 n^{-\kappa}\right) \\
&\leq 10n^3 \exp\{-\frac{1}{36}\epsilon^8 n\} + 4\exp\{-\frac{2}{3}c_3^2\epsilon^6 n\} \\
&+ 4\exp\left\{-\frac{2}{27}c_4^2\epsilon^6\delta^2(1+c_4 n^{-\kappa})^{-2}n^{1-2\kappa}\right\} \\
&+ 2.5n^3 \exp\left\{-\frac{1}{9}c_4^2\delta^2\epsilon^8(3+3c_4 n^{-\kappa}+c_4 n^{-\kappa}\delta)^{-2}n^{1-2\kappa}\right\} \\
&+ 4\exp\{-\frac{2}{27}c_5^2\epsilon^6 n^{1-2\kappa}\} \\
&+ 2.5n^3 \exp\{-\frac{1}{9}c_5^2\epsilon^8(3+c_5 n^{-\kappa})^{-2}n^{1-2\kappa}\}.
\end{aligned} \tag{A.1.13}$$

We now show the second statement of Theorem 1.

$$\hat{\mathcal{M}} = \{k : |\widehat{VUS}_k - 1/6| \geq \gamma_n\}$$

where $\gamma_n = cn^{-\kappa}$ with $c < c_0$. By Condition 2,

$$\begin{aligned}
1 - P(\mathcal{M}_* \subset \hat{\mathcal{M}}) &= P(\exists l \in \mathcal{M}_*, |\widehat{VUS}_l - 1/6| < \gamma_n) \\
&\leq P(\max_{l\in\mathcal{M}_*}|\widehat{VUS}_l - VUS| \geq (c_0-c)n^{-\kappa}) \\
&\leq sP(|\widehat{VUS} - VUS| \geq (c_0-c)n^{-\kappa}).
\end{aligned} \tag{A.1.14}$$

Therefore,

$$P(\mathcal{M}_* \subset \hat{\mathcal{M}}) \geq 1 - sP(|\widehat{VUS} - VUS| \geq (c_0-c)n^{-\kappa}).$$

See Equation A.1.13, $s = |\mathcal{M}_*|$, the cardinality of $\mathcal{M}_*$.

A.2 Proof of Theorem 2

Assume $\sum_{l=1}^{p} |VUS_l - 1/6| = O(n^{\xi})$. Based on Condition 3 that $\min_{l \in \mathcal{M}^*} |VUS_l - 1/6| \geq c_0 n^{-\kappa}$ for some $0 < \kappa < 1/2$ and $c_0 > 0$, the cardinality of the set $\{l : |VUS_l - 1/6| \geq c_0 n^{-\kappa}\}$ is at most $O(n^{\xi+\kappa})$.

On the set $\{\max_{1 \leq l \leq p} |\widehat{VUS}_l - VUS_l| \leq c_0 n^{-\kappa}\}$, the number of $\{l : |\widehat{VUS}_l - 1/6| > 2c_0 n^{-\kappa}\}$ is no bigger than the number of $\{l : |VUS_l - 1/6| > c_0 n^{-\kappa}\}$. Therefore, the number of $\{l : |\widehat{VUS}_l - 1/6| > 2c_0 n^{-\kappa}\}$ is smaller than or equal to $O(n^{\xi+\kappa})$ on $\{\max_{1 \leq l \leq p} |\widehat{VUS}_l - VUS_l| \leq c_0 n^{-\kappa}\}$.

