Published in final edited form as:

Stat Med. 2017 November 10; 36(25): 4041–4049. doi:10.1002/sim.7414.

On comparing two correlated C indices with censored survival data

Xiaoxia Han^a, Yilong Zhang^b, and Yongzhao Shao^{a,*†}

^aDepartments of Population Health, New York University School of Medicine

^bMerck Research Laboratories, Rahway, NJ

Abstract

As new biomarkers and risk prediction procedures are in rapid development, it is of great interest to develop valid methods for comparing predictive power of two biomarkers or risk score systems. Harrell's C statistic has been routinely used as a global adequacy assessment of a risk score system, and the difference of two Harrell's C statistics as a test statistic has been suggested in recent literature for comparison of predictive power of two biomarkers for censored outcome. In this study, we showed that such a test can have severely inflated type I errors as the difference between the two Harrell's C statistics does not converge to zero under the null hypothesis of equal predictive power measured by concordance probabilities, as illustrated by two counterexamples and corresponding numerical simulations. We further investigate a necessary and sufficient condition under which the difference of two Harrell's C statistics converges to zero under the null hypothesis.

Keywords

C index; concordance probability; censored data; predictive models; discriminatory power

1. Introduction

As new biomarkers and diagnostic/prognostic procedures are in rapid development, methods to evaluate and compare performance of two biomarkers or two given scoring algorithms are of great current interest. In particular, the discriminatory accuracy for predicting binary outcomes is often compared using the receiver operating characteristic (ROC) curve, e.g. the area under the ROC curve (AUC) which is a measure of concordance probability [1, 2]. In the case of testing added value of new biomarkers to an existing predictive model, Pepe *et al.* [3] have shown that under some regularity conditions the hypothesis of no added values for nested models is equivalent to the hypothesis that the parameters corresponding to the added new predictors equal to 0. As follow up, many experts have advocated using likelihood ratio test for the significance of the added predictors instead of direct testing of performance measures such as differences between the correlated AUCs. Clearly, the usefulness of this

*Correspondence to: Yongzhao Shao, Ph.D., Departments of Population Health, New York University School of Medicine. 650 First Avenue, Room 538, New York, NY 10016, USA.

†Yongzhao.Shao@nyumc.org

approach depend crucially on the assumption of correctly-specified nested models. Importantly, correct model specification is frequently challenging in applications, e.g. for censored survival outcomes where specifying a correct model for the unknown censoring distribution is often practically impossible. Therefore, it is of interest to study nonparametric procedures [4, 5, 6, 7]. Additionally, in many applications, it is important to compare two non-nested biomarkers or predictive models. For example, Wieand *et al.* [8] investigated a case-control study for pancreatic cancer to determine which of the two serum biomarkers, a cancer antigen CA-125 and a carbohydrate antigen CA-19-9, can better distinguish cases from control. Etzioni *et al.* [9] analyzed serum levels of prostate specific antigen (PSA) for prostate cancer and compared the predictive performance of two different risk scores based on measures of PSA, namely total serum PSA and the ratio of free to total PSA. For Alzheimer's disease (AD), AD patients tend to have increased values of cerebrospinal fluid (CSF) tau protein than the normal controls [10, 11, 12], and recent studies have shown that there is a significant elevation of plasma tau in AD patients [13, 14, 15]. Therefore, comparing performance of the invasive and expensive CSF tau protein [16, 17] to the non-invasive plasma tau protein is of great interest for diagnosing Alzheimer's disease (AD) [18]. In short, it is important to investigate nonparametric procedures to compare performance for two correlated non-nested biomarkers or predictive models. For predicting a binary outcomes, nonparametric comparison of the AUCs between two ROCs of correlated risk scores X and Y can often be made using DeLong's test [4]. However, nonparametric comparison of predictive accuracy for survival outcomes is much more complicated in the presence of censoring as will be discussed next.

For a randomly selected subject, let T , D and X denote event time, censoring time, and predictive score, respectively. Without much loss of generality, for easy exposition, we assume no tied observations in T , D and X . Let $\tilde{T} = \min(T, D)$ denote the censored observed time and $\delta = I(T < D)$ denote event indicator under right censoring. Harrell's C statistic has been routinely used as a global adequacy assessment of a risk score system. More specifically, for independently observed data $(\tilde{T}_i, \delta_i, X_i)$, $i = 1, \dots, n$, Harrell's C statistic [19] is defined as

$$\hat{C}_{TX} = \frac{\sum_{i \neq j} \delta_i I(\tilde{T}_i < \tilde{T}_j) I(X_i > X_j)}{\sum_{i \neq j} \delta_i I(\tilde{T}_i < \tilde{T}_j)}. \quad (1)$$

Note that $\hat{C}_{TX}$ is commonly regarded as an estimate for the following concordance probability [20]

$$C_{TX} = pr(X_1 > X_2 | T_1 < T_2), \quad (2)$$

where (T_1, X_1) and (T_2, X_2) are bivariate observations from a pair of randomly selected independent subjects. The concordance index C_{TX} is a useful global assessment of the predictive power of a risk score system X . Unfortunately, $\hat{C}_{TX}$ is generally a biased estimator of C_{TX} and its bias depends on the unknown censoring distribution [5]. More precisely,

instead of converging to the concordance probability C_{TX} , it is known that Harrell's C Statistic $\hat{C}_{TX}$ converges to the following B_{TX} :

$$B_{TX} = pr(X_1 > X_2 | T_1 < T_2, T_1 \leq D_1 \wedge D_2), \quad (3)$$

which clearly depends on the distributions of censoring time D_1 and D_2 . Note that without censoring, we have $D_1 \wedge D_2 = \infty$, thus $C_{TX} = B_{TX}$. That is, Harrell's C statistic $\hat{C}_{TX}$ converges to the concordance probability C_{TX} in the absence of right censoring.

