Lifetime Data Analysis (2022) 28:1–22
https://doi.org/10.1007/s10985-021-09539-z

# Prognostic accuracy for predicting ordinal competing risk outcomes using ROC surfaces

**Song Zhang[1] · Yang Qu[2] · Yu Cheng[3] · Oscar L. Lopez[4] · Abdus S. Wahed[1]**

Received: 5 November 2020 / Accepted: 8 October 2021 / Published online: 22 November 2021


## Abstract

Many medical conditions are marked by a sequence of events in association with continuous changes in biomarkers. Few works have evaluated the overall accuracy of a biomarker in predicting disease progression. We thus extend the concept of receiver operating characteristic (ROC) surface and the volume under the surface (VUS) from multi-category outcomes to ordinal competing-risk outcomes that are also subject to noninformative censoring. Two VUS estimators are considered. One is based on the definition of the ROC surface and obtained by integrating the estimated ROC surface. The other is an inverse probability weighted U estimator that is built upon the equivalence of the VUS to the concordance probability between the marker and sequential outcomes. Both estimators have nice asymptotic results that can be derived using counting process techniques and U-statistics theory. We illustrate their good practical performances through simulations and applications to two studies of cognition and a transplant dataset.

**Keywords** Concordance probability · Correct classification probability · Discriminative capability · Disease progression · Inverse probability of censoring weighting

✉ Yu Cheng
yucheng@pitt.edu

1 Department of Biostatistics, University of Pittsburgh, Pennsylvania, PA 15260, USA

2 Department of Statistics, University of Pittsburgh, Pennsylvania, PA 15260, USA

3 Departments of Statistics and Biostatistics, University of Pittsburgh, Pennsylvania, PA 15260, USA

4 Departments of Neurology, Psychiatry, and Clinical and Translational Sciences, University of Pittsburgh, Pennsylvania, PA 15260, USA

## 1 Introduction

In biomedical studies, it is often of interest to measure a biomarker's predictive power for future events. Accurate prognostic analysis can help clinicians screen high-risk subjects, which in turn will lead to timely and efficient therapeutic interventions and reduced mortality and morbidity. A time-dependent ROC curve was proposed by Heagerty et al. (2000) to extend diagnostic accuracy analysis from a binary outcome to a typical survival outcome by summarizing sensitivity and specificity at a specific time $t_0$. Assuming a lower biomarker value is associated with a worse outcome, the area under the ROC curve (AUC) can be interpreted as a concordant probability that for a randomly selected pair of a case (i.e., developing the event of interest by $t_0$) and a control (no event by $t_0$), the biomarker value from the case is lower than that from the control.

Competing risk censoring commonly arises in studies where subjects are at risk for multiple failures, and one failure precludes or alters the occurrence probability of the others (Gooley et al. 1999). For instance, in a liver transplant study that we discuss later on, transplant and death without transplant are two competing events. With competing risk censoring, a popular method for measuring the cumulative probability of the target event by a specific time is the cumulative incidence function (CIF) (Prentice et al. 1978; Kalbfleisch and Prentice 2002). It has been widely employed for its intuitive probability interpretation and non-parametric identifiability.

Expanding the ROC concept to competing risk censoring, Saha and Heagerty (2010) proposed estimating the sensitivity with cumulative cases accruing to a fixed time and estimating the specificity among "healthy" control subjects (i.e., those without developing any event yet). Following similar definitions of sensitivity and specificity, Zheng et al. (2012) evaluated prognostic accuracy with multiple covariates using both Cox (1959) and more flexible Scheike et al. (2008) models. Blanche et al. (2013) and Wolbers et al. (2014) used nonparametric inverse probability of censoring weighting (IPCW) to derive the estimators of AUCs and their asymptotic properties. All these analyses compare cases from each cause to the event-free controls one at a time. One limitation of such an approach is that there is no overall predictive accuracy assessment across all events simultaneously, because different cases are considered for each specific cause in separate ROC analyses. In contrast, Shi et al. (2014) evaluated the improved accuracy of new markers for competing outcomes by defining "controls" that combine both event-free subjects and those subjects who have developed competing events. Though the definition of the "control" group is in line with the augmented "at-risk" set in Fine and Gray (1999), it may not be ideal if subjects with competing events are very different from those with no event. The existing methods either have to force unnatural grouping of competing events with the healthy state or provide evaluation of only one specific cause each time against the healthy state. It would be of interest to evaluate a biomarker's discriminatory power on competing risk outcomes simultaneously, analogous to an overall assessment of group differences in addition to pairwise group comparisons in a multi-group comparison setting.

Medical conditions often manifest a natural ordinal disease status. For example, in the MYHAT and AD studies, disease progresses through several sequential stages: healthy, MCI/dementia, and then death. An important aim in clinical practices is to

characterize a sequence of progression based on continuous changes in a prognostic biomarker. With multi-level progressive status, Obuchowski (2005) pointed out that simply dichotomizing an ordinal outcome led to an upward bias in diagnostic testing using the ROC curve. Mossman (1999) introduced the concept of multi-dimensional ROC surface and Volume Under the ROC Surface (VUS) to evaluate discriminatory accuracy of two diagnostic tests for three disease categories. The variance of Mossman (1999)'s VUS estimator was derived by Dreiseitl et al. (2000) based on the theory of U-statistics. Li and Fine (2008) applied the concept of VUS to unordered multilevel categorical outcomes and further expanded multi-way ROC analysis and the summary statistics of Hypervolume Under the ROC Manifold (HUM). Li and Zhou (2009) studied the VUS using nonparametric and semiparametric methods and developed asymptotic properties of the estimators. Wu and Chiang (2013) showed through a rigorous proof that HUM is directly related to an explicit U-estimator.

However, there is no well-developed method to assess the global accuracy in a biomarker's prediction of which stage of disease progression a subject would land on by a specific time. Thus, we propose to utilize the concept of the ROC surface and the VUS to demonstrate the prognostic accuracy of a biomarker for competing risk outcomes. As in the cognitive studies here we focus on competing events with a natural order in severity as in the congitive studies and introduce the concept of the ROC surface and the VUS from two perspectives. One is to employ the building blocks of an ROC surface, namely correct classification probabilities (CCPs), to derive the VUS. The second is to measure a concordance probability between a biomarker and ordinal competing risk outcomes. This extension is certainly nontrivial. As the event status changes over time, the ROC surface and the VUS are now time dependent. Moreover, right censoring (e.g., administrative censoring) is often present in competing risk data and may lead to an indeterminable event status at a time of interest. Thus, the development of the ROC surface and the VUS for competing risk outcomes requires additional methodology to handle "missing" event status. We will show through simulations that ignoring indeterminable censored observations could lead to substantial bias in estimating the VUS.

## 2 ROC surface for ordinal competing risk outcomes

### 2.1 Notation

Without loss of generality, we now consider the prognostic accuracy of one single biomarker in predicting two ordered competing risk outcomes, called a cause-1 event and a cause-2 event. The cause-1 event is assumed to be a worse medical condition as compared to healthy controls. Let $Y$ denote a prognostic biomarker where lower values correspond to worse medical conditions. Let $T$ be the time to any ordinal competing events, and $\epsilon = 1, 2$ be the corresponding cause of failure. Even though the disease in general progresses from mild cognitive impairment to a more severe form and then to death, a subject may not be followed continuously and the first detected event could be dementia without MCI, or a subject may die without dementia due to some other disease processes. In the MYHAT and AD studies, as participants underwent annual

assessment for cognition, we combined MCI and dementia together and simply refer to as "impaired" or "impairment." If both competing events (e.g., impairment and death) have occurred at a fixed time $t_0$, the more severe progression time is recorded as $T$ and $\epsilon$ is set to be 1. We then define disease status $D(t_0)$ as:

$$\begin{cases} D\,(t_0) = 1, \text{ if } T \le t_0, \epsilon = 1, \\ D\,(t_0) = 2, \text{ if } T \le t_0, \epsilon = 2, \\ D\,(t_0) = 0, \text{ if } T > t_0. \end{cases}$$

In practice, there may be administrative censoring $C$ on top of competing risks. Thus, we observe $X = \min(T, C)$ and the combined cause indicator $\eta = I\,(T \le C)\,\epsilon$, where $I\,(\cdot)$ is an indicator function. The observed data consist of i.i.d replicates $\{(Y_i, X_i, \eta_i),\ i = 1, \ldots, n\}$.