The size of $\hat{\mathcal{M}}$ is the size of $\{l : |\widehat{VUS}_l - 1/6| > cn^{-\kappa}\}$ for some $c < c_0$.

$$
\begin{aligned}
P(|\hat{\mathcal{M}}| \leq O(n^{\xi+\kappa})) &\geq P(\{\max_{1 \leq l \leq p} |\widehat{VUS}_l - VUS_l| \leq \frac{1}{2} cn^{-\kappa}\}) \\
&\geq 1 - pP(|\widehat{VUS}_l - VUS_l| \geq \frac{1}{2} cn^{-\kappa}).
\end{aligned}
$$

Appendix B Supplement to Chapter 3

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_L)$ denote the unknown parameters in CIF functions for L causes, which include the coefficients and the non-parametric part. The true value of $\boldsymbol{\theta}$ is denoted as $\boldsymbol{\theta}_0$ and the estimated value is denoted as $\hat{\boldsymbol{\theta}}$. Dependence of CIFs on covariates $\mathbf{Z}$ are skipped for now and will be shown whenever needed. We also define the following quantities:

$$I_{ijk} = I(X_i \leq t_0, \eta_i = 1, X_j \leq t_0, \eta_j = 2, X_k > t_0)$$

$$Q_0(t_0) = P(T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0)$$

$$\hat{Q}_0(t_0) = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} \frac{I_{ijk}}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)}$$

$$C_{ijk}(F(t_0|\boldsymbol{\theta})) = I(\max F_1(t_0|\boldsymbol{\theta}) = F_{1i}(t_0|\boldsymbol{\theta}), \max F_2(t_0|\boldsymbol{\theta}) = F_{2j}(t_0|\boldsymbol{\theta}), \max S = F_k(t_0|\boldsymbol{\theta}))$$

$$\hat{Q}(F) = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq k} \frac{C_{ijk}(F) I_{ijk}}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)}$$

$$Q(F) = E[C_{ijk}(F) I(T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0)]$$

We show the asymptotic normality of $\widehat{\text{ExC}^*}$, then $\widehat{\text{ExC}^*} = \frac{\hat{Q}(F(\hat{\theta}))}{\hat{Q}_0(t_0)}$. Asymptotic normality of $\widehat{\text{ExC}}$ can be shown similarly with a different C_{ijk}. To show the asymptotic normality, we make the following assumptions:

Condition 1 There exists $\delta > 0$, such that $P(T_i \leq t_0, \epsilon_i = 1, T_j \leq t_0, \epsilon_j = 2, T_k > t_0) > \delta$.

Condition 2 $\sqrt{n}(\hat{\boldsymbol{\theta}}_l - \boldsymbol{\theta}_{l0}) = \frac{1}{\sqrt{n}} \sum_{s=1}^{n} \mathbb{I}_{ls} + o_p(1)$, where $\mathbb{I}_{ls}$ is the influence function with mean 0.

Condition 3 $\frac{\partial F_l(t_0|\boldsymbol{\theta}_l, \mathbf{Z})}{\partial \boldsymbol{\theta}_l}$ is uniformly bounded in $\mathbf{Z}$.

We use the martingale representation $\frac{\hat{G}(t_0)}{G(t_0)} - 1 = -\frac{1}{n} \sum_{s=1}^{n} \int_0^{t_0} \frac{dM_{C_s}(u)}{S_X(u)} + o_p(1)$ and the Taylor expansion

$$\frac{C}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} =$$
$$\frac{C}{G(X_i)G(X_j)G(t_0)}[1 - (\frac{\hat{G}(X_i)}{G(X_i)} - 1) - (\frac{\hat{G}(X_j)}{G(X_j)} - 1) - (\frac{\hat{G}(t_0)}{G(t_0)} - 1)] + o_p(1).$$

