Biostatistics (2019), **0**, 0, *pp.* 1–29
doi:10.1093/biostatistics/IDI_NRI_CompetingRisk_final

Quantifying Diagnostic Accuracy Improvement of New Biomarkers for Competing Outcomes

ZHENG WANG

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

YU CHENG*

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

yucheng@pitt.edu

ERIC C. SEABERG

Department of Epidemiology, Johns Hopkins University, Baltimore, MD 21202, USA

JAMES T. BECKER

Departments of Psychiatry, Neurology, and Psychology, University of Pittsburgh, Pittsburgh, PA 15260, USA

SUMMARY

The net reclassification improvement (NRI) and the integrated discrimination improvement (IDI) were originally proposed to characterize accuracy improvement in predicting a binary outcome, when new biomarkers are added to regression models. These two indices have been extended from dichotomous outcomes to multi-categorical and survival outcomes. Working on an AIDS study where the onset of cognitive impairment is competing risks censored by death, we extend the NRI and the IDI to competing risk outcomes, by using cumulative incidence functions to quantify cumulative risks of competing events, and adopting the definitions of the two indices for multi-category outcomes. The "missing" category due to

*To whom correspondence should be addressed.

independent censoring is handled through inverse probability weighting. Various competing risks models are considered, such as the Fine and Gray, multistate, and multinomial logistic models. Estimation methods for the NRI and the IDI from competing risks data are presented. The inference for the NRI is constructed based on asymptotic normality of its estimator, and the bias-corrected and accelerated bootstrap procedure is applied for the IDI inference. Simulations demonstrate that the proposed inferential procedures perform very well. The Multicenter AIDS Cohort Study is used to illustrate the practical utility of the extended NRI and IDI for competing risks outcomes.

Key words: Cumulative incidence function; Fine and Gray's model; Integrated discrimination improvement; Multinomial logistic model; Multistate model; Net reclassification improvement.

1. INTRODUCTION

For clinicians, introducing a new biomarker into a statistical model may change the risks associated with various outcomes of interest, and subsequently may influence treatment decisions. Risk prediction algorithms using statistical modeling are among the most popular tools to evaluate significance of biomarkers. Although effect size and statistical significance are important, they do not provide direct information on the contribution of new biomarkers to diagnostic accuracy. For the latter, we are interested in evaluating the improvement in correctly "classifying" patients into several outcome categories, such as dementia, death and "nonevent," with the additional information from new biomarkers. In contrast, risk prediction algorithms typically attempt to predict the risks associated with each outcome in the course of time.

To investigate accuracy improvement over the course of variable additions for binary outcomes, the commonly used Receiver Operating Characteristic (ROC) curve and its corresponding Area Under the Curve (AUC) were shown to be insensitive to detecting the added values of new markers (Greenland and O'Malley, 2005; Pepe *and others*, 2004; Ware, 2006), and novel indicators were developed to complement the AUC measure (Pencina *and others*, 2008), such as the net reclassification improvement (NRI) and the

integrated discrimination improvement (IDI). The NRI is the improvement in classification rates of disease categories by the "new" model which incorporates additional markers over those by the "old" model without the additional markers. On the other hand, the IDI quantifies the improvement in the integrated sensitivity minus that of specificity over all possible cutoff values, from the model without new biomarkers to the model with new biomarkers. Both indices have become popular in medical fields and been extended from categorical outcomes to survival outcomes (Pencina *and others*, 2011; Uno *and others*, 2013).

However, there are few works in quantifying accuracy improvement for competing risks outcomes. Shi *and others* (2014) were among the first to consider accuracy improvement for competing risks, where the population is divided into two groups at a fixed time point – the "disease" group including subjects who have developed the event of interest, and the "healthy" group including those who have not had any event and those who have experienced competing events. Such a definition of the "healthy" group, which is in line with the augmented "at-risk" set in a popular regression model by Fine and Gray (1999) for competing risks data, is reasonable if competing events are not of interest, and those who have developed competing events are more or less similar to those who have not failed yet. However, there are many situations where we would like to separate subjects with competing events from those without any events. As an example, the Multicenter AIDS Cohort Study (MACS) involves two endpoints, death and dementia, where the age of dementia onset may be competing-risk censored by death. When the dementia onset is of concern, it does not seem appropriate to group those subjects who died with those who were alive and stayed healthy. Ideally they could be treated as separate categories in evaluation of accuracy improvement.

Li *and others* (2013) proposed reclassification statistics for assessing improvements in diagnostic accuracy for multi-level outcomes. In this paper, we specifically consider how the definitions of the NRI and the IDI for multi-category outcomes can be extended to the competing risks setting. The detailed definitions are given in Sections 2.2 and 2.3 for two competing risks outcomes. One issue with estimating the adapted NRI and IDI is that independent censoring often occurs in additional to competing risks censoring, and a subject's disease status may not be determinable if this subject was censored before the

time of interest. As detailed in Sections 2.2 and 2.3, the "missingness" due to censoring can be overcome by using the method of inverse probability of censoring weighting.

Demler *and others* (2017) have evaluated the feasibility of establishing U-statistics theory under different assumptions for changes in the NRI and the IDI. If the models are under the alternative, both the NRI and the IDI are non-degenerate and variance estimators based on the U-statistics theory should work, though some adjustments are needed for the IDI. The bootstrap technique is valid under this situation. On the other hand, if the models are under the null, and the models compared by the NRI and the IDI are nested, both the NRI and the IDI are degenerate and the theoretical formulas for estimating their variance do not apply. In evaluating the accuracy improvement associated with new biomarkers, we are comparing the "new" model with the additional variables and the "old" model without them. Since these two models are nested, the degeneracy of the NRI and the IDI under the null should be and can be remedied, as suggested by Demler *and others* (2017).

Though the focus of this project is to evaluate diagnostic accuracy, it remains crucial to select a proper regression model to distinguish all survival outcomes and identify covariate effects on each outcome at different time points. In this work, we adopted three models, the proportional hazards regression, Fine-Gray's model (Fine and Gray, 1999) and the multinomial logistic risk regression model (Gerds *and others*, 2012). Three simulation designs were considered in Section 3, one for each of these three models. Robustness of the NRI and the IDI towards model mis-specification was also examined. In Section 4, we applied both NRI and IDI estimators to the MACS data for assessing whether including a new biomarker, CD4 cell count, would improve predictive ability over the old model. Some discussions are given in Section 5.

2. METHODS

2.1 *Notation*

In a competing-risk setting, there are two or more types of events. To simplify the notation, only two types are considered here, which are denoted as $\epsilon = 1, 2$, though the proposed methods can be naturally

extended to more than two competing events. Let T be the time to first event from either type. With two competing events, we can define three categories according to their disease status at a fixed time point t_0. For the i-th subject, if $T_i \le t_0$ and $\epsilon_i = 1$, the subject belongs to the first category; if $T_i \le t_0$ and $\epsilon_i = 2$, the subject belongs to the second category; otherwise the subject is in the third category of being healthy. In practice there is often independent censoring C. Hence $X = \min(T; C)$ and the combined cause indicator $\eta = I(T \le C)\epsilon$ are observed. Let z_1, a p-dimension vector, denote conventional predictors and let z_2, a q-dimension vector, denote new biomarkers. The data consist of $\{X_i, \eta_i, \boldsymbol{z}_{i1}, \boldsymbol{z}_{i2} | i = 1, ..., n\}$. In the sequel we denote the "old" model with conventional markers as $\mathscr{M}_1$ and the "new" model with both conventional and new markers as $\mathscr{M}_2$.

An extension of the NRI in Li *and others* (2013) for a K-level categorical outcome D is:

$$S = \sum_{k=1}^{K} \omega_k P\{\hat{p}_k(\mathscr{M}_2) = \max \hat{\boldsymbol{p}}(\mathscr{M}_2), \hat{p}_k(\mathscr{M}_1) \neq \max \hat{\boldsymbol{p}}(\mathscr{M}_1) | D = k\}$$
$$- \sum_{k=1}^{K} \omega_k P\{\hat{p}_k(\mathscr{M}_2) \neq \max \hat{\boldsymbol{p}}(\mathscr{M}_2), \hat{p}_k(\mathscr{M}_1) = \max \hat{\boldsymbol{p}}(\mathscr{M}_1) | D = k\},$$

where ω_k is some weight function for the k-th category of the outcome, and $\sum_k \omega_k = 1$, and $\hat{p}_k(\mathscr{M}_m)$ is the estimated probability of the outcome being from the k-th category based on the model m for $m = 1, 2$. When there are only two categories $K = 2$, and the weights are $\omega_k = 1/2$ for $k = 1, 2$, then the S is equivalent to the NRI given in Pencina *and others* (2008). Li *and others* (2013) also proposed an extension of the IDI based on the relationship between the IDI and the increase in the coefficient of determination R^2 from the "old" multinomial logistic model to the "new" one with additional markers. That is, $R = \sum_{k=1}^{K} \omega_k \{R_k^2(\mathscr{M}_2) - R_k^2(\mathscr{M}_1)\}$, where ω_k is again a weight function for the k-th category of the outcome, and $R_k^2(\mathscr{M}_m)$ is the coefficient of determination from $\mathscr{M}_m$, $m = 1, 2$. Again when $K = 2$ and $\omega_k = 1/2$, the multi-category IDI reduces to the original IDI in Pencina *and others* (2008).