Let X and Y be two biomarkers or risk prediction procedures for a subject with event time T . A question of interest is to compare predictive powers between X and Y , i.e., to test the following null hypothesis

$$H_0: C_{TX} = C_{TY}. \quad (4)$$

For testing the above hypothesis, Antolini *et al.* [21] proposed a jackknife based methodology to compare two correlated C indices. Uno *et al.* [5] and Liu *et al.* [6] independently investigated test procedures based on inverse-probability-of-censoring weighted (IPCW) C statistic defined as

$$\tilde{C}_{TX} = \frac{\sum_{i \neq j} \delta_i \{\hat{G}(X_i)\}^{-2} I(\tilde{T}_i < \tilde{T}_j) I(X_i > X_j)}{\sum_{i \neq j} \delta_i \{\hat{G}(X_i)\}^{-2} I(\tilde{T}_i < \tilde{T}_j)}, \quad (5)$$

where $\hat{G}(\cdot)$ is the Kaplan-Meier estimator for the censoring distribution. The IPCW C statistic is consistent to C_{TX} when censoring is assumed to be noninformative, i.e., the survival time T is independent of the censoring time D [5, 6]. However, the survival time T is generally not independent of the censoring time D in practice. Researchers commonly adopt the conditional independent censoring assumption that event time is independent of censoring time conditional on covariates [22]. For example, subjects in a study are frequently lost to follow-up differentially according to the severity of the disease. Moreover, this IPCW based approach is computationally intensive since it involves resampling based evaluation of the variance of the test statistic as well as p -values. Specifically, let $\tilde{C} = \tilde{C}_{TX} - \tilde{C}_{TY}$ denote the difference between two IPCW-based C statistics, Uno *et al.* [5] and Liu *et al.* [6] both proposed a Wald type test statistic that require estimation of the standard error of $\tilde{C}$, $se(\tilde{C})$. To calculate the standard error of the difference between two IPCW C-statistics, Uno *et al.* [5] used a perturbation resampling method, while Liu *et al.* [6] provided an asymptotic variance formula and estimated it empirically.

Liu *et al.* [6] also numerically compared the IPCW method to the test procedure based on the difference between two Harrell's C statistics, i.e., $\hat{C} = \hat{C}_{TX} - \hat{C}_{TY}$. Interestingly, their simulation results indicated that the $\hat{C}$ based procedure produced negligible bias in estimating $C = C_{TX} - C_{TY}$. Recently, for testing $H_0: C_{TX} = C_{TY}$, Kang *et al.* [7] proposed

a formal nonparametric approach using $\hat{C}$ as a test statistic and developed an analytical form of the standard error of $\hat{C}$. Compared with the resampling procedure, the testing procedure proposed by Kang *et al.* [7] is computational efficient. In their simulation study, they simulated two predictive scores X and Y from bivariate normal distribution under independent censoring. Under this setting, their simulation results indicated that the $\hat{C}$ led to negligible bias, had satisfactory performance in terms of Type I error,. Kang *et al.* [7] also provided a publicly accessible R package `compareC` to implement their test procedure. However, since Harrell's C statistic $\hat{C}_{TX}$ is well known to be a biased estimate of the concordance probability C_{TX} . In general, $\hat{C} = \hat{C}_{TX} - \hat{C}_{TY}$ may not have mean zero even asymptotically under the null hypothesis $H_0 : C_{TX} = C_{TY}$.

In this paper, we address two important questions regarding the comparison of two correlated C indices. 1) Does the test procedure based on the difference of Harrell's C statistics have valid type I error in general? 2) If not, what is the necessary and sufficient condition under which $\hat{C} = \hat{C}_{TX} - \hat{C}_{TY}$ converges to zero under the null hypothesis $H_0 : C_{TX} = C_{TY}$. The rest of this paper is organized as follows. In section 2, we provide a change point Cox proportional hazard (PH) model and a threshold model as two practical counterexamples, in which the difference of two Harrell's C statistics (i.e., $\hat{C}$) based approach can have severely inflated Type I error. We also compare Kang *et al.*'s with the IPCW method. Moreover, we investigate the necessary and sufficient condition under which $\hat{C}$ converges to zero under the null hypothesis $H_0 : C_{TX} = C_{TY}$. The paper is concluded in Section 3.

2. Results

Under $H_0 : C_{TX} = C_{TY}$, for the test procedure based on $\hat{C}$ to be valid, it is necessary that $\hat{C}$ is consistent to 0. It is equivalent to $B_{TX} - B_{TY} = 0$, since $\hat{C}_{TX}$ converges to B_{TX} and $\hat{C}_{TY}$ converges to B_{TY} . However, the condition $B_{TX} - B_{TY} = 0$ under $H_0 : C_{TX} = C_{TY}$ does not always hold. In this section, we first discuss two counterexamples and then provide the necessary and sufficient condition for $B_{TX} - B_{TY} = 0$ under $H_0 : C_{TX} = C_{TY}$.