First for the denominator $\hat{Q}_0(t_0)$,

$$
\begin{aligned}
&\sqrt{n}(\hat{Q}_0 - Q_0) \\
&= \frac{\sqrt{n}}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \left\{ \frac{I_{ijk}}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - Q_0 \right\} \\
&= \frac{\sqrt{n}}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \left\{ \frac{I_{ijk}}{G(X_i)G(X_j)G(t_0)} \times \right. \\
&\left[1 - (\frac{\hat{G}(X_i)}{G(X_i)} - 1) - (\frac{\hat{G}(X_j)}{G(X_j)} - 1) - (\frac{\hat{G}(t_0)}{G(t_0)} - 1) \right] \\
&\left. + o_p(1) - Q_0 \right\} \\
&= \frac{\sqrt{n}}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \left\{ \frac{I_{ijk}}{G(X_i)G(X_j)G(t_0)} - Q_0 + o_p(1) \right\} \\
&+ \frac{\sqrt{n}}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \frac{I_{ijk}}{G(X_i)G(X_j)G(t_0)} \times \\
&\left\{ \frac{1}{n} \sum_{s=1}^{n} \int_0^{X_i} \frac{dM_{C_s}(u)}{S_X(u)} + \frac{1}{n} \sum_{s=1}^{n} \int_0^{X_j} \frac{dM_{C_s}(u)}{S_X(u)} + \frac{1}{n} \sum_{s=1}^{n} \int_0^{t_0} \frac{dM_{C_s}(u)}{S_X(u)} + o_p(1) \right\} \\
&= \frac{\sqrt{n}}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \left\{ \frac{I_{ijk}}{G(X_i)G(X_j)G(t_0)} - Q_0 + o_p(1) \right\} \\
&+ \frac{\sqrt{n}}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \frac{I_{ijk}}{G(X_i)G(X_j)G(t_0)} \times \\
&\left\{ \frac{1}{n} \sum_{s=1}^{n} \int_0^{t_0} [1 + I(X_i \geq u) + I(X_j \geq u)] \frac{dM_{C_s}(u)}{S_X(u)} + o_p(1) \right\} \\
&= \frac{\sqrt{n}}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \left\{ \frac{I_{ijk}}{G(X_i)G(X_j)G(t_0)} - Q_0 + o_p(1) \right\} \\
&+ \frac{1}{\sqrt{n}} \sum_{s=1}^{n} \int_0^{t_0} \left\{ \frac{1}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \frac{I_{ijk}[1 + I(X_i \geq u) + I(X_j \geq u)]}{G(X_i)G(X_j)G(t_0)} \right\} \frac{dM_{C_s}(u)}{S_X(u)} + o_p(1) \\
&= \frac{\sqrt{n}}{n(n-1)(n-2)} \sum_{i\neq j\neq k} \left\{ \frac{I_{ijk}}{G(X_i)G(X_j)G(t_0)} - Q_0 + o_p(1) \right\} \\
&+ \frac{1}{\sqrt{n}} \sum_{s=1}^{n} \int_0^{t_0} E\left[\frac{I_{ijk}\{1 + I(X_i \geq u) + I(X_j \geq u)\}}{G(X_i)G(X_j)G(t_0)S_X(u)} \right] dM_{C_s}(u) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{s=1}^{n} h_{1s} + \frac{1}{\sqrt{n}} \sum_{s=1}^{n} \int_0^{t_0} h_{2s} dM_{C_s}(u) + o_p(1),
\end{aligned}
$$

where we use projection of U-statistic to the first sum to get the last inequality:

$$\begin{aligned}
\hat{U} &= \sum_{s=1}^{n} E(U|X_s, \eta_s, \mathbf{Z}_s) \\
&= \sum_{s=1}^{n} \frac{(n-1)(n-2)}{n(n-1)(n-2)} \left[E\{\frac{I(X_s \le t_0, \eta_s = 1) I_j I_k}{G(X_i)G(X_j)G(t_0)} | X_s, \eta_s, \mathbf{Z}_s\} \right. \\
&+ E\{\frac{I_i I(X_s \le t_0, \eta_s = 2) I_k}{G(X_i)G(X_j)G(t_0)} | X_s, \eta_s, \mathbf{Z}_s\} \\
&+ \left. E\{\frac{I_i I_j I(X_s > t_0)}{G(X_i)G(X_j)G(t_0)} | X_s, \eta_s, \mathbf{Z}_s\} - 3Q_0 \right] \\
&= \frac{1}{n} \sum_{s=1}^{n} h_{1s}
\end{aligned}$$

and

$$\begin{aligned}
h_{1s} = E \left[\frac{I(X_s \le t_0, \eta_s = 1) I_j I_k}{G(X_i)G(X_j)G(t_0)} + \frac{I_i I(X_s \le t_0, \eta_s = 2) I_k}{G(X_i)G(X_j)G(t_0)} + \frac{I_i I_j I(X_s > t_0)}{G(X_i)G(X_j)G(t_0)} \right. \\
\left. -3Q_0 | X_s, \eta_s, \mathbf{Z}_s \right],
\end{aligned}$$

$$h_{2s} = E \left[\frac{I_{ijk}\{1 + I(X_i \ge u) + I(X_j \ge u)\}}{G(X_i)G(X_j)G(t_0)S_X(u)} \right].$$

