Research Article

Statistics in Medicine

Received 7 June 2013, Accepted 29 October 2014 Published online 17 November 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6370

Comparing two correlated *C* indices with right-censored survival outcome: a one-shot nonparametric approach

Le Kang,[a,b*†] Weijie Chen[b*†], Nicholas A. Petrick[b] and Brandon D. Gallas[b]

The area under the receiver operating characteristic curve is often used as a summary index of the diagnostic ability in evaluating biomarkers when the clinical outcome (truth) is binary. When the clinical outcome is right-censored survival time, the *C* index, motivated as an extension of area under the receiver operating characteristic curve, has been proposed by Harrell as a measure of concordance between a predictive biomarker and the right-censored survival outcome. In this work, we investigate methods for statistical comparison of two diagnostic or predictive systems, of which they could either be two biomarkers or two fixed algorithms, in terms of their *C* indices. We adopt a *U*-statistics-based *C* estimator that is asymptotically normal and develop a nonparametric analytical approach to estimate the variance of the *C* estimator and the covariance of two *C* estimators. A *z*-score test is then constructed to compare the two *C* indices. We validate our one-shot nonparametric method via simulation studies in terms of the type I error rate and power. We also compare our one-shot method with resampling methods including the jackknife and the bootstrap. Simulation results show that the proposed one-shot method provides almost unbiased variance estimations and has satisfactory type I error control and power. Finally, we illustrate the use of the proposed method with an example from the Framingham Heart Study. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: *C* index; bootstrap; concordance; hypothesis testing; jackknife; type I error; power

1. Introduction

Methods for assessing and comparing diagnostic performance are of increasing importance as new biomarkers and high-throughput molecular diagnostics are in rapid development. When a diagnostic test for a binary outcome, for example, non-diseased and diseased, is based on an observed variable that lies on the ordinal scale, the receiver operating characteristic (ROC) curves and the area under the ROC curves (AUC) are commonly used diagnostic accuracy measures [1–5].

A number of methods have been proposed to estimate the AUC and its variance [2, 6–8]. There have also been parametric and nonparametric methods developed to compare correlated AUCs. Metz *et al.* [9] proposed a bivariate binormal model. Hanley and McNeil [10] provided a table that converts the observed correlations in diagnostic scores between two modalities into a correlation between two AUCs. DeLong *et al.* [11] presented a nonparametric method for comparing correlated AUCs on the basis of a structural components variance estimate following the work of Sen [12] and the asymptotic normality of the *U*-statistic AUC estimator.

Despite the prevailing use of AUC as a summary index of the ROC curve in the context of dichotomous outcomes [13–15], it has limitations in evaluating and comparing biomarkers with censored survival outcomes. Essentially, ROC analysis requires a reference standard that categorizes subjects into diseased and

[a]*Department of Biostatistics, School of Medicine, Virginia Commonwealth University, Richmond, VA, U.S.A.*
[b]*Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD, U.S.A.*
[*]*Correspondence to:*
Le Kang, VCU Medical Center Department of Biostatistics, 830 East Main Street, Richmond, VA 23298-0032, U.S.A.
Weijie Chen, US Food and Drug Administration Division of Imaging, Diagnostics, and Software Reliability, 10903 New Hampshire Avenue, Silver Spring, MD 20993-0002, U.S.A.
[†]*E-mail: LKang@vcu.edu; Weijie.Chen@fda.hhs.gov*

non-diseased populations. However, the survival outcome, that is, time to event, is typically continuous rather than binary. Additionally, the conventional ROC analysis cannot handle censored outcome data.

There have been generalizations of ROC analysis that can overcome these constraints. The time-dependent ROC curves [16–18] have been proposed in thinking ROC curves as a varying function of time t. Intuitively, the censored outcome can be dichotomized given time point t, for example, being 1 if a subject has died prior to time t and 0 otherwise. Correspondingly, the area under the time-dependent ROC curves has been considered [19].

Smith *et al.* [20] generalized the area under the empirical ROC curve to a concordance measure that allows for polytomous ordinal patient outcomes. Obuchowski [21] proposed a concordance-type measure of diagnostic accuracy that differs from Smith *et al.* in how it deals with ties in the reference standard. While these methods accommodate polytomous and continuous data, neither handles censoring of clinical outcomes such as those in survival analysis.

The overall C index, motivated as an extension of AUC to survival analysis, has been proposed by Harrell *et al.* [22, 23]. It is a conditional concordance probability measure between a survival outcome that is possibly right censored and a predictive-score variable, which can represent a measured biomarker or a composite-score output from an algorithm that combines multiple biomarkers. Various popular extensions of the C index have been proposed in the literature since then [19, 24, 25]. Pencina *et al.* [26] studied these different C statistics systematically and concluded that the C index proposed by Harrell *et al.* [22, 23] is the most appropriate in capturing the discriminating ability of a predictive variable to separate those with longer event-free survival from those with shorter event-free survival within some time horizon of interest.

Here we consider a dataset collected on a cohort of n subjects. The actual right-censored survival time of each subject is denoted as X_i with censoring indicator δ_i ($i = 1, 2, \ldots, n$), where $\delta_i = 1$ if an event of interest occurred (e.g., death) and 0 if censored. Together with two predictive variables denoted as Y and Z, the collected data can be arranged in a matrix form with each row representing observations for a subject

$$\begin{pmatrix} X & \delta & Y & Z \\ \hline X_1 & \delta_1 & Y_1 & Z_1 \\ X_2 & \delta_2 & Y_2 & Z_2 \\ \vdots & \vdots & \vdots & \vdots \\ X_n & \delta_n & Y_n & Z_n \end{pmatrix}. \tag{1}$$

Upon randomly drawing a pair of subjects, Harrell *et al.* [22] defined the overall C index between the right-censored survival time X and the predictive score Y (or Z) as the probability that the subject with the higher values of Y (or Z) had the longer survival time X, given that the order of two survival times can be validly inferred. Values of C near 0.5 indicate that the predictive score is no better than tossing a coin in determining which subject will live longer, while values of C near 0 or 1 indicate that the score, lower or higher, virtually always determines which subject has a better prognosis [22].

For two predictive tests performed on the same cohort of patients, it is essential to account for the correlated nature of the data to make a formal statistical comparison in terms of the C index. Pencina and D'Agostino [27] and Nam and D'Agostino [28] investigated asymptotic variance estimations as well as confidence interval constructions for a single C. Antolini *et al.* [29] developed methods for comparing two correlated C indices, in which a consistent estimator for the variance of the difference of two C indices was derived via the jackknife approach. In this article, we develop a one-shot estimator that does not require resampling for the variance of the difference of two C indices. A z-score test is then constructed to statistically compare the two C indices. For comparison purposes, we also consider a bootstrap resampling approach [30, 31].

The rest of our article is organized as follows. In Section 2, we discuss the C index in detail and present the relationship between the C index and the generalized Kendall's tau. In Section 3, we present the proposed variance estimator and the test statistic for the difference between two correlated C indices. Moreover, the procedures for carrying out the jackknife approach [29] and the bootstrap resampling approach are also described. We present extensive simulation studies in Section 4 for validating the performance of the proposed test statistic and compare that with the jackknife and the bootstrap approach in terms of type I error rate and statistical power. In Section 5, we apply the proposed method to an example from the Framingham Heart Study data [32] to compare the ability of biomark-

ers in predicting heart-disease-free survival. Finally, a broader discussion on assessing and comparing diagnostic/prognostic accuracy is presented in Section 6.

2. The C index

In this section, we provide a review of the historical development of the prediction probability P_K of Smith *et al.* [20], the C index [22, 27], and their relationships with various versions of Kendall's tau. We show under a unified framework that the C index is a linear function of our generalized Kendall's tau. Because of this relation, the generalized Kendall's tau will serve as a vehicle for making inference about the C index.

2.1. Outcomes with no censoring

First, assume that the survival time X is actually observed without any censoring, that is, $\delta_i = 1$ $(i = 1, 2, \ldots, n)$. Following Smith *et al.* [20], upon randomly drawing a pair of subjects, say (i, j), $i \neq j$, we may have five types of pairs between the survival time X and the predictive score Y,

(1) a *concordance* with probability $\Pi_c = P(X_i < X_j \text{ and } Y_i < Y_j \quad \text{or} \quad X_i > X_j \text{ and } Y_i > Y_j)$;
(2) a *discordance* with probability $\Pi_d = P(X_i < X_j \text{ and } Y_i > Y_j \quad \text{or} \quad X_i > X_j \text{ and } Y_i < Y_j)$;
(3) an X-only tie with probability $\Pi_{tX} = P(X_i = X_j \text{ and } Y_i > Y_j \quad \text{or} \quad X_i = X_j \text{ and } Y_i < Y_j)$;
(4) a Y-only tie with probability $\Pi_{tY} = P(X_i < X_j \text{ and } Y_i = Y_j \quad \text{or} \quad X_i > X_j \text{ and } Y_i = Y_j)$;
(5) a joint tie in both X and Y with probability $\Pi_{tXY} = P(X_i = X_j \text{ and } Y_i = Y_j)$.

These five possibilities for a random pair are comprehensive and mutually exclusive, and therefore

$$\Pi_c + \Pi_d + \Pi_{tX} + \Pi_{tY} + \Pi_{tXY} = 1.$$

Recall that our interest is to assess the ability of the predictive scores in predicting survival. Smith *et al.* [20] considered Kim's measure [33] $d_{X \cdot Y}$,

$$d_{X \cdot Y} = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d + \Pi_{tY}} = \frac{\Pi_c - \Pi_d}{1 - \Pi_{tX} - \Pi_{tXY}}, \tag{2}$$

which is the probability of a concordance minus the probability of a discordance, both conditioned on the occurrence of distinct values of outcome X, for quantifying the degree of relationship between X and Y. The subscript for Kim's measure $d_{X \cdot Y}$ indicates the intent to predict X from Y. Smith *et al.* [20] showed that a prediction probability P_K (related to Kim's measure),

$$P_K = \frac{1}{2}\left(d_{X \cdot Y} + 1\right) = \frac{\Pi_c + \frac{1}{2}\Pi_{tY}}{\Pi_c + \Pi_d + \Pi_{tY}}, \tag{3}$$

which is the probability of a concordance plus one half the probability of a predictive-score-only (Y-only) tie, both conditioned on distinct values of state or outcome X, is a direct generalization of the trapezoidal AUC to polytomous ordinal patient outcomes. When the outcome X is dichotomous, P_K reduces to the trapezoidal AUC.