### 2.2 ROC surface

Analogous to sensitivity and specificity that describe the accuracy level by specifying a series of cutpoints along with a continuous classifier for a binary outcome, we need two cutpoints $(c_1, c_2) \in \mathcal{R}^2$ from $Y$ with $c_1 \le c_2$ for the two competing events, where we assign a subject to Class 1 if their biomarker $Y \le c_1$, to Class 2 if $c_1 < Y \le c_2$, and to Class 3, otherwise. Correct classification probabilities (CCPs) are then defined for subjects experiencing a cause-1 event, a cause-2 event, or none of the events at a given time $t_0$ as follows:

$$\begin{aligned} CCP_1 &= P\,(Y_i \le c_1 \mid T_i \le t_0, \epsilon_i = 1) = F_{Y|1}(c_1), \\ CCP_2 &= P\,(c_1 < Y_i \le c_2 \mid T_i \le t_0, \epsilon_i = 2) = F_{Y|2}(c_2) - F_{Y|2}(c_1), \\ CCP_3 &= P\,(Y_i > c_2 \mid T_i > t_0) = 1 - F_{Y|3}(c_2), \end{aligned}$$

where $F_{Y|d}(y) = P(Y \le y|D(t_0) = d), d = 1, 2, 3$, are conditional cumulative distribution functions (CDFs) of $Y$, given that a subject is in disease status $d$. Similar to the ROC curve that characterizes the full spectrum of sensitivity and specificity in a two-dimensional space, the plot of $(CCP_1, CCP_2, CCP_3)$ at all possible values of $(c_1, c_2)$ generates a three-dimensional ROC surface for three-category time-dependent outcomes. Following Li and Zhou (2009), the ROC surface is defined by expressing $CCP_2$ as a function of $CCP_1$ and $CCP_3$:

$$Q(u, v) = \begin{cases} F_{Y|2}\{F_{Y|3}^{-1}(1-u)\} - F_{Y|2}\{F_{Y|1}^{-1}(v)\} & \text{if } F_{Y|1}^{-1}(v) \le F_{Y|3}^{-1}(1-u), \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The VUS is then defined as:

$$\text{VUS} = \int_0^1 \int_0^1 Q(u, v) du dv. \quad (2)$$

### 2.3 VUS as a concordance index

According to Mossman (1999), Dreiseitl et al. (2000) and Wu and Chiang (2013), the VUS corresponds to a concordance measure between a biomarker and sequential competing risk outcomes. For three randomly selected subjects 1, 2, 3 that $T_1 \leq t_0, \epsilon_1 = 1, T_2 \leq t_0, \epsilon_2 = 2$, and $T_3 > t_0$ with the corresponding biomarkers $Y_1$, $Y_2$ and $Y_3$, we can show that

$$\text{VUS} = P(Y_1 < Y_2 < Y_3 \mid T_1 \leq t_0, \epsilon_1 = 1, T_2 \leq t_0, \epsilon_2 = 2, T_3 > t_0). \tag{3}$$

## 3 Estimation

### 3.1 Nonparametric estimation of the ROC surface and the VUS

Nonparametric methods are used to estimate $F_{Y|d}(y) = P(Y \leq y | D(t_0) = d), d = 1, 2, 3$ for any pair of $(y, t_0)$. In terms of estimating the conditional distribution of $Y$ given the subjects experiencing either the cause-1 event or the cause-2 event by $t_0$, we have

$$F_{Y|k}(y) = \frac{P(Y_i \leq y, T_i \leq t_0, \epsilon_i = k)}{P(T_i \leq t_0, \epsilon_i = k)},$$

where $k = 1, 2$. The numerator is a bivariate CIF, and the denominator is a univariate CIF. We adopt the bivariate CIF estimator in Cheng et al. (2007), $\hat{F}_{Y,T}(t_0)$, to estimate the bivariate probability, which includes the completely observed $Y$ as a special case. The univariate CIF, $F_k(t_0) = P(T \leq t_0, \epsilon = k)$, is estimated by the standard nonparametric estimator $\hat{F}_k(t_0)$ (Kalbfleisch and Prentice 2002). Also, we formulate the conditional distribution of $Y$ for those subjects having no events by $t_0$ in terms of bivariate and univariate survival functions

$$F_{Y|3}(y) = \frac{P(Y_i \leq y, T_i > t_0)}{P(T_i > t_0)} = \frac{P(T_i > t_0) - P(Y_i > y, T_i > t_0)}{P(T_i > t_0)} = \frac{S_T(t_0) - S_{Y,T}(y, t_0)}{S_T(t_0)},$$

which can be estimated by using the bivariate survival estimator $\hat{S}_{Y,T}(y, t_0)$ (Dabrowska 1989) and the univariate survival estimator $\hat{S}_T(t_0)$ (Kaplan and Meier 1958). Now plugging those estimators $\hat{F}_{Y|d}, d = 1, 2, 3$ into the definition of the three-dimensional ROC surface, we have

$$\hat{Q}(u, v) = \hat{F}_{Y|2}\{\hat{F}^{-1}_{Y|3}(1 - u)\} - \hat{F}_{Y|2}\{\hat{F}^{-1}_{Y|1}(v)\}, \tag{4}$$

if $\hat{F}^{-1}_{Y|1}(v) \leq \hat{F}^{-1}_{Y|3}(1 - u)$. The resulting estimated ROC surface is an increasing step function jumping at observed event times. The estimated VUS is constructed as

$$\widehat{\text{VUS}}(t_0) = \int_0^1 \int_0^1 \hat{Q}(u, v) du dv. \tag{5}$$

We approximate $\widehat{\text{VUS}}(t_0)$ by summing up the volumes of rectangular prisms whose lengths, widths, and heights correspond to $CCP_1$, $CCP_2$, and $CCP_3$ at varying thresholds of $(c_1, c_2)$.

### 3.2 Concordance definition based estimation of VUS

Without independent right censoring, the disease status by a fixed time $D(t_0)$ would be observed for each subject in the sample. A U-type VUS estimator can be constructed by randomly selecting three subjects, one from each disease status, and computing the concordance probability between the biomarker and the outcomes. However, in the presence of independent censoring, there are four possible scenarios for the $i$th subject:

$$\begin{aligned} I\{X_i \le t_0, \eta_i = 1\} &= I\{T_i \le t_0, \epsilon_i = 1, C_i \ge T_i\} \\ I\{X_i \le t_0, \eta_i = 2\} &= I\{T_i \le t_0, \epsilon_i = 2, C_i \ge T_i\} \\ I\{X_i > t_0\} &= I\{T_i > t_0, C_i > t_0\} \\ I\{X_i \le t_0, \eta_i = 0\} &= I\{C_i \le t_0, T_i > C_i\}. \end{aligned}$$

The disease status $D(t_0)$ is determinable for the first three scenarios, but not for the fourth one. To adjust for "missing" disease status due to independent censoring before $t_0$, we adopt the idea of inverse probability of censoring weighting (IPCW). The IPCW is used to weight the observed subjects who have developed the cause-1 or cause-2 event by $t_0$, inversely proportional to their probabilities of being observed at the times of occurrence, and weight those survivors without any events at $t_0$ with the inverse probability of being censored at $t_0$. For more information on the IPCW, we refer the readers to Van der Laan and Robins (2003). Let $G(t) = P(C > t)$ be the survival function of censoring and $\hat{G}(t)$ be the Kaplan-Meier estimator of $G(t)$. Following that the Kaplan-Meier estimator is consistent (Andersen et al. 1993), $I\{X_i \le t_0, \eta_i = k\}/\hat{G}(X_i)$ is an asymptotically unbiased estimator of $P(T_i \le t_0, \epsilon_i = k)$ for $k = 1, 2$ and $I\{X_i > t_0\}/\hat{G}(t_0)$ is an asymptotically unbiased estimator of $P(T_i > t_0)$. Hence, we propose the following IPCW U-type estimator of the VUS:

$$\widehat{\text{VUS}}(t_0) = \frac{\sum_i \sum_{j\neq i} \sum_{k\neq i,j} \frac{I(X_i \le t_0, \eta_i = 1, X_j \le t_0, \eta_j = 2, X_k > t_0, Y_i < Y_j < Y_k)}{\hat{G}(X_i-)\hat{G}(X_j-)\hat{G}(t_0)}}{\sum_i \sum_{j\neq i} \sum_{k\neq i,j} \frac{I(X_i \le t_0, \eta_i = 1, X_j \le t_0, \eta_j = 2, X_k > t_0)}{\hat{G}(X_i-)\hat{G}(X_j-)\hat{G}(t_0)}}. \quad (6)$$

### 3.3 Adaption of ties in the biomarkers

We often encounter ties in a biomarker when its values are rounded to the nearest integers. To handle ties in the biomarker, we can easily modify $\widehat{\text{VUS}}(t_0)$ in Eq. (6) by substituting $I(Y_i < Y_j < Y_k)$ with $I(Y_i < Y_j < Y_k) + \frac{1}{2}I(Y_i < Y_j = Y_k) + \frac{1}{2}I(Y_i = Y_j < Y_k) + \frac{1}{6}I(Y_i = Y_j = Y_k)$ similar to the ideas used in Wang and Cheng

(2014). With the linear interpolation in the biomarker, uniform consistency and weak convergence for $\widehat{\text{VUS}}(t_0)$ still hold, following similar arguments in Section 4.2.

To appreciate how ties affect the ROC surface, let us first carry out a geometry exercise on the ROC curve for a binary outcome. We observe that the ROC curve is a step function when there are no ties. Either sensitivity or specificity changes when the ordered biomarker moves from one value to the next, as a unique value of the biomarker is associated with either a case or a control. However, in the presence of tied scores in the biomarker, when a single tied score is associated with both cases and controls, it leads to changes in both sensitivity and specificity as the biomarker moves to the next value, causing a sloped line segment in the ROC curve. Thus, the AUC can be under or overestimated, depending on how tied scores affect the concordance between the underneath true biomarker value and the outcome. This exercise can be carried over to our proposed VUS. In practice, it is reasonable to assume that ties appear in a random pattern without any systematic trend, e.g., due to rounding. Therefore, $\widehat{\text{VUS}}$ is robust against tied scores in a biomarker. Later we will show through simulations that the proposed $\widehat{\text{VUS}}(t_0)$ and $\widehat{\text{VUS}}(t_0)$ can both well handle ties in a biomarker.