2.1. Counterexamples

Cox PH model is widely used in survival analysis, and it assumes the hazard ratio (HR) does not depend on time. However, in many applications, a change in HR occurs frequently at some time point during followup. For example, HR may change after a major operation such as organ transplantation or when a treatment does not exert its effect until a time lag passes. Specifically, Liang *et al.* [23] developed a change point variant of the Cox PH model to investigate relative risks varying with early versus late onset of disease while incorporating other covariates. Here we consider a similar change point Cox PH model with two covariates X and Y . The hazard ratio of X does not change over time, while the hazard ratio of Y is different before and after a change point τ . We assume an exponential baseline hazard with constant hazard rate λ_0 , and thus the hazard function is written as

$$\lambda(t|X, Y) = \lambda_0 \exp(\beta_1 X + \beta_2(t)Y), \quad (6)$$

where

$$\beta_2(t) = \beta_{20}I(t \leq \tau) + \beta_{21}I(t > \tau), \text{ and } \beta_{20} \neq \beta_{21}.$$

Under this setting, risk score X has effect β_1 , which does not change during the follow up time while risk score Y has effect β_{20} before time τ and β_{21} after time τ .

We conduct simulations studies to investigate the nonparametric test procedure proposed by Kang et. al. [7] using their package `compareC`. For the change point model (6), covariates X and Y are both generated from Uniform[0, 1]. Parameters in hazard function are $\lambda_0 = 0.01$, $\beta_1 = 3$, $\beta_{20} = 0.1$, and $\beta_{21} = 15$. The change point $\tau = 8.59$ is selected such that $C_{TX} = C_{TY} = 0.655$. We consider both exponential censoring $\exp(\lambda)$ and uniform censoring $U(0, c_2)$. Censoring parameter λ and c_2 are chosen to achieve the desired censoring proportions. We repeat the simulation 5000 times and calculate the proportion of rejection being observed in these simulation experiments. The simulation results are summarized in Table 1.

As shown in Table 1, simulation results indicate that, without censoring, test procedure based on $\hat{C}$ works reasonably well. More specifically, both $\hat{C}_{TX}$ and $\hat{C}_{TY}$ are close to the true value 0.655, and the proportions of reject H_0 are close to the nominal values in the absence of censoring. When there is censoring, it is no surprise that both $\hat{C}_{TX}$ and $\hat{C}_{TY}$ are biased, since Harrell's C statistic depends on censoring distribution. Note that the biases for $\hat{C}_{TX}$ and $\hat{C}_{TY}$ are not the same, and thus $\hat{C}$ is not always close to zero when there is censoring. When censoring proportion is 20%, the bias is greater than 0.04, and when censoring proportion increase to 50%, the bias is as large as 0.12. Moreover, the bias does not decrease as sample size increases. In addition, for sample size $n=100$, the rejection proportions at 5% level is far from 5%, i.e. 40% for exponential censoring and higher than 50% for uniform censoring. As sample size increase to $n=500$, the rejection proportions at 5% level are almost 100% under both censoring distribution.

In addition to the above Cox change-point HR model (6), an additional counterexample is the threshold model. In this case, C_{TX} and C_{TY} can be the same, but the value of B_{TX} and B_{TY} can be different since censoring time can have different effects on $\hat{C}_{TX}$ and $\hat{C}_{TY}$. Specifically, for some fixed time τ , we consider two predictive scores X and Y (conditioning on T) generated from a mixture of bivariate normal distribution, i.e.,

$$(X, Y) | T \sim I(T > \tau) \mathcal{N}(\mu, \Sigma_1) + I(T < \tau) \mathcal{N}(\mu, \Sigma_2), \quad (7)$$

where

$$\mu = (u_0 \log(T), u_0 \log(T)), \quad \Sigma_1 = \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} \sigma_2^2 & \rho \\ \rho & \sigma_1^2 \end{pmatrix}.$$

The parameter ρ controls correlation between X and Y conditional on T . When $\sigma_2^2 < \sigma_1^2$, biomarker X has higher prognostic accuracy for early event, while biomarker Y has higher prognostic accuracy for later event.

For the threshold model (7), in the simulation, event time T is generated from the standard exponential distribution (i.e. with hazard rate 1). Two risk scores X and Y are generated from a mixture of bivariate normal distribution described in equation (7), with parameters $\mu_0 = 2$, $\rho = 0$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 0.25$. The cut point $\tau = 0.718$ is selected such that $C_{TX} = C_{TY} = 0.803$. We consider both exponential censoring $\exp(\lambda)$ and uniform censoring $U(0, c_2)$. The censoring parameters λ and c_2 are chosen to achieve the desired censoring proportions. We repeat the simulation 5000 times and calculate the proportion of rejection being observed in these simulation experiments. Simulation results are summarized in Table 2. Briefly, the simulation results are similar to the change point model. These counterexamples indicate that Kang *et al.*'s [7] test procedure based on $\hat{C}$ under right censoring can lead to inflated Type I error and non-negligible bias. Therefore, the suggested test procedure based on $\hat{C}$ is invalid in general.

In addition, we conducted additional simulation studies to compare the one-shot non-parametric approach [7] and the IPCW method [6] for the threshold model and change point model. The set-up for simulation parameters is the same for change point model and threshold model as described above. We consider the exponential censoring distribution, $\exp(\lambda)$, and values of the censoring parameter λ are chosen to achieve the desired censoring proportions. We repeat the simulation 5000 times and calculate the proportion of rejection being observed in these simulation experiments. Tables 3 and 4 summarized the simulation results for sample size $n=100, 200$ and 400 . Generally, the biases of both methods increase as censoring proportion increases, while the IPCW method produces a relative smaller bias compared to the one-shot nonparametric approach. For the one-shot nonparametric approach, the expected Type I error is seriously inflated as censoring proportion increases in both threshold model and change point model. For the IPCW method, the expected Type I error first becomes conservative and then becomes inflated as censoring proportion increases in threshold model, while the Type I error for the IPCW method becomes too conservative as censoring proportion increases in the change point model (6). Moreover, the simulation results for various sample sizes indicated that the adverse impact of censoring pattern on both methods could not be reduced by increasing sample size.