2.2 *Net reclassification improvement for competing outcomes*

Without loss of generality, we consider competing outcomes with three categories. For model $\mathscr{M}_m, m = 1, 2$, define $p_k(\mathscr{M}_m, t_0) = P(T \le t_0, \epsilon = k | \mathscr{M}_m)$, for $k = 1, 2$, and $p_3(\mathscr{M}_m, t_0) = P(T > t_0 | \mathscr{M}_m)$. A well-

calibrated regression model such as the multi-state (Cheng et al., 1998), Fine and Gray (1999), and Gerds (2012) models can be used. For each subject i, we obtain the estimators $\hat{p}_{1i}(\mathscr{M}_m, t_0)$, $\hat{p}_{2i}(\mathscr{M}_m, t_0)$, $\hat{p}_{3i}(\mathscr{M}_m, t_0)$, and $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \hat{p}_3)$. The NRI defined for multi-category outcomes can thus be extended to the competing-risk setting at any $t_0 > 0$:

$$\begin{aligned} S(t_0) = &\sum_{k=1}^{K} \omega_k P\{\hat{p}_k(\mathscr{M}_2, t_0) = \max \hat{\boldsymbol{p}}(\mathscr{M}_2, t_0), \hat{p}_k(\mathscr{M}_1, t_0) \neq \max \hat{\boldsymbol{p}}(\mathscr{M}_1, t_0) | D = k\} \\ &- \sum_{k=1}^{K} \omega_k P\{\hat{p}_k(\mathscr{M}_2, t_0) \neq \max \hat{\boldsymbol{p}}(\mathscr{M}_2, t_0), \hat{p}_k(\mathscr{M}_1, t_0) = \max \hat{\boldsymbol{p}}(\mathscr{M}_1, t_0) | D = k\}. \end{aligned} \tag{2.1}$$

One complication in estimating $S(t_0)$ with censored competing risks data is that not every subject status is available. For example, some subjects may have been censored before t_0, and hence their disease status cannot be determined. Therefore, those subjects whose disease status can be decided based on the observed pair $(X_i; \eta_i)$ should be properly weighted to account for those subjects with "missing" disease status due to censoring. Thus, we propose the following estimator of the NRI at any time point t_0 as:

$$\hat{S}(t_0) = \omega_1 \frac{\sum_{i=1}^{n} (h_{i,1}^{+}(t_0) - h_{i,1}^{-}(t_0))}{\sum_{i=1}^{n} I\{X_i \leq t_0, \eta_i = 1\} / \hat{G}(X_i-)} + \omega_2 \frac{\sum_{i=1}^{n} (h_{i,2}^{+}(t_0) - h_{i,2}^{-}(t_0))}{\sum_{i=1}^{n} I\{X_i \leq t_0, \eta_i = 2\} / \hat{G}(X_i-)} + \omega_3 \frac{\sum_{i=1}^{n} (h_{i,31}^{+}(t_0) - h_{i,3}^{-}(t_0))}{\sum_{i=1}^{n} I\{X_i > t_0\} / \hat{G}(t_0)},$$

$$h_{i,k}^{+}(t_0) = I\{\hat{p}_{1i}(\mathscr{M}_2, t_0) = \max \hat{\mathbf{p}}(\mathscr{M}_2, t_0), \hat{p}_{1i}(\mathscr{M}_1, t_0) \neq \max \hat{\mathbf{p}}(\mathscr{M}_1, t_0), X_i \leq t_0, \eta_i = k\} / G(X_i-), k = 1, 2,$$

$$h_{i,k}^{-}(t_0) = I\{\hat{p}_{1i}(\mathscr{M}_2, t_0) \neq \max \hat{\mathbf{p}}(\mathscr{M}_2, t_0), \hat{p}_{1i}(\mathscr{M}_1, t_0) = \max \hat{\mathbf{p}}(\mathscr{M}_1, t_0), X_i \leq t_0, \eta_i = k\} / G(X_i-), k = 1, 2,$$

$$h_{i,3}^{+}(t_0) = I\{\hat{p}_{3i}(\mathscr{M}_2, t_0) = \max \hat{\mathbf{p}}(\mathscr{M}_2, t_0), \hat{p}_{3i}(\mathscr{M}_1, t_0) \neq \max \hat{\mathbf{p}}(\mathscr{M}_1, t_0), X_i > t_0\} / G(t_0),$$

$$h_{i,3}^{-}(t_0) = I\{\hat{p}_{3i}(\mathscr{M}_2, t_0) \neq \max \hat{\mathbf{p}}(\mathscr{M}_2, t_0), \hat{p}_{3i}(\mathscr{M}_1, t_0) = \max \hat{\mathbf{p}}(\mathscr{M}_1, t_0), X_i > t_0\} / G(t_0),$$

where $\omega_k, k = 1, 2, 3$, are weight functions for the three disease categories and can be simply set to be $1/3$ if there is no prior on the categories, and $\hat{G}$ is the Kaplan-Meier estimator of the censoring survival function.

We now establish the asymptotic normality of $\hat{S}(t_0)$, starting with introducing more notation. Let $Q_X(t) = P(X_i > t)$, and define the martingale of the censoring time C as $M_{C_i}(t) = I\{\eta_i = 0, X_i \leq t\} - \int_0^t I\{X_i \geq u\} d\Lambda_C(u)$, where $\Lambda_C(\cdot)$ is the cumulative hazard function of C. For $k = 1, 2$ and the third "healthy" category, we define $f_{i,k}(t_0) = I\{X_i \leq t_0, \eta_i = k\} / G(X_i-)$ and $f_{i,3}(t_0) = I\{X_i > t_0\} / G(t_0)$. For $k = 1, 2, 3$, define $h_k^{+/-}(t_0) = E h_{i,k}^{+/-}(t_0)$, $f_k(t_0) = E f_{i,k}(t_0)$, where the expectation is with respect to T, C, and covariates Z. Let $\hat{M}_C$ be the estimator defined by plugging in the usual Nelson-Aalen estimator of

the cumulative hazard function of the censoring time C and let $\hat{h}^{+/-}_{i,k}$, $\hat{f}_{i,k}$ be defined by plugging in the Kaplan-Meier estimator $\hat{G}(\cdot)$, if applicable. Define

$$\hat{h}^{+/-}_{k}(t_0) = \frac{1}{n}\sum_{i=1}^{n}\hat{h}^{+/-}_{i,k}(t_0), \quad \hat{f}_k(t_0) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{i,k}(t_0), \quad \hat{S}(t_0) = \sum_{k=1}^{3}\omega_k\frac{\hat{h}^{+}_{k}(t_0) - \hat{h}^{-}_{k}(t_0)}{\hat{f}_k(t_0)}.$$

Since $G(t_0)$ in $h^{+/-}_{i,3}(t_0)$ and $f_{i,3}(t_0)$ will be canceled out, we redefine $h^{+/-}_{i,3}(t_0)$ and $f_{i,3}(t_0)$ by multiplying $G(t_0)$. Following P. *and others* (2013), the Martingale representation of the Kaplan-Meier estimator of the censoring survival function entails that: $\sup_t \left| \sqrt{n}(\hat{G}(t) - G(t)) + \frac{G(t)}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t \frac{dM_{C_i}(u)}{Q_X(u)} \right| = o_p(1)$. By Taylor's expansion, $\sup_{t_0} \left| \sqrt{n}(\hat{h}^{+}_{k}(t_0) - h^{+}_{k}(t_0)) - \left[\frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} h^{+}_{i,k}(t_0)\{1 + \int_0^{X_i}\frac{dM_{C_j}(u)}{Q_X(u)}\} - h^{+}_{k}(t_0) \right] \right| = o_p(1)$, for $k = 1, 2$. Similar results hold for $\hat{h}^{-}_{k}(t_0), k = 1, 2$. Moreover, $\sup_{t_0} \left| \sqrt{n}(\hat{f}_k(t_0) - f_k(t_0)) - \left[\frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} f_{i,k}(t_0)\{1 + \int_0^{X_i}\frac{dM_{C_j}(u)}{Q_X(u)}\} - F_k(t_0) \right] \right| = o_p(1)$. When $k = 3$, again we have $\sup_{t_0} |\sqrt{n}(\hat{h}^{+}_{3}(t_0) - h^{+}_{3}(t_0)) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(h^{+}_{i,3}(t_0) - h^{+}_{3}(t_0))| = 0$ and $\sup_{t_0} |\sqrt{n}(\hat{f}_3(t_0) - f_3(t_0)) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(f_{i,3}(t_0) - f_3(t_0))| = 0$. Then $\hat{S}(t_0)$ can be further formulated using Taylor's expansion:

$$\sup_t |\sqrt{n}(\hat{S}(t_0) - S(t_0)) - \frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Psi_{ij}(t_0)| = o_p(1),$$

where

$$\begin{aligned}
\Psi_{ij}(t_0) &= \sum_{k=1}^{2}\omega_k\{\frac{IF_{h^{+}_{k}(t_0)} - IF_{h^{-}_{k}(t_0)}}{f_k(t_0)} - \frac{h^{+}_{k}(t_0) - h^{-}_{k}(t_0)}{f^2_k(t_0)}IF_{f_k(t_0)}\} \\
&= \sum_{k=1}^{2}\omega_k\{[h^{+}_{i,k}(t_0) - h^{-}_{i,k}(t_0) - \frac{f_{i,k}(t_0)}{f_k(t_0)}(h^{+}_{k}(t_0) - h^{-}_{k}(t_0))](1 + \int_0^{X_i}\frac{dM_{C_j}(u)}{Q_X(u)})\}/f_k(t_0) \\
&+ \omega_3\{h^{+}_{i,3}(t_0) - h^{-}_{i,3}(t_0) - \frac{f_{i,3}(t_0)}{f_3(t_0)}(h^{+}_{3}(t_0) - h^{-}_{3}(t_0))\}/f_3(t_0),
\end{aligned}$$

and IF denotes the influence function of each estimator respectively. By Hájek's projection principle, the following Hoeffding's decomposition holds:

$$\frac{\sqrt{n}}{n(n-1)}\sum_{i\neq j}^{n}\Psi_{ij}(t_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}IF(X_i, \eta_i, t_0) + o_p(1).$$

Given Martingale's properties, we also know that $E[IF(X_i, \eta_i, t_0)] = 0$. Let $\hat{M}_C$ be the estimator by plugging in the usual Nelson-Aalen estimator of the cumulative hazard function of the censoring time C. $\hat{h}^{+/-}_{k}$, $\hat{F}_k$, $\hat{h}^{+/-}_{i,k}$, and $\hat{f}_{i,k}$ were defined as above. $\hat{Q}_X$ is the estimate of Q using the empirical distribution of X.

Plugging in these estimators to estimate Ψ_{ij}, we compute $IF(X_i, \eta_i, t_0)$ as

$$\hat{IF}(X_i, \eta_i, t_0) = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j \neq i}^{n} [\hat{\Psi}_{ij}(t_0) + \hat{\Psi}_{ji}(t_0)]. \tag{2.2}$$

2.3 *Integrated discrimination improvement for competing outcomes*

We first define the time-dependent IDI for competing risks outcomes by adapting its definition for multi-category outcomes as following:

$$R(t_0) = \sum_{k=1}^{K} \omega_k \{R_k^2(\mathscr{M}_2, t_0) - R_k^2(\mathscr{M}_1, t_0)\}, \tag{2.3}$$

where ω_k are again some weight functions. The estimation of the IDI at time t_0 involves the evaluation of $R_k^2(\mathscr{M}_m, t_0)$, which is the proportion of variability in the k-th category that is explained by model $\mathscr{M}_m$, for $m = 1, 2$ and $k = 1, 2$. Without any covariates, we estimate the probability of falling into the k-th category by $\hat{\pi}_k(t_0)$, where $\hat{\pi}_k(t_0) = \hat{n}_k(t_0)/(\hat{n}_1(t_0) + \hat{n}_2(t_0) + \hat{n}_3(t_0))$, with $\hat{n}_k(t_0) = \sum_{i=1}^{n} I\{X_i \leq t_0, \eta_i = k\}/\hat{G}(X_i-), k = 1, 2$ and $\hat{n}_3(t_0) = \sum_{i=1}^{n} I\{X_i > t_0\}/\hat{G}(t_0)$. Hence the variance without any model is $\hat{\pi}_k(t_0)(1-\hat{\pi}_k(t_0))$. With model $\mathscr{M}_m, m = 1, 2$, the variance can be estimated by $\frac{1}{n}\sum_{i=1}^{n} \{\hat{p}_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)}\}^2$, where $\overline{\hat{p}_k(\mathscr{M}_m, t_0)} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{ki}(\mathscr{M}_m, t_0)$. Therefore, we propose the following estimator of the IDI at time t_0:

$$\hat{R}(t_0) = \sum_{k=1}^{3} \frac{\omega_k}{n\hat{\pi}_k(t_0)\{1 - \hat{\pi}_k(t_0)\}} \sum_{i=1}^{n} [\{\hat{p}_{ki}(\mathscr{M}_2, t_0) - \overline{\hat{p}_k(\mathscr{M}_2, t_0)}\}^2 - \{\hat{p}_{ki}(\mathscr{M}_1, t_0) - \overline{\hat{p}_k(\mathscr{M}_1, t_0)}\}^2].$$

Define $\pi_k = E\hat{n}_k/n$, for $k = 1, 2, 3$. Analogous to the arguments in Section 2.2, we have

$$\sup_{t_0} \left| \sqrt{n}(\hat{\pi}_k(t_0) - \pi_k(t_0)) - \{\frac{\sqrt{n}}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} n_{i,k}(t_0)(1 + \int_0^{X_i} \frac{dM_{C_j}(u)}{Q_X(u)}) - \pi_k(t_0)\} \right| = o_p(1), \quad k = 1, 2,$$

$$\sup_{t_0} \left| \sqrt{n}(\hat{\pi}_3(t_0) - \pi_3(t_0)) - \{\frac{\sqrt{n}}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} n_{i,3}(t_0)(1 + \int_0^{t_0} \frac{dM_{C_j}(u)}{Q_X(u)}) - \pi_3(t_0)\} \right| = o_p(1).$$

Let $Q_{ij,k}(t_0) = n_{i,k}(t_0)\left(1 + \int_0^{X_i} \frac{dM_{C_j}(u)}{Q_X(u)}\right) - \pi_k(t_0), k = 1, 2$, and $Q_{ij,3}(t_0) = n_{i,3}(t_0)\left(1 + \int_0^{t_0} \frac{dM_{C_j}(u)}{Q_X(u)}\right) - \pi_3(t_0)$. When covariates are involved, the variance that is explained by model $\mathscr{M}_m, m = 1, 2$, is given as

$\hat{D}_{k(m)} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{p}_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)}\}^2$. $\hat{D}_{k(m)}$ can be rewritten as

$$\begin{aligned}\hat{D}_{k(m)} &= \frac{1}{n}\sum_{i=1}^{n}\{\hat{p}_{ki}(\mathscr{M}_m, t_0) - p_{ki}(\mathscr{M}_m, t_0) + p_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)}\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}[\{\hat{p}_{ki}(\mathscr{M}_m, t_0) - p_{ki}(\mathscr{M}_m, t_0)\}^2 + \{p_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)}\}^2 \\ &+ 2\{\hat{p}_{ki}(\mathscr{M}_m, t_0) - p_{ki}(\mathscr{M}_m, t_0)\}\{p_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)}\}].\end{aligned}$$

Let $D_{k(m)} = E\hat{D}_{k(m)}$. By taking Taylor's expansion, it's easy to get the asymptotic linear representation for $\hat{D}_{k(m)}$. For $k = 1, 2, 3$:

$$\begin{aligned}\sup_{t_0}|\sqrt{n}(\hat{D}_{k(m)} - D_{k(m)}) - \frac{\sqrt{n}}{n}\sum_{i=1}^{n}[2(p_{ki}(\mathscr{M}_l, t_0) - p_k(\mathscr{M}_m, t_0))IF_{\hat{p}_{ki}} \\ +(p_{ki}(\mathscr{M}_m, t_0) - p_k(\mathscr{M}_m, t_0))^2 - D_{k(m)}]| = o_p(1), \quad m = 1, 2,\end{aligned}$$

where $IF_{\hat{p}_{ki}}$ is the influence function that is specific to the estimated CIF from a particular competing risks model, and will be discussed again in the following paragraph. Denote $B_{ki(m)}(t_0) = (\hat{p}_{ki}(\mathscr{M}_m, t_0) - p_k(\mathscr{M}_m, t_0))^2 - D_{k(m)}$. By Taylor's expansion:

$$\sup_{t_0}\left|\sqrt{n}(\hat{R}(t_0) - R(t_0)) - \frac{\sqrt{n}}{n(n-1)}\sum_{i\neq j}^{n}\Psi_{ij}^{**}(t_0)\right| = o_p(1),$$

where

$$\Psi_{ij}^{**}(t_0) = \sum_{k=1}^{3}\omega_k\{\frac{B_{ki(2)} - B_{ki(1)}}{\pi_k(t_0)(1-\pi_k(t_0))} + \frac{Q_{ij,k}(t_0)(D_{k(2)} - D_{k(1)})(2\pi_k(t_0) - 1)}{\pi_k^2(t_0)(1-\pi_k(t_0))^2}\}.$$

By Hájek's projection principle, the following Hoeffding decomposition holds:

$$\frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Psi_{ij}^{**}(t_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}IF^{**}(X_i, \eta_i, M_i, t_0) + o_p(1),$$

where

$$IF^{**}(X_i, \eta_i, M_i, t_0) = E(\Psi_{ij}^{**}(t_0) + \Psi_{ji}^{**}(t_0)|(X_i, \eta_i, M_i))$$

and $E[IF^{**}(X_i, \eta_i, M_i, t_0)] = 0$. Using the same procedure from the previous proof, we can also estimate $IF^{**}(X_i, \eta_i, M_i, t_0)$ using the sample.