## 4 Asymptotic properties and inference

### 4.1 Consistency and weak convergence of of $\widehat{VUS}(t_0)$

The $\widehat{\text{VUS}}(t_0)$ involves the estimation of the conditional CDF of the biomarker $Y$ given a disease status. Bayes' principle allows formulating the conditional distribution $F_{Y|1}(y)$ as

$$F_{Y|1}(y) = P(Y \leq y \mid D(t_0) = 1) = \frac{P\left(T_i \leq t_0, \epsilon_i = 1 \mid Y_i \leq y\right) P\left(Y_i \leq y\right)}{P(T_i \leq t_0, \epsilon_i = 1)},$$

where $P(T_i \leq t_0, \epsilon_i = 1 \mid Y_i \leq y)$ is the conditional cause-1 CIF given that $\{Y_i \leq y\}$. The same idea also applies to $F_{Y|2}(y)$ and $F_{Y|3}(y)$. Counting process and martingale theories (Kalbfleisch and Prentice 2002) were utilized to derive the influence functions of the Kaplan-Meier estimator and nonparametric estimators of CIF. $\mathbb{I}_{F_{Y|d}(y)}$, the influence functions of $F_{Y|d}(y)$ for $d = 1, 2, 3$, were developed through Taylor's expansion at the marginal and the conditional CIFs (or survival function). We further derive the asymptotic linear representation of the ROC surface $\mathbb{I}_{Q(u,v)}$ by applying Hadamard's differentiability and the functional-$\delta$ method to quantile functions and compound functions, resulting in the following theorem on the asymptotic properties of $\widehat{\text{VUS}}(t_0)$:

**Theorem 1** *Let* $v_1 > \inf\{u : F_1(u) > 0, F_2(u) > 0\}$ *and* $v_2 < \sup\{u : S_X(u) > 0\}$. $\widehat{\text{VUS}}(t_0)$ *is uniformly consistent for* $t_0 \in [v_1, v_2]$, *and has the asymptotic linear representation:*

$$n^{1/2}\{\widehat{\text{VUS}}(t_0) - \text{VUS}(t_0)\} = n^{-1/2} \sum_{i=1}^{n} \mathbb{I}_{\widehat{\text{VUS}}(t_0)} + o_p(1),$$

*where* $\mathbb{I}_{\widehat{\text{VUS}}(t_0)}$ *is the influence function of* $\widehat{\text{VUS}}(t_0)$.

The detailed proof of Theorem 1 and the influence function $\mathbb{I}_{\widehat{\text{VUS}}(t_0)}$ are given in the Supplementary Material. The variance of $\widehat{\text{VUS}}(t_0)$ is estimated as

$$\hat{\sigma}^{\,2}_{\widehat{\text{VUS}}(t_0)} = n^{-1}\sum_{i=1}^{n}\hat{\mathbb{I}}^{2}_{\widehat{\text{VUS}}(t_0)}.$$

Given that $\widehat{\text{VUS}}(t_0)$ is asymptotically normal at any fixed $t_0$,

$$\left\{\widehat{\text{VUS}}(t_0) - z_{1-\alpha/2}\hat{\sigma}_{\widehat{\text{VUS}}(t_0)}n^{-1/2}, \widehat{\text{VUS}}(t_0) + z_{1-\alpha/2}\hat{\sigma}_{\widehat{\text{VUS}}(t_0)}n^{-1/2}\right\},$$

where $z_{1-\alpha/2}$ is the corresponding standard normal quantile, provides a Wald-type $(1-\alpha)$ confidence interval.

### 4.2 Consistency and weak convergence of $\widetilde{\text{VUS}}(t_0)$

The IPCW is used to account for missing data due to independent right-censoring. Again by the consistency of the Kaplan-Meier estimator $\hat{G}(t)$, coupled with Slutsky's theorem, as $n \to \infty$, $n^{-1}\sum_{i=1}^{n} I(X_i \le t_0, \eta = 1)/\hat{G}(X_i)$ converges to

$$\begin{aligned} E\left\{\frac{I(X_i \le t_0, \eta_i = 1)}{G(X_i)}\right\} &= E\left\{E\left[\frac{I(T_i \le t_0, \epsilon_i = 1)I(T_i \le C_i)}{G(X_i)} \mid X_i, \eta_i\right]\right\} \\ &= E\left\{I(T_i \le t_0, \epsilon_i = 1)\left[\frac{E\{I(T_i \le C_i) \mid X_i, \eta_i\}}{G(X_i)}\right]\right\} \\ &= P(T_i \le t_0, \epsilon_i = 1), \end{aligned}$$

in probability. Analogously, $n^{-1}\sum_{j=1}^{n} I(X_j \le t_0, \eta_j = 2)/\hat{G}(X_j)$ and $n^{-1}\sum_{k=1}^{n} I(X_k > t_0)/\hat{G}(t_0)$ converge to $P(T_j \le t_0, \epsilon_j = 2)$ and $P(T_k > t_0)$, respectively. By the functional $\delta$-method, we can show that $\widetilde{\text{VUS}}(t_0)$ converges to $P(Y_i < Y_j < Y_k \mid T_i \le t_0, \epsilon_i = 1, T_j \le t_0, \epsilon_j = 2, T_k > t_0)$, given the observed data $(X_i, X_j, X_k, \eta_i, \eta_j, \eta_k, Y_i, Y_j, Y_k)$.

Weak convergency of $\widetilde{\text{VUS}}$ can be established using counting process techniques and the theory of U-statistics. More specifically, we adapted the proof from Hung and Chiang (2010) for typical survival data without competing risks. For a fixed $t_0$, we have

$$\frac{\hat{G}(t_0)}{G(t_0)} - 1 = -n^{-1}\sum_{i=1}^{n}\int_0^{t_0}\frac{dM_{C_i}(u)}{S(u)} + o_p(1), \tag{7}$$

where $M_{C_i}(t_0) = I(X_i \le t_0, \eta_i = 0) - \int_0^{t_0} I(X_i \ge t_0)d\Lambda_C(u)$, and $\Lambda_C(\cdot)$ is a cumulative hazard function of the censoring time $C$. $\Lambda_C(\cdot)$ can be estimated by the Nelson-Aalen estimator $\hat{\Lambda}_C(\cdot)$, and $S(t_0) = P(X > t_0)$ can be estimated by the Kaplan-Meier estimator $\hat{S}(t_0)$.

**Theorem 2** *Given $C$ is independent of $T$ and $t_0 \in [v_1, v_2]$ as in Theorem 1, we have*

$$n^{1/2}\{\widehat{\mathrm{VUS}}(t_0) - \mathrm{VUS}(t_0)\} = n^{-1/2}\sum_{i=1}^{n} \mathbb{I}_{\widehat{\mathrm{VUS}}(t_0)} + o_p(1),$$

*where $\mathbb{I}_{\widehat{\mathrm{VUS}}(t_0)}$, given in Eq. (9), is the influence function of $\widehat{\mathrm{VUS}}(t_0)$.*

***Proof*** We first define $\tilde{\mathbb{A}}_{ijk} = I(X_i \le t_0, \eta_i = 1, X_j \le t_0, \eta_j = 2, X_k > t_0, Y_i > Y_j > Y_k)/\{G(X_i)G(X_j)G(t_0)\}$. Let $\mathbb{A} = E(\tilde{\mathbb{A}}_{ijk})$ and derive $\hat{\mathbb{A}}_{ijk}$ by plugging in the Kaplan-Meier estimator $\hat{G}(\cdot)$ for $G(\cdot)$ in $\tilde{\mathbb{A}}_{ijk}$. Let $\mathbb{C}_{ijk} = I(X_i \le t_0, \eta_i = 1, X_j \le t_0, \eta_j = 2, X_k > t_0, Y_i > Y_j > Y_k)$. By Taylor's expansion, we have $\mathbb{C}/\{\hat{G}(X_i)\hat{G}(X_j)\hat{G}(t_0)\} =$

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

By Eq. (7) and the estimation theory of U-statistics, we have $n^{1/2}(\hat{\mathbb{A}} - \mathbb{A}) =$

$$\frac{n^{1/2}}{\binom{n}{6}} \sum_{i\neq j\neq k\neq p\neq q\neq r} \left[\tilde{\mathbb{A}}_{ijk}\left\{1 + \int_0^{X_i} \frac{dM_{C_p}(u)}{S(u)} + \int_0^{X_j} \frac{dM_{C_q}(u)}{S(u)} + \int_0^{t_0} \frac{dM_{C_r}(u)}{S(u)}\right\} - \mathbb{A}\right] + o_p(1),$$

where $\binom{n}{m}$ is the number of combinations in choosing an $m$-element subset from $n$ objects.

