

Received 12 October 2009, Accepted 22 October 2010 Published online 13 January 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4154

On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data

Hajime Uno,^{a,b,*†} Tianxi Cai,^b Michael J. Pencina,^c
Ralph B. D'Agostino^d and L. J. Wei^b

For modern evidence-based medicine, a well thought-out risk scoring system for predicting the occurrence of a clinical event plays an important role in selecting prevention and treatment strategies. Such an index system is often established based on the subject's 'baseline' genetic or clinical markers via a working parametric or semi-parametric model. To evaluate the adequacy of such a system, C-statistics are routinely used in the medical literature to quantify the capacity of the estimated risk score in discriminating among subjects with different event times. The C-statistic provides a global assessment of a fitted survival model for the continuous event time rather than focussing on the prediction of t -year survival for a fixed time. When the event time is possibly censored, however, the population parameters corresponding to the commonly used C-statistics may depend on the study-specific censoring distribution. In this article, we present a simple C-statistic without this shortcoming. The new procedure consistently estimates a conventional concordance measure which is free of censoring. We provide a large sample approximation to the distribution of this estimator for making inferences about the concordance measure. Results from numerical studies suggest that the new procedure performs well in finite sample. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: AUC; Cox's proportional hazards model; Framingham risk score; ROC

1. Introduction

For modern clinical medicine, risk prediction procedures are valuable tools for disease prevention and management. Pioneered by the Framingham study, risk score systems have been established for assessing individual risks of developing cardiovascular (CV) diseases, cancer, or many other conditions within a certain time period [1–4]. A key component in the assessment of risk algorithm performance is its ability to distinguish subjects who will develop an event ('cases') from those who will not ('controls'). This concept, known as discrimination, has been well studied and quantified for binary outcomes using measures such as the estimated area under the Receiver Operating Characteristics (ROC) curve (AUC), which is also referred to as a 'C-statistic' [5]. Such a statistic is an estimated conditional probability that for any pair of 'case' and 'control', the predicted risk of an event is higher for the 'case' [6].

If the primary response variable is the time to a certain event, the aforementioned procedure for binary outcomes can be used to quantify the ability of the risk score system to differentiate cases from controls at a time point t . If one is not interested in a particular time point, a standard concordance measure may be used to evaluate the overall performance of the risk scoring system. Specifically, let T be the event time, Z be a $p \times 1$ covariate vector, and $g(Z)$ be an estimated risk score for subjects

^aDepartment of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Boston, MA, U.S.A.

^bDepartment of Biostatistics, Harvard University, Boston, MA, U.S.A.

^cDepartment of Biostatistics, Boston University, Boston, MA, U.S.A.

^dDepartment of Mathematics and Statistics, Boston University, Boston, MA, U.S.A.

*Correspondence to: Hajime Uno, Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, U.S.A.

†E-mail: huno@jimmy.harvard.edu

with Z . There is a large class of measures to quantify how well the risk score $g(Z)$ predicts the distribution of T or a function thereof. A good review paper is given by Korn and Simon [7] or more recently by Hielscher *et al.* [8]. Such prediction measures can be classified into two broad classes, one based on explicit loss functions between the risk score and the survival time and the other based on rank correlations between these two quantities. The C-statistic proposed by Harrell *et al.* [9–11] is essentially a rank-correlation measure, motivated by Kendall's tau for censored survival data [12]. A critical issue for rank-correlation methods is how to order survival times in the presence of censoring. Brown *et al.* [12] used all observations and assigned probability scores to pairs in which ordering is not obvious due to censoring, based on the pooled Kaplan–Meier estimate for T . However, the score based on the pooled Kaplan–Meier estimate may not be appropriate when the covariates are associated with T . Alternative forms of C-statistic considered in [9–11, 13] use only so-called ‘useable’ pairs and calculate the proportion of concordant pairs among them. However, such C-statistics estimate population parameters that may depend on the current study-specific censoring distribution. In this article, we propose a modified C-statistic which is consistent for a population concordance measure that is free of censoring.