2.2. Survival outcomes with Type I censoring assuming continuous predictive score

Pencina and D'Agostino [27] assumed a continuous distribution of the predictive score Y, that is, $P(Y_i = Y_j) = 0$. This eliminates Π_{tY} and Π_{tXY}, and in this case, Kim's measure $d_{X \cdot Y}$ in Equation (2) is simplified to

$$d'_{X \cdot Y} = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d} = \frac{\Pi_c - \Pi_d}{1 - \Pi_{tX}}, \tag{4}$$

which is the modified Kendall's tau presented by Pencina and D'Agostino [27].

They only considered the ties caused by equal survival times X of subjects who remained in the study to the end T_{final} without developing the event of interest. This type of censoring has been defined as Type I censoring [34]. Under this assumption, they showed that Harrell's definition of the C index [22] can be expressed in terms of functions of the probability of a concordance and the probability of a discordance,

$$C_{XY} = P\left(X_i < X_j \text{ and } Y_i < Y_j \quad \text{or} \quad X_i > X_j \text{ and } Y_i > Y_j | X_i \neq X_j\right) = \frac{\Pi_c}{\Pi_c + \Pi_d}.$$

Comparing the preceding equation with Equation (3), we can connect the dots and relate C_{XY} back to AUC via P_K [20].

Pencina and D'Agostino also showed that a linear relationship exists between the overall C and Kim's measure $d'_{X \cdot Y}$ (their modified Kendall's tau) in Equation (4) as

$$C_{XY} = \frac{1}{2}\left(d'_{X \cdot Y} + 1\right).$$

2.3. *Survival outcomes with random right-censoring without assuming continuous predictive score*

The censoring type that Pencina and D'Agostino [27] considered is Type I censoring. In addition, they assumed no ties in predictive scores. To overcome these limitations, we develop a general C index estimator that allows for various right-censoring (e.g., random dropouts) in survival time and ties in predictive scores.

Define *sign* and *csign* (*sign* with *censoring*) functions as follows,

$$sign\left(Y_i, Y_j\right) = I\left(Y_i \geqslant Y_j\right) - I\left(Y_i \leqslant Y_j\right), \tag{5}$$

$$csign\left(X_i, \delta_i, X_j, \delta_j\right) = I\left(X_i \geqslant X_j\right)\delta_j - I\left(X_i \leqslant X_j\right)\delta_i, \tag{6}$$

where $I(\cdot)$ is the indicator function. It can be verified that *sign* and *csign* functions take values in $\{-1, 0, 1\}$. The order of two survival times X_i and X_j can be unambiguously determined if and only if $csign\left(X_i, \delta_i, X_j, \delta_j\right) \neq 0$. Notice that here we may determine the order even when $X_i = X_j$ if only one of the two is censored. The observation that the higher value of Y corresponds to the longer survival time X is mathematically equivalent to $csign\left(X_i, \delta_i, X_j, \delta_j\right) sign\left(Y_i, Y_j\right) = 1$. Under such setting, we formulate the general C index as a conditional probability between the survival time X and the predictive score Y as

$$\begin{aligned} C^g_{XY} = {} & P\left(csign\left(X_i, \delta_i, X_j, \delta_j\right) sign\left(Y_i, Y_j\right) = 1 | csign\left(X_i, \delta_i, X_j, \delta_j\right) \neq 0\right) \\ & + \frac{1}{2} P\left(sign\left(Y_i, Y_j\right) = 0 | csign\left(X_i, \delta_i, X_j, \delta_j\right) \neq 0\right). \end{aligned} \tag{7}$$

If we categorize randomly drawn pairs into the following mutually exclusive types using our *sign* and *csign* functions,

(1) a generalized *concordance* with probability $\Pi^g_c = P(csign\left(X_i, \delta_i, X_j, \delta_j\right) sign\left(Y_i, Y_j\right) = 1)$;
(2) a generalized *discordance* with probability $\Pi^g_d = P(csign\left(X_i, \delta_i, X_j, \delta_j\right) sign\left(Y_i, Y_j\right) = -1)$;
(3) a generalized X-only tie with probability $\Pi^g_{tX} = P(csign\left(X_i, \delta_i, X_j, \delta_j\right) = 0, sign\left(Y_i, Y_j\right) \neq 0)$;
(4) a generalized Y-only tie with probability $\Pi^g_{tY} = P(csign\left(X_i, \delta_i, X_j, \delta_j\right) \neq 0, sign\left(Y_i, Y_j\right) = 0)$;
(5) a generalized joint tie in both X and Y with probability $\Pi^g_{tXY} = P(csign\left(X_i, \delta_i, X_j, \delta_j\right) = 0, sign\left(Y_i, Y_j\right) = 0)$,

we can then express C^g_{XY}, analogous to P_K in Equation (3) , in terms of the probabilities defined previously,

$$C^g_{XY} = \frac{\Pi^g_c + \frac{1}{2}\Pi^g_{tY}}{\Pi^g_c + \Pi^g_d + \Pi^g_{tY}} = \frac{\Pi^g_c + \frac{1}{2}\Pi^g_{tY}}{1 - \Pi^g_{tX} - \Pi^g_{tXY}}.$$

Define our generalized Kendall's tau $\tau^g_{XY} = \dfrac{\Pi^g_c - \Pi^g_d}{\Pi^g_c + \Pi^g_d + \Pi^g_{tY}} = \dfrac{\Pi^g_c - \Pi^g_d}{1 - \Pi^g_{tX} - \Pi^g_{tXY}}$, which is analogous to $d_{X \cdot Y}$ in Equation (2), the linear relationship holds immediately between C^g_{XY} and τ^g_{XY},

$$C^g_{XY} = \frac{1}{2}\left(\tau^g_{XY} + 1\right). \tag{8}$$

Notice that C^g_{XY} and τ^g_{XY} deal with censoring, while P_K and $d_{X \cdot Y}$ do not. At the same time, realizing that

$$\tau^g_{XY} = \frac{\Pi^g_c - \Pi^g_d}{1 - \Pi^g_{tX} - \Pi^g_{tXY}} = \frac{E\left[csign\left(X_i, \delta_i, X_j, \delta_j\right) sign\left(Y_i, Y_j\right)\right]}{E\left[csign\left(X_i, \delta_i, X_j, \delta_j\right)^2\right]},$$

we can construct an estimator for our generalized Kendall's tau, on the basis of which we obtain a point estimator of the C index and make statistical comparison between two C indices.

2.4. A point estimate of the C index based on a random sample

Define $t_{ijXY} = csign\left(X_i, \delta_i, X_j, \delta_j\right) sign\left(Y_i, Y_j\right)$, $t^*_{ijXX} = csign\left(X_i, \delta_i, X_j, \delta_j\right)^2$, the sample estimate

$$t_{XY} = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} t_{ijXY}$$

is a U-statistic-based estimator for $E\left[csign\left(X_i, \delta_i, X_j, \delta_j\right) sign\left(Y_i, Y_j\right)\right]$, that is, the numerator of τ^g_{XY}. Similarly,

$$t^*_{XX} = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} t^*_{ijXX}$$

is a U statistic for the denominator of τ^g_{XY}. Thus, a plug-in estimator for the general C^g_{XY} based on the sample is

$$\hat{C}^g_{XY} = \frac{1}{2}\left(\frac{t_{XY}}{t^*_{XX}} + 1\right).$$

3. The proposed method for comparing two correlated C indices

Because the sample $\hat{C}^g_{XY}$ is a continuous function of U statistics [35], it can be shown that $\hat{C}^g_{XY}$ is asymptotically normal [36]. Consequently, the difference between two correlated sample C indices is also asymptotically normal, provided that the asymptotically bivariate normal distribution of two C indices does not degenerate. Given another predictive score Z, we can obtain

$$\begin{aligned} \mathrm{var}\left(\hat{C}^g_{XY} - \hat{C}^g_{XZ}\right) &= \mathrm{var}\left(\hat{C}^g_{XY}\right) + \mathrm{var}\left(\hat{C}^g_{XZ}\right) - 2\mathrm{cov}\left(\hat{C}^g_{XY}, \hat{C}^g_{XZ}\right) \\ &= \frac{1}{4}\left[\mathrm{var}\left(\frac{t_{XY}}{t^*_{XX}}\right) + \mathrm{var}\left(\frac{t_{XZ}}{t^*_{XX}}\right) - 2\mathrm{cov}\left(\frac{t_{XY}}{t^*_{XX}}, \frac{t_{XZ}}{t^*_{XX}}\right)\right]. \end{aligned} \tag{9}$$