Next we define $\tilde{\mathbb{B}}_{ijk} = \frac{I(X_i \le t_0, \eta_i = 1, X_j \le t_0, \eta_j = 2, X_k > t_0)}{G(X_i)G(X_j)G(t_0)}$, and similarly $\mathbb{B}$ and $\hat{\mathbb{B}}$. The expression for $n^{1/2}(\hat{\mathbb{B}} - \mathbb{B})$ is the same as $n^{1/2}(\hat{\mathbb{A}} - \mathbb{A})$ except for substituting $\hat{\mathbb{A}}_{ijk}$ and $\mathbb{A}$ with $\hat{\mathbb{B}}_{ijk}$ and $\mathbb{B}$ respectively. Applying Taylor's expansion again yields $\frac{\hat{\mathbb{A}}}{\hat{\mathbb{B}}} - \frac{\mathbb{A}}{\mathbb{B}} = \frac{(\hat{\mathbb{A}} - \mathbb{A}) - \frac{\mathbb{A}}{\mathbb{B}}(\hat{\mathbb{B}} - \mathbb{B})}{\mathbb{B}} + o_P(1)$. Assembling the aforementioned results, we derive

$$\sup_{t_0} \left| n^{1/2}\{\widehat{\mathrm{VUS}}(t_0) - \mathrm{VUS}(t_0)\} - \frac{n^{1/2}}{\binom{n}{6}} \sum_{i\neq j\neq k\neq p\neq q\neq r} \Psi_{ijkpqr}(t_0) \right| = o_p(1), \quad (8)$$

where

$$\Psi_{ijkpqr}(t_0) = \frac{1}{\mathbb{B}}\left(\tilde{\mathbb{A}}_{ijk}\left\{1 + \int_0^{X_i} \frac{dM_{C_p}(u)}{S(u)} + \int_0^{X_j} \frac{dM_{C_q}(u)}{S(u)} + \int_0^{t_0} \frac{dM_{C_r}(u)}{S(u)}\right\} - \mathbb{A} - \frac{\mathbb{A}}{\mathbb{B}}\left[\tilde{\mathbb{B}}_{ijk}\left\{1 + \int_0^{X_i} \frac{dM_{C_p}(u)}{S(u)} + \int_0^{X_j} \frac{dM_{C_q}(u)}{S(u)} + \int_0^{t_0} \frac{dM_{C_r}(u)}{S(u)}\right\} - \mathbb{B}\right]\right).$$

Applying Hájek's projection theory (Van Der Vaart 1998) to this U statistic with a 6-degree kernel function, we have

$$\frac{n^{1/2}}{\binom{n}{6}} \sum_{i\neq j\neq k\neq p\neq q\neq r} \Psi_{ijkpqr}(t_0) = n^{-1/2} \sum_{i=1}^{n} \mathbb{I}_{\widetilde{\text{VUS}}}(X_i, \eta_i, Y_i, t_0) + o_p(1),$$

where

$$\begin{aligned}\mathbb{I}_{\widetilde{\text{VUS}}}(X_i, \eta_i, Y_i, t_0) = E\{&\Psi_{ijkpqr}(t_0) + \Psi_{jikpqr}(t_0) + \Psi_{jkipqr}(t_0) + \Psi_{jkpiqr}(t_0) \\ &+\Psi_{jkpqir}(t_0) + \Psi_{jkpqri}(t_0) \mid (X_i, \eta_i, Y_i)\}. \end{aligned} \tag{9}$$

As $E[\mathbb{I}_{\widetilde{\text{VUS}}}(X_i, \eta_i, Y_i, t_0)] = 0$, the variance of $\widetilde{\text{VUS}}$ can be estimated by $\hat{\sigma}^2_{\widetilde{\text{VUS}}(t_0)} = 1/n \sum_{i=1}^{n} \hat{\mathbb{I}}_{\widetilde{\text{VUS}}}(X_i, \eta_i, Y_i, t_0)^2$. These conditional expectations can be estimated by their corresponding sample means. However, the implementation of this variance estimator in R is not trivial in order to avoid slow loops of summing over six indices. We have used matrix operations and will make our R code available once the article is accepted. An asymptotic $(1-\alpha)$ confidence interval of $\text{VUS}(t_0)$ is $\left\{\widetilde{\text{VUS}}(t_0) - z_{1-\alpha/2}\hat{\sigma}_{\widetilde{\text{VUS}}(t_0)}n^{-1/2},\ \widetilde{\text{VUS}}(t_0) + z_{1-\alpha/2}\hat{\sigma}_{\widetilde{\text{VUS}}(t_0)}n^{-1/2}\right\}$.

## 5 Simulation studies

In this section, we conducted simulations to assess the performance of our proposed two VUS estimators. The following simulation strategy was adopted. We first generated a continuous biomarker variable through a uniform distribution as $Y \sim \text{Uniform}(0, U]$, and specified two points $(c_1, c_2)$, $0 < c_1 < c_2 < U$, to break $Y$ into three subsets $(0, c_1], (c_1, c_2], (c_2, U]$. We then generated 3 subsets with an equal number of pairs $(T_d, \epsilon_d)$, where $T_d \sim e^{-(\beta_d+\gamma_d)}$ is the time to the first event, and $\epsilon_d = I(\phi_d = 1) + 2I(\phi_d = 0)$ is the type of the event with $\phi_d \sim \text{Bernoulli}(\beta_d/(\beta_d+\gamma_d)), d = 1, 2, 3$. The parameters $\beta_d$ and $\gamma_d$ determine how cause-1 and cause-2 events are associated with the biomarker $Y$ through the piecewise hazard functions $\lambda_1(t) = \exp\{\beta_1 I(Y \le c_1) + \beta_2 I(c_1 < Y \le c_2) + \beta_3 I(Y > c_2)\}$, and $\lambda_2(t) = \exp\{\gamma_1 I(Y \le c_1) + \gamma_2 I(c_1 < Y \le c_2) + \gamma_3 I(Y > c_2)\}$, where $\lambda_1$ and $\lambda_2$ are the cause-specific hazard functions for cause-1 and cause-2 events. We let $T = I(0 < Y \le c_1)T_1 + I(c_1 < Y \le c_2)T_2 + I(c_2 < Y \le U)T_3$ and $\epsilon = I(0 < Y \le c_1)\epsilon_1 + I(c_1 < Y \le c_2)\epsilon_2 + I(c_2 < Y \le U)\epsilon_3$.

We evaluated the performance of the estimators under various scenarios, including three settings with different predictive powers of the biomarker to the competing events. In all simulations, we assigned $U = 20$ and let $c_1 = 5, c_2 = 10$. The first setting represented a strong association between the two by defining $\beta_1 = 50$ and $\gamma_1 = 3$ when $Y \in (0, 5]$; $\beta_2 = 3$ and $\gamma_2 = 50$ when $Y \in (5, 10]$; and $\beta_3 = 1$ and $\gamma_3 = 1$ when $Y \in (10, 20]$. Here we allow the possibility of both events across all marker values. By assigning a much larger value of $\beta_1$ as compared to $\gamma_1$, those individuals with marker values $Y$ in the interval $(0, 5]$ are more likely to experience the

cause-1 event (the most severe state). In contrast, those with marker values $Y \in (5, 10]$ are more likely to experience the cause 2 event (the less severe state) because $\beta_2$ is much smaller than $\gamma_2$. For those with marker values above 10, both $\beta_3$ and $\gamma_3$ are small and thus they are more likely not to experience any event (the healthy state) by a certain time point. The second setting simulated a moderate predictive power by setting $\beta_1 = 8.0$, $\gamma_1 = 3.5$, $\beta_2 = 3.5$, $\gamma_2 = 3.0$, $\beta_3 = 1.0$ and $\gamma_3 = 2.0$. The third one was for the null where $\beta_d = 3$ and $\gamma_d = 3$, $d = 1, 2, 3$. For each strength of association, we selected three different $t_0$s at which 25-30%, 65-70%, and 75-85% subjects experienced the events. Independent censoring times $C$ were generated from Uniform distributions with censoring rates of 15%, 30% or 50%. We chose sample size=150, 300 with 1000 simulated datasets for each combination of scenarios.

We computed both $\widetilde{\text{VUS}}(t_0)$ and $\widehat{\text{VUS}}(t_0)$ for each simulated dataset, and compared their performance with an existing nonparametric VUS estimator proposed by Li and Zhou (2009), which is denoted as $\widehat{\text{VUS}}_{NP}(t_0)$ here to distinguish it from our proposed estimators. However, $\widehat{\text{VUS}}_{NP}(t_0)$ has not been designed for survival outcomes and cannot deal with censoring. As a result, we removed all observations censored by $t_0$ when estimating $\widehat{\text{VUS}}_{NP}(t_0)$, since the disease status $D(t_0)$ is indeterminable for this particular group. The IPCW adjusted standard error $\hat{\sigma}_{\widehat{\text{VUS}}}$ was calculated based on Eq. (9). For $\widehat{\text{VUS}}(t_0)$, its variance can be estimated from the influence function in Theorem 4.1. However, the evaluation of the influence function is rather complicated. Bootstrap with 250 replications was used instead to estimate standard error $\hat{\sigma}_{\widehat{\text{VUS}}}$. For $\widehat{\text{VUS}}_{NP}(t_0)$, $\hat{\sigma}_{\widehat{\text{VUS}}_{NP}}$ was estimated also based on 250 bootstrap samples.