2.2. A necessary and sufficient condition

After introducing two counterexamples, we next investigate a necessary and sufficient condition under which $\hat{C}$ converges to zero under the null for censored survival outcomes.

Proposition 1—Under the null hypothesis $H_0 : C_{TX} = C_{TY}$, if $p_T(X_1 > X_2, T_1 < T_2)$ and $p_T(Y_1 > Y_2, T_1 < T_2)$ are both non-zero, for a pair of multivariate independent observations (X_1, Y_1, T_1, D_1) and (X_2, Y_2, T_2, D_2) , a necessary and sufficient condition for $B_{TX} = B_{TY}$ is

$$pr(T_1 < D_1 \wedge D_2 | X_1 > X_2, T_1 < T_2) = pr(T_1 < D_1 \wedge D_2 | Y_1 > Y_2, T_1 < T_2). \quad (8)$$

Proof 1—Since $C_{TX} = C_{TY}$, i.e., $pr(X_1 > X_2 | T_1 < T_2) = pr(Y_1 > Y_2 | T_1 < T_2)$, by Bayes theorem, we have

$$pr(X_1 > X_2, T_1 < T_2) = pr(Y_1 > Y_2, T_1 < T_2). \quad (9)$$

Using Bayes theorem again, we have $B_{TX} = B_{TY}$, i.e., $pr(X_1 > X_2 | T_1 < T_2, T_1 < D_1 \wedge D_2) = pr(Y_1 > Y_2 | T_1 < T_2, T_1 < D_1 \wedge D_2)$ holds if and only if

$$pr(T_1 < D_1 \wedge D_2, X_1 > X_2, T_1 < T_2) = pr(T_1 < D_1 \wedge D_2, Y_1 > Y_2, T_1 < T_2). \quad (10)$$

By bayes theorem and equation (9), equation (8) holds if and only if equation (10) holds.

Given the above counterexamples, the applications of the test by Kang et al. [7] to real data cannot be properly justified unless a checkable sufficient condition for $B_{TX} = B_{TY}$ under H_0 $C_{TX} = C_{TY}$ is identified. Thus the above necessary and sufficient condition for $B_{TX} = B_{TY}$ under H_0 $C_{TX} = C_{TY}$ will be crucial when applying the test by Kang et al. [7].

Unfortunately, in practical applications, it seems extremely hard to check the above necessary and sufficient condition, i.e. equality of the two concordance probabilities in Proposition 1 equation (8). Next, we provide a sufficient condition in Proposition 2 which is not easy to check in practical applications, but it can be used to properly understand the simulation results of Liu *et al.* [6] and Kang *et al.* [7].

Proposition 2—If X and Y conditional on T and D have the same distribution, then

$$pr(T_1 < D_1 \wedge D_2, X_1 > X_2, T_1 < T_2) = pr(T_1 < D_1 \wedge D_2, Y_1 > Y_2, T_1 < T_2). \quad (11)$$

Proof 2—Let $X | T, D$ and $Y | T, D$ denote the distribution of X and Y conditional on T and D , respectively. Suppose $X | T, D$ and $Y | T, D$ have the same distribution. For a pair of multivariate independent observations (X_1, Y_1, T_1, D_1) and (X_2, Y_2, T_2, D_2) , we then have $(X_1, T_1, D_1), (X_2, T_2, D_2), (Y_1, T_1, D_1)$ and (Y_2, T_2, D_2) follow the same distributions. Due to independence, the distribution of $(X_1, T_1, D_1, X_2, T_2, D_2)$ and $(Y_1, T_1, D_1, Y_2, T_2, D_2)$ are the same. Therefore we have (11).

Remark 1—Propositions 1 and 2 provide theoretical insight for proper understanding of the simulation results of Liu *et al.* [6] and Kang *et al.* [7]. In both of their simulation studies, independent censoring was assumed. Moreover, $(X, Y) | T$ were simulated from bivariate normal distribution with mean μ and covariance matrix Σ , where

$$\mu = \log(T) \times (1, 1), \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Obviously, the conditional distribution $X|T, D$ and $Y|T, D$ are both normal with mean $\log(T)$ and variance 1, and thus the condition in Proposition 2 is satisfied. Under the null hypothesis, equation (8) holds by equation (9) and bayes theorem. Therefore, it is not surprising that both simulation studies showed negligible bias for $\hat{C}$ in Liu *et al.* [6] and Kang *et al.* [7] because the if and only if condition in Proposition 1 is satisfied.

Remark 2—Equation (8) does not necessarily hold in general as shown in the two counterexamples in which $X|T$ and $Y|T$ do not have the same distribution under the null hypothesis. Specifically, in the change point model (6), it is easy to show that

$$pr(t|X, Y) = \exp \{ - \exp(\lambda_0 t \exp(\beta_1 X + \beta_2(t)Y)) \}, \quad (12)$$

Since the marginal distribution of X and Y are both $U[0, 1]$ distribution, we have

$$pr(t|X) = \int_0^1 pr(t|X, y)dy; \quad pr(t|Y) = \int_0^1 pr(t|x, Y)dx. \quad (13)$$

It is clear that $pr(t|X)$ does not equal to $pr(t|Y)$ from equation (12). By Bayes theorem and independent censoring, $X|T, D$ and $Y|T, D$ do not have the same distribution.