More specifically, for two independent copies $\{(T_1, g(Z_1))', (T_2, g(Z_2))'\}$ of $(T, g(Z))'$, a commonly used concordance measure is

$$C = \text{pr}(g(Z_1) > g(Z_2) | T_2 > T_1). \quad (1)$$

When T is subject to right censoring, as discussed in Heagerty and Zheng [14] one would typically consider a modified C_τ with a fixed, prespecified follow-up period $(0, \tau)$, where

$$C_\tau = \text{pr}(g(Z_1) > g(Z_2) | T_2 > T_1, T_1 < \tau). \quad (2)$$

Estimation of (1) or (2) when the event time may be censored, however, is not straightforward [11, 13, 15]. The estimator for C or C_τ proposed by Heagerty and Zheng [14] is derived under a proportional hazards model. If this semi-parametric working model is not correctly specified, the resulting estimator may be biased. A popular nonparametric C-statistic for estimating C or C_τ was proposed by Harrell *et al.* [9–11] and extensively studied by Pencina and D'Agostino [13]. Note that this generalization is a weighted area under an ‘incident/dynamic’ ROC curve [14, 16] with weights depending on the study-specific censoring distribution.

When the study individuals have different follow-up times, the C-statistic studied by Harrell *et al.* [9–11] converges to an association measure that involves the study censoring distribution. In this article, under the general random censorship assumption, we provide a simple nonparametric estimator for the concordance measure in (2), which is free of censoring. Furthermore, we study the large sample properties of the new estimation procedure. Our proposal is illustrated with two real examples. The performance of the new proposal under various practical settings is also examined via a simulation study. Note that Gönen and Heller [17] proposed a method for censored survival data to estimate $\text{pr}(T_2 > T_1 | g(Z_1) > g(Z_2))$, which is also a concordance measure and has a similar form to (1), but we do not focus on this type of measures in this paper.

2. Inference procedures for degree of association between event times and estimated risk scores

In this section, we consider a non-trivial case that at least one component of the covariate vector Z is continuous. For the survival time T , let D be the corresponding censoring variable. Assume that D is independent of T and Z . Let $\{(T_i, Z_i, D_i), i = 1, \dots, n\}$ be n independent copies of $\{(T, Z, D)\}$. For the i th subject, we only observe (X_i, Z_i, Δ_i) , where $X_i = \min(T_i, D_i)$, and Δ_i equals 1 if $X_i = T_i$ and 0 otherwise.

Suppose that we fit the data with a *working* parametric or semi-parametric regression model, for example, a standard Cox proportional hazards model [18]:

$$\Lambda_Z(t) = \Lambda_0(t) \exp(\beta' Z), \quad (3)$$

where $\Lambda_Z(\cdot)$ is the cumulative hazard function for subjects with covariate vector Z , $\Lambda_0(\cdot)$ is the unknown baseline cumulative hazard function and β is the unknown $p \times 1$ parameter vector. Let the maximum partial likelihood estimator for β be denoted by $\hat{\beta}$. Note that even when the model (3) is not correctly

specified, under a rather mild non-separable condition that there does not exist vector ζ such that $\text{pr}(T_1 > T_2 | \zeta' Z_1 < \zeta' Z_2) = 1$, $\hat{\beta}$ converges to a constant vector, say, β_0 , as $n \rightarrow \infty$. This stability property is important for deriving the new inference procedure.

For a pair of future patients with covariate vectors $\{Z_k^0, k=1, 2\}$ and the potential survival times $\{T_k^0, k=1, 2\}$, their corresponding risk scores are $\{\hat{\beta}' Z_k^0, k=1, 2\}$. To evaluate this risk score system, one may use the concordance measure discussed in Section 1

$$C_n = \text{pr}(\hat{\beta}' Z_1^0 > \hat{\beta}' Z_2^0 | T_1^0 < T_2^0),$$

where the probability is evaluated with respect to the data, and (T_1^0, Z_1^0) and (T_2^0, Z_2^0) . Note that C_n depends on the sample size. Let the limit of C_n be denoted by

$$C = \text{pr}(\beta_0' Z_1^0 > \beta_0' Z_2^0 | T_1^0 < T_2^0). \quad (4)$$

Now, since the support of the censoring variable D is usually shorter than that of the failure time T , the tail part of the estimated survival function of T is rather unstable. Therefore, we consider a truncated version of C in (4), that is

$$C_\tau = \text{pr}(\beta_0' Z_1^0 > \beta_0' Z_2^0 | T_1^0 < T_2^0, T_1^0 < \tau), \quad (5)$$

where τ is a prespecified time point such that $\text{pr}(D > \tau) > 0$.