3.1. The Delta method for variance estimators

On the basis of the multivariate Delta method [37], the variance and covariance terms in Equation (9) can be approximated as

$$\mathrm{var}\left(\frac{t_{XY}}{t^*_{XX}}\right) \approx \begin{pmatrix} \frac{1}{t^*_{XX}} & -\frac{t_{XY}}{t^{*2}_{XX}} \end{pmatrix} \begin{bmatrix} \mathrm{var}\left(t_{XY}\right) & \mathrm{cov}\left(t^*_{XX}, t_{XY}\right) \\ \mathrm{cov}\left(t^*_{XX}, t_{XY}\right) & \mathrm{var}\left(t^*_{XX}\right) \end{bmatrix} \begin{pmatrix} \frac{1}{t^*_{XX}} & -\frac{t_{XY}}{t^{*2}_{XX}} \end{pmatrix}^T, \tag{10}$$

$$\operatorname{var}\left(\frac{t_{XZ}}{t^*_{XX}}\right) \approx \begin{pmatrix} \frac{1}{t^*_{XX}} & -\frac{t_{XZ}}{t^{*2}_{XX}} \end{pmatrix} \begin{bmatrix} \operatorname{var}\left(t_{XZ}\right) & \operatorname{cov}\left(t^*_{XX}, t_{XZ}\right) \\ \operatorname{cov}\left(t^*_{XX}, t_{XZ}\right) & \operatorname{var}\left(t^*_{XX}\right) \end{bmatrix} \begin{pmatrix} \frac{1}{t^*_{XX}} & -\frac{t_{XZ}}{t^{*2}_{XX}} \end{pmatrix}^T, \tag{11}$$

$$\operatorname{cov}\left(\frac{t_{XY}}{t^*_{XX}}, \frac{t_{XZ}}{t^*_{XX}}\right) \approx \begin{pmatrix} \frac{1}{t^*_{XX}} & -\frac{t_{XY}}{t^{*2}_{XX}} \end{pmatrix} \begin{bmatrix} \operatorname{cov}\left(t_{XY}, t_{XZ}\right) & \operatorname{cov}\left(t^*_{XX}, t_{XY}\right) \\ \operatorname{cov}\left(t^*_{XX}, t_{XZ}\right) & \operatorname{var}\left(t^*_{XX}\right) \end{bmatrix} \begin{pmatrix} \frac{1}{t^*_{XX}} & -\frac{t_{XZ}}{t^{*2}_{XX}} \end{pmatrix}^T. \tag{12}$$

The details are provided in the Appendix. Now the problem boils down to obtaining estimates for the variance and covariance matrix terms in Equations (10)–(12). In general, in addition to the random variables X (with censoring) and Y (without censoring), given another pair of random variables U (reference variable with censoring) and Z (predictive score without censoring), we present the following lemma providing an unbiased estimator for $\operatorname{cov}\left(t_{XY}, t_{UZ}\right)$, which will be used for obtaining variance and covariance estimates in Equations (10)–(12) and thus the variance estimate of $\hat{C}^g_{XY} - \hat{C}^g_{XZ}$ in Equation (9).

Lemma 3.1
An unbiased estimator for $\operatorname{cov}\left(t_{XY}, t_{UZ}\right)$ is

$$\widehat{\operatorname{cov}}\left(t_{XY}, t_{UZ}\right) = \frac{4\sum_i \left(\sum_j t_{ijXY} \sum_{j'} t_{ij'UZ}\right) - 2\sum_i \sum_j t_{ijXY} t_{ijUZ} - \frac{2(2n-3)}{n(n-1)} \sum_i \sum_j t_{ijXY} \sum_{i'} \sum_{j'} t_{i'j'UZ}}{n(n-1)(n-2)(n-3)}.$$

The proof is given in the Appendix.

From Lemma 3.1, we may obtain unbiased estimators for terms $\operatorname{var}\left(t^*_{XX}\right)$, $\operatorname{var}(t_{XY})$, $\operatorname{var}(t_{XZ})$, $\operatorname{cov}\left(t^*_{XX}, t_{XY}\right)$, $\operatorname{cov}\left(t^*_{XX}, t_{XZ}\right)$, and $\operatorname{cov}(t_{XY}, t_{XZ})$ in Equations (10–12) by replacing subscripts. For example,

$$\widehat{\operatorname{var}}\left(t^*_{XX}\right) = \frac{4\sum_i \left(\sum_j t^*_{ijXX}\right)^2 - 2\sum_i \sum_j t^{*2}_{ijXX} - \frac{2(2n-3)}{n(n-1)} \left(\sum_i \sum_j t^*_{ijXX}\right)^2}{n(n-1)(n-2)(n-3)}.$$

See the Appendix for more details. Putting all the variance and covariance estimates together, $\widehat{\operatorname{var}}\left(\hat{C}^g_{XY} - \hat{C}^g_{XZ}\right)$ is acquired.

In view of the preceding results, we form a z score on the basis of the large sample approach [38] for testing the null hypothesis $H_0 : C^g_{XY} = C^g_{XZ}$. The test statistic is

$$z = \frac{\hat{C}^g_{XY} - \hat{C}^g_{XZ}}{\sqrt{\widehat{\operatorname{var}}\left(\hat{C}^g_{XY} - \hat{C}^g_{XZ}\right)}},$$

and we would reject the null hypothesis if we observe $|z| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the upper $\alpha/2$ quantile from a standard normal distribution.

Antolini *et al.* [29] considered the jackknife approach to estimate $\operatorname{var}\left(\hat{C}^g_{XY} - \hat{C}^g_{XZ}\right)$. The procedure lies in leaving out one observation, say, individual i, at a time from the sample set and recomputing the difference between two C indices on the basis of remaining $n-1$ observations, that is, $\hat{C}^{g\,(-i)}_{XY} - \hat{C}^{g\,(-i)}_{XZ}$. An estimate for the variance of the two C's difference can be calculated from this new set of replicates $\left\{\hat{C}^{g\,(-i)}_{XY} - \hat{C}^{g\,(-i)}_{XZ}\right\}$, $i = 1, 2, \ldots, n$. Antolini *et al.* [29] showed that the jackknife estimate of variance converges to the true value asymptotically.

Alternatively, we can utilize the bootstrap resampling approach [30, 31] for variance estimations when testing the difference between two correlated C indices. The procedure is as follows. We resample the original data matrix shown in Equation (1) by row (individuals) with replacement and compute the estimated $\hat{C}^g_{XY} - \hat{C}^g_{XZ}$ on the basis of the resampled data. Repeat this step $B = 500$ times and each time we obtain an estimate of the difference between $\hat{C}^g_{XY}$ and $\hat{C}^g_{XZ}$. The sample variance of the bootstrap replicates of $\hat{C}^g_{XY} - \hat{C}^g_{XZ}$ is considered a bootstrap variance estimate for $\operatorname{var}\left(\hat{C}^g_{XY} - \hat{C}^g_{XZ}\right)$. A z score can be similarly formed for testing the same hypothesis.

4. Simulation studies

We conducted extensive simulations in validating the performance of our proposed method in terms of type I error rate and power. We also compared our method with the jackknife and the bootstrap methods.

The survival time data is generated from either an exponential or a Weibull distribution. The exponential distribution assumes a fixed failure rate and has been widely used in continuous survival time modeling [34, 39]. The Weibull distribution [34], which is a generalization to the exponential distribution by allowing the failure rate to vary over the time, was also selected because it was found to fit the Kaplan–Meier survival curve in the Framingham Heart Study example as shown in Figure 1. We used both distributions in our simulations. Of course, it should be expected that our proposed method is nonparametric and thus should work for any distribution.

We consider non-informative random censoring for each subject. The observed right-censored survival time X is recorded as the minimum of censoring time and survival time. The event indicator δ is 1 if the right-censored survival time is indeed the true survival time or 0 if the right-censored survival time is censoring time. In our simulation, the censoring time is generated using exponential distributions with various failure rates such that the censoring percentage is controlled at certain levels, ranging from 0% to 50%. Determining the failure rate parameters given the censoring percentage is achieved by iterative approximations based on sufficiently large sample sizes.

Conditioned on true survival times, two predictive scores Y and Z for each subject are simulated from bivariate normal distributions with different means and varying correlations. For simplicity, the standard deviations of two scores are fixed at 1. The means are set such that the true $\Delta c = C_{XY} - C_{XZ}$ is 0.00, 0.05, or 0.10, respectively. The individual values of C_{XY} and C_{XZ} are given in Tables I–III and Supporting information Tables S1–S3. Again, determining mean parameters given the target C index value is achieved by iterative approximations based on sufficiently large sample sizes. It is important to keep in mind that the individual C index does depend on the censoring distribution [40], and the individual C indices reported in the tables are calculated approximately on the basis of sufficiently large sample sizes, conditional on true survival times (without censoring). However, as we will show later in this section, the $\Delta c = C_{XY} - C_{XZ}$ does not change with varying correlations and censoring percentages.

The first scenario with $\Delta c = 0.00$ corresponds to the null hypothesis, whereas the latter two correspond to the alternative. The null hypothesis that there is no difference between two correlated C is rejected if $|z| > z_{1-\alpha/2}$, where $\alpha = 0.05$ for all simulations. For each configuration with different sample size n, censoring percentage, bivariate normal correlation, and true Δc, we sample data from the designated distributions and apply our method, the jackknife method, and the bootstrap to perform hypothesis testing.

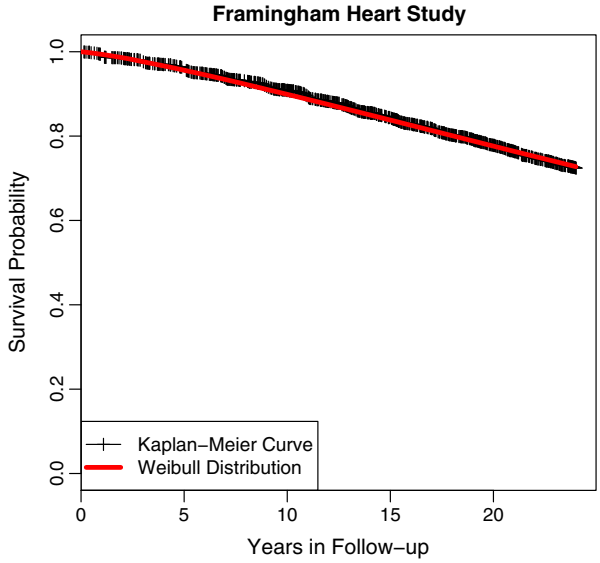

Figure 1. Empirical survival function based on the Framingham Heart Study. A vertical tick represents loss of a patient (censoring). The Weibull distribution with $\lambda = 60$, $k = 1.25$ was found to fit these survival data quite well.

Table I. Bias and RMSE given as a percent of the true variance for all three variance estimators of Δc. This case study used an exponential model for survival data with 20% data censored and the correlation between two scores set to 0.50.