In the threshold model, from equation (7),

$$X|T \sim I(T > \tau)\mathcal{N}(u_0 \log(T), \sigma_1^2) + I(T < \tau)\mathcal{N}(u_0 \log(T), \sigma_2^2),$$

$$Y|T \sim I(T > \tau)\mathcal{N}(u_0 \log(T), \sigma_2^2) + I(T < \tau)\mathcal{N}(u_0 \log(T), \sigma_1^2).$$

Given $\sigma_1^2 > \sigma_2^2$, $X|T, D$ and $Y|T, D$ do not have the same distribution. Moreover, we have shown that $B_{XT} \neq B_{YT}$ for both of the two examples via simulation studies in the previous subsection.

3. Discussion

Prediction modeling is a mainstay of statistical practice. It is of great current interest due to the promise of highly predictive biomarkers recently identified through imaging and genetic/genomic technologies. Recently, based on simulation results, Kang *et al.* [7] proposed a promising nonparametric test procedure based on the difference of two Harrell's C statistics and further developed an R package `compareC` to implement their approach. However, in general, the difference of two correlated Harrell's C statistics may not converges to zero

under the null hypothesis, thus may lead to severely inflated type I error for the proposed test. Thus, the validity of this approach is not without question in general as indicated by the counterexamples we provided in this paper. We further investigate a necessary and sufficient condition under which the difference of two Harrell's C statistics converges to zero. This result provides theoretical insight for the understanding of the simulation results of Liu *et al.* [6] and Kang *et al.* [7].

As a cautionary note, we would like to emphasize that there is no currently available practical strategy to verify the necessary and sufficient conditions we investigated in this short paper. Thus, given the counterexamples presented in this paper, the use of the test proposed in Kang *et al.* [7] is not justifiable before new practicable ways are identified to verify such conditions. An alternative approach to hypothesis testing is the use of confidence intervals for the difference of two correlated Harrell's C indices based on $\hat{C}$. Because Harrell's C statistic has an explicit formula and is very easy to compute, it is straightforward to calculate the standard error $se(\hat{C})$ and confidence intervals using bootstrap as in Uno *et al.* [5] and Liu *et al.* [6] for the difference of the modified C statistics $\tilde{C}$. However, both the standard error $se(\hat{C})$ and the corresponding confidence interval depend on the unknown censoring distributions that are frequently different for different studies. Thus, despite easy implementation using available statistical packages and algorithms, practical applications of such confidence intervals for the difference of Harrell's C indices should be avoided or accompanied with warnings because of the unclear validity/reliability and difficult interpretation of such confidence intervals.

Acknowledgments

This work was supported in part by NIH grants P30 AG0851, 2P30 CA16087 and 5P30 ES00260.

References

1. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*. 1982; 143(1):29–36. [PubMed: 7063747]
2. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*. 2005; 92(4):965–970.
3. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. *Statistics in medicine*. 2013; 32(9):1467–1482. [PubMed: 23296397]
4. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988:837–845. [PubMed: 3203132]
5. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei L. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*. 2011; 30(10):1105–1117. [PubMed: 21484848]
6. Liu X, Jin Z, Graziano JH. Comparing paired biomarkers in predicting quantitative health outcome subject to random censoring. *Statistical methods in medical research*. 2012 0962280212460 434.
7. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated c indices with right-censored survival outcome: a one-shot nonparametric approach. *Statistics in medicine*. 2015; 34(4):685–703. [PubMed: 25399736]
8. Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*. 1989; 76(3):585–592.

9. Etzioni R, Pepe M, Longton G, Hu C, Goodman G. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making*. 1999; 19(3):242–251. [PubMed: 10424831]
10. Hock C, Golombowski S, Naser W, Müller-Spahn F. Increased levels of τ protein in cerebrospinal fluid of patients with alzheimer's disease: Correlation with degree of cognitive impairment. *Annals of neurology*. 1995
11. Sunderland T, Mirza N, Putnam KT, Linker G, Bhupali D, Durham R, Soares H, Kimmel L, Friedman D, Bergeson J, et al. Cerebrospinal fluid β -amyloid 1–42 and tau in control subjects at risk for alzheimers disease: The effect of apoe ϵ 4 allele. *Biological psychiatry*. 2004; 56(9):670–676. [PubMed: 15522251]
12. Kester MI, Goos JD, Teunissen CE, Benedictus MR, Bouwman FH, Wattjes MP, Barkhof F, Scheltens P, van der Flier WM. Associations between cerebral small-vessel disease and alzheimer disease pathology as measured by cerebrospinal fluid biomarkers. *JAMA neurology*. 2014; 71(7): 855–862. [PubMed: 24818585]
13. Chiu MJ, Yang SY, Horng HE, Yang CC, Chen TF, Chieh JJ, Chen HH, Chen TC, Ho CS, Chang SF, et al. Combined plasma biomarkers for diagnosing mild cognition impairment and alzheimers disease. *ACS chemical neuroscience*. 2013; 4(12):1530–1536. [PubMed: 24090201]
14. Zetterberg H, Wilson D, Andreasson U, Minthon L, Blennow K, Randall J, Hansson O. Plasma tau levels in alzheimer's disease. *Alzheimer's research & therapy*. 2013; 5(2):1.
15. Chiu MJ, Chen YF, Chen TF, Yang SY, Yang FPG, Tseng TW, Chieh JJ, Chen JCR, Tzen KY, Hua MS, et al. Plasma tau as a window to the brain: negative associations with brain volume and memory function in mild cognitive impairment and early alzheimer's disease. *Human brain mapping*. 2014; 35(7):3132–3142. [PubMed: 24129926]
16. Amorim JA, de Barros MVG, Valença MM. Post-dural (post-lumbar) puncture headache: risk factors and clinical features. *Cephalalgia*. 2012 0333102412453 951.
17. Van Crevel H, Hijdra A, De Gans J. Lumbar puncture and the risk of herniation: when should we first perform ct? *Journal of neurology*. 2002; 249(2):129–137. [PubMed: 11985377]
18. Olsson B, Lautner R, Andreasson U, Öhrfelt A, Portelius E, Bjerke M, Hölttä M, Rosén C, Olsson C, Strobel G, et al. Csf and blood biomarkers for the diagnosis of alzheimer's disease: a systematic review and meta-analysis. *The Lancet Neurology*. 2016; 15(7):673–684. [PubMed: 27068280]
19. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*. 1984; 3(2):143–152. [PubMed: 6463451]
20. Pencina MJ, D'Agostino RB. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine*. 2004; 23(13): 2109–2123. [PubMed: 15211606]
21. Antolini L, Nam BH, D'Agostino RB. Inference on correlated discrimination measures in survival analysis: a nonparametric approach. *Communications in Statistics-Theory and Methods*. 2004; 33(9):2117–2135.
22. Tsiatis, A. *Semiparametric theory and missing data*. Springer Science & Business Media; 2007.
23. Liang KY, Self SG, Liu X. The cox proportional hazards model with change point: An epidemiologic application. *Biometrics*. 1990:783–793. [PubMed: 2242414]