It follows from an 'inverse probability weighting' technique as employed in Cheng *et al.* [19] for dealing with a completely different problem in survival analysis that C_τ can be consistently, nonparametrically estimated by

$$\hat{C}_\tau = \frac{\sum_{i=1}^n \sum_{j=1}^n \Delta_i \{\hat{G}(X_i)\}^{-2} I(X_i < X_j, X_i < \tau) I(\hat{\beta}' Z_i > \hat{\beta}' Z_j)}{\sum_{i=1}^n \sum_{j=1}^n \Delta_i \{\hat{G}(X_i)\}^{-2} I(X_i < X_j, X_i < \tau)}, \quad (6)$$

where $I(\cdot)$ is the indicator function and $\hat{G}(\cdot)$ is the Kaplan–Meier estimator for the censoring distribution $G(t) = \text{pr}(D > t)$. Heuristically, the consistency of the above estimator follows from the fact that as $n \rightarrow \infty$, the denominator of (6) divided by n^2 converges to

$$\begin{aligned} E\{\Delta_1 \{G(X_1)\}^{-2} I(X_1 < X_2, X_1 < \tau)\} &= E[E\{I(D_1 \wedge D_2 > T_1)\{G(T_1)\}^{-2} I(T_1 < T_2, T_1 < \tau) | T_1\}] \\ &= E[E\{I(D_1 > T_1) I(D_2 > T_1)\{G(T_1)\}^{-2} I(T_1 < T_2, T_1 < \tau) | T_1\}]. \end{aligned}$$

Note that $E\{I(D > T_1)\{G(T_1)\}^{-1} | T_1\} = 1$ in the above equation, where the expectation is taken with respect to D . Thus, the denominator and the numerator of (6) divided by n^2 converge to $\text{pr}(T_1^0 < T_2^0, T_1^0 < \tau)$ and $\text{pr}(\beta_0' Z_1^0 > \beta_0' Z_2^0, T_1^0 < T_2^0, T_1^0 < \tau)$, respectively.

In Appendix A, we show that

$$W = n^{\frac{1}{2}}(\hat{C}_\tau - C_\tau)$$

is asymptotically normal with mean 0. Moreover, in the Appendix, we show how to use a perturbation-resampling method to approximate the distribution of W . Specifically, we show that the asymptotic distribution of \tilde{W} given in (A3) is the same as that of W . The realizations of \tilde{W} can be generated easily by simulating a large number, M , of random samples from, for instance, the unit exponential. Inferences about C_τ can then be made via the normal approximation to the distribution of \hat{C}_τ and these realizations of \tilde{W} . For instance, a two-sided 0.95 confidence interval for C_τ would be $\hat{C}_\tau \pm 1.96n^{-1/2}\sigma$, where σ^2 is the standard sample variance or a robust version thereof based on the above M realizations of \tilde{W} . It is important to note that the C-statistic proposed by Harrell *et al.* [11] is

$$\frac{\sum_{i \neq j} \Delta_i I(X_i < X_j) I(\hat{\beta} Z_i > \hat{\beta} Z_j)}{\sum_{i \neq j} \Delta_i I(X_i < X_j)}, \quad (7)$$

which converges to a censoring-dependent quantity,

$$\text{pr}(\beta_0' Z_1^0 > \beta_0' Z_2^0 | T_1^0 < T_2^0, T_1^0 \leq D_1^0 \wedge D_2^0).$$

Pencina and D'Agostino [13] formulated an alternative form of C-statistic

$$\frac{\sum_{i \neq j} \Delta_i I(X_i < X_j, X_i < \tau) I(\hat{\beta} Z_i > \hat{\beta} Z_j)}{\sum_{i \neq j} \Delta_i I(X_i < X_j, X_i < \tau)} \quad (8)$$

for various τ . The limiting value of this statistic,

$$\text{pr}(\beta_0' Z_1^0 > \beta_0' Z_2^0 | T_1^0 < T_2^0, T_1^0 \leq D_1^0 \wedge D_2^0, T_1^0 < \tau) \quad (9)$$

also involves the censoring distribution.

When there are two competing survival regression models, say A and B, one may compare the overall predictive performances of these models based on their C-statistics. Specifically, let $C_\tau^{(A)}$ and $C_\tau^{(B)}$ be C_τ for Model A and Model B, respectively. Let $\xi = C_\tau^{(A)} - C_\tau^{(B)}$. Then, a consistent estimator for ξ is $\hat{\xi} = \hat{C}_\tau^{(A)} - \hat{C}_\tau^{(B)}$, where $\hat{C}_\tau^{(A)}$ and $\hat{C}_\tau^{(B)}$ are the corresponding C-statistics (6). Note that these two C-statistics are correlated. It follows from a similar argument for deriving the large sample properties of \hat{C}_τ for a single model, that the distribution of

$$W_\xi = n^{\frac{1}{2}}(\hat{\xi} - \xi)$$

can be approximated by a normal with mean zero. Its variance can be obtained via the above perturbation-resampling method. The details of this normal approximation are given in Appendix B. To make inference about ξ , a two-sided 0.95 confidence interval is $\hat{\xi} \pm 1.96 n^{-1/2} \hat{\sigma}_{W_\xi}$, where $\hat{\sigma}_{W_\xi}^2$ denotes the estimated variance of W_ξ .