However, the IDI estimator relies on the estimated probabilities from a particular competing-risk model, and asymptotic variance will change if another model is used. Some competing-risk models have

well-defined influence functions $IF_{\hat{p}_{ki}}$, while others do not have explicit forms. Thus, we propose to use a bias-corrected and accelerated (BCa) bootstrap procedure to obtain confidence intervals for the IDI, which correct the skewness and the bias of the bootstrap distribution (Efron, 1987; Efron and Tibshirani, 1994).

3. SIMULATION STUDIES

In practice, we usually do not know the "right" model, and there is a chance that we could pick a reasonable yet incorrect model for our data. Thus, we need to evaluate the impact of model choices on the performance of accuracy improvement evaluation with new biomarkers included, and it is important that the extended NRI and IDI are relatively robust against model mis-specification. Here, we first designed three different sets of data with respect to three popular competing-risk models, including multi-state, Fine and Gray, and multinomial logistic models, to examine the proposed estimators for the extended NRI and IDI in competing risks settings. Three covariates were used in all three designs, where Z_1 and Z_3 were generated from standard normal distribution, truncated at ± 3.5 to prevent extreme values, and Z_2 was generated from a Bernoulli (0.7) distribution. The three cases of data were simulated as follows:

Case 1. We simulated the event time from a Weibull model with three covariates,

$$\log(T) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \sigma W,$$

where W was generated from the standard extreme value distribution. This error distribution gives the proportional hazard interpretations for all covariates. We set $\beta_0 = 2.5$, $\beta_1 = 0.05$, $\beta_2 = -0.05$, $\beta_3 = 0.15$ and $\sigma = 0.2$. Since the coefficient for the new marker β_3 is three times the size of the coefficients β_1 and β_2 for two conventional predictors, we expect that the "new" model including Z_1, Z_2 and Z_3 would have improved predictive ability over the "old" model that only uses Z_1 and Z_2. The cause indicators, $k = 1, 2$, were generated with equal probability. The censoring time was simulated from a uniform $[2, 31]$ distribution, resulting in about 30% censoring, and from a uniform $[1, 21]$ with 50% censoring.

Case 2. We used a simulation design similar to the one proposed by Fine and Gray (1999) in this case.

The subdistribution for cause 1 is defined by

$$F_1(t|\mathbf{Z}) = 1 - [1 - p\{1 - \exp((t/20)^5)\}]^{\exp(\beta_{11}Z_1+\beta_{12}Z_2+\beta_{13}Z_3)},$$

with a mass of $1-p$ when t is at ∞ and all covariates are zeros. When a uniform random number exceeds $F_1(\infty|\mathbf{Z})$, subjects are assumed to experience the cause 2 event with the conditional probability

$$P(T \le t|\epsilon = 2, \mathbf{Z}) = 1 - \exp(-\exp(\beta_{21}Z_1 + \beta_{22}Z_2 + \beta_{23}Z_3)(t/20)^5).$$

We set $\beta_{11} = 0.2$, $\beta_{12} = -0.5$, $\beta_{13} = 1$, $\beta_{21} = 0.02$, $\beta_{22} = -0.05$, $\beta_{23} = 0.1$, and $p = 0.65$. Including Z_3 in the model, besides Z_1 and Z_2, is expected to improve prediction over the one not including Z_3. The censoring distribution follows uniform $[10, 37.5]$ with 30% censoring and uniform $[7, 29.5]$ with 50% censoring.

Case 3. We considered a multinomial logistic regression model as suggested by Gerds *and others* (2012) in this case. Define $F_k(t, \boldsymbol{Z}) = P(T \le t, \epsilon = k|\boldsymbol{Z})$, $k = 1, 2$. For cause k, logistic-transformed probabilities were set as

$$\log\{F_k(t)/(1 - F_1(t) - F_2(t))\} = \mu(t) + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3, \quad t > 0,$$

where $\mu(t)$ was set to be $t + 11$, $\beta_1 = 0.5$, $\beta_2 = 0.5$ and $\beta_3 = 1$. Since β_3 is twice the size of β_1 and β_2, we expect the new model including Z_3 would have a better predictive ability than the old one using only Z_1 and Z_2. The event time was simulated by inverting the survival probability, and cause indicators were assigned with equal probabilities. Independent censoring time was simulated from a uniform $[0, 32]$ distribution for 30% censoring and from a uniform $[0, 29.2]$ distribution for 50% censoring.

In light of Demler *and others* (2017), we also want to examine the robustness of the inferential procedures for the NRI and the IDI under the null, where adding the new biomarker into "old" model doesn't improve the predictive ability. Thus, we consider the following three scenarios:

Case 4. Similar to *Case 1*, we set $\beta_0 = 2.5$, $\beta_1 = 0.25$, $\beta_2 = -0.05$, $\beta_3 = 0$ and $\sigma = 0.2$. Censoring time was simulated from uniform [3, 32] for 30% censoring, and from uniform [1, 21] for 50% censoring.

Case 5. Similar to *Case 2*, we set $\beta_{11} = 0.2$, $\beta_{12} = -0.5$, $\beta_{13} = 0$, $\beta_{21} = 0.02$, $\beta_{22} = -0.05$, $\beta_{23} = 0$, and

$p = 0.65$. The censoring distributions followed a uniform $[10, 39]$ distribution for the 30% censoring case and from uniform $[7, 30]$ for 50% censoring.

Case 6. Similar to *Case 3*, we set $\beta_1 = 3$, $\beta_2 = 1$ and $\beta_3 = 0$. Independent censoring time was simulated from uniform [0, 32] for 30% censoring and from uniform [0, 29.2] for 50% censoring.

Under *Cases 4, 5* and *6*, we expect that the predictive ability of the "new" model would not be improved. Thus the true NRI and IDI are zero. For each simulation case, we generated 1,000 samples of size 400, and applied all three models, Cox's proportional hazard model, Gerds' multinomial logistic risk regression, and Fine-Gray's subdistribution hazard model.Cox regression and Fine-Gray's model estimate the CIF for each cause separately, while Gerds' model estimates CIFs for both causes simultaneously. We built confidence intervals (CIs) for the NRI based on (2.2) and CIs for the IDI using biased-corrected accelerated (BCa) bootstrapping. Simulation was run in `R`, where packages `survival`, `riskRegression` and `cmprsk` were used for competing risks modeling, and package `boot` was used for BCa bootstrapping. The simulation results for the NRI and the IDI from cases 1, 2 and 3 under 30% censoring, in which model predictive ability should improve with the "new" marker, are shown in Table 1 and Table 2, respectively. Table 3 and Table 4 summarize the simulation results for the NRI and the IDI under cases 4, 5 and 6 with 30% censoring, when the added covariate does not improve prediction accuracy.

From Table 1 and Table 2, we first notice that, for both NRI and IDI, estimated $\hat{S}$ and $\hat{R}$ on average are very close to true values S and R, when the correct model for a specific data design. The average standard deviations of the estimated NRIs based on formula (2.2) approximate the empirical standard errors closely. The 95% CIs based on asymptotic normality and estimated standard deviation cover the true values about 95% times, though the coverage rates are a bit lower than 95% at time 13. One possible reason is that, at time 13, simulated event times are more likely to be censored, and due to the use of approximation from Taylor's expansion, our formula-based asymptotic variance could underestimate the true variance of the proposed NRI estimators in this situation. Nevertheless, when models are specified correctly, the results are very good in general. Similar to the NRI, IDI estimators are close to their true values when models are

correctively specified, average bootstrap standard deviations are comparable to empirical standard errors, and coverage rates are around 95% using BCa bootstrap CIs. As the censoring rate increases from 30% to 50% (results shown in Tables S1 and S2 in the supplementary material), standard errors of both NRI and IDI estimators increase but similar coverage rates are observed.