For $\sqrt{n}(\hat{Q}(F(\hat{\boldsymbol{\theta}})) - Q(F(\boldsymbol{\theta}_0)))$, we use decomposition

$$\sqrt{n}(\hat{Q}(F(\hat{\boldsymbol{\theta}})) - Q(F(\boldsymbol{\theta}_0))) = \sqrt{n}(\hat{Q}(F(\hat{\boldsymbol{\theta}})) - Q(F(\hat{\boldsymbol{\theta}}))) + \sqrt{n}(Q(F(\hat{\boldsymbol{\theta}})) - Q(F(\boldsymbol{\theta}_0))).$$

From Condition 2 and Condition 3, we have $\sqrt{n}(F_l(t_0|\hat{\boldsymbol{\theta}}_l, \mathbf{Z}) - F_l(t_0|\boldsymbol{\theta}_0, \mathbf{Z})) = \frac{1}{\sqrt{n}} \sum_{s=1} \Psi_s + o_p(1)$, where $\Psi_s = \left. \frac{\partial F_l(t_0|\boldsymbol{\theta}_l, \mathbf{Z})}{\partial \boldsymbol{\theta}_l} \right|_{\boldsymbol{\theta}_l = \boldsymbol{\theta}_{l0}} \mathbb{I}_{ls}$. Let Γ be the Hadamard derivative of Q at $F(\boldsymbol{\theta}_0)$, then

$$\sqrt{n}(Q(F(\hat{\boldsymbol{\theta}})) - Q(F(\boldsymbol{\theta}_0))) = \frac{1}{\sqrt{n}} \sum_{s=1}^{n} \Gamma(\Psi_s) + o_p(1).$$

For $\sqrt{n}(\hat{Q}(F(\hat{\boldsymbol{\theta}})) - Q(F(\hat{\boldsymbol{\theta}})))$, similar to the denominator,

$$\begin{aligned}
&\sqrt{n}(\hat{Q}(F(\hat{\boldsymbol{\theta}})) - Q(F(\hat{\boldsymbol{\theta}}))) \\
&= \frac{\sqrt{n}}{n(n-1)(n-2)} \sum_{i \ne j \ne k} \left\{ \frac{I_{ijk} C_{ijk}(F(\hat{\boldsymbol{\theta}}))}{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)} - Q(F(\hat{\boldsymbol{\theta}})) \right\} \\
&= \frac{1}{\sqrt{n}} \sum_{s=1}^{n} h_{3s} + \frac{1}{\sqrt{n}} \sum_{s=1}^{n} \int_0^{t_0} h_{4s} dM_{C_s}(u) + o_p(1),
\end{aligned}$$

$$h_{3s} = E\left[\frac{I(X_s \le t_0, \eta_s = 1) I_j I_k C_{sjk}}{G(X_i)G(X_j)G(t_0)} | X_s, \eta_s, \mathbf{Z}_s\right] + E\left[\frac{I_i I(X_s \le t_0, \eta_s = 2) I_k C_{isk}}{G(X_i)G(X_j)G(t_0)} | X_s, \eta_s, \mathbf{Z}_s\right] + E\left[\frac{I_i I_j I(X_s > t_0) C_{ijs}}{G(X_i)G(X_j)G(t_0)} | X_s, \eta_s, \mathbf{Z}_s\right] - 3Q(F(\hat{\boldsymbol{\theta}})),$$

$$h_{4s} = E\left[\frac{I_{ijk} C_{ijk}}{G(X_i)G(X_j)G(t_0)S_X(u)}\{1 + I(X_i \ge u) + I(X_j \ge u)\}\right].$$