3. Numerical studies

3.1. Examples

First, we illustrate the proposed procedures with two data sets. The first one is from the Framingham Heart Study. For this data set, there were 3087 study participants whose baseline covariate vectors Z 's were obtained at their study entry times between 1991 and 1995. Here, each Z consists of age, gender, smoking status (SMK), total cholesterol (TC), HDL cholesterol (HDL), systolic blood pressure (SBP), and use of medication for high-blood pressure (TxBP). These individuals were then followed until 2006. Here, the event time T is the first time that the subject experienced any of the following CV events including coronary death, myocardial infarction, coronary insufficiency, angina pectoris, fatal and non-fatal stroke, intermittent claudication, or congestive heart failure. For this data set, there are 377 such events observed during the entire follow-up period, and 282 of which occurred in the first 10 years. The Kaplan–Meier estimates for the survival distributions of both the event time T and the censoring time D are given in Figure 1. Note that most study subjects were followed for more than 10 years, but less than 13 years.

We fitted the data with a Cox proportional hazards model (3). The resulting risk score $\hat{\beta}' Z^0$ is

$$0.54 \times (\text{AGE}/10) - 0.41 \times I(\text{Male}) + 0.53 \times I(\text{SMK} = \text{Yes}) + 0.40 \times (\text{TC}/10^2 [\text{mg/dL}]) - 0.21 \\ \times (\text{HDL}/10 [\text{mg/dL}]) + 0.15 \times (\text{SBP}/10 [\text{mmHg}]) + 0.33 \times I(\text{TxBP} = \text{Yes}).$$

In Table I, we present point and 0.95 interval estimates of C_τ for various τ . When $\tau = 8, 10, 12$ (years), our results are very similar to those based on the conventional C-index (7) procedure with a point estimate of 0.75 and a 0.95 confidence interval of (0.73, 0.77). Note that all the τ -specific conventional C-statistics (8) give us similar point and interval estimates. When $\tau = 14$, our estimated standard error for the new C-statistic is markedly larger than that of the conventional method. In this case, study subjects did not have similar follow-up times and it is known that the existing methods in the literature may not work well [13]. Note that all the results reported in Table I were based on $M = 500$ independent realizations of a random sample with $n = 3087$ from the unit exponential for (A3).

For $\tau = 10$ (years), we also evaluated the incremental value of HDL by comparing the model containing all risk factors described above (Model A) with the model containing risk factors other than HDL (Model B). Results of the Cox regression analysis were given in Table II. The point estimate for $\xi = C_\tau^{(A)} - C_\tau^{(B)}$ was 0.01, and corresponding 0.95 confidence interval was $(-0.00, 0.02)$. It is interesting

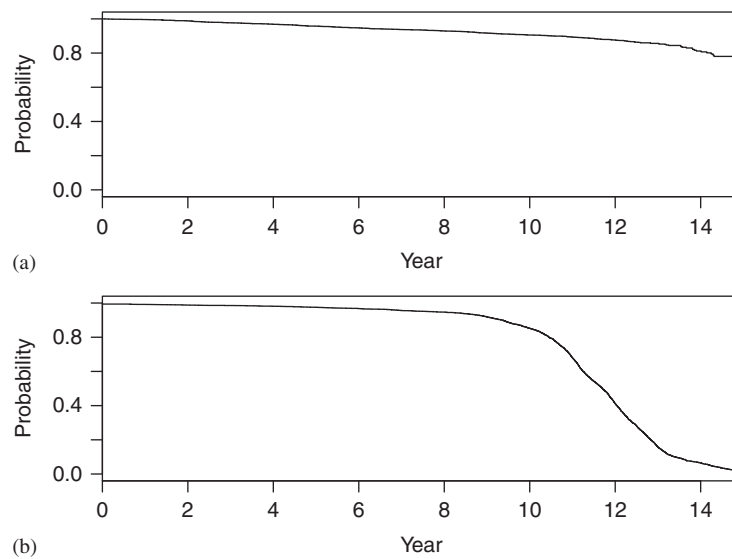


Figure 1. Estimates for survival functions for CV events and censoring variables with Framingham study data: (a) Kaplan–Meier curve for CV events and (b) Kaplan–Meier curve for the censoring.