When the model is mis-specified, the NRI seems to be a more robust measure than the IDI towards model mis-specification, in the sense that coverage rates of the NRI from the misspecified models are in general satisfactory, and NRI estimators are close to true values obtained from correct models. This follows that the NRI uses ranks of estimated probabilities instead of probabilities themselves, which are used in the IDI estimation. However, despite the appealing interpretation of covariate effects on cumulative incidence functions, Fine and Gray's model does not guarantee the sum of all cause probabilities is equal to one. So, proposed standard deviation estimation for the NRI often underestimates.The underestimation is worse for the IDI estimators when Fine and Gray is mis-used in predicting event probabilities.

Similar conclusions are observed from Table 3 and Table 4. Even though the true underlying data are from the null and both NRI and IDI are degenerate, the probabilities of covering zero are high for both the NRI's formula-based CIs and the IDI's BCa bootstrap CIs. This applies no matter whether the model is correctly specified. Our methods are robust against degenerate cases and model mis-specification. Demler *and others* (2017) suggested to "un-nest" the models by including independent weak predictors in both models such that they are not nested any more. However, by doing so, bias would be introduced into the IDI estimation, which might lead to lower coverage rate of CIs. Thus, we chose to simply use the BCa bootstrapping procedure instead. Results for the null hypothesis under 50% censoring are summarized in Tables S3-S4 in the supplemental material. The same patterns and robustness are observed.

4. Application to the Multicenter AIDS Cohort Study

We applied the NRI and IDI methods to data obtained by the MACS. It is an ongoing study of homosexual and bisexual men at risk for or infected with HIV, recruited from four institutions in Baltimore, Chicago,

Pittsburgh and Los Angeles (Kingsley *and others*, 1987; Kaslow *and others*, 1987). The data used for this analysis were gathered between April 2,1984 and April 8, 2017. Each participant underwent a clinical examination semi-annually, and neuropsychological testing approximately every two years (however, see Miller *and others* (1990); Becker *and others* (2014) for details) until they drop out of the study voluntarily or die. The current analysis utilizes the data from a substudy of the legacy effect of HIV on cognitive impairment, which contains 2,783 HIV seropositive men (Farinpour *and others*, 2003).

Individuals with HIV disease have historically been at risk for cognitive impairment. The MACS measured cognitive functions over time with a battery of neuropsychological (NP) tests which were summarized by T scores in six cognitive domains: working memory & attention, learning, motor speed & coordination, executive functioning, speed of information processing, and memory. We adopted the Multivariate Normative Comparison (MNC) method to define abnormality in cognition as in Su *and others* (2015) and Wang *and others* (2019). Time to impairment was defined as the interval between study entry and the first visit where the six domain scores were deemed abnormal by the MNC method. Those subjects who were impaired at their first visit were excluded from the analyses. If a subject died after the last complete NP visit and no cognitive impairment was detected, his time to impairment was competing-rick censored by death. Otherwise, subjects were censored at their last visit.

In the presence of competing-risk censoring, techniques such as Cox regression, Gerds' model, and Fine-Gray's model can be used to identify potential risk factors affecting cognition after the onset of HIV infection. However, these methods do not directly quantify the relative importance a factor is in predicting who might develop impairment, who might die, or who might be alive and disease free after a fixed time interval. Here we apply the NRI and the IDI treating CD4+ cell count as the "new" biomarker (with both linear and quadratic terms to account for nonlinearity when modeling cognitive impairment) to examine whether the inclusion of this variable will yield a better prediction. In the Legacy substudy (Popov *and others*, 2019), three other predictors – age, center for epidemiologic studies depression scale, and recruitment cohort (before or after 2001) – were found to be significantly related to cognitive impairment

and were treated here as conventional predictors. Final analysis included 1,972 seropositive subjects who had at least one visit with complete cognitive tests and the information on four predictors.

Within this subsample, 553 men were classified with cognitive impairment using the MNC method (28.0%), 597 died during follow-up without any cognitive impairment (30.3%), and 822 were censored by the "end" (at the data freeze) of the study (41.7%). Time to event or time in the study ranged from 5 months to 33 years. We examined the performance of CD4+ cell count and its quadratic transformation as the "new" biomarkers in predicting health status at 10 and 12 years since the start of the study with a proportional hazard model, Fine-Gray's model, and Gerds' model, using both NRI and IDI. The two events, cognitive impairment and death without cognitive impairment, were again modeled separately with Cox's model or Fine-Gray's model, and they were modeled simultaneously in the Gerds model. Based on the predicted probabilities of both events that were calculated from the three models, we computed the values of the NRI and IDI. For IDI, 10,000 bootstrap samples were used to produce 95% BCa bootstrap CIs. In order to select the most suitable regression model, we also computed Brier scores (Gerds and Schumacher, 2006) for all three models under consideration. The results are summarized in Table 5.

In Table 5 we can see that the estimated NRI and IDI and their 95% CIs are comparable across the three different models. This is consistent with the robust nature of the NRI/IDI estimators under model misspecification that was observed in our simulation studies. Among the three models, Cox regression has the lowest Brier scores for both events at 10 and 12 years, suggesting that Cox regression is the most suitable competing risks model for our data. Moreover, from the Cox-Snell residual plots shown in Fig. 1, we can see that Cox regression provides a good overall fit to the MACS data, as the cumulative hazards of Cox-Snell residuals for both events go through a straight line with slope 1.

From the Cox regression model the estimated NRIs at 10 and 12 years since the start of the study are .042 and .079 with 95% CIs [.027, .056] and [.062, .096] respectively. The estimated IDIs are .049 and .060 with 95% BCa CIs [.039, .065] and [.048, .078]. Because the 95% CIs of both NRI and IDI do not include zero, we conclude that including the CD4+ cell counts in competing risks models increases the accuracy

of predicting cognitive impairment and death after 10 and 12 years in the study. More specifically, the probabilities of correctly predicting health status (impairment, death, or neither) for a subject after 10 and 12 years of observation improves by 4.2% and 7.9%, by simply incorporating CD4+ all counts with its quadratic transformation into the model. Also, the variability explained by the predictive model is increased by 4.9% and 6.0% for events at 10 and 12 years with the addition of the CD4+ counts.

However, some participants withdrew from the legacy study and died many years afterwards. If a subject died more than 4 years after his last NP visit, he may have experienced cognitive impairment between his last NP visit and death. As a sensitivity analysis, we censored such subjects four years after their last NP visit, assuming cognition stayed relatively stable over two consecutive NP visits (about 4 years as scheduled). In this way, 553 men were classified with cognitive impairment using the MNC method (28.0%), 425 died within 4 years after the last NP visit without any cognitive impairment (21.6%), and 994 were censored either at their last study visit or 4 years following the last NP exam, whichever was first (50.4%). Using the Cox regression model, the estimated NRIs at 10 and 12 years since the start of the study are .100 and .106 with 95% confidence intervals [.082, .118] and [.088, .124] respectively. The estimated IDIs are .095 and .100 with 95% BCa confidence intervals of [.084, .137] and [.086, .143]. Again, these findings suggest that including CD4+ cell counts in competing risks models can increase prediction accuracy of death and cognitive impairment after 10 and 12 years in the study.

5. DISCUSSION

We have demonstrated here the good practical performance of the extended NRI and IDI in competing risks settings. Although a CI for the IDI can be efficiently constructed based on the asymptotic linear representation for a well-studied regression model, the BCa bootstrap method serves as a flexible alternative when a model is relatively new and its theoretical properties are less known. When the added variables have no effect on the events and models to be compared are nested, Demler *and others* (2017) showed that the theory based on U-statistics fails. Still, the CIs for the NRI based on asymptotic normality and the

BCa bootstrap CIs for the IDI seem to have satisfactory coverage as demonstrated by simulations.

In this work we have considered three reasonable competing risks models. However, one can use any other semiparametric or parametric models such as Scheike *and others* (2008) and Cheng (2009). Although certain robustness of the NRI and IDI against model mis-specification has been observed, it remains important to select a proper predictive model before examining diagnostic accuracy improvement over the course of variables' addition. Metrics, such as the Brier score, are useful in choosing the most appropriate model for the data.

Competing endpoints are common in biomedical research, although they are often neglected in analysis. The extended NRI and IDI for competing events provide alternative and straightforward interpretations of the importance of new biomarkers on top of conventional factors. They also serve as more unifying metrics than model coefficients such as hazards ratio or odds ratio, since the latter depend on the types and the scales of covariates. Moreover, this is in line with recent debate about moving away from statistical significance of 0.05 level (Wasserstein *and others*, 2019). Instead of simply looking at p values for the added variables in a regression model, one can assess the contribution of additional risk factors in prediction through interval estimates of the IDI and NRI. Thus, the extended NRI and IDI for multiple competing endpoints might be useful in screening and selecting covariates in high dimensional settings.