To assess the performance under tied scores in a biomarker, we generated tied scores by rounding $Y \in (0, 5]$ or $(10, 20]$ to the nearest 0.1 and rounding $Y \in (5, 10]$ to the nearest 0.2. This design emulated real clinical studies in which some ranges of values occur more frequently than others. We also tested the scenarios without ties for sample sizes 150 and 300. However, for the sake of space, only the results from tied data with $n = 300$ are reported here, and results from untied data are provided in the Supplementary Materials. For biomarkers with ties, we compared the estimates with the true VUS, which was obtained based on a large sample without censoring (n=300,000), and report the biases $B_{\widetilde{\text{VUS}}}$ for $\widetilde{\text{VUS}}(t_0)$, $B_{\widehat{\text{VUS}}}$ for $\widehat{\text{VUS}}(t_0)$, and $B_{\widehat{\text{VUS}}_{NP}}$ for $\widehat{\text{VUS}}_{NP}(t_0)$ in Table 1. $ESE_{\widetilde{\text{VUS}}}$ (empirical standard error), $ASE_{\widetilde{\text{VUS}}}$ (average of model-based standard errors), and $CP_{\widetilde{\text{VUS}}}$ (coverage probability) for $\widetilde{\text{VUS}}(t_0)$ are also reported in Table 1 at different time points, based on 1000 samples of size 300 with 15%, 30% or 50% censoring. For $\widehat{\text{VUS}}(t_0)$ and $\widehat{\text{VUS}}_{NP}(t_0)$, we report in Table 1 their empirical standard errors $ESE_{\widehat{\text{VUS}}}$ and $ESE_{\widehat{\text{VUS}}_{NP}}$, average bootstrap standard errors $BSE_{\widehat{\text{VUS}}}$ and $BSE_{\widehat{\text{VUS}}_{NP}}$, and their coverage probabilities $CP_{\widehat{\text{VUS}}}$ and $CP_{\widehat{\text{VUS}}_{NP}}$. The average frequencies of cause-1 event, cause-2 event, survivors and censored subjects are reported in Supplementary Material.

Table 1 suggests that both $\widetilde{\text{VUS}}$ and $\widehat{\text{VUS}}$ perform well especially when the censoring rate is not too high. Their mean values are barely different from true values, and their coverage rates are close to the nominal level 0.95. The model-based standard error, the bootstrap standard error, and the empirical standard errors from the two estimators all agree well with each other. In general, the standard errors increase with the increasing censoring rate noticeably at later time points. At 50% censoring, $\widetilde{\text{VUS}}$ still has good

performance, while $\widehat{\text{VUS}}$ is slightly less robust to the increased censoring rate with a larger bias when the predictive power is strong. Compared to the two proposed estimators, the existing estimator $\widehat{\text{VUS}}_{NP}(t_0)$ cannot handle censored outcomes well, and has much larger bias at later time points where more observations are censored with $D(t_0)$ indeterminable.

When there is no association between a biomarker and outcomes, the means of the two proposed estimators barely deviate from the true value 0.167 and the coverage rates are very close to 95%, which ensures the accurate type I error. Interestingly, when we look closely at the two VUS estimates from each simulated dataset (data not shown here), the discrepancies between the two are more apparent under moderate association or non-informative association than under strong association, even though their means are always very close to each other across different associations. For example, the discrepancy between the two estimates from 1000 individual datasets that we simulated with 30% censoring can be as high as 26% in moderate association at $t_0 = 0.3$, and the maximum discrepancy between the two estimates is 18% for the no association case at $t_0 = 0.3$. By contrast, the discrepancy is only as much as 9% under the strong association at $t_0 = 0.05$.

Averaging across 1000 simulations, we observe similar satisfactory performances of the two estimators for untied data with $n = 300$ and $n = 150$, and the two proposed estimators had overall better performance than $\widehat{\text{VUS}}_{NP}(t_0)$; see Table 1 and 2 in the Supplementary Materials. Both proposed methods establish good behaviors even when $n = 150$, with standard errors quite close to $2^{1/2}$ times those when $n = 300$, indicating that these standard errors go to 0 as $n \to \infty$.

## 6 Applications

### 6.1 Prediction of cognitive impairment and death in MYHAT

We applied our proposed methods to the MYHAT data. Beginning in 2006, a random community sample with normal cognitive functioning to mild cognitive impairment was recruited in three age strata 65-74, 75-84, and 85+ years. Participants were followed prospectively for up to 9 annual visits at the time of data analysis for cognitive decline and onset of dementia. Their cognitive status was evaluated at baseline and follow-up visits using Clinical Dementia Rating (CDR). We classified CDR $\geq 0.5$ as the occurrence of cognitive impairment or dementia, and treated death as the worse competing event. Our analysis included 1,412 participants who were cognitively normal at baseline and completed their cognitive test battery in follow-ups.

We were interested in testing whether five cognitive domain scores of attention, executive function, language, memory, and visuospatial at baseline predict subsequent cognitive impairment and death within a 5- or 7-year window. The cognitive scores were standardized by subtracting the age-and-sex adjusted population means and dividing by their standard deviations. According to the age-stratified design, our analysis was conducted for each of the three age groups. The frequencies of observed numbers of deaths, cognitive impairment, event-free survivors, and censored participants at the selected $t_0$ of 5 years were reported in the Supplementary Material, where

**Table 1** Simulation results of the three VUS estimators for tied scores in biomarker (n=300); $\widetilde{\text{VUS}}$ is the concordance estimator, $\widehat{\text{VUS}}$ is the surface integrated estimator, and $\widehat{\text{VUS}}_{NP}$ is an existing estimator ignoring indeterminable cases; $B$ denotes bias, $ESE$ is the empirical standard error, $ASE$ is the asymptotic error, $BSE$ is the standard error based on 250 bootstrap samples, and $CP$ is the empirical coverage of 95% confidence intervals

| | | | $\widetilde{\text{VUS}}$ | | | | $\widehat{\text{VUS}}$ | | | | $\widehat{\text{VUS}}_{NP}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $t_0$ | Cen (%) | VUS | $B_{\widetilde{\text{VUS}}}$ | $ESE_{\widetilde{\text{VUS}}}$ | $ASE_{\widetilde{\text{VUS}}}$ | $CP_{\widetilde{\text{VUS}}}(\%)$ | $B_{\widehat{\text{VUS}}}$ | $ESE_{\widehat{\text{VUS}}}$ | $BSE_{\widehat{\text{VUS}}}$ | $CP_{\widehat{\text{VUS}}}(\%)$ | $B_{\widehat{\text{VUS}}_{NP}}$ | $ESE_{\widehat{\text{VUS}}_{NP}}$ | $BSE_{\widehat{\text{VUS}}_{NP}}$ | $CP_{\widehat{\text{VUS}}_{NP}}(\%)$ |
| *Strong predicative power* | | | | | | | | | | | | | | |
| 0.01 | 15 | 0.542 | 0.002 | 0.030 | 0.034 | 97.1 | −0.003 | 0.032 | 0.037 | 96.7 | 0.006 | 0.031 | 0.037 | 96.2 |
| 0.05 | 15 | 0.790 | 0.003 | 0.035 | 0.035 | 95.2 | 0.003 | 0.037 | 0.036 | 94.1 | 0.035 | 0.033 | 0.034 | 78.1 |
| 0.20 | 15 | 0.760 | 0.001 | 0.030 | 0.035 | 97.6 | −0.011 | 0.034 | 0.037 | 96.1 | −0.027 | 0.047 | 0.040 | 87.0 |
| 0.01 | 30 | 0.542 | −0.001 | 0.033 | 0.036 | 96.6 | −0.005 | 0.033 | 0.037 | 96.6 | 0.006 | 0.032 | 0.037 | 96.4 |
| 0.05 | 30 | 0.790 | −0.001 | 0.038 | 0.037 | 93.6 | −0.006 | 0.041 | 0.040 | 94.3 | 0.037 | 0.034 | 0.035 | 77.0 |
| 0.20 | 30 | 0.760 | 0.000 | 0.034 | 0.038 | 96.2 | −0.016 | 0.037 | 0.039 | 95.9 | −0.15 | 0.102 | 0.084 | 60.5 |
| 0.01 | 50 | 0.542 | 0.001 | 0.036 | 0.040 | 96.3 | −0.004 | 0.037 | 0.042 | 97.4 | 0.006 | 0.036 | 0.041 | 96.4 |
| 0.05 | 50 | 0.790 | −0.003 | 0.049 | 0.050 | 95.1 | −0.017 | 0.071 | 0.069 | 92.1 | 0.043 | 0.040 | 0.043 | 78.6 |
| 0.20 | 50 | 0.760 | 0.003 | 0.053 | 0.055 | 94.2 | −0.036 | 0.079 | 0.079 | 91.8 | −0.163 | 0.156 | 0.152 | 87.6 |