Using delta method, we have

$$\begin{aligned}
&\sqrt{n}(\widehat{\mathrm{ExC}^*} - \mathrm{ExC}^*) = \sqrt{n}\left(\frac{\hat{Q}(F(\hat{\boldsymbol{\theta}}))}{\hat{Q}_0(t_0)} - \frac{Q(F(\boldsymbol{\theta}_0))}{Q_0(t_0)}\right) \\
&= \frac{\sqrt{n}(\hat{Q}(F(\hat{\boldsymbol{\theta}})) - Q(F(\boldsymbol{\theta}_0))) - \mathrm{ExC}^* \sqrt{n}(\hat{Q}_0(t_0) - Q_0(t_0))}{Q_0(t_0)} + o_p(1) \\
&= Q_0^{-1}(t_0)\{\sqrt{n}(\hat{Q}(F(\hat{\boldsymbol{\theta}})) - Q(F(\hat{\boldsymbol{\theta}}))) + \sqrt{n}(Q(F(\hat{\boldsymbol{\theta}})) - Q(F(\boldsymbol{\theta}_0)))\} \\
&+ Q_0^{-1}(t_0) \cdot \mathrm{ExC}^* \cdot \sqrt{n}(\hat{Q}_0 - Q_0) + o_p(1) \\
&= Q_0^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{s=1}^{n}\Gamma(\Psi_s) + \frac{1}{\sqrt{n}}\sum_{s=1}^{n} h_{3s} + \frac{1}{\sqrt{n}}\sum_{s=1}^{n}\int_0^{t_0} h_{4s} dM_{C_s}(u)\right\} \\
&+ Q_0^{-1} \cdot \mathrm{ExC}^* \cdot \left\{\frac{1}{\sqrt{n}}\sum_{s=1}^{n} h_{1s} + \frac{1}{\sqrt{n}}\sum_{s=1}^{n}\int_0^{t_0} h_{2s} dM_{C_s}(u)\right\} + o_p(1).
\end{aligned}$$

By functional central limit theorem, $\sqrt{n}(\widehat{\mathrm{ExC}^*} - \mathrm{ExC}^*)$ is asymptotically normal.

Bibliography

D. Bitouzé, B. Laurent, and P. Massart. A dvoretzky-kiefer-wolfowitz type inequality for the kaplan-meier estimator. *Annales de L'Institut Henri Poincare Section Physique Theorique*, 35:735–763, 11 1999. doi: 10.1016/S0246-0203(99)00112-0.

Paul Blanche, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397, 2013. doi: https://doi.org/10.1002/sim.5958. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5958`.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2. URL `https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml`.

Jan Budczies and Daniel Kosztyla. *cancerdata: Development and validation of diagnostic tests from high-dimensional molecular data: Datasets*, 2021. R package version 1.30.0.

E. Rolland Dickson, Patricia M. Grambsch, Thomas R. Fleming, Lloyd D. Fisher, and Alice Langworthy. Prognosis in primary biliary cirrhosis: Model for decision making. *Hepatology*, 10(1):1–7, 1989. doi: https://doi.org/10.1002/hep.1840100102. URL `https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep.1840100102`.

Maomao Ding, Jing Ning, and Ruosha Li. Evaluation of competing risks prediction models using polytomous discrimination index. *Canadian Journal of Statistics*, 49(3):731–753, 2021. doi: https://doi.org/10.1002/cjs.11583. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.11583`.

Stephan Dreiseitl, Lucila Ohno-Machado, and Michael Binder. Comparing three-class diagnostic tests by three-way roc analysis. *Medical Decision Making*, 20(3):323–331, 2000. doi: 10.1177/0272989X0002000309. URL `https://doi.org/10.1177/0272989X0002000309`. PMID: 10929855.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987. doi: 10.1080/01621459.1987.10478410. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1987.10478410`.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5): 849–911, 2008.