			Delta + U[b]		Bootstrap		Jackknife	
Effect size	Sample size n	True variance[a] ($\times 10^4$)	Relative bias[c] (%)	Relative RMSE[d] (%)	Relative bias[c] (%)	Relative RMSE[d] (%)	Relative bias[c] (%)	Relative RMSE[d] (%)
$\Delta c = 0.00$	50	27.9	−0.29	33	14	36	11	36
$C_{XY} = 0.6$	65	21.2	−0.38	28	10	31	8	30
$C_{XZ} = 0.6$	80	16.8	−0.19	25	9	28	7	27
	95	14.1	−0.26	23	7	25	5	24
$\Delta c = 0.05$	50	29.8	−0.26	32	13	35	11	35
$C_{XY} = 0.6$	65	22.8	−0.34	27	9	30	7	29
$C_{XZ} = 0.55$	80	18.0	−0.13	25	8	27	6	26
	95	15.1	−0.24	22	7	24	5	24
$\Delta c = 0.10$	50	32.2	−0.35	31	12	33	10	33
$C_{XY} = 0.6$	65	24.6	−0.45	26	8	28	7	28
$C_{XZ} = 0.5$	80	19.5	−0.26	24	7	26	6	25
	95	16.4	−0.39	21	6	23	5	22

[a]True variance is estimated from 100,000 MC samples.
[b]Delta + U refers to the proposed variance estimator, which is a U statistic applied to a multivariate Delta method.
[c]Relative Bias (%) = Bias/TrueVariance ×100%.
[d]Relative RMSE (%) = RMSE/TrueVariance ×100%.

Table II. Bias and RMSE given as a percent of the true variance for all three variance estimators of Δc. This case study used an exponential model for survival data with 20% data censored and the correlation between two scores set to 0.95.

			Delta + U[b]		Bootstrap		Jackknife	
Effect size	Sample size n	True variance[a] ($\times 10^4$)	Relative bias[c] (%)	Relative RMSE[d] (%)	Relative bias[c] (%)	Relative RMSE[d] (%)	Relative bias[c] (%)	Relative RMSE[d] (%)
$\Delta c = 0.00$	50	3.6	−0.34	44	44	67	27	55
$C_{XY} = 0.6$	65	2.6	−0.41	38	35	55	21	46
$C_{XZ} = 0.6$	80	2.1	−0.44	34	31	48	18	40
	95	1.7	−0.18	31	27	43	16	36
$\Delta c = 0.05$	50	4.8	−0.31	44	38	61	24	53
$C_{XY} = 0.6$	65	3.5	−0.36	38	30	50	18	44
$C_{XZ} = 0.55$	80	2.7	−0.37	34	26	45	15	39
	95	2.3	−0.16	31	22	40	13	35
$\Delta c = 0.10$	50	7.7	−0.52	44	27	53	18	49
$C_{XY} = 0.6$	65	5.7	−0.44	38	21	45	13	42
$C_{XZ} = 0.5$	80	4.5	−0.51	34	17	40	11	37
	95	3.7	−0.35	32	15	36	9	34

[a]True variance is estimated from 100,000 MC samples.
[b]Delta + U refers to the proposed variance estimator, which is a U statistic applied to a multivariate Delta method.
[c]Relative bias (%) = Bias/TrueVariance ×100%.
[d]Relative RMSE (%) = RMSE/TrueVariance ×100%.

We repeat the Monte Carlo (MC) simulation $nsim = 100,000$ times to calculate the proportion of rejection being observed in these simulation experiments. This fraction is the empirical type I error rate under the null hypothesis and the empirical statistical power under the alternative.

Figures 2 and 3 show the proportion of rejection under various settings with exponential survival data and the correlation between two predictive scores set to 0.50 and 0.95, respectively. From Figures 2 and 3, we see that the observed error rate of our method is nominally higher than the expected rate at small

Table III. Bias ($\times 10^4$) for the difference estimator of two C indices. This case study used an exponential model for survival data.

		Correlation = 0.50 Censoring Percentage				Correlation = 0.95 Censoring Percentage			
Effect size	Sample size n	0%	10%	20%	50%	0%	10%	20%	50%
$\Delta c = 0.00$	50	−0.8653	−0.8514	−0.8428	−0.8757	−0.8505	−0.8622	−0.8430	−0.8682
$C_{XY} = 0.6$	65	−0.7375	−0.7564	−0.7354	−0.7397	−0.7504	−0.7619	−0.7567	−0.7412
$C_{XZ} = 0.6$	80	−0.6400	−0.6457	−0.6575	−0.6552	−0.6486	−0.6537	−0.6584	−0.6395
	95	−0.5833	−0.5658	−0.5743	−0.5557	−0.5673	−0.5740	−0.5802	−0.5691
$\Delta c = 0.05$	50	−0.7888	−0.7996	−0.7951	−0.8097	−0.7918	−0.8009	−0.7890	−0.7959
$C_{XY} = 0.6$	65	−0.6594	−0.6469	−0.6412	−0.6508	−0.6446	−0.6477	−0.6396	−0.6419
$C_{XZ} = 0.55$	80	−0.5485	−0.5507	−0.5452	−0.5443	−0.5373	−0.5419	−0.5334	−0.5392
	95	−0.4288	−0.4319	−0.4379	−0.4187	−0.4229	−0.4307	−0.4199	−0.4263
$\Delta c = 0.10$	50	−0.7185	−0.7296	−0.7177	−0.7163	−0.7179	−0.7223	−0.7162	−0.7135
$C_{XY} = 0.6$	65	−0.5719	−0.5888	−0.5627	−0.5767	−0.5680	−0.5757	−0.5673	−0.5692
$C_{XZ} = 0.5$	80	−0.4404	−0.4427	−0.4536	−0.4397	−0.4468	−0.4529	−0.4441	−0.4437
	95	−0.3603	−0.3693	−0.3734	−0.3561	−0.3683	−0.3718	−0.3668	−0.3655

Above observed with exponential survival data based on 100,000 samples; values in the table equal original bias $\times 10^4$.

sample sizes, for example, $n = 50$ (detailed numbers can be found in the Supporting information Table S1). The inflation of the type I error rate is more obvious with more data censored, which can be explained by the fact that censoring reduces the effective amount of data for calculating the statistics. Nevertheless, the results indicate that the observed error rate of our method converges steadily to the expected rate as n increases. Furthermore, with increased correlations between two predictive scores Y and Z, the type I error rate is more close to the nominal level.

In contrast, the observed error rates of the jackknife and the bootstrap methods are further from the nominal level in the conservative direction. With 20% survival data censored and a correlation of 0.95 between two predictive variables, the observed error rates of the jackknife and the bootstrap could go as low as 0.0291 and 0.0184 at $n = 50$, respectively. Even with sample size $n = 95$, the observed error rates of the jackknife and the bootstrap are 0.0346 and 0.0271, respectively. As n increases, the error rates of the jackknife and the bootstrap methods would converge to the expected rate. Yet, it is evident that with increased correlations between two predictive scores Y and Z, the type I error rate is further away from the nominal level.

Under the alternative hypothesis, the proposed method has larger power than either the jackknife or the bootstrap method, which is consistent with the conservativeness we have just observed. For instance, when the effect size $\Delta c = 0.05$, with 20% exponential survival data censored and the correlation of 0.95 between two predictive variables, the observed power of our method is 0.7969 at $n = 65$, compared with 0.7388 and 0.6979, for the jackknife and the bootstrap method, respectively. Table S2 in the Supporting information showing the proportion of rejection under various settings with Weibull survival data tells similar trends in the type I error rate and power.

To further understand the differences of observed type I error rate and power among different methods, we empirically assess the bias and the root mean square error (RMSE) for all three variance estimators. The true variance for the difference between two correlated C indices is estimated by the sample variance of 100,000 MC sample estimates of the difference. For every MC sample, each of our proposed method, the jackknife, and the bootstrap method, would give an estimate of the variance for the difference. The mean and the variance, as well as the bias and the RMSE, for each variance estimator can be obtained across 100,000 MC samples.

Tables I and II show the bias and the RMSE given as a percent of the true variance associated with each variance estimator with the correlation between two predictive scores set to 0.50 and 0.95, respectively. It demonstrates that our proposed variance estimator is almost unbiased across a large range of correlations even with a small to moderate sample size (downward bias < 1% for all the simulations). By contrast, the variance estimator based on the bootstrap or the jackknife is substantially biased upwards, and the relative