**Table 1** continued

| $t_0$ | Cen (%) | VUS | $\widetilde{\text{VUS}}$ | | | | $\widehat{\text{VUS}}$ | | | | $\widehat{\text{VUS}}_{NP}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $B_{\widetilde{\text{VUS}}}$ | $ESE_{\widetilde{\text{VUS}}}$ | $ASE_{\widetilde{\text{VUS}}}$ | $CP_{\widetilde{\text{VUS}}}(\%)$ | $B_{\widehat{\text{VUS}}}$ | $ESE_{\widehat{\text{VUS}}}$ | $BSE_{\widehat{\text{VUS}}}$ | $CP_{\widehat{\text{VUS}}}(\%)$ | $B_{\widehat{\text{VUS}}_{NP}}$ | $ESE_{\widehat{\text{VUS}}_{NP}}$ | $BSE_{\widehat{\text{VUS}}_{NP}}$ | $CP_{\widehat{\text{VUS}}_{NP}}(\%)$ |
| *Moderate predictive power* | | | | | | | | | | | | | | |
| 0.05 | 15 | 0.296 | 0.001 | 0.037 | 0.037 | 94.1 | 0.003 | 0.038 | 0.038 | 95.2 | 0.006 | 0.038 | 0.039 | 95.7 |
| 0.20 | 15 | 0.355 | 0.000 | 0.033 | 0.034 | 96.2 | 0.001 | 0.035 | 0.034 | 94.7 | 0.003 | 0.035 | 0.035 | 95.1 |
| 0.30 | 15 | 0.382 | 0.001 | 0.038 | 0.038 | 94.2 | −0.002 | 0.040 | 0.040 | 95.3 | −0.027 | 0.049 | 0.046 | 88.6 |
| 0.05 | 30 | 0.296 | −0.001 | 0.038 | 0.039 | 95.7 | 0.002 | 0.038 | 0.038 | 95.1 | 0.006 | 0.039 | 0.040 | 96.2 |
| 0.20 | 30 | 0.355 | 0.000 | 0.037 | 0.037 | 95.0 | 0.002 | 0.039 | 0.038 | 94.1 | 0.001 | 0.040 | 0.040 | 94.9 |
| 0.30 | 30 | 0.382 | 0.002 | 0.044 | 0.044 | 95.1 | −0.001 | 0.046 | 0.045 | 94.1 | −0.071 | 0.063 | 0.058 | 71.1 |
| 0.05 | 50 | 0.296 | 0.001 | 0.041 | 0.039 | 94.2 | 0.003 | 0.041 | 0.042 | 95.3 | 0.007 | 0.041 | 0.043 | 96.8 |
| 0.20 | 50 | 0.355 | −0.002 | 0.052 | 0.051 | 94.1 | −0.001 | 0.057 | 0.055 | 93.2 | −0.024 | 0.061 | 0.060 | 89.2 |
| 0.30 | 50 | 0.382 | −0.003 | 0.074 | 0.068 | 90.6 | −0.011 | 0.074 | 0.071 | 91.3 | −0.117 | 0.072 | 0.081 | 63.7 |
| *No predictive power* | | | | | | | | | | | | | | |
| 0.05 | 15 | 0.167 | −0.001 | 0.030 | 0.029 | 94.1 | 0.004 | 0.030 | 0.030 | 94.7 | 0.036 | 0.025 | 0.034 | 93.0 |
| 0.20 | 15 | 0.167 | −0.001 | 0.024 | 0.024 | 94.6 | 0.003 | 0.025 | 0.025 | 94.9 | 0.031 | 0.018 | 0.026 | 91.7 |
| 0.30 | 15 | 0.167 | −0.001 | 0.029 | 0.029 | 94.0 | 0.000 | 0.030 | 0.029 | 94.7 | 0.031 | 0.023 | 0.031 | 94.4 |
| 0.05 | 30 | 0.167 | 0.003 | 0.031 | 0.030 | 93.8 | 0.006 | 0.031 | 0..031 | 94.5 | 0.037 | 0.025 | 0.035 | 93.9 |
| 0.20 | 30 | 0.167 | 0.000 | 0.027 | 0.027 | 94.5 | 0.003 | 0.027 | 0.027 | 94.7 | 0.032 | 0.021 | 0.029 | 93.5 |
| 0.30 | 30 | 0.167 | −0.003 | 0.033 | 0.034 | 94.6 | 0.001 | 0.034 | 0.033 | 94.1 | 0.033 | 0.029 | 0.038 | 95.8 |
| 0.05 | 50 | 0.167 | 0.000 | 0.032 | 0.031 | 93.2 | 0.006 | 0.033 | 0.033 | 94.6 | 0.038 | 0.027 | 0.037 | 93.8 |
| 0.20 | 50 | 0.167 | 0.000 | 0.039 | 0.037 | 93.2 | 0.002 | 0.038 | 0.037 | 93.8 | 0.038 | 0.029 | 0.040 | 95.5 |
| 0.30 | 50 | 0.167 | −0.003 | 0.060 | 0.054 | 85.9 | −0.005 | 0.054 | 0.052 | 91.9 | 0.043 | 0.044 | 0.061 | 95.8 |

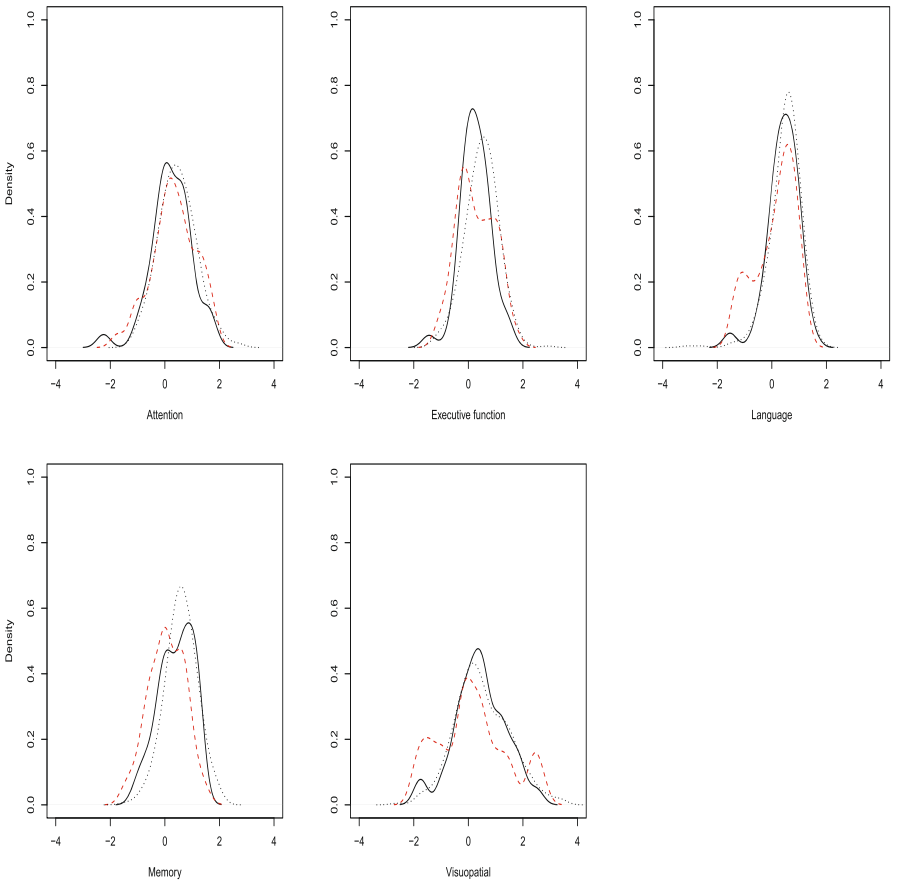


**Fig. 1** IPCW-weighted density plots of cognitive scores in the MYHAT young Group; the solid line is for those who died, the dash line is for those who had impairment, and the dotted line is for those who had experienced neither event in five years

82% of participants in the younger age group survived without dementia 5 years after the study entry, and the survivor rates were 59% and 33% in the middle and older age groups.

In Fig. 1, we present the estimated density plots for the five domain scores for those who experienced cognitive impairment (dash lines), those who died (solid lines), and those who were alive and cognitively normal (dotted lines) after 5 years of follow-up in the young age group. To handle missingness due to right censoring before 5 years, we adopted the IPCW method by weighting domain scores for each disease status as in Eq. (6), and then applied the Kernel smoothing method to generate the density plots using the R function "density." The three density curves are in general close to each other, indicating poor separability of the disease progression in five years based on those domain scores at baseline.

We next computed the VUSs using the two methods, as well as two AUC estimates for each of the competing events (death and cognitive impairment) compared

**Table 2** $VUS$ and $AUCs$ of 5 cognitive test scores at $t_0$= 5 years. $\text{AUC}_1$ compares deaths to healthy controls, and $\text{AUC}_2$ compares subjects with cognitive impairment to healthy controls

| Age group | Cognitive test | $\widehat{\text{VUS}}$ ($\hat{\sigma}_{\widehat{\text{VUS}}}$) | $\widehat{\text{VUS}}$ ($\hat{\sigma}_{\widehat{\text{VUS}}}$) | $\text{AUC}_1$ ($\hat{\sigma}_{\text{AUC}_1}$) | $\text{AUC}_2$ ($\hat{\sigma}_{\text{AUC}_2}$) |
|---|---|---|---|---|---|
| | Attention | 0.190 (0.030) | 0.170 (0.032) | 0.581 (0.048) | 0.538 (0.054) |
| | Executive | 0.176 (0.030) | 0.151 (0.027) | 0.623 (0.044) | 0.594 (0.056) |
| Young | Language | 0.168 (0.030) | 0.160 (0.033) | 0.547 (0.049) | 0.636 (0.051) |
| | Memory | 0.184 (0.036) | 0.181 (0.037) | 0.572 (0.051) | 0.722 (0.046) |
| | Visuosptial | 0.139 (0.026) | 0.146 (0.029) | 0.510 (0.048) | 0.591 (0.058) |
| | Attention | 0.194 (0.022) | 0.195 (0.021) | 0.571 (0.036) | 0.556 (0.031) |
| | Executive | 0.253 (0.027) | 0.233 (0.026) | 0.659 (0.035) | 0.671 (0.029) |
| Middle | Language | 0.199 (0.024) | 0.175 (0.022) | 0.589 (0.035) | 0.637 (0.029) |
| | Memory | 0.234 (0.028) | 0.221 (0.026) | 0.656 (0.036) | 0.750 (0.028) |
| | Visuosptial | 0.216 (0.024) | 0.216 (0.031) | 0.622 (0.037) | 0.623 (0.032) |
| | Attention | 0.276 (0.037) | 0.272 (0.033) | 0.684 (0.048) | 0.619 (0.051) |
| | Executive | 0.303 (0.039) | 0.273 (0.038) | 0.690 (0.047) | 0.648 (0.050) |
| Old | Language | 0.228 (0.034) | 0.225 (0.034) | 0.633 (0.050) | 0.633 (0.051) |
| | Memory | 0.255 (0.036) | 0.248 (0.037) | 0.730 (0.045) | 0.695 (0.048) |
| | Visuosptial | 0.209 (0.033) | 0.247 (0.038) | 0.612 (0.053) | 0.665 (0.053) |