Jianqing Fan, Yang Feng, and Yichao Wu. *High-dimensional variable selection for Cox's proportional hazards model*, volume Volume 6 of *Collections*, pages 70–86. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2010.

Jason P. Fine and Robert J. Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999. doi: 10.1080/01621459.1999.10474144. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10474144.

Thomas A. Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040, 2006. doi: https://doi.org/10.1002/bimj.200610301. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200610301.

Thomas A. Gerds, Thomas H. Scheike, and Per K. Andersen. Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine*, 31(29):3921–3930, 2012. doi: https://doi.org/10.1002/sim.5459. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5459.

Thomas A. Gerds, Per K. Andersen, and Michael W. Kattan. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine*, 33(18):3191–3203, 2014. doi: https://doi.org/10.1002/sim.6152. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6152.

Ted A. Gooley, Wendy Leisenring, John Crowley, and Barry E. Storer. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine*, 18(6):695–706, 1999. doi: https://doi.org/10.1002/(SICI)1097-0258(19990330)18:6⟨695::AID-SIM60⟩3.0.CO;2-O. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2819990330%2918%3A6%3C695%3A%3AAID-SIM60%3E3.0.CO%3B2-O.

Anders Gorst-Rasmussen and Thomas Scheike. Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):217–245, 2013.

S. R. Haile, J.-H. Jeong, X. Chen, and Y. Cheng. A 3-parameter gompertz distribution for survival data with competing risks, with an application to breast cancer data. *Journal of Applied Statistics*, 43(12):2239–2253, 2016. doi: 10.1080/02664763.2015.1134450. URL https://doi.org/10.1080/02664763.2015.1134450.

Jr Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982. ISSN 0098-7484. doi: 10.1001/jama.1982.03320430047030. URL https://doi.org/10.1001/jama.1982.03320430047030.

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

030814.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2005.030814.x.

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. doi: https://doi.org/10.1111/j.0006-341X.2000.00337.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2000.00337.x.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

Hyokyoung G. Hong, Xuerong Chen, David C. Christiani, and Yi Li. Integrated powered density: Screening ultrahigh dimensional covariates with survival outcomes. *Biometrics*, 74(2):421–429, 2018.

Jong-Hyeon Jeong and Jason P. Fine. Parametric regression on cumulative incidence function. *Biostatistics*, 8(2):184–196, 04 2006. ISSN 1465-4644. doi: 10.1093/biostatistics/kxj040. URL https://doi.org/10.1093/biostatistics/kxj040.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Jialiang Li and Jason P. Fine. ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics*, 9(3):566–576, 02 2008. ISSN 1465-4644. doi: 10.1093/biostatistics/kxm050. URL https://doi.org/10.1093/biostatistics/kxm050.

Jialiang Li and Xiao-Hua Zhou. Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *Journal of Statistical Planning and Inference*, 139(12):4133–4142, 2009. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2009.05.043. URL https://www.sciencedirect.com/science/article/pii/S0378375809001694.

Jialiang Li, Qi Zheng, Limin Peng, and Zhipeng Huang. Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics*, 72(4):1145–1154, 2016.

Shuiyun Lu, Xiaolin Chen, Sheng Xu, and Chunling Liu. Joint model-free feature screening for ultra-high dimensional semi-competing risks data. *Computational Statistics & Data Analysis*, 147:106942, 2020. ISSN 0167-9473.

Douglas Mossman. Three-way rocs. *Medical Decision Making*, 19(1):78–89, 1999. doi: 10.1177/0272989X9901900110. URL https://doi.org/10.1177/0272989X9901900110. PMID: 9917023.

Christos T. Nakas and Constantin T. Yiannoutsos. Ordered multiple-class roc analysis with continuous measurements. *Statistics in Medicine*, 23(22):3437–3449, 2004. doi: https://doi.org/10.1002/sim.1917. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1917.

Jing Pan, Yuan Yu, and Yong Zhou. Nonparametric independence feature screening for ultrahigh-dimensional survival data. *Metrika*, 81(7):821–847, 2018. ISSN 1435-926X.