Change point Cox PH simulation results. Covariates X and Y are generated from Uniform[0, 1]. Parameters in hazard function are $\lambda_0 = 0.01$, $\beta_1 = 3$, $\beta_{20} = 0.1$, and $\beta_{21} = 15$. The change point $\tau = 8.59$ is selected such that $C_{TX} = C_{TY} = 0.655$ based on 100,000 simulations. Parameters λ for exponential censoring and c_2 for uniform (0, c_2) censoring are chosen to achieve the desired censoring proportions.

Table 1

Censoring distribution / n	% censor	$\hat{C}_{TX}$	$\hat{C}_{TY}$	$\hat{C}_{TX} - \hat{C}_{TY}$	% Reject H_0 at $\alpha = 5\%$	% Reject H_0 at $\alpha = 10\%$
Exponential n=100	0%	0.655	0.655	0.000	5.6%	10.7%
	10%	0.661	0.642	0.020	7.6%	13.3%
	20%	0.669	0.627	0.042	12.7%	20.3%
	33%	0.679	0.604	0.075	24.6%	35.4%
	50%	0.694	0.575	0.119	41.8%	54.2%
n=200	0%	0.655	0.655	0.000	5.5%	10.3%
	10%	0.661	0.642	0.019	8.9%	15.7%
	20%	0.668	0.627	0.041	19.0%	27.9%
	33%	0.678	0.604	0.075	41.5%	53.6%
	50%	0.693	0.571	0.123	70.7%	80.9%
n=500	0%	0.655	0.655	0.000	4.7%	9.8%
	10%	0.661	0.642	0.019	12.6%	20.8%
	20%	0.668	0.627	0.041	37.1%	50.3%
	33%	0.679	0.604	0.075	79.3%	87.6%
	50%	0.694	0.569	0.125	97.7%	99.1%
Uniform n=100	10%	0.661	0.641	0.020	8.1%	13.8%
	20%	0.669	0.625	0.044	13.1%	21.4%
	33%	0.679	0.602	0.077	26.9%	37.5%
	50%	0.697	0.568	0.129	51.5%	62.9%
	10%	0.661	0.642	0.019	8.8%	15.3%
n=200	20%	0.669	0.626	0.043	20.7%	30.3%
	33%	0.68	0.602	0.077	44.7%	56.6%
	50%	0.698	0.563	0.135	80.1%	87.7%
	10%	0.661	0.642	0.019	12.7%	21.1%
	20%	0.669	0.626	0.042	40.7%	52.9%

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Censoring distribution / n	% censor	$\hat{\mathcal{C}}_{TX}$	$\hat{\mathcal{C}}_{TY}$	$\hat{\mathcal{C}}_{TX} - \hat{\mathcal{C}}_{TY}$	% Reject H_0 at $\alpha = 5\%$	% Reject H_0 at $\alpha = 10\%$
	33%	0.679	0.603	0.077	81.6%	89.1%
	50%	0.697	0.562	0.135	99.3%	99.7%

Table 2

Threshold model simulation results. Event time T is generated from exponential distribution with rate 1. Two risk scores X and Y are generated from a mixture of bivariate normal distribution described in equation (7), with parameters $\lambda_0 = 2$, $\rho = 0$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 0.25$. The cut point $\tau = 0.718$ is selected such that $C_{TX} = C_{TY} = 0.803$ based on 100,000 simulations. Parameters λ for exponential censoring and c_2 for uniform $(0, c_2)$ censoring are chosen to achieve the desired censoring proportions.