to the event-free survivors (Table 2). The AUCs were derived using the R package "timeROC." The VUS estimates for the 5 cognitive test scores are all relatively small, which are consistent with the results from the AUCs. Consistent with what we have observed in Fig. 1, we conclude that the five domain scores at baseline exhibit poor predictive power in discriminating the sequence of the competing events in five years. Nevertheless, VUSs allow the measurement of the global concordance between the continuous cognitive test scores and the sequence of cognitive impairment and death. Moreover, the estimated VUSs from the two approaches are reasonably close and most discrepancies are less than one standard deviation. This is consistent with what we have discovered in simulations in which the two estimators differ more for weaker association. We repeated the analyses at $t_0 = 7$ years and observed similar results (not reported here). Interestingly, we observe stronger associations between baseline domain scores and disease progression in older groups at both time points, despite that the predictive power is generally weak.

### 6.2 Prediction of cognitive impairment and death in ADRC

We applied our proposed method to a dataset from the Alzheimer Disease Research Center (ADRC). AD is a neurodegenerative brain disease that causes progressive deterioration of episodic memory and a global decline of cognitive functions. Different from the MYHAT study, here our interest is in how a paticular risk factor may predict the progression from healthy state to cognitive impairment (cause 2) and further to death (cause 1), at different time points. This analysis included 390 subjects who were free of MCI/dementia at baseline. We used age at baseline as an example and evaluated

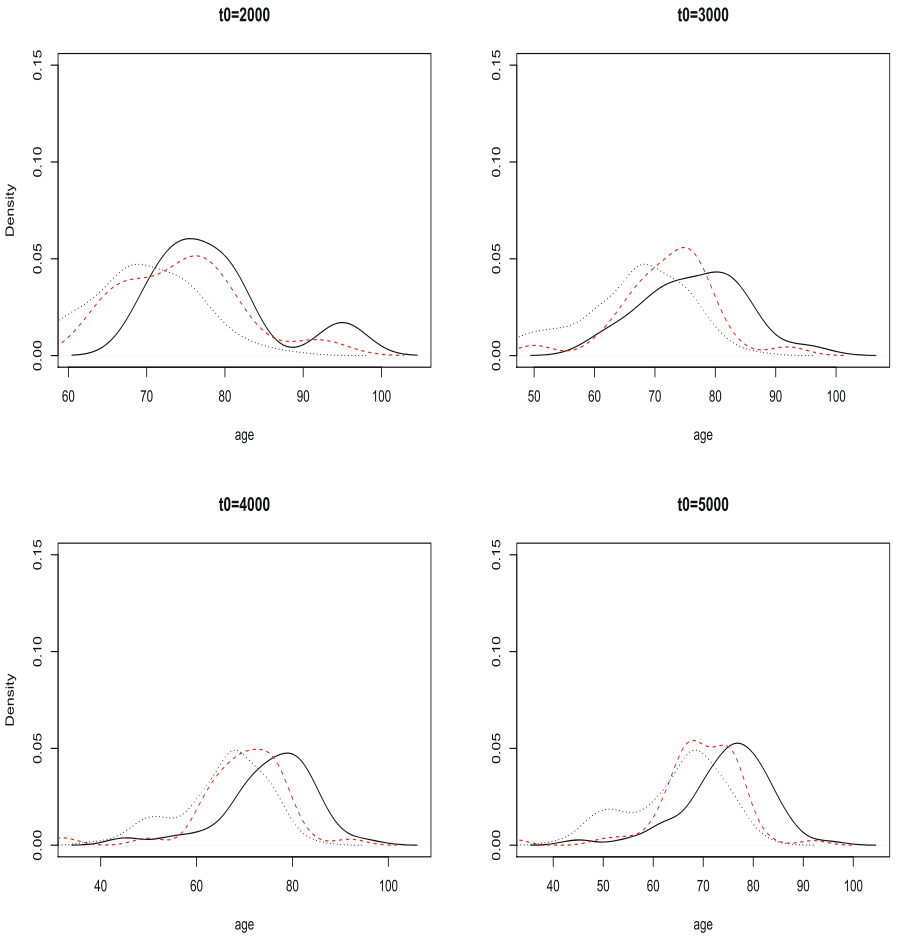


**Fig. 2** IPCW-weighted distributions of the baseline age in the ADRC study; the solid line is for those who died, the dash line is for those who had cognitive impairment, and the dotted line is for those who had experienced neither event

its prognostic accuracy at $t_0 = 2000, 3000, 4000, 5000$ days through estimated VUSs and AUCs.

To appreciate how age is associated with cognitive impairment and death over time, in Fig. 2 we present IPCW density plots of baseline age at each time point. We denoted subjects who died by black lines, those who developed MCI/dementia by dashed lines and those who survived without MCI/dementia by dotted lines. There are noticeable separations in the densities of the three classes at each time point, with older participants more likely to fall into severe categories.

We then computed VUSs using the two proposed methods with 150-age (subtracting age from 150) as the marker value to be consistent with our assumption of smaller marker values associating with more severe disease status. The AUCs were estimated for each competing event as compared to event-free survivors using the R package

**Table 3** VUS and AUC of age at $t_0$= 2000, 3000, 4000, 5000 days. $\text{AUC}_1$ compares deaths to healthy controls and $\text{AUC}_2$ compares MCI/dementia to healthy controls

| $t_0$ (days) | $\widehat{\text{VUS}}$ ($\hat{\sigma}_{\widehat{\text{VUS}}}$) | $\widehat{\text{VUS}}$ ($\hat{\sigma}_{\widehat{\text{VUS}}}$) | $\text{AUC}_1$ ($\hat{\sigma}_{\text{AUC}_1}$) | $\text{AUC}_2$ ($\hat{\sigma}_{\text{AUC}_2}$) |
|---|---|---|---|---|
| 2000 | 0.387 (0.087) | 0.321 (0.098) | 0.839 (0.055) | 0.712 (0.068) |
| 3000 | 0.379 (0.063) | 0.361 (0.066) | 0.794 (0.057) | 0.657 (0.059) |
| 4000 | 0.391 (0.047) | 0.390 (0.053) | 0.796 (0.049) | 0.633 (0.052) |
| 5000 | 0.409 (0.041) | 0.448 (0.053) | 0.807 (0.040) | 0.647 (0.046) |

**Table 4** VUS and AUC of bilirubin at $t_0$= 1000, 1500, 2000, 2500, 3000 days. $\text{AUC}_1$ compares deaths to healthy controls and $\text{AUC}_2$ compares liver transplant to healthy controls

| $t_0$ (days) | $\widehat{\text{VUS}}$ ($\hat{\sigma}_{\widehat{\text{VUS}}}$) | $\widehat{\text{VUS}}$ ($\hat{\sigma}_{\widehat{\text{VUS}}}$) | $\text{AUC}_1$ ($\hat{\sigma}_{\text{AUC}_1}$) | $\text{AUC}_2$ ($\hat{\sigma}_{\text{AUC}_2}$) |
|---|---|---|---|---|
| 1000 | 0.486 (0.045) | 0.486 (0.047) | 0.823 (0.027) | 0.799 (0.050) |
| 1500 | 0.510 (0.039) | 0.499 (0.042) | 0.856 (0.022) | 0.780 (0.049) |
| 2000 | 0.513 (0.039) | 0.502 (0.038) | 0.864 (0.022) | 0.827 (0.040) |
| 2500 | 0.416 (0.046) | 0.411 (0.047) | 0.820 (0.028) | 0.816 (0.560) |
| 3000 | 0.390 (0.046) | 0.384 (0.047) | 0.805 (0.031) | 0.822 (0.058) |

"timeROC." The results are summarized in Table 3. Age is a clear risk factor for cognitive impairment with AUCs nearly 0.8 or above at all time points. By contrast, its effect on death is less prominent.

It is worth pointing out that higher values of AUCs do not necessarily translate into higher values of VUSs, for example at $T = 2000$ vs. at $T = 5000$, and vice versa. The strength of VUS depends on the association of the marker with each competing event and whether the relationships are ordinal. This is analogous to the associations of a covariate with ordinal outcomes in a regression setting – the overall covariate effect may be stronger or weaker than individual effects depending on the ordinal relationship. There are more clear ordinal separations at later time points as shown in Fig. 2, which correspond to larger values of VUS. Across all time points, the overall effects of baseline age in separating cognitive impairment, death and healthy controls appear moderately strong.