Mengjiao Peng. *Analysis of Complex Survival Data Subject to Semi-competing Risks*. PhD thesis, Nanyang Technological University, Singapore, 2019.

Margaret Sullivan Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

P. Saha and P. J. Heagerty. Time-dependent predictive accuracy in the presence of competing risks. *Biometrics*, 66(4):999–1011, 2010. doi: https://doi.org/10.1111/j.1541-0420.2009.01375.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2009.01375.x.

Thomas H. Scheike, Mei-Jie Zhang, and Thomas A. Gerds. Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95(1):205–220, 02 2008. ISSN 0006-3444. doi: 10.1093/biomet/asm096. URL https://doi.org/10.1093/biomet/asm096.

Rotraut Schoop, Jan Beyersmann, Martin Schumacher, and Harald Binder. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112, 2011. doi: https://doi.org/10.1002/bimj.201000073. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201000073.

Brian K. Scurfield. Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, 40(3):253–269, 1996. ISSN 0022-2496. doi: https://doi.org/10.1006/jmps.1996.0024. URL https://www.sciencedirect.com/science/article/pii/S0022249696900243.

Haiwen Shi, Yu Cheng, and Jong-Hyeon Jeong. Constrained parametric model for simultaneous inference of two cumulative incidence functions. *Biometrical Journal*, 55(1):82–96, 2013. doi: https://doi.org/10.1002/bimj.201200011. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201200011.

Rui Song, Wenbin Lu, Shuangge Ma, and X. Jessie Jeng. Censored rank independence screening for high-dimensional survival data. *Biometrika*, 101(4):799–814, 10 2014. ISSN 0006-3444.

Ben Van Calster, Vanya Van Belle, Yvonne Vergouwe, Dirk Timmerman, Sabine Van Huffel, and Ewout W. Steyerberg. Extending the c-statistic to nominal polytomous outcomes: the polytomous discrimination index. *Statistics in Medicine*, 31(23):2610–2626, 2012. doi: https://doi.org/10.1002/sim.5321. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5321.

Marc J. van de Vijver, Yudong D. He, Laura J. van 't Veer, Hongyue Dai, Augustinus A.M. Hart, Dorien W. Voskuil, George J. Schreiber, Johannes L. Peterse, Chris Roberts, Matthew J. Marton, Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Leonie Delahaye, Tony van der Velde, Harry Bartelink, Sjoerd Rodenhuis, Emiel T. Rutgers, Stephen H. Friend, and René Bernards. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002. doi: 10.1056/NEJMoa021967. URL https://doi.org/10.1056/NEJMoa021967. PMID: 12490681.

Marcel Wolbers, Paul Blanche, Michael T. Koller, Jacqueline C. M. Witteman, and Thomas A. Gerds. Concordance for prognostic models with competing risks. *Biostatistics*, 15(3):526–539, 02 2014. ISSN 1465-4644. doi: 10.1093/biostatistics/kxt059. URL https://doi.org/10.1093/biostatistics/kxt059.

Cai Wu and Liang Li. Quantifying and estimating the predictive accuracy for censored time-to-event data with competing risks. *Statistics in Medicine*, 37(21):3106–3124, 2018. doi: https://doi.org/10.1002/sim.7806. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7806.

Song Zhang, Yang Qu, Yu Cheng, Oscar L. Lopez, and Abdus S. Wahed. Prognostic accuracy for predicting ordinal competing risk outcomes using ROC surfaces. *Lifetime Data Analysis*, 2021. ISSN 1572-9249. doi: 10.1007/s10985-021-09539-z. URL https://doi.org/10.1007/s10985-021-09539-z.

Sihai Dave Zhao and Yi Li. Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105(1):397 – 411, 2012. ISSN 0047-259X.

Yingye Zheng, Tianxi Cai, Yuying Jin, and Ziding Feng. Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics*, 68(2):388–396, 2012. doi: https://doi.org/10.1111/j.1541-0420.2011.01671.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2011.01671.x.