SOFTWARE

Software in the form of R code, together with a sample input data set and complete documentation is available on request from the corresponding author (yucheng@pitt.edu).

SUPPLEMENTARY MATERIAL

Supplementary material is available online at `http://biostatistics.oxfordjournals.org`.

ACKNOWLEDGMENTS

We are grateful for the computational resources for this paper provided by PittGrid from the University of Pittsburgh (`www.pittgrid.pitt.edu`). Data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS). MACS (Principal Investigators): Johns Hopkins University Bloomberg School of Public Health (Joseph Margolick, Todd Brown), U01-AI35042; Northwestern University (Steven Wolinsky), U01-AI35039; University of California, Los Angeles (Roger Detels, Otoniel Martinez-Maza, Otto Yang), U01-AI35040; University of Pittsburgh (Charles Rinaldo, Lawrence Kingsley, Jeremy Martinson), U01-AI35041; the Center for Analysis and Management of MACS, Johns Hopkins University Bloomberg School of Public Health (Lisa Jacobson, Gypsyamber D'Souza), UM1-AI35043. The MACS is funded primarily by the National Institute of Allergy and Infectious Diseases (NIAID), with additional co-funding from the National Cancer Institute (NCI), the National Institute on Drug Abuse (NIDA), and the National Institute of Mental Health (NIMH). Targeted supplemental funding for specific projects was also provided by the National Heart, Lung, and Blood Institute (NHLBI), and the National Institute on Deafness and Communication Disorders (NIDCD). MACS data collection is also supported by UL1-TR001079 (JHU ICTR) from the National Center for Advancing Translational Sciences (NCATS) a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Additional support was provided by the Johns Hopkins University Center for AIDS Research (P30AI094189). The contents of this publication are solely the responsibility of the authors and do not represent the official views of the National Institutes of Health (NIH), Johns Hopkins ICTR, or NCATS. The MACS website is located at `http://aidscohortstudy.org/`. *Conflict of Interest*: None declared.

FUNDING

This work was partially supported by the NIA grant R01 AG034852 (to Becker) and NSF DMS 1916001 (to Cheng).

REFERENCES

AALEN, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6**(4), 701–726.

ANDERSEN, P. K., ABILDSTROM, S. Z. AND ROSTHØJ, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research* **11**(2), 203–215.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. AND KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

BECKER, J.T., KINGSLEY, L.A., MOLSBERRY, S., REYNOLDS, S., ARONOW, A., LEVINE, A.J., MARTIN, E., MILLER, E.N., MUNRO, C.A., RAGIN, A., SACKTOR, N. *and others*. (2014). Cohort Profile: Recruitment cohorts in the neuropsychological substudy of the Multicenter AIDS Cohort Study. *International Journal of Epidemiology* **44**(5), 1506–1516.

BEYERSMANN, J., SCHUMACHER, M. AND ALLIGNOL, A. (2012). *Competing risks and multistate models with R*. New York: Springer.

BRESLOW, N. AND CROWLEY, J. (1974, 05). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.* **2**(3), 437–453.

CAI, T., PEPE, M. S., ZHENG, Y., LUMLEY, T. AND JENNY, N. S. (2006). The sensitivity and specificity of markers for event times. *Biostatistics* **7**(2), 182–197.

CHENG, S. C., FINE, JASON P. AND WEI, L. J. (1998). Prediction of cumulative incidence function under the proportional hazards model. *Biometrics* **54**(1), 219–228.

CHENG, Y. (2009). Modeling cumulative incidences of dementia and dementia-free death using a novel three-parameter logistic function. *The International Journal of Biostatistics* **5**(1), 1557–4679.

CHENG, Y. AND FINE, J. P. (2012). Cumulative incidence association models for bivariate competing risks data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(2), 183–202.

CHENG, Y., FINE, J. P. AND K., MICHAEL R. (2007). Nonparametric association analysis of bivariate competing-risks data. *Journal of the American Statistical Association* **102**(480), 1407–1415.

CHENG, Y. AND LI, J. (2015). Time-dependent diagnostic accuracy analysis with censored outcome and censored predictor. *Journal of Statistical Planning and Inference* **156**, 90 – 102.

CHIPMAN, J. AND BRAUN, D. (2017). Simpson's paradox in the integrated discrimination improvement. *Statistics in Medicine* **36**(28), 4468–4481.

COOK, N. R., DEMLER, O. V. AND PAYNTER, N. P. (2017). Clinical risk reclassification at 10 years. *Statistics in Medicine* **36**(28), 4498–4502.

COX, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological* **34**, 187–220.

DEMLER, O. V., PENCINA, M. J., COOK, N. R. AND D'AGOSTINO, R. B. (2017). Asymptotic distribution of δAUC, NRIs, and IDI based on theory of U-statistics. *Statistics in medicine* **36**(21), 3334–3360.

DREISEITL, S., OHNO-MACHADO, L. AND BINDER, M. (2000). Comparing three−class diagnostic tests by three−way ROC analysis. *Medical Decision Making* **20**(3), 323–331.

EDWARDS, D. C., METZ, C. E. AND KUPINSKI, M. A. (2004, July). Ideal observers and optimal roc hypersurfaces in n-class classification. *IEEE Transactions on Medical Imaging* **23**(7), 891–895.

EFRON, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**(397), 171–185.

EFRON, BRADLEY AND TIBSHIRANI, ROBERT J. (1994). *An introduction to the bootstrap*. Boca Raton: CRC press.

FARINPOUR, R., MILLER, E.N., P., SATZ, SELNES, O.A., COHEN, B.A., BECKER, J.T., SKOLASKY, R.L. AND VISSCHER, B.R. (2003). Psychosocial risk factors of HIV morbidity and mortality: Findings

from the Multicenter AIDS Cohort Study (MACS). *Journal of Clinical and Experimental Neuropsychology* **25**(5), 654–670.

FINE, J. P. AND GRAY, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**(446), 496–509.

FOUCHER, Y., GIRAL, M., SOULILLOU, J. P. AND DAURES, J. P. (2010). Time−dependent ROC analysis for a three−class prognostic with application to kidney transplantation. *Statistics in Medicine* **29**(30), 3079–3087.

GERDS, T.A. AND SCHUMACHER, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal* **48**, 1029–1040.

GERDS, T. A., SCHEIKE, T. H. AND ANDERSEN, P. K. (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine* **31**(29), 3921–3930.

GOOLEY, T. A., LEISENRING, W., CROWLEY, J. AND STORER, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine* **18**(6), 695–706.

GREENLAND, P. AND O'MALLEY, P. G. (2005). When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase a2 and c-reactive protein for stroke risk. *Archives of Internal Medicine* **165**(21), 2454–2456.

HEAGERTY, P. J., LUMLEY, T. AND PEPE, M. S. (2000). Time−dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**(2), 337–344.

HEAGERTY, P. J. AND ZHENG, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics* **61**(1), 92–105.

HUNG, H. AND CHIANG, C. (2010). Estimation methods for time−dependent auc models with survival data. *Canadian Journal of Statistics* **38**(1), 8–26.

INÁCIO, V., TURKMAN, A. A., NAKAS, C. T. AND ALONZO, T. A. (2011). Nonparametric bayesian estimation of the three-way receiver operating characteristic surface. *Biometrical Journal* **53**(6), 1011–1024.

KALBFLEISCH, J. D. AND PRENTICE, R. L. (2002). *The statistical analysis of Failure time data*, 2nd edition. New York: JohnWiley & Sons.

KAPLAN, E. L. AND MEIER, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**(282), 457–481.

KASLOW, R.A., OSTROW, D.G., DETELS, R., PHAIR, J.P., POLK, F.B. AND RINALDO, JR. C.R. (1987). The multicenter AIDS cohort study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology* **126**(2), 310–318.

KINGSLEY, L., KASLOW, R., RINALDO, C. J. R., DETRE, K., ODAKA, N. AND VANRADEN, M. (1987). Risk factors for seroconversion to human immunodeficiency virus among male homosexuals. *The Lancet* **329**(8529).

KLEIN, J. P. (2006). Modelling competing risks in cancer studies. *Statistics in Medicine* **25**, 1015–1034.

KLEIN, J. P. AND MOESCHBERGER, M. L. (2003). *Survival analysis: techniques for censored and truncated data*. New York: Springer.

LI, J. AND FINE, J. P. (2008). ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics* **9**(3), 566–576.

LI, J., JIANG, B. AND FINE, J. P. (2013). Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. *Biostatistics* **14**(2), 382–394.

LI, J. AND ZHOU, X. (2009). Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *Journal of Statistical Planning and Inference* **139**(12), 4133 – 4142.