### 6.3 Prediction of transplant and death in PBC

As the MYHAT and ADRC data are not generally available to the public, we next applied our methods to a well-known public dataset from the Mayo Clinical trial of Primary Biliary Cirrhosis (PBC) of the liver so that the analysis code can be made available to the readers upon the publication of the manuscript. The PBC trial was conducted between 1974 and 1984, and the analysis included 418 patients who met eligible criteria with complete data. Bilirubin is a prognostic biomarker for PBC since cirrhosis causes high bilirubin level. In the PBC data, bilirubin was heavily tied especially at levels below 2 mg/dL. We reversed bilirubin values using a reciprocal function to be consistent with our definition of the VUS. The aim of the analysis was to deter-

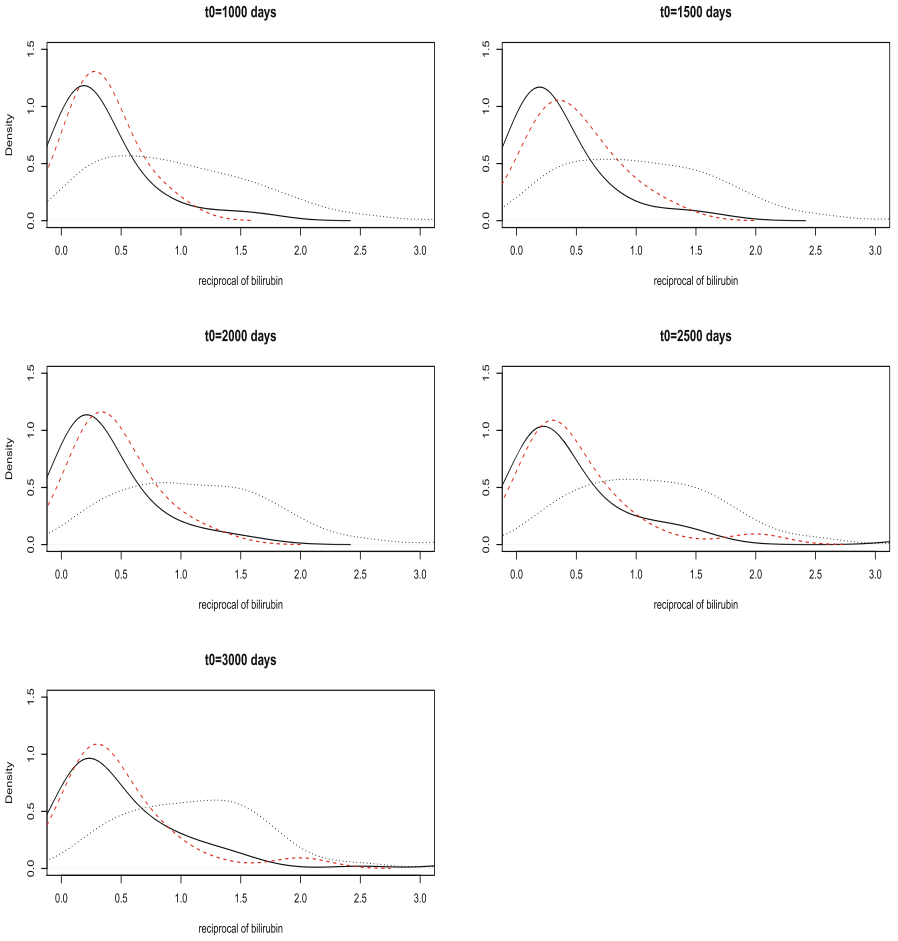


**Fig. 3** IPCW-weighted distributions of the reciprocal of bilirubin in the PBC study; the solid line is for those who died, the dash line is for those who had transplant, and the dotted line is for those who had experienced neither event

mine the time window by which bilirubin could reasonably predict liver transplant (cause-2) and death (cause-1). Since liver transplants typically took place after 500 days of follow-up while death occurred from day 41, we picked 5 time points as $t_0 = 1000, 1500, 2000, 2500, 3000$ days. Based on the estimated VUSs and AUCs (in Table 4), bilirubin presented an overall predictive power on both events across those selected times, with highest discriminative power during 4–5.5 years. Consistently, the three IPCW-adjusted density curves of the reciprocal bilirubin values for those who died, had liver transplant, or survived without transplant were most separated at $t_0 = 1500$ days; see Fig. 3 for details. As expected, the estimates were closer than the ones in the MYHAT study, as the associations between bilirubin and competing events were generally stronger.

## 7 Discussion

In this article we introduced the concept of the ROC surface and the associated volume under the ROC surface (VUS) in measuring prognostic accuracy of a continuous biomarker to competing risks outcomes, where the samples were also subject to random censoring. Our methods can be applied to both competing and semi-competing data. In our PBC data example we considered standard competing risks data where the occurrence of one event prevents the onset of the other events, i.e., liver transplant and death before transplant. In the MYHAT and ADRC examples, death can occur after cognitive impairment but not vice versa. Such semi-competing risk data occur more naturally with disease progression. Our proposed methods can be readily applied to semi-competing risk data by looking at the most severe state that an individual has been at a particular time point.

One important task in clinical decision making is to identify those subjects who will develop a certain severity in their disease progression by a specific time window, thus providing targets for better prevention or treatment. The VUS allows projecting a patient's prognostic biomarker onto three-stage disease progression and provides a complementary global discriminatory metric for assessing ordered events simultaneously. Because of the ordinal assumption, it remains important to examine weighted density plots and individual ROCs and their associated AUCs for distinct events to fully appreciate the relationship between a biomarker and competing events.

The VUS metric can be extended to higher dimensional outcomes as the HUM for the ROC manifold (Li and Fine 2008). We can estimate the HUM as a U-type statistic, as well as by integrating CCPs from multiple dimensions. Large sample properties such as consistency and weak convergence still hold for the estimated HUM, and the discussions on handling tied scores in a biomarker can be readily carried over to HUM estimators.

Currently we included a single biomarker to predict a sequence of competing events. In the MYHAT study, we evaluated discriminatory power of the five cognitive test scores on cognitive impairment and death stratified by age. It may be more informative in treating age as a covariate rather than a stratifying factor. To achieve this goal, we can fit semiparametric regression models as in Zheng et al. (2012). Moreover, if competing events are not sequential, the biomarker should not be a scalar, but rather a vector (Obuchowski 2005). Subjects are assigned to the disease status category with the highest probability given the vector of risk factors, which is estimated through some regression models. The VUS can then be defined as the concordance between the assigned disease category and the true disease status. Therefore, the extension of the VUS to nominal competing outcomes is technically feasible. This will be a topic of future research.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10985-021-09539-z.

**Acknowledgements** The authors thank Drs. Mary Ganguli and Joyce Chang for providing the MYHAT data. Cheng was partially supported by the U.S. National Science Foundation (DMS 1916001). This research was supported in part by the University of Pittsburgh Center for Research Computing through the resources provided.

## Declarations

**Conflict of interest** The authors declare that they have no conflicts of interest.

## References

Andersen PK, Borgan O, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer, New York

Blanche P, Dartigues JF, Jacqmin-Gadda H (2013) Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med 32(30):5381–5397

Cheng Y, Fine JP, Kosorok MR (2007) Nonparametric analysis of bivariate competing risks data. J Am Stat Assoc 102:1407–1416

Cox DR (1959) The analysis of exponentially distributed life-times with two types of failure. J Roy Stat Soc B 21:411–421

Dabrowska DM (1989) Kaplan-Meier estimate on the plane: Weak convergence, LIL, and the bootstrap. J Multivar Anal 29:308–325

Dreiseitl S, L O, Binder M, (2000) Comparing three-class diagnostic tests by three-way ROC analysis. Med Decis Making 20(3):323–331

Fine JP, Gray RJ (1999) A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc 94:496–509

Gooley TA, Leisenring W, Crowley J, Storer BE (1999) Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Stat Med 18:695–706

Heagerty P, Lumley T, Pepe MS (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56:337–344

Hung H, Chiang CT (2010) Estimation methods for time-dependent auc models with survival data. Can J Stat 38(1):8–26

Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, 2nd edition. Wiley, New York

Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Van der Laan MJ, Robins JM (2003) Unified methods for censored longitudinal data and causality. Springer Science & Business Media, Berlin

Li J, Fine JP (2008) ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. Biostatistics 9(3):566–576

Li J, Zhou XH (2009) Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. J Stat Plann Inference 139(12):4133–4142

Mossman D (1999) Three-way ROCs. Med Decis Making 19(1):78–89

Obuchowski NA (2005) Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary. Acad Radiol 12(9):1198–1204

Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE (1978) The analysis of failure time data in the presence of competing risks. Biometrics 12:737–751

Saha P, Heagerty PJ (2010) Time-dependent predictive accuracy in the presence of competing risks. Biometrics 66:999–1011

Scheike TH, Zhang MJ, Gerds TA (2008) Predicting cumulative incidence probability by direct binomial regression. Biometrika 95(1):205–220

Shi H, Cheng Y, Li J (2014) Assessing diagnostic accuracy improvement for survival or competing-risk censored outcomes. Can J Stat 42(1):109–125

Van Der Vaart A (1998) Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press

Wang H, Cheng Y (2014) Piecewise cause-specific association analyses of multivariate untied or tied competing risks data. Int J Biostat 10:197–220

Wolbers M, Blanche P, Koller MT, Witteman JCM, Gerds TA (2014) Concordance for prognostic models with competing risks. Biostatistics 15(3):526–539

Wu Y, Chiang C (2013) Optimal receiver operating characteristic manifolds. J Math Psychol 57(5):237–248

Zheng Y, Cai T, Jin Y, Feng Z (2012) Evaluating prognostic accuracy of biomarkers under competing risk. Biometrics 68:388–396