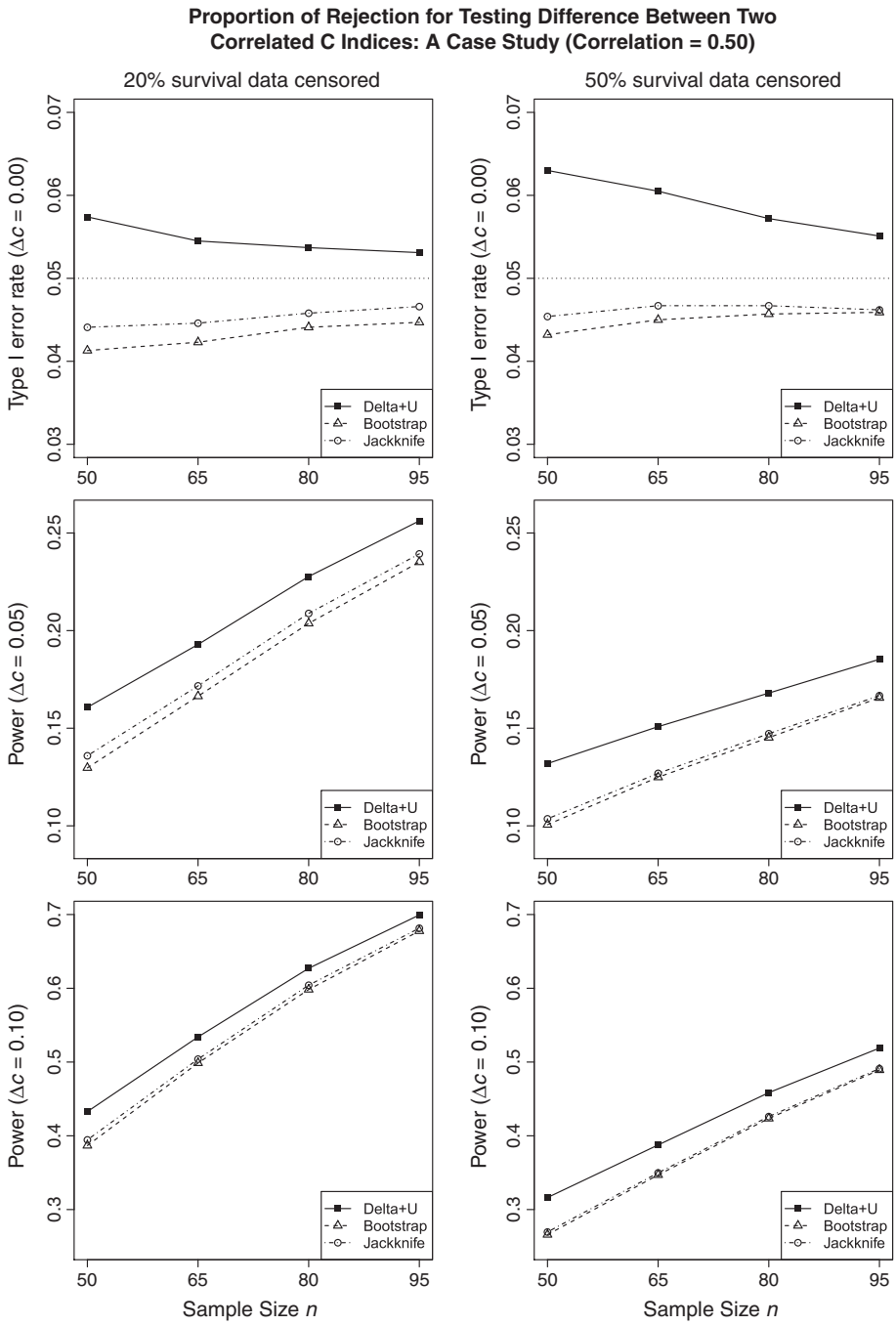

Figure 2. Proportion of rejection for testing the difference between two correlated *C* indices. This case study used an exponential model for survival data and the correlation between two scores set to 0.50. Detailed numbers can be found in the Supporting information Table S1.

bias and RMSE increase, while the correlation is higher. With sufficiently large sample sizes, all three methods converge in theory. These results explain conservative type I error rate and underpower with the jackknife and the bootstrap methods. To gain more insights on the variance estimators, we present the sampling distributions of all the variance estimates in the 100,000 MC experiments for the three variance estimators corresponding to Tables I and II in Supporting information Figures S1–S6. Again, under both the null and the alternative, our proposed variance estimator is almost unbiased, while the bootstrap and the jackknife are substantially biased upwards.

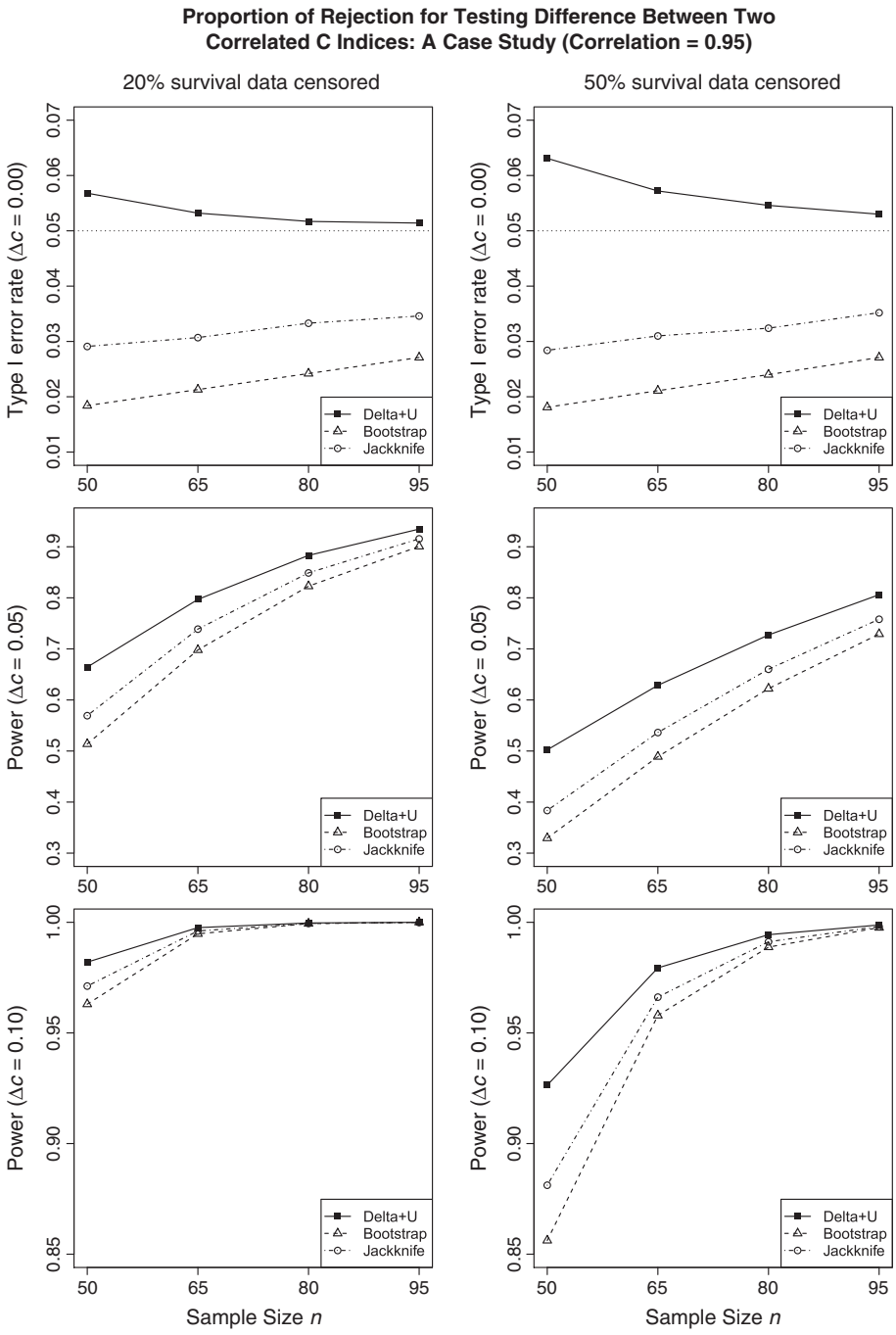

Figure 3. Proportion of rejection for testing the difference between two correlated *C* indices. This case study used an exponential model for survival data and the correlation between two scores set to 0.95. Detailed numbers can be found in the Supporting information Table S1.

We mention in Section 2.3 that our formulation of the general *C* index can accommodate ties in predictive scores. To evaluate the robustness of our method to ties in predictive scores, we carried out simulations similar to those preceding except using discrete predictive scores that have ties. We use numerical rounding as a way to generate data with ties. Specifically, the simulated bivariate normal continuous scores are rounded to one decimal place as well as to the nearest integers, which results in approximately 40% and 90% of the subjects having tied score value with another subject, respectively. Then, similar to the MC simulations with the continuous data presented previously, we

apply our proposed method, the jackknife, and the bootstrap method to perform hypothesis testing. The results with the discrete predictive scores are generally the same as those with the continuous scores, indicating robustness of our method to data with ties. See Table S3 in the Supporting information for details.

Last but not least, as we point out in the beginning of this section, it is known that the estimate of a single C index is affected by the percentage of censored observations [40], but to what extent the difference between two C indices is affected by censoring remains a concern. To this end, we evaluate the effect of censoring and the correlation between two predictive scores on the estimation of the difference between two correlated C indices via a simulation study. The result summarizing the empirical bias estimate is in Table III. From Table III, although there are fluctuations in empirical bias estimates, there is no evidence that the differences, all less than 0.0001, between two C indices are affected by censoring. By contrast, the bias associated with a single C index estimate may vary from 0.012 to 0.084, when the censoring percentage is increased from 10% to 50% within one of our simulation studies. The results indicate that when comparing two correlated C indices conditional on the same censored survival outcome, the biases for each C index caused by censoring are canceled out. It is also observed that the correlation does not have an effect on the estimation of the difference between two correlated C indices. This is within our expectation, as the correlation usually does not affect the marginal difference but does impact the variance/covariance terms.

5. Analysis of Framingham Heart Study data example

The Framingham Heart Study is a long-term prospective study of the etiology of cardiovascular disease in a population of free living subjects in the community of Framingham, Massachusetts [32]. It was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified risk factors and their joint effects. Much of the now-common knowledge concerning heart disease, such as the effects of diet, exercise, and medications, is based on this study [41–43].

The study began in 1948, and 5209 subjects were initially enrolled in the study. Participants have been examined since the inception of the study, and all subjects are continuously followed through regular surveillance for cardiovascular outcomes. Clinic examination data have included cardiovascular disease risk factors and markers of disease such as blood pressure, blood chemistry, lung function, smoking history, health behaviors, ECG tracings, echocardiography, and medication use. Through regular surveillance of area hospitals, participant contact, and death certificates, the Framingham Heart Study reviews and adjudicates events for the occurrence of angina pectoris, myocardial infarction, heart failure, cerebrovascular disease, and death [32]. Part of the dataset in this study is publicly available [44], and we use this subset of the data (4434 participants) to demonstrate the application of our proposed method.

In this application, the event time of interest is the first time that a subject had a cardiovascular disease event (angina pectoris, myocardial infarction, coronary insufficiency, or fatal coronary heart disease) during the interval of baseline to end of follow-up. The Kaplan–Meier estimates for the survival distributions of the event time and the censoring are given in Figure 1.

We consider baseline measurements (PERIOD = 1) serum total cholesterol (TOTCHOL, mg/dL), body mass index (BMI, weight in kilograms/height meters squared), systolic blood pressure (SYSBP, mm Hg) and diastolic blood pressure (DIABP, mm Hg) as predictive variables for the event-free survival time. These are common risk factors for cardiovascular diseases [45–47]. We conducted testing to determine if a certain biomarker has better prognosis than others in terms of the C index.