Table I. Point estimates (Est.), standard error estimates (SE), and 0.95 confidence intervals (CI) for C_τ with Framingham study data.						
τ	New method			Conventional		
	Est.	SE	CI	Est.	SE	CI
8	0.76	0.02	(0.73, 0.79)	0.76	0.01	(0.73, 0.79)
10	0.75	0.01	(0.72, 0.78)	0.75	0.01	(0.73, 0.78)
12	0.75	0.01	(0.72, 0.77)	0.75	0.01	(0.73, 0.78)
14	0.75	0.02	(0.70, 0.80)	0.75	0.01	(0.73, 0.78)
∞	NA	NA	NA	0.75	0.01	(0.73, 0.77)

Table II. Estimates of regression parameters for Cox's models with Framingham study data.						
	Model A (with HDL)			Model B (without HDL)		
	Est.*	SE [†]	p [‡]	Est.	SE	p
Age/10 (yrs)	0.54	0.07	0.00	0.53	0.07	0.00
Gender= male	−0.41	0.12	0.00	−0.63	0.11	0.00
SMK=yes	0.53	0.13	0.00	0.57	0.12	0.00
TC/10 ² (mg/dL)	0.40	0.15	0.01	0.34	0.15	0.03
SBP/10 (mmHg)	0.15	0.03	0.00	0.16	0.03	0.00
TxBP=yes	0.33	0.12	0.01	0.40	0.12	0.00
HDL/10 (mg/dL)	−0.21	0.04	0.00	—	—	—
C-index with $\tau=10$ (0.95 confidence interval)						
Conventional		0.75 (0.73, 0.78)			0.74 (0.72, 0.77)	
New method		0.75 (0.72, 0.78)			0.74 (0.71, 0.77)	

*Estimate.

[†]Standard error estimate.

[‡]p-Value.

to note that although HDL is statistically highly significant with respect to the risk of CV disease in the first 10 years of follow-up, its incremental value evaluated via C-statistics is quite modest.

The second data set for illustration is from a recent cancer study [20]. One primary goal of the study was to evaluate the prognostic value of a new gene signature in predicting the time to death or metastasis for breast cancer patients. An important clinical implications for establishing a risk score system is to identify future patients who may benefit from adjuvant systemic, but potentially toxic, therapies. The

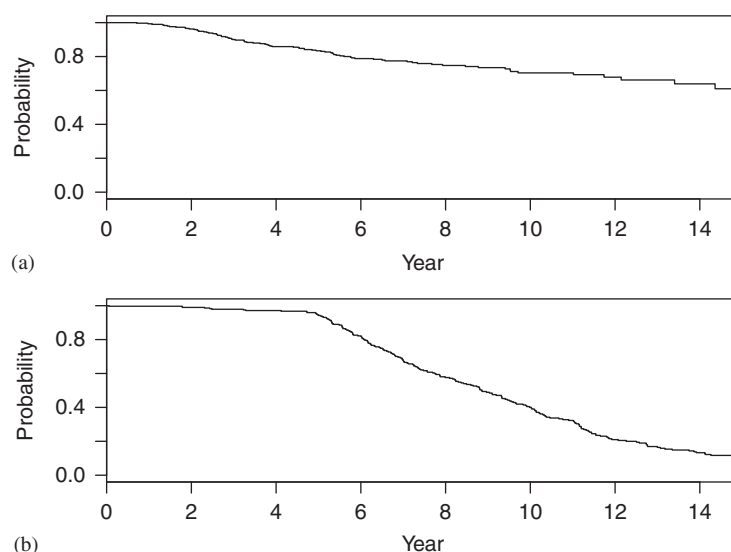


Figure 2. Estimates for survival functions for death and censoring variables with breast cancer data: (a) Kaplan–Meier curve for death and (b) Kaplan–Meier curve for the censoring.

details of participants in the study are given in van't Veer *et al.* [21] and van de Vijver *et al.* [22]. The data set for analysis consists of 295 breast cancer patient records from the Netherlands Cancer Institute. Our goal is to evaluate the prediction of patient survival based on the patients' baseline covariates consisting of the new gene score and other conventional markers including age and estrogen receptor status. The Kaplan–Meier curves for the event time and the censoring time are given in Figure 2. Note that at Year 15, the survival rate is 0.61, the size of the risk set is 19, and there were no deaths beyond this time point (http://microarray-pubs.stanford.edu/wound_NKI/explore.html).