Censoring distribution / n	% censor	$\hat{C}_{TX}$	$\hat{C}_{TY}$	$\hat{C}_{TX} - \hat{C}_{TY}$	% Reject H_0 at $\alpha = 5\%$	% Reject H_0 at $\alpha = 10\%$
Exponential n=100	0%	0.802	0.803	-0.001	5.5%	10.7%
	10%	0.815	0.804	0.011	6.9%	13.1%
	20%	0.830	0.806	0.024	10.3%	17.3%
	33%	0.851	0.809	0.042	20.2%	29.0%
	50%	0.883	0.818	0.065	39.2%	51.5%
n=200	0%	0.802	0.802	0.000	4.7%	10.0%
	10%	0.815	0.803	0.012	8.0%	13.8%
	20%	0.830	0.805	0.025	16.0%	25.4%
	33%	0.851	0.809	0.043	34.7%	45.7%
	50%	0.883	0.817	0.067	66.9%	76.8%
n=500	0%	0.802	0.803	-0.001	5.2%	10.0%
	10%	0.816	0.804	0.012	10.9%	17.9%
	20%	0.831	0.806	0.025	30.8%	42.5%
	33%	0.852	0.809	0.043	69.1%	79.3%
	50%	0.884	0.818	0.066	96.6%	98.5%
Uniform n=100	10%	0.815	0.803	0.012	7.0%	12.5%
	20%	0.829	0.804	0.025	10.7%	18.1%
	33%	0.847	0.806	0.041	19.3%	28.4%
	50%	0.879	0.811	0.067	42.0%	54.0%
n=200	10%	0.815	0.803	0.012	7.9%	14.0%
	20%	0.829	0.804	0.024	15.0%	23.4%
	33%	0.847	0.806	0.041	31.7%	43.9%
	50%	0.879	0.812	0.067	68.4%	79.1%
	10%	0.815	0.803	0.012	11.1%	18.5%

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Censoring distribution / n	% censor	$\hat{C}_{TX}$	$\hat{C}_{TY}$	$\hat{C}_{TX} - \hat{C}_{TY}$	% Reject H_0 at $\alpha = 5\%$	% Reject H_0 at $\alpha = 10\%$
	20%	0.829	0.804	0.024	29.2%	40.4%
	33%	0.847	0.806	0.041	64.4%	74.7%
	50%	0.879	0.811	0.068	97.2%	98.7%

Table 3

Compare one-shot nonparametric method and IPCW method. Change point Cox PH model: covariates X and Y are generated from Uniform[0, 1]. Parameters in hazard function are $\lambda_0 = 0.01$, $\beta_1 = 3$, $\beta_{20} = 0.1$, and $\beta_{21} = 15$. The change point $\tau = 8.59$ is selected such that $C_{TX} = C_{TY} = 0.655$ based on 100,000 simulations. The values of parameter λ for exponential censoring is chosen to achieve the desired censoring proportions. Threshold model: event time T is generated from exponential distribution with the rate $\lambda = 1$. Two risk scores X and Y are generated from a mixture of bivariate normal distribution described in equation (7), with parameters $\mu_0 = 2$, $\rho = 0$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 0.25$. The cut point $\tau = 0.718$ is selected such that $C_{TX} = C_{TY} = 0.803$ based on 100,000 simulations. Parameters λ for exponential censoring is chosen to achieve the desired censoring proportions. Sample size $n=100$.

	% Censor	The one-shot nonparametric method					The IPCW method				
		$\hat{C}_{TX}$	$\hat{C}_{TY}$	$\hat{C}_{TX} - \hat{C}_{TY}$	% Reject H_0 at $\alpha = 5\%$	% Reject H_0 at $\alpha = 10\%$	$\hat{C}_{TX}$	$\hat{C}_{TY}$	$\hat{C}_{TX} - \hat{C}_{TY}$	% Reject H_0 at $\alpha = 5\%$	% Reject H_0 at $\alpha = 10\%$
		n=100									
Change point model	0%	0.657	0.660	-0.003	5.0%	9.8%	0.661	0.659	0.002	4.6%	9.1%
	10%	0.663	0.647	0.016	6.4%	11.8%	0.661	0.659	0.002	4.2%	8.3%
	20%	0.669	0.633	0.036	10.5%	18.0%	0.661	0.658	0.002	3.6%	7.3%
	33%	0.68	0.612	0.068	21.9%	31.6%	0.661	0.659	0.002	2.7%	5.7%
	50%	0.695	0.582	0.113	39.4%	51.6%	0.661	0.658	0.003	1.7%	4.1%
	65%	0.710	0.563	0.147	47.2%	59.5%	0.663	0.657	0.006	1.3%	3.1%
Threshold model	80%	0.720	0.565	0.156	36.3%	48.4%	0.682	0.656	0.027	1.2%	2.7%
	0%	0.802	0.802	0.000	5.4%	10.2%	0.803	0.802	0.001	4.9%	9.8%
	10%	0.816	0.803	0.013	7.2%	12.6%	0.803	0.802	0.001	4.5%	8.8%
	20%	0.830	0.805	0.025	10.7%	18.0%	0.803	0.802	0.001	3.8%	7.5%
	33%	0.852	0.808	0.044	20.9%	30.9%	0.805	0.802	0.003	3.3%	6.9%
	50%	0.884	0.818	0.066	40.0%	52.4%	0.809	0.803	0.006	2.1%	5.2%
	65%	0.918	0.834	0.083	57.1%	69.7%	0.821	0.803	0.018	4.7%	8.0%
	80%	0.955	0.865	0.090	65.5%	77.8%	0.863	0.802	0.061	10.2%	17.6%