The estimated sample values of C are 0.4018, 0.4021, 0.3651, and 0.3938 for TOTCHOL, BMI, SYSBP and DIABP, respectively. Note that the C values are less than 0.5, which suggests that lower values, not higher values, of the biomarkers correlate with longer event-free survival time. The difference between two values of C from SYSBP and DIABP is 0.0287 with a p-value < 0.0001 using our proposed method. Although the difference of 0.0003 between two C indices from TOTCHOL and BMI is quite small, it reports a p-value of 0.0388. That is to say, SYSBP is better than DIABP, and TOTCHOL might be slightly better than BMI in predicting event-free survival time. Our finding is consistent with the result of Stamler *et al.* that systolic blood pressure relates more strongly to all cardiovascular risk than diastolic blood pressure [48].

6. Discussion

In this article, we investigated the problem of assessing the performance of biomarkers or classifiers in predicting right-censored survival outcomes in terms of the C index. We developed a nonparametric approach for comparing two correlated C indices for two tests (biomarkers or classifiers) performed on a common cohort of subjects. Specifically, we derived analytical one-shot estimators for the variance of the C index estimator and the covariance between two C indices. The one-shot estimators require no resampling procedure and thus are computationally efficient. These provided the necessary recipes for using the z score test to statistically compare two tests in regard to their C indices. Our extensive simulations showed that our proposed approach had satisfactory performance in terms of type I error control and statistical power with moderate to large sample sizes. We also showed that our approach compares favorably with resampling methods such as the jackknife method [29] and the bootstrap method as it provides an almost unbiased variance estimator for the difference estimate between two correlated C indices. Finally, we applied our method to the Framingham Heart Study data for an application in the problem of biomarker comparison for survival prognosis.

Given the negligible bias of our proposed variance estimator across a large range of correlations and sample sizes shown in the simulation studies, it is clear that the proposed variance estimator is not impacted by the correlations or the sample sizes. On the contrary, the bias and the RMSE associated with the bootstrap or the jackknife variance estimator increase significantly when the correlations are higher, indicating that the accuracy of the bootstrap or the jackknife variance estimator heavily depends on the correlations. Meanwhile, our simulation results indicate that the type I error rate of our proposed test is not impacted much by the correlations either, as compared with the increasing conservativeness for the test based on the bootstrap or the jackknife variance estimator when the correlations are increasingly higher. In short, the relative gain of our proposed method as compared with the resampling methods when the correlations are higher is not because our method depends on the correlations but because of the dependence of the resampling methods on the correlations; that is, they perform a lot worse when the correlations increase. By theory, our proposed method appropriately accounts for the correlations in estimating the variance of the difference between two correlated C indices, and thus we do not expect the performance to depend on varying correlations.

Linking up with the fact that our proposed variance estimator works quite well across a large range of correlations and sample sizes, the inflated type I error rate at small sample sizes observed in our results indicates that the only limitation of our proposed method is the use of the standard normal distribution for the test statistic at very small sample sizes. The test statistic for the comparison of two general C indices is not quite normal at small sample sizes, because we are estimating the variance; we do not know the population variance parameter exactly. Consequently, the distribution of the test statistic is wider. It may be possible to account for this by using the t distribution with an appropriate estimate of the degree of freedom. This will be addressed in our future work.

In the meantime, the seemingly fine controlled type I error rate for the z score test based on the bootstrap or the jackknife variance estimator when the correlations are moderate is in fact caused by the compromise of two types of errors, (1) an overestimate of the true variance of $\Delta c = C_{XY} - C_{XZ}$ and (2) an underestimate of the critical value based on standard normal distribution. These two types of errors sometimes happen to cancel out so the type I error rate looks just fine. However, when the correlations are much higher, the error caused by the overestimation of the true variance of Δc would outweigh the error caused by using an underestimated critical value on the basis of the standard normal distribution, and thus we observe extreme conservativeness. We suggest that caution be taken before using the z score test based on the bootstrap or the jackknife variance estimator as people may not realize that errors are being mixed in producing p-values.

Besides the C index we have considered in this article, various accuracy measures have been suggested to assess the ability of a predictive score Y in predicting the censored survival time X. Some examples of these include proportion of explained variation [49–51], integrated Brier score [52], time-dependent ROC measure [16–19], to name a few. Specifically, by choosing appropriate weight functions $w(t)$, it can be shown that a weighted average of the area under an incident time-dependent ROC curve at time t, $\mathrm{AUC}_t = \int \mathrm{ROC}_t(u)du$, where $\mathrm{ROC}_t(u) = \mathrm{TPR}_t\{\mathrm{FPR}_t^{-1}(u)\}, \mathrm{TPR}_t(y) = P(Y \geqslant y | X = t), \mathrm{FPR}_t(y) = P(Y \geqslant y | X > t)$, is equivalent to the C index [17].

Our method for comparing two correlated C indices is analogous to the role of the method of DeLong *et al.* [11] for comparing two correlated AUCs with binary disease outcomes. We reiterate that predictive scores in this article can be measured values of biomarkers, composite patient scores output

from algorithms combining multiple biomarkers, predicted survival times, or predicted probabilities of survival until any fixed time point based on mathematical models. We account for ties in the predictive scores. This is more robust than existing methods that assume continuous scores because ties in predictive scores can occur in many practical applications. For instance, ties occur naturally in categorical predictive variables, and they may even occur in theoretically continuous variables as a result of categorization or discretization.

We note that our method cannot deal with the problem of limit of detection (LoD) that may occur in biomarker measurements. Vexler *et al.* [53] considered comparing the correlated AUCs of diagnostic biomarkers whose measurements are subject to a limit of detection. Their approach may be extended to the C index metric for dealing with both censored outcome and LoD in predictive scores. It is also possible to extend our method by properly modifying the *csign* function in Equations (5)–(6) to deal with LoD. It remains interesting future work for such extensions and comparisons.

In recent years, many investigators, for example, [54, 55], have noticed empirically that the method of DeLong *et al.* based on U statistics [11] for testing two correlated AUCs resulting from nested models (reduced versus full model) often produces a nonsignificant result in assessing the incremental value when a corresponding Wald test or likelihood ratio test from the underlying regression model is significant. We have investigated this interesting problem in a recent paper [56]. The reported "problems" of the method of DeLong *et al.* is essentially because the method is misused (i.e., used in a fashion that the method is not designed for) by training and testing the models using the same dataset. The test by DeLong *et al.* is designed for comparing two *fixed* models that are tested on a common dataset that is independent of the training set. The variance estimator of DeLong *et al.* does not incorporate the variance caused by the model training process, and therefore, fitting the full model and the reduced model based on the training dataset and then testing two correlated AUCs based on the *same* dataset would violate the assumption of the method of DeLong *et al.* and thus lead to an incorrect variance estimate. Similarly, for our proposed method, when the predictive scores are outputs from an algorithm that combines multiple biomarkers, our method assumes that the algorithm is trained and fixed and ready to be tested on an independent test dataset. In other words, our method cannot be used to compare two resubstitution C indices that are obtained by training and testing the algorithm using the same dataset.

A supplementary R package to implement the methods described in this article is available for download and ready to use at *http://code.google.com/p/assessment-of-classifiers/*. An example demo on how to use the software can be found in the package.

Appendix A: The multivariate Delta method for variance estimators

Following the formula of Casella and Berger [37], we have

$$\begin{aligned}
\operatorname{var}\left(\frac{t_{XY}}{t^*_{XX}}\right) &\approx \left[\frac{\partial}{t_{XY}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]^2 \operatorname{var}\left(t_{XY}\right) + \left[\frac{\partial}{t^*_{XX}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]^2 \operatorname{var}\left(t^*_{XX}\right) \\
&\quad + 2\left[\frac{\partial}{t_{XY}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t^*_{XX}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right] \operatorname{cov}\left(t_{XY}, t^*_{XX}\right) \\
&= \left(\frac{1}{t^*_{XX}} \quad -\frac{t_{XY}}{t^{*2}_{XX}}\right) \begin{bmatrix} \operatorname{var}\left(t_{XY}\right) & \operatorname{cov}\left(t^*_{XX}, t_{XY}\right) \\ \operatorname{cov}\left(t^*_{XX}, t_{XY}\right) & \operatorname{var}\left(t^*_{XX}\right) \end{bmatrix} \left(\frac{1}{t^*_{XX}} \quad -\frac{t_{XY}}{t^{*2}_{XX}}\right)^T,
\end{aligned}$$

$$\begin{aligned}
\operatorname{var}\left(\frac{t_{XZ}}{t^*_{XX}}\right) &\approx \left[\frac{\partial}{t_{XZ}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]^2 \operatorname{var}\left(t_{XZ}\right) + \left[\frac{\partial}{t^*_{XX}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]^2 \operatorname{var}\left(t^*_{XX}\right) \\
&\quad + 2\left[\frac{\partial}{t_{XZ}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t^*_{XX}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right] \operatorname{cov}\left(t_{XZ}, t^*_{XX}\right) \\
&= \left(\frac{1}{t^*_{XX}} \quad -\frac{t_{XZ}}{t^{*2}_{XX}}\right) \begin{bmatrix} \operatorname{var}\left(t_{XZ}\right) & \operatorname{cov}\left(t^*_{XX}, t_{XZ}\right) \\ \operatorname{cov}\left(t^*_{XX}, t_{XZ}\right) & \operatorname{var}\left(t^*_{XX}\right) \end{bmatrix} \left(\frac{1}{t^*_{XX}} \quad -\frac{t_{XZ}}{t^{*2}_{XX}}\right)^T,
\end{aligned}$$