We fit the data with two working Cox's proportional hazards models. The covariate vector Z for the first model consists of only age and estrogen receptor (ER: positive or negative). The resulting risk score $\hat{\beta}'Z$ is

$$-0.47 \times (\text{AGE}/10) - 1.13 \times I(\text{ER} = \text{positive}).$$

In the second model, we included the gene score variable (GS), age and ER. The risk score with the second model is

$$3.28 \times (\text{GS}) - 0.62 \times (\text{AGE}/10) - 0.70 \times I(\text{ER} = \text{positive}).$$

In Table III, we report the point and interval estimates of C_τ for both models with various τ . Our standard error estimates tend to be larger than those from the conventional C-statistic especially for the case when $\tau = 15$. To examine the performance of the new proposal, we conducted a simulation study. The details are given in the next subsection.

3.2. Simulation study

To examine the performance of the new and conventional methods, we conducted a numerical study under various practical settings. As described in Section 2, our proposed inference procedure does not require the fitted model to be correct. Here, we used a *working* Cox regression model to obtain an estimated risk score, and make inferences about C_τ regardless of the adequacy of the *working* model. For instance, in one of the simulation settings, we considered two cases with survival time T generated from a Weibull regression model for the first case and from a log-normal regression model for the second case. For both models, we let $Z = (\text{AGE}, \text{ER}, \text{GS})'$. We also considered three kinds of censoring: (a) type I censoring without staggered entry; (b) random censoring which is independent of survival time and covariates; and (c) random censoring which is independent of survival time conditional on the covariates. Note that in theory, the existing C-statistics may be valid only for (a). Our procedure is valid for (a) or (b). Setting (c) is explored to examine whether the proposed procedures are sensitive

Table III. Point estimates (Est.), standard error estimates (SE), and 0.95 confidence intervals (CI) for C_τ with breast cancer data.						
τ	New method			Conventional		
	Est.	SE	CI	Est.	SE	CI
<i>Without gene score model</i>						
6	0.66	0.04	(0.58, 0.74)	0.68	0.04	(0.60, 0.76)
8	0.65	0.04	(0.58, 0.73)	0.68	0.04	(0.60, 0.75)
10	0.65	0.04	(0.57, 0.72)	0.67	0.04	(0.60, 0.74)
15	0.62	0.05	(0.53, 0.71)	0.67	0.04	(0.60, 0.74)
∞^*	NA	NA	NA	0.67	0.04	(0.60, 0.74)
<i>With gene score model</i>						
6	0.75	0.03	(0.69, 0.82)	0.76	0.03	(0.69, 0.82)
8	0.75	0.03	(0.68, 0.81)	0.75	0.03	(0.69, 0.81)
10	0.74	0.03	(0.68, 0.80)	0.75	0.03	(0.69, 0.81)
15	0.70	0.05	(0.60, 0.79)	0.75	0.03	(0.69, 0.81)
∞^*	NA	NA	NA	0.75	0.03	(0.69, 0.81)

*Note that no event has been seen after $t = 15$ in the breast cancer data set. All usable pairs in the data set are already utilized with $\tau = 15$ for calculation of C-statistics. Thus, the estimates for $\tau = \infty$ and those for $\tau = 15$ are identical. With this data set, the censoring probability at $t = 15$ was 0.89.

to violation of the independent censoring assumption, as suggested by the reviewers. We chose $\tau = 10$ and 15, and $n = 100, 150, 200$, and 300 for each simulation setting.

First, we simulated $Z = (\text{AGE}, \text{ER}, \text{GS})'$, utilizing the breast cancer data. In particular, we generated ER status (positive or negative) from its empirical distribution, and then generated the corresponding (AGE, GS) from a multivariate normal distribution, $MVN(\mu_{ER}, \Sigma_{ER})$, where μ_{ER} and Σ_{ER} are the empirical mean and variance-covariance matrix of (AGE, GS) conditioning on ER, respectively. For each given value of Z , we generated T from the aforementioned Weibull regression model for case (I) and the log-normal regression model for case (II). We mimicked the breast cancer study to create the *true* models for our simulation. Specifically, the regression coefficients estimates were obtained by fitting the observed breast cancer data for each model.

With regard to generating censoring time, for censoring type (a), the censoring time D was set as $\tau + 0.1$ for all subjects. For censoring type (b), we fitted the breast cancer data with a one-sample Weibull distribution (with two unknown parameters) to generate D from it. For censoring type (c), we fitted the breast cancer data with a Weibull regression model with the covariate $Z = (\text{AGE}, \text{ER}, \text{GS})'$. About 45 per cent of the subjects were censored by Year 10 and 70 per cent by Year 15, which was similar to those from the breast cancer study.