Table 4

Compare one-shot nonparametric method and IPCW method. Change point Cox PH model: covariates X and Y are generated from Uniform[0, 1]. Parameters in hazard function are $\lambda_0 = 0.01$, $\beta_1 = 3$, $\beta_{20} = 0.1$, and $\beta_{21} = 15$. The change point $\tau = 8.59$ is selected such that $C_{TX} = C_{TY} = 0.655$ based on 100,000 simulations. The values of parameter λ for exponential censoring is chosen to achieve the desired censoring proportions. Threshold model: event time T distribution with the rate $\lambda = 1$. Two risk scores X and Y are generated from a mixture of bivariate normal distribution described in equation (7), with parameters $\mu_0 = 2$, $\rho = 0$, $\sigma_1^2 = 9$ and $\sigma_2^2 = 0.25$. The cut point $\tau = 0.718$ is selected such that $C_{TX} = C_{TY} = 0.803$ based on 100,000 simulations. Parameters λ for exponential censoring is chosen to achieve the desired censoring proportions. Sample size $n=200$ and 400.

		The one-shot nonparametric method					The IPCW method				
% Censor	$\hat{C}_{TX}$	$\hat{C}_{TY}$	$\hat{C}_{TX} - \hat{C}_{TY}$	% Reject H_0 at $\alpha = 5\%$	% Reject H_0 at $\alpha = 10\%$	$\hat{C}_{TX}$	$\hat{C}_{TY}$	$\hat{C}_{TX} - \hat{C}_{TY}$	% Reject H_0 at $\alpha = 5\%$	% Reject H_0 at $\alpha = 10\%$	
n=200											
Change point model	0%	0.656	0.659	-0.003	5.1%	9.9%	0.660	0.658	0.002	5.0%	9.9%
	10%	0.662	0.646	0.016	6.5%	12.8%	0.660	0.658	0.002	4.8%	8.9%
	20%	0.668	0.631	0.037	15.8%	25.7%	0.660	0.657	0.003	3.9%	7.6%
	33%	0.679	0.609	0.070	37.6%	51.2%	0.660	0.658	0.002	3.5%	7.1%
	50%	0.695	0.576	0.119	68.5%	78.0%	0.661	0.657	0.003	2.3%	5.5%
	65%	0.709	0.551	0.158	80.0%	87.5%	0.661	0.656	0.005	2.0%	4.0%
80%	0.718	0.545	0.173	69.4%	80.5%	0.664	0.653	0.010	1.0%	2.1%	
Threshold model											
0%	0.803	0.802	0.001	4.8%	10.3%	0.803	0.802	0.001	4.5%	9.9%	
10%	0.817	0.803	0.013	8.0%	14.2%	0.803	0.802	0.001	3.9%	8.9%	
20%	0.831	0.805	0.026	16.3%	25.2%	0.804	0.802	0.002	3.4%	7.7%	
33%	0.852	0.808	0.044	36.9%	48.8%	0.804	0.802	0.002	3.1%	6.5%	
50%	0.885	0.817	0.068	68.9%	79.1%	0.807	0.803	0.004	2.6%	5.6%	
65%	0.919	0.834	0.085	87.6%	92.7%	0.815	0.803	0.013	4.0%	7.0%	
80%	0.956	0.866	0.090	94.3%	97.2%	0.849	0.799	0.050	12.6%	20.7%	
n=400											
Change point model	0%	0.656	0.656	0.000	4.8%	9.6%	0.660	0.655	0.005	5.9%	10.8%
	10%	0.662	0.643	0.019	11.2%	18.4%	0.660	0.655	0.006	4.9%	9.5%
	20%	0.668	0.628	0.041	30.4%	42.2%	0.660	0.655	0.005	4.2%	8.4%

	The one-shot nonparametric method						The IPCW method				
	% Censor	$\hat{C}_{TX}$	$\hat{C}_{TY}$	$\hat{C}_{TX} - \hat{C}_{TY}$	% Reject H_0 at $\alpha = 5\%$	% Reject H_0 at $\alpha = 10\%$	$\hat{C}_{TX}$	$\hat{C}_{TY}$	$\hat{C}_{TX} - \hat{C}_{TY}$	% Reject H_0 at $\alpha = 5\%$	% Reject H_0 at $\alpha = 10\%$
	33%	0.679	0.605	0.074	61.3%	71.3%	0.660	0.655	0.006	3.4%	7.5%
	50%	0.694	0.570	0.124	54.8%	57.4%	0.660	0.654	0.006	3.6%	7.2%
	65%	0.709	0.541	0.168	37.5%	38.3%	0.661	0.654	0.007	2.3%	5.7%
	80%	0.719	0.532	0.187	55.3%	57.2%	0.662	0.653	0.010	0.8%	1.7%
Threshold model	0%	0.803	0.802	0.000	5.7%	10.1%	0.803	0.802	0.000	5.6%	9.6%
	10%	0.816	0.803	0.013	10.3%	17.7%	0.803	0.803	0.000	5.0%	9.2%
	20%	0.830	0.805	0.025	26.0%	37.5%	0.803	0.803	0.000	4.4%	8.6%
	33%	0.852	0.809	0.044	60.2%	71.5%	0.803	0.803	0.001	2.9%	6.1%
	50%	0.884	0.817	0.067	92.7%	95.8%	0.805	0.803	0.002	2.7%	5.4%
	65%	0.919	0.834	0.085	99.3%	99.7%	0.811	0.803	0.007	3.0%	5.8%
	80%	0.956	0.866	0.090	99.9%	100.0%	0.838	0.800	0.038	14.6%	20.8%