$$
\begin{aligned}
\operatorname{cov}\left(\frac{t_{XY}}{t^*_{XX}}, \frac{t_{XZ}}{t^*_{XX}}\right) \approx & \underbrace{\left[\frac{\partial}{t_{XY}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t_{XY}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]\operatorname{var}\left(t_{XY}\right)}_{=0} + \underbrace{\left[\frac{\partial}{t_{XZ}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t_{XZ}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]\operatorname{var}\left(t_{XZ}\right)}_{=0} \\
& + \left[\frac{\partial}{t^*_{XX}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t^*_{XX}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]\operatorname{var}\left(t^*_{XX}\right) \\
& + \left[\frac{\partial}{t_{XY}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t_{XZ}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]\operatorname{cov}\left(t_{XY}, t_{XZ}\right) + \underbrace{\left[\frac{\partial}{t_{XZ}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t_{XY}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]\operatorname{cov}\left(t_{XY}, t_{XZ}\right)}_{=0} \\
& + \left[\frac{\partial}{t_{XY}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t^*_{XX}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]\operatorname{cov}\left(t_{XY}, t^*_{XX}\right) + \underbrace{\left[\frac{\partial}{t^*_{XX}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t_{XY}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]\operatorname{cov}\left(t_{XY}, t^*_{XX}\right)}_{=0} \\
& + \underbrace{\left[\frac{\partial}{t_{XZ}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t^*_{XX}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]\operatorname{cov}\left(t_{XZ}, t^*_{XX}\right)}_{=0} + \left[\frac{\partial}{t^*_{XX}}\left(\frac{t_{XY}}{t^*_{XX}}\right)\right]\left[\frac{\partial}{t_{XZ}}\left(\frac{t_{XZ}}{t^*_{XX}}\right)\right]\operatorname{cov}\left(t_{XZ}, t^*_{XX}\right) \\
= & \begin{pmatrix} \frac{1}{t^*_{XX}} & -\frac{t_{XY}}{t^{*2}_{XX}} \end{pmatrix} \begin{bmatrix} \operatorname{cov}\left(t_{XY}, t_{XZ}\right) & \operatorname{cov}\left(t^*_{XX}, t_{XY}\right) \\ \operatorname{cov}\left(t^*_{XX}, t_{XZ}\right) & \operatorname{var}\left(t^*_{XX}\right) \end{bmatrix} \begin{pmatrix} \frac{1}{t^*_{XX}} & -\frac{t_{XZ}}{t^{*2}_{XX}} \end{pmatrix}^T.
\end{aligned}
$$

Appendix B: Covariance estimation of Kendall's τ

Given random variables X (with censoring) and Y (without censoring), Kendall's τ_{XY} quantifies the concordance between ordinal relations on two variables,

$$\tau_{XY} = E\left[csign\left(X_i, \delta_i, X_j, \delta_j\right) sign\left(Y_i, Y_j\right)\right]$$

for a pair of subjects $(i, j), i \neq j$. Define $t_{ijXY} = csign\left(X_i, \delta_i, X_j, \delta_j\right) sign\left(Y_i, Y_j\right)$, the sample estimate for τ_{XY} is

$$t_{XY} = \frac{1}{n(n-1)} \sum_i \sum_{j \neq i} t_{ijXY},$$

and it is easy to verify that t_{XY} is an unbiased estimator for τ_{XY}.

In general, given another pair of random variables U (with censoring) and Z (without censoring),

$$\operatorname{cov}\left(t_{XY}, t_{UZ}\right) = E\left[t_{XY} t_{UZ}\right] - \tau_{XY}\tau_{UZ}.$$

Note that $t_{XY}t_{UZ} = \frac{1}{n^2(n-1)^2} \sum_i \sum_j \sum_{i'} \sum_{j'} t_{ijXY} t_{i'j'UZ}$, where i, j, i', and j' represent sampled individuals,

$$
\begin{aligned}
E\left(t_{XY}t_{UZ}\right) &= \frac{1}{n^2(n-1)^2} \sum_i \sum_{j \neq i} \sum_{i'} \sum_{j' \neq i'} E\left[t_{ijXY} t_{i'j'UZ}\right] \\
&= \frac{1}{n^2(n-1)^2} \sum_i \sum_{j \neq i} \left\{ E\left[t_{ijXY} t_{ijUZ}\right] + \sum_{\substack{j' \neq j \\ j' \neq i}} E\left[t_{ijXY} t_{ij'UZ}\right] + \sum_{\substack{i' \neq i \\ i' \neq j}} E\left[t_{ijXY} t_{i'jUZ}\right] + \sum_{i' \neq i} \sum_{\substack{j' \neq j \\ j' \neq i'}} E\left[t_{ijXY} t_{i'j'UZ}\right] \right\} \\
&= \frac{1}{n(n-1)} E\left[t_{ijXY} t_{ijUZ}\right] + \frac{n-2}{n(n-1)} E\left[t_{ijXY} t_{ij'UZ}\right] + \frac{n-2}{n(n-1)} E\left[t_{ijXY} t_{i'jUZ}\right] + \frac{(n-1)(n-2)+1}{n(n-1)} E\left[t_{ijXY} t_{i'j'UZ}\right].
\end{aligned}
$$

For each term, the coefficient is simply obtained by counting. Realizing that $t_{ijXY} = t_{jiXY}$ and $t_{ijXY}t_{i'jUZ} = t_{jiXY}t_{ji'UZ} = t_{ijXY}t_{ij'UZ}$,

$$E\left(t_{XY}t_{UZ}\right) = \frac{2}{n(n-1)}E\left[t_{ijXY}t_{ijUZ}\right] + \frac{4(n-2)}{n(n-1)}E\left[t_{ijXY}t_{ij'UZ}\right] + \frac{(n-2)(n-3)}{n(n-1)}E\left[t_{ijXY}t_{i'j'UZ}\right],$$

with each term corresponding to two subscripts, one subscript, and no subscript in common, respectively, regardless of order. Therefore,

$$\begin{aligned}\operatorname{cov}\left(t_{XY}, t_{UZ}\right) &= \frac{2}{n(n-1)}\left[\operatorname{cov}\left(t_{ijXY}, t_{ijUZ}\right) + \tau_{XY}\tau_{UZ}\right] \\ &\quad + \frac{4(n-2)}{n(n-1)}\left[\operatorname{cov}\left(t_{ijXY}, t_{ij'UZ}\right) + \tau_{XY}\tau_{UZ}\right] \\ &\quad + \frac{(n-2)(n-3)}{n(n-1)}\tau_{XY}\tau_{UZ} - \tau_{XY}\tau_{UZ} \\ &= \frac{2}{n(n-1)}\operatorname{cov}\left(t_{ijXY}, t_{ijUZ}\right) + \frac{4(n-2)}{n(n-1)}\operatorname{cov}\left(t_{ijXY}, t_{ij'UZ}\right).\end{aligned}$$

It is of interest to obtain an unbiased estimate for $\operatorname{cov}\left(t_{XY}, t_{UZ}\right)$. We could directly apply U statistics to the preceding two terms, but the forms would be awkward to program (sums would not run over all samples). Here, we present similar results with Cliff and Charlin [57] but correct their estimators and then show the unbiasedness. The estimators presented in the following are easier to implement.

Lemma B.1
An unbiased estimator for $\operatorname{cov}\left(t_{XY}, t_{UZ}\right)$ is

$$\widehat{\operatorname{cov}}\left(t_{XY}, t_{UZ}\right) = \frac{4\sum_i\left(\sum_j t_{ijXY}\sum_{j'} t_{ij'UZ}\right) - 2\sum_i\sum_j t_{ijXY}t_{ijUZ} - \frac{2(2n-3)}{n(n-1)}\sum_i\sum_j t_{ijXY}\sum_{i'}\sum_{j'} t_{i'j'UZ}}{n(n-1)(n-2)(n-3)}.$$

Proof B.1
Taking expectation of each term in the numerator, we have

$$\begin{aligned}E\left[4\sum_i\left(\sum_j t_{ijXY}\sum_{j'} t_{ij'UZ}\right)\right] &= 4n(n-1)E\left(t_{ijXY}t_{ijUZ}\right) + 4n(n-1)(n-2)E\left(t_{ijXY}t_{ij'UZ}\right) \\ E\left[2\sum_i\sum_j t_{ijXY}t_{ijUZ}\right] &= 2n(n-1)E\left(t_{ijXY}t_{ijUZ}\right) \\ E\left[\frac{2(2n-3)}{n(n-1)}\sum_i\sum_j t_{ijXY}\sum_{i'}\sum_{j'} t_{i'j'UZ}\right] &= 2(2n-3)E\left[t_{ijXY}\sum_i\sum_j t_{ijUZ}\right] \\ &= 2(2n-3)\left[2E\left(t_{ijXY}t_{ijUZ}\right) + 4(n-2)E\left(t_{ijXY}t_{ij'UZ}\right)\right. \\ &\quad \left. + (n-2)(n-3)E(t_{ijXY}t_{i'j'UZ})\right]\end{aligned}$$

Combine all preceding terms,

$$\begin{aligned}
E\left[\widehat{\mathrm{cov}}\left(t_{XY}, t_{UZ}\right)\right] &= \frac{2(n-2)(n-3)E\left(t_{ijXY}t_{ijUZ}\right)}{n(n-1)(n-2)(n-3)} \\
&\quad + \frac{4(n-2)^2(n-3)E\left(t_{ijXY}t_{ij'UZ}\right)}{n(n-1)(n-2)(n-3)} \\
&\quad - \frac{2(2n-3)(n-2)(n-3)E\left(t_{ijXY}t_{i'j'UZ}\right)}{n(n-1)(n-2)(n-3)} \\
&= \frac{2(n-2)(n-3)\left[\mathrm{cov}\left(t_{ijXY}, t_{ijUZ}\right) + \tau_{XY}\tau_{UZ}\right]}{n(n-1)(n-2)(n-3)} \\
&\quad + \frac{4(n-2)^2(n-3)\left[\mathrm{cov}\left(t_{ijXY}, t_{ij'UZ}\right) + \tau_{XY}\tau_{UZ}\right]}{n(n-1)(n-2)(n-3)} \\
&\quad - \frac{2(2n-3)(n-2)(n-3)\tau_{XY}\tau_{UZ}}{n(n-1)(n-2)(n-3)} \\
&= \frac{2}{n(n-1)}\mathrm{cov}\left(t_{ijXY}, t_{ijUZ}\right) + \frac{4(n-2)}{n(n-1)}\mathrm{cov}\left(t_{ijXY}, t_{ij'UZ}\right) \\
&= \mathrm{cov}\left(t_{XY}, t_{UZ}\right)
\end{aligned}$$