For all simulation configurations, we used a Cox regression model with $Z = (\text{AGE}, \text{ER}, \text{GS})'$ as the *working* model to analyze each simulated data set regardless of whether the true model is the Weibull regression model or the log-normal regression model. Note that this model is correct under the aforementioned Weibull regression and incorrect under the log-normal regression model. For each configuration, we used the monte carlo method to calculate the true values of C_τ for the risk score obtained by fitting the *working* Cox model. Specifically, we generated one million data points of (T, Z) from each model. The *true* value for $C_{10} = 0.724$ and $C_{15} = 0.718$ when the Weibull regression model was true; $C_{10} = 0.727$ and $C_{15} = 0.710$ when the log-normal regression model was true.

For each setting, we then simulated 1000 independent realizations of $\{(X_i, \Delta_i, Z_i)', i = 1, \dots, n\}$. We fitted each simulated data set with the *working* Cox regression model and constructed a 0.95 confidence interval for C_τ given in (5). We also constructed the 0.95 confidence interval based on the conventional C-index given by Pencina and D'Agostino [13]. With the 1000 iterations, we obtained empirical coverage probabilities, average lengths of intervals, biases, and root mean square errors (rMSE).

Table IV shows the results for the case where the *working* model is correct. When $\tau = 15$ and $n = 100$ with independent censoring (b), for example, the coverage probabilities of the new and conventional methods are 0.944 and 0.876, respectively. The coverage probability of the new method is 0.946 and still close to the nominal level of 0.95 even under censoring type (c). Table V shows the results for the case where the fitted *working* model is not correct. For $\tau = 15$ and $n = 100$, the coverage probabilities of the new method are 0.947 and 0.945 under independent and conditionally independent censoring, respectively. On the other hand, those of the conventional method are 0.771 and 0.774, respectively.

Table IV. Coverage probabilities and average length of 0.95 confidence intervals, bias, and root mean square error (rMSE) for C_τ with $\tau = 10$ and 15, based on 1000 of iterations for a *working* Cox's model with (AGE, ER, GS), when the true model is (I) the Weibull regression model with AGE, ER, and GS.

τ	Censoring	N	Coverage		Ave. length		Bias		rMSE	
			*	†	*	†	*	†	*	†
10	(a) Degenerate	100	0.951	0.913	0.214	0.176	0.008	0.008	0.047	0.047
		150	0.945	0.917	0.162	0.145	0.008	0.008	0.038	0.038
		200	0.948	0.934	0.136	0.126	0.005	0.005	0.033	0.033
		300	0.942	0.934	0.108	0.103	0.003	0.003	0.027	0.027
	(b) Independent	100	0.948	0.896	0.241	0.196	0.011	0.015	0.053	0.055
		150	0.952	0.903	0.184	0.163	0.007	0.010	0.043	0.044
		200	0.942	0.914	0.152	0.141	0.009	0.012	0.037	0.039
		300	0.954	0.938	0.122	0.118	0.003	0.007	0.029	0.030
	(c) Cond. independent	100	0.948	0.892	0.241	0.196	0.011	0.015	0.053	0.055
		150	0.951	0.898	0.184	0.163	0.007	0.010	0.043	0.044
		200	0.946	0.916	0.152	0.142	0.009	0.012	0.037	0.039
		300	0.954	0.937	0.122	0.118	0.003	0.007	0.029	0.030
15	(a) Degenerate	100	0.951	0.912	0.174	0.149	0.006	0.006	0.040	0.040
		150	0.945	0.925	0.134	0.123	0.007	0.007	0.032	0.032
		200	0.956	0.938	0.113	0.107	0.003	0.003	0.027	0.027
		300	0.956	0.944	0.091	0.087	0.002	0.002	0.022	0.022
	(b) Independent	100	0.944	0.876	0.271	0.191	0.011	0.019	0.059	0.055
		150	0.952	0.896	0.207	0.159	0.007	0.015	0.048	0.045
		200	0.932	0.892	0.172	0.138	0.007	0.017	0.042	0.040
		300	0.950	0.926	0.137	0.115	0.003	0.012	0.033	0.031
	(c) Cond. independent	100	0.946	0.872	0.271	0.192	0.011	0.020	0.059	0.056
		150	0.952	0.894	0.206	0.159	0.007	0.015	0.048	0.045
		200	0.942	0.881	0.173	0.139	0.008	0.017	0.043	0.040
		300	0.951	0.925	0.137	0.115	0.004	0.012	0.033	0.031

*New method.

†Conventional.