MILLER, E. N., SEINES, O. A., MCARTHUR, J. C., SATZ, P., BECKER, J. T., COHEN, B. A., SHERIDAN, K., MACHADO, A. M., GORP, W.G. VAN AND VISSCHER, B. (1990). Neuropsychological performance in HIV-1-infected homosexual men. *Neurology* **40**(2), 197–197.

MOSSMAN, D. (1999). Three-way ROCs. *Medical Decision Making* **19**(1), 78–89.

P., BLANCHE, J., DARTIGUES AND H., JACQMIN−GADDA. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine* **32**, 5381–5397.

PENCINA, M. J., D' AGOSTINO, R. B., D' AGOSTINO, R. B. AND VASAN, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**(2), 157–172.

PENCINA, M. J., D'AGOSTINO, R. B. AND STEYERBERG, E. W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine* **30**(1), 11–21.

PENCINA, M. J., FINE, J. P. AND D'AGOSTINO, R. B. (2017*a*). Discrimination slope and integrated discrimination improvement − properties, relationships and impact of calibration. *Statistics in Medicine* **36**(28), 4482–4490.

PENCINA, M. J., STEYERBERG, E. W. AND D'AGOSTINO, R. B. (2017*b*). Net reclassification index at event rate: properties and relationships. *Statistics in Medicine* **36**(28), 4455–4467.

PEPE, M. S., JANES, H., LONGTON, G., LEISENRING, W. AND NEWCOMB, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* **159**(9), 882–890.

PEPE, M. S., ZHENG, Y., JIN, Y., HUANG, Y., PARIKH, C. R. AND LEVY, W. C. (2008). Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis* **14**(1), 86–113.

POPOV, M., MOLSBERRY, S.A., LECCI, F., JUNKER, B., KINGSLEY, L.A., LEVINE, A., MARTIN, E., MILLER, E., MUNRO, C.A., RAGIN, A., SEABERG, E., SACKTOR, N. *and others*. (2019). Brain structural correlates of trajectories to cognitive impairment in men with and without hiv disease. *Brain Imaging and Behavior*.

PRENTICE, R. L., KALBFLEISCH, J. D., PETERSON, A. V., FLOURNOY, N., FAREWELL, V. T. AND BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34**(4), 541–554.

SAHA, P. AND HEAGERTY, P. J. (2010). Time−dependent predictive accuracy in the presence of competing risks. *Biometrics* **66**(4), 999–1011.

SCHEIKE, T. H., ZHANG, M. AND GERDS, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika* **95**(1), 205–220.

SHI, H., CHENG, Y. AND LI, J. (2014). Assessing diagnostic accuracy improvement for survival or competing−risk censored outcomes. *Canadian Journal of Statistics* **42**(1), 109–125.

SMEDEN, M. AND MOONS, K. G. M. (2017). Event rate net reclassification index and the integrated discrimination improvement for studying incremental value of risk markers. *Statistics in Medicine* **36**(28), 4495–4497.

SU, T., SCHOUTEN, J., GEURTSEN, G.J., WIT, F.W., STOLTE, I.G., PRINS, M., PORTEGIES, P., CAAN, MWA, REISS, P., MAJOIE, C.B. *and others*. (2015). Multivariate normative comparison, a novel method for more reliably detecting cognitive impairment in hiv infection. *AIDS* **29**(5), 547–557.

TSIATIS, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences* **72**(1), 20–22.

UNO, H., TIAN, L., CAI, T., KOHANE, I. S. AND WEI, L. J. (2013). A unified inference procedure for

a class of measures to assess improvement in risk prediction systems with survival data. *Statistics in Medicine* **32**(14), 2430–2442.

VAN DER VAART, A. (1998). *Asymptoptic statistics*. Cambridge: Cambridge University Press.

WANG, Z., MOLSBERRY, S.A., CHENG, Y., KINGSLEY, L., LEVINE, A.J., MARTIN, E., MUNRO, C.A., A., RAGIN, L.H., RUBIN, SACKTOR, N., SEABERG, E. *and others*. (2019). Cross-sectional analysis of cognitive function using multivariate normative comparisons in men with HIV disease. *AIDS* **33**.

WARE, J. H. (2006). The limitations of risk factors as prognostic tools. *New England Journal of Medicine* **355**(25), 2615–2617.

WASSERSTEIN, R.L., SCHIRM, A.L AND LAZAR, N.A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician* **73:sup1**, 1–19.

WU, Y. AND CHIANG, C. (2013). Optimal receiver operating characteristic manifolds. *Journal of Mathematical Psychology* **57**(5), 237 – 248. Special Issue: A Discussion of Publication Bias and the Test for Excess Significance.

ZHANG, Y. AND LI, J. (2011). Combining multiple markers for multi−category classification: An roc surface approach. *Australian & New Zealand Journal of Statistics* **53**(1), 63–78.

ZHENG, Y., CAI, T., JIN, Y. AND FENG, Z. (2012). Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics* **68**(2), 388–396.

ZHENG, Y. AND HEAGERTY, P. J. (2004). Semiparametric estimation of timeâĂŘdependent roc curves for longitudinal marker data. *Biostatistics* **5**(4), 615–632.

ZHENG, Y. AND HEAGERTY, P. J. (2007). Prospective accuracy for longitudinal markers. *Biometrics* **63**(2), 332–341.

Table 1. *Simulation details for the NRI when the added covariate improves predictability (30% censoring). Results from correct models are given in bold. True mean S was calculated using 1,000 samples with size 1,000. 1,000 samples with size 400 each was used to calculate the sample mean $\hat{S}$ and empirical standard error $SE_{\hat{S}}$. Mean of estimated standard deviations $SD_{\hat{S}}$ was outputted by formula provided. Coverage rate $CR_{\hat{S}}$=(count of true NRI entering the intervals* $[\hat{S}-1.96SD_{\hat{S}}, \hat{S}+1.96SD_{\hat{S}}]$*)/1,000.*

Data Design		Cox Regression			Fine Gray			Gerds		
		$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$
Weibull (*Case 1*)	S	**.114**	**.100**	**.126**	.114	.100	.126	.114	.100	.126
	$\hat{S}$	**.112**	**.105**	**.125**	.068	.105	.100	.108	.104	.121
	$SE_{\hat{S}}$	**.027**	**.030**	**.031**	.027	.029	.033	.025	.030	.031
	$SD_{\hat{S}}$	**.024**	**.030**	**.030**	.016	.028	.026	.024	.030	.030
	$CR_{\hat{S}}$	**.935**	**.944**	**.943**	.346	.926	.754	.930	.945	.937
		$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$
Fine Gray (*Case 2*)	S	.109	.127	.123	**.109**	**.127**	**.123**	.109	.127	.123
	$\hat{S}$	.114	.131	.127	**.115**	**.128**	**.126**	.121	.132	.126
	$SE_{\hat{S}}$	.035	.033	.028	**.038**	**.035**	**.030**	.035	.033	.030
	$SD_{\hat{S}}$	.032	.029	.025	**.033**	**.032**	**.026**	.032	.028	.025
	$CR_{\hat{S}}$	.919	.911	.902	**.911**	**.918**	**.909**	.907	.886	.873
		$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$
Gerds (*Case 3*)	S	.209	.189	.169	.209	.189	.169	**.209**	**.189**	**.169**
	$\hat{S}$	.208	.193	.176	.065	.143	.165	**.205**	**.186**	**.172**
	$SE_{\hat{S}}$	.027	.031	.031	.029	.028	.029	**.028**	**.031**	**.031**
	$SD_{\hat{S}}$	.025	.027	.030	.015	.020	.027	**.026**	**.027**	**.030**
	$CR_{\hat{S}}$	.924	.897	.933	0	.435	.933	**.929**	**.920**	**.937**