□

In the case of the variances and covariances involving a shared variable,

$$\widehat{\mathrm{var}}\left(t^*_{XX}\right) = \frac{4\sum_i\left(\sum_j t^*_{ijXX}\right)^2 - 2\sum_i\sum_j t^{*2}_{ijXX} - \frac{2(2n-3)}{n(n-1)}\left(\sum_i\sum_j t^*_{ijXX}\right)^2}{n(n-1)(n-2)(n-3)},$$

$$\widehat{\mathrm{var}}(t_{XY}) = \frac{4\sum_i\left(\sum_j t_{ijXY}\right)^2 - 2\sum_i\sum_j t^2_{ijXY} - \frac{2(2n-3)}{n(n-1)}\left(\sum_i\sum_j t_{ijXY}\right)^2}{n(n-1)(n-2)(n-3)},$$

$$\widehat{\mathrm{cov}}\left(t^*_{XX}, t_{XY}\right) = \frac{4\sum_i\left(\sum_j t^*_{ijXX}\sum_{j'} t_{ij'XY}\right) - 2\sum_i\sum_j t^*_{ijXX}t_{ijXY} - \frac{2(2n-3)}{n(n-1)}\sum_i\sum_j t^*_{ijXX}\sum_{i'}\sum_{j'} t_{i'j'XY}}{n(n-1)(n-2)(n-3)},$$

$$\widehat{\mathrm{cov}}\left(t_{XY}, t_{XZ}\right) = \frac{4\sum_i\left(\sum_j t_{ijXY}\sum_{j'} t_{ij'XZ}\right) - 2\sum_i\sum_j t_{ijXY}t_{ijXZ} - \frac{2(2n-3)}{n(n-1)}\sum_i\sum_j t_{ijXY}\sum_{i'}\sum_{j'} t_{i'j'XZ}}{n(n-1)(n-2)(n-3)}.$$

Acknowledgements

This project was supported in part by an appointment to the Research Participation Program at the Center for Devices and Radiological Health administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and the US Food and Drug Administration. Le Kang was partly supported by the National Heart, Lung, and Blood Institute grant R01HL113697 and by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development grant R01HD060913. The authors are indebted to Dr. Frank Samuelson and Dr. Adam Wunderlich for helpful discussions. The opinions expressed are those of the authors and not necessarily those of the Editors. The authors thank the Associated Editor and three anonymous referees for constructive comments that greatly improved the article.

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.
2. Metz CE. Basic principles of ROC analysis. In *Seminars in Nuclear Medicine*, Vol. 8, no. 4. WB Saunders, 1978; 283–298.
3. Swets J, Pickett R. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press: New York, 1982.
4. Pepe M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: USA, 2004.
5. Zhou X-H, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*, Vol. 712. John Wiley & Sons, 2011.
6. Dorfman D, Alf E. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating-method data. *Journal of Mathematical Psychology* 1969; **6**(3):487–496.
7. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**(4):387–415.
8. Zou K, Hall W, Shapiro D. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* 1998; **16**(19):2143–2156.
9. Metz C, Wang P, Kronman H. A new approach for testing the significance of differences between ROC curves measured from correlated data. In *Information Processing in Medical Imaging*, Vol. 8. The Hague: The Netherlands, 1984; 286–295.
10. Hanley J, McNeil B. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; **148**(3):839–843.
11. DeLong E, DeLong D, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**(3):837–845.
12. Sen PK. On some convergence properties of *U*-statistics. *Calcutta Statistical Association Bulletin* 1960; **10**(1):1–18.
13. Murphy J, Berwick D, Weinstein M, Borus J, Budman S, Klerman G. Performance of screening and diagnostic tests: application of receiver operating characteristic analysis. *Archives of General Psychiatry* 1987; **44**(6):550–555.
14. Greiner M, Pfeiffer D, Smith R. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine* 2000; **45**(1):23–41.
15. Zou K, O'Malley A, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation* 2007; **115**(5):654–657.
16. Heagerty PJ, Lumley T, Pepe MS. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics* 2000; **56**(2):337–344.
17. Heagerty PJ, Zheng Y. Survival model predictive accuracy and roc curves. *Biometrics* 2005; **61**(1):92–105.
18. Cai T, Pepe MS, Zheng Y, Lumley T, Jenny NS. The sensitivity and specificity of markers for event times. *Biostatistics* 2006; **7**(2):182–197.
19. Chambless LE, Diao G. Estimation of time-dependent area under the roc curve for long-term risk prediction. *Statistics in Medicine* 2006; **25**(20):3474–3486.
20. Smith W, Dutton R, Smith N. A measure of association for assessing prediction accuracy that is a generalization of non-parametric ROC area. *Statistics in Medicine* 1996; **15**(11):1199–1215.
21. Obuchowski N. An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. *Statistics in Medicine* 2005; **25**(3):481–493.
22. Harrell F, Jr., Califf R, Pryor D, Lee K, Rosati R. Evaluating the yield of medical tests. *JAMA: The Journal of the American Medical Association* 1982; **247**(18):2543–2546.
23. Harrell F, Lee K, Califf R, Pryor D, Rosati R. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* 1984; **3**(2):143–152.
24. 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.
25. Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; **92**(4):965–970.
26. Pencina MJ, D'Agostino RB, Song L. Quantifying discrimination of Framingham risk functions with different survival *C* statistics. *Statistics in Medicine* 2012; **31**(15):1543–1553.
27. Pencina M, D'Agostino R. 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.
28. Nam B, D'Agostino R. Discrimination index, the area under the ROC curve. *Goodness-of-Fit Tests and Model Validity* 2002:267–279.
29. 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.
30. Efron B. Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 1979; **7**(1):1–26.
31. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. London: Chapman & Hall, 1993.
32. Kannel WB. The Framingham Study. *British Medical Journal* 1976; **2**(6046):1255–1255.
33. Kim JO. Predictive measures of ordinal association. *American Journal of Sociology* 1971; **76**(5):891–907.
34. Klein J, Moeschberger M. *Survival Analysis: Techniques for Censored and Truncated Data*. Spring-Verlag: New York, 1997.
35. Hoeffding W. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* 1948; **19**(3):293–325.
36. Noether G. *Elements of Nonparametric Statistics*, Vol. 4. Wiley: New York, 1967.
37. Casella G, Berger RL. *Statistical Inference*. Duxbury Press: CA, 2001.
38. Marx M, Larsen R. *Introduction to Mathematical Statistics and Its Applications*. Pearson/Prentice Hall, 2006.
39. Klein J, Goel P. *Survival Analysis: State of the Art*. Springer, 1992.

40. Koziol JA, Jia Z. The concordance index C and the Mann–Whitney parameter $\Pr(x > y)$ with randomly censored data. *Biometrical Journal* 2009; **51**(3):467–474.
41. Dawber TR, Meadors GF, Moore FE, Jr. Epidemiological approaches to heart disease: the Framingham Study. *American Journal of Public Health and the Nations Health* 1951; **41**(3):279–286.
42. Castelli W. Epidemiology of coronary heart disease: the Framingham Study. *The American Journal of Medicine* 1984; **76**(2):4–12.
43. Ho KK, Pinsky JL, Kannel WB, Levy D. The epidemiology of heart failure: the Framingham Study. *Journal of the American College of Cardiology* 1993; **22**(4):A6–A13.
44. *Framingham Heart Study, a Project of National Heart, Lung and Blood Institute and Boston University*, 2013. Available from: http://www.framinghamheartstudy.org/ [Accessed on 9 January 2013].
45. Neaton JD, Wentworth D. Serum cholesterol, blood pressure, cigarette smoking, and death from coronary heart disease overall findings and differences by age for 316099 white men. *Archives of Internal Medicine* 1992; **152**(1):56–64.
46. Eckel RH, Krauss RM. American heart association call to action: obesity as a major risk factor for coronary heart disease. *Circulation* 1998; **97**(21):2099–2100.
47. Kannel WB, Gordon T, Schwartz MJ. Systolic versus diastolic blood pressure and risk of coronary heart disease: the Framingham Study. *The American Journal of Cardiology* 1971; **27**(4):335–346.
48. Stamler J, Stamler R, Neaton JD. Blood pressure, systolic and diastolic, and cardiovascular risks: US population data. *Archives of Internal Medicine* 1993; **153**(5):598–615.
49. Korn EL, Simon R. Measures of explained variation for survival data. *Statistics in Medicine* 1990; **9**(5):487–503.
50. Henderson R. Problems and prediction in survival-data analysis. *Statistics in Medicine* 1995; **14**(2):161–184.
51. Schemper M, Stare J. Explained variation in survival analysis. *Statistics in Medicine* 1996; **15**(19):1999–2012.
52. Begg CB, Cramer LD, Venkatraman ES, Rosai J. Comparing tumour staging and grading systems: a case study and a review of the issues, using thymoma as a model. *Statistics in Medicine* 2000; **19**(15):1997–2014.
53. Vexler A, Liu A, Eliseeva E, Schisterman EF. Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection. *Biometrics* 2008; **64**(3):895–903.
54. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Medical Research Methodology* 2011; **11**(1):13.
55. Demler OV, Pencina MJ, D'Agostino RB. Misuse of delong test to compare AUCs for nested models. *Statistics in Medicine* 2012; **31**(23):2577–2587.
56. Chen W, Samuelson FW, Gallas BD, Kang L, Sahiner B, Petrick N. On the assessment of the added value of new predictive biomarkers. *BMC Medical Research Methodology* 2013; **13**(1):98.
57. Cliff N, Charlin V. Variances and covariances of Kendall's tau and their estimation. *Multivariate Behavioral Research* 1991; **26**(4):693–707.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.