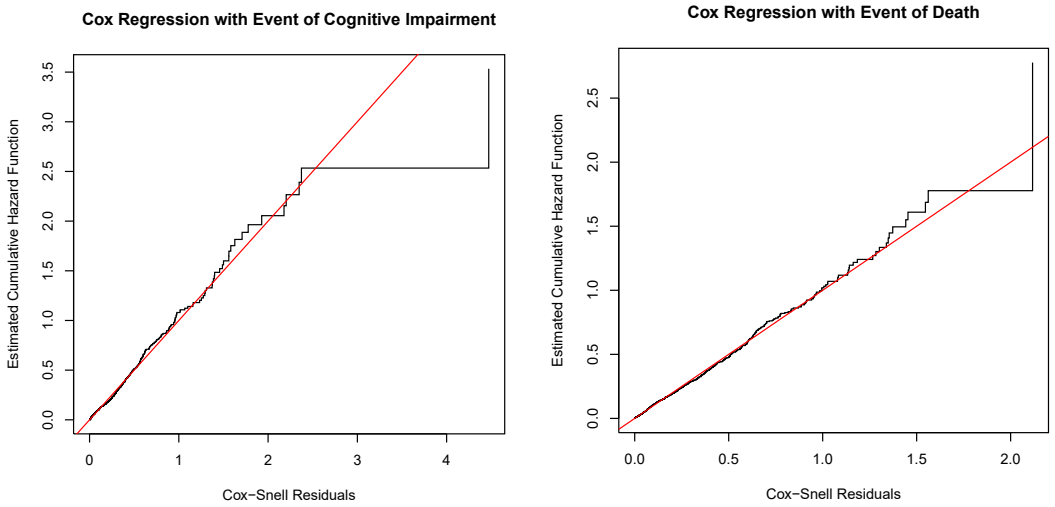

Fig. 1: Cox-Snell Residual Plots for the MACS Data with Cox Regression

Table 2. *Simulation details for the IDI when the added covariate improves predictability (30% censoring). Results from correct models are given in bold. True mean R was calculated using 1,000 samples with size 1,000. 1,000 samples with size 400 was used to calculate the sample mean $\hat{R}$ and empirical standard error $SE_{\hat{R}}$. Each sample was bootstrapped 1,000 times, and the mean of 1,000 bootstrap standard deviations is denoted as $BSD_{\hat{R}}$. Coverage rate $CR_{\hat{R}}$=(count of true IDI entering the 95% BCa bootstrap intervals)/1,000.*

Data Design		Cox Regression			Fine Gray			Gerds		
		$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$
Weibull (*Case 1*)	R	**.109**	**.106**	**.099**	.109	.106	.099	.109	.106	.099
	$\hat{R}$	**.110**	**.108**	**.100**	.018	.025	.034	.101	.092	.079
	$SE_{\hat{R}}$	**.016**	**.016**	**.015**	.005	.006	.008	.017	.016	.016
	$BSD_{\hat{R}}$	**.017**	**.016**	**.016**	.007	.008	.010	.018	.017	.016
	$CR_{\hat{R}}$	**.953**	**.952**	**.946**	0	0	.008	.935	.873	.810
		$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$
Fine Gray (*Case 2*)	R	.151	.158	.165	**.151**	**.158**	**.165**	.151	.158	.165
	$\hat{R}$	.128	.137	.144	**.148**	**.155**	**.162**	.035	.045	.055
	$SE_{\hat{R}}$	.020	.021	.022	**.020**	**.021**	**.021**	.014	.015	.016
	$BSD_{\hat{R}}$	.021	.022	.023	**.021**	**.021**	**.022**	.015	.017	.017
	$CR_{\hat{R}}$	.661	.730	.781	**.946**	**.946**	**.950**	.012	.002	.002
		$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$
Gerds (*Case 3*)	R	.288	.271	.251	.288	.271	.251	**.288**	**.271**	**.251**
	$\hat{R}$	.266	.251	.234	.034	.045	.058	**.289**	**.272**	**.252**
	$SE_{\hat{R}}$	.026	.024	.023	.007	.008	.010	**.020**	**.018**	**.018**
	$BSD_{\hat{R}}$	.027	.024	.023	.010	.011	.012	**.023**	**.020**	**.019**
	$CR_{\hat{R}}$	.512	.606	.732	0	0	0	**.947**	**.958**	**.949**

Table 3. *Simulation details for the NRI when added covariate does not improve predictability (30% censoring). Results from correct models are given in bold. 1,000 samples with size 400 each was used to calculate the sample mean $\hat{S}$, and empirical standard error $SE_{\hat{S}}$. Mean of estimated standard deviations $SD_{\hat{S}}$ was outputted by formula provided. Coverage rate $CR_{\hat{S}}$=(count of true NRI entering the intervals $[\hat{S}-1.96SD_{\hat{S}}, \hat{S}+1.96SD_{\hat{S}}]$)/1,000.*

Data Design		Cox Regression			Fine Gray			Gerds		
		$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$
Weibull (*Case 4*)	$\hat{S}$	**.005**	**.005**	**.005**	.005	.004	.003	.004	.004	.004
	$SE_{\hat{S}}$	**.015**	**.015**	**.015**	.013	.015	.015	.015	.015	.016
	$SD_{\hat{S}}$	**.013**	**.014**	**.014**	.010	.013	.015	.014	.014	.015
	$CR_{\hat{S}}$	**.915**	**.919**	**.939**	.845	.918	.940	.931	.949	.935
		$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$
Fine Gray (*Case 5*)	$\hat{S}$	.005	.004	.002	**.005**	**.004**	**.002**	.006	.004	.002
	$SE_{\hat{S}}$	.016	.015	.012	**.015**	**.015**	**.011**	.016	.016	.011
	$SD_{\hat{S}}$	.014	.014	.010	**.013**	**.013**	**.010**	.014	.014	.010
	$CR_{\hat{S}}$	.915	.934	.891	**.913**	**.913**	**.897**	.916	.919	.918
		$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$
Gerds (*Case 6*)	$\hat{S}$	.005	.005	.005	.004	.007	.005	**.005**	**.006**	**.006**
	$SE_{\hat{S}}$	.016	.016	.016	.010	.015	.016	**.018**	**.017**	**.016**
	$SD_{\hat{S}}$	.015	.015	.015	.006	.011	.014	**.015**	**.015**	**.015**
	$CR_{\hat{S}}$	.915	.922	.933	.709	.856	.907	**.769**	**.783**	**.786**

Table 4. *Simulation details for the IDI when added covariate does not improve predictability (30% censoring). Results from correct models are given in bold. 1,000 samples with size 400 each was used to calculate the sample mean $\hat{R}$ and empirical standard error $SE_{\hat{R}}$. Each sample was bootstrapped 1,000 times, and mean of 1,000 bootstrap standard deviations is denoted as $BSD_{\hat{R}}$. Coverage rate $CR_{\hat{R}}$=(count of true IDI entering the 95% BCa bootstrap intervals)/1,000.*

Model		Cox Regression			Fine Gray			Gerds		
		$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$
Weibull (*Case 4*)	$\hat{R}$	**.002**	**.002**	**.002**	.002	.002	.002	.002	.002	.002
	$SE_{\hat{R}}$	**.002**	**.003**	**.003**	.003	.003	.003	.002	.002	.002
	$BSD_{\hat{R}}$	**.004**	**.004**	**.005**	.004	.005	.005	.004	.004	.004
	$CR_{\hat{R}}$	**.962**	**.947**	**.953**	.897	.903	.909	.955	.957	.955
		$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$
Fine Gray (*Case 5*)	$\hat{R}$	.002	.003	.003	**.002**	**.002**	**.002**	.0004	.0007	.0009
	$SE_{\hat{R}}$	.003	.003	.003	**.002**	**.002**	**.003**	.0006	.0008	.001
	$BSD_{\hat{R}}$	.004	.005	.005	**.004**	**.004**	**.004**	.001	.001	.002
	$CR_{\hat{R}}$	.864	.857	.842	**.776**	**.773**	**.755**	.865	.872	.870
		$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$
Gerds (*Case 6*)	$\hat{R}$	.003	.003	.003	.002	.002	.002	**.002**	**.002**	**.002**
	$SE_{\hat{R}}$	.003	.003	.003	.003	.003	.004	**.002**	**.002**	**.003**
	$BSD_{\hat{R}}$	.005	.005	.005	.004	.005	.005	**.004**	**.004**	**.004**
	$CR_{\hat{R}}$	.953	.927	.932	.910	.907	.905	**.960**	**.943**	**.935**

Table 5. *NRI and IDI results for the MACS data at times 10 and 12 years. Competing risk censoring by death occurred when subjects died without cognitive impairment.*

Model	Time	NRI	IDI	Brier Score with Event	
				Cognitive Impairment	Death
Cox Regression	t=10	.042 [.027, .056]	.049 [.039, .065]	.089	.112
	t=12	.079 [.062, .096]	.060 [.048, .078]	.098	.128
Gerds	t=10	.018 [.007, .028]	.050 [.040, .068]	.090	.116
	t=12	.065 [.049, .081]	.055 [.044, .074]	.098	.137
Fine and Gray	t=10	.020 [.011, .029]	.030 [.022, .041]	.090	.116
	t=12	.073 [.058, .087]	.038 [.028, .051]	.100	.133

Preflight Results

Document Overview

Title:
Author:
Creator: TeX
Producer: MiKTeX pdfTeX-1.40.15

Preflight Information

Profile: Convert to PDF/A-1b
Version: Qoppa jPDFPreflight v2020R2.01
Date: Jun 5, 2021 4:01:00 PM

Legend: (X) - Can NOT be fixed by PDF/A-1b conversion.
(!X) - Could be fixed by PDF/A-1b conversion. User chose to be warned in PDF/A settings.

Page 6 Results
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 7 Results
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 8 Results
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.