In general, we find that the new interval estimation procedure performs well with relatively small sample size, regardless of the adequacy of the fitted model. Furthermore, the procedure is not sensitive to violation of the covariate independent censoring assumption with empirical coverage levels practically identical to their nominal counterparts under various situations. The conventional method proposed by Pencina and D'Agostino [13] performs well when the censoring distribution is a degenerate distribution (Type I censoring without staggered entry).

4. Remarks

In this article, we show that the overall performance of a parametric or semi-parametric model in predicting the subject-level survival over an interval $(0, \tau)$ can be evaluated via a simple, unbiased estimation procedure for C_τ . Based on our numerical study, we find that the new estimation procedure is robust with respect to the choice of τ . However, if the prespecified τ is 'too' large such that very few events were observed or very few study subjects were followed beyond this time point, the standard error estimate for \hat{C}_τ can be quite large, reflecting a high degree of uncertainty of our inferences about C_τ . For this case, one should cautiously utilize the fitted model for prediction over this large time interval.

There are various C-statistics proposed in the literature. With the same technique utilized in this article, one may modify these statistics accordingly so that they estimate concordance measures which are free of the study-specific censoring distribution. The computer code for implementing the new inference procedure can be downloaded from (<http://bcb.dfci.harvard.edu/~huno>).

For the new proposal, we assume that the censoring distribution is independent of the covariates. This assumption is not unreasonable in a well-executed clinical study, especially when there are no competing risks for observing the endpoint (for example, death or a composite clinical endpoint) and the event times are possibly censored mainly due to administrative censoring. When the covariate vector can be discretized, our new procedure can be modified easily using stratified Kaplan–Meier estimates for the censoring. When some of the covariates are continuous, one needs to assume a model for

Table V. Coverage probabilities and average length of 0.95 confidence intervals, bias, and root mean square error (rMSE) for C_τ with $\tau = 10$ and 15, based on 1000 of iterations for a *working* Cox's model with (AGE, ER, GS), when the true model is (II) the log-normal regression model with AGE, ER, and GS.

τ	Censoring	N	Coverage		Ave. length		Bias		rMSE	
			*	†	*	†	*	†	*	†
10	(a) Degenerate	100	0.947	0.909	0.201	0.165	0.009	0.009	0.045	0.045
		150	0.949	0.923	0.153	0.137	0.007	0.007	0.037	0.037
		200	0.951	0.931	0.128	0.119	0.006	0.006	0.031	0.031
		300	0.953	0.945	0.102	0.097	0.005	0.005	0.025	0.025
	(b) Independent	100	0.947	0.831	0.230	0.179	0.013	0.028	0.050	0.056
		150	0.950	0.858	0.176	0.150	0.007	0.024	0.041	0.046
		200	0.947	0.850	0.146	0.130	0.007	0.023	0.036	0.041
		300	0.926	0.822	0.116	0.106	0.006	0.023	0.030	0.036
	(c) Cond. independent	100	0.945	0.837	0.230	0.179	0.013	0.028	0.050	0.056
		150	0.942	0.859	0.175	0.150	0.008	0.024	0.042	0.046
		200	0.950	0.853	0.146	0.130	0.007	0.023	0.036	0.041
		300	0.928	0.813	0.116	0.106	0.006	0.023	0.030	0.036
15	(a) Degenerate	100	0.953	0.921	0.170	0.144	0.006	0.006	0.039	0.039
		150	0.960	0.935	0.131	0.119	0.004	0.004	0.031	0.031
		200	0.960	0.945	0.111	0.103	0.003	0.003	0.026	0.026
		300	0.949	0.944	0.088	0.085	0.002	0.002	0.022	0.022
	(b) Independent	100	0.947	0.771	0.263	0.175	0.011	0.043	0.060	0.063
		150	0.949	0.789	0.201	0.146	0.008	0.039	0.046	0.054
		200	0.943	0.751	0.168	0.127	0.007	0.038	0.041	0.050
		300	0.944	0.683	0.135	0.104	0.004	0.038	0.033	0.047
	(c) Cond. independent	100	0.945	0.774	0.264	0.175	0.011	0.043	0.060	0.063
		150	0.944	0.788	0.200	0.146	0.009	0.039	0.046	0.055
		200	0.937	0.743	0.168	0.127	0.007	0.038	0.041	0.050
		300	0.939	0.682	0.135	0.104	0.005	0.038	0.033	0.047

*New method.

†Conventional.

censoring distribution and use the resulting estimated distribution of the censoring to construct a new C-statistic. In theory, the validity of the resulting estimators depends on the adequacy of the model. On the other hand, based on the results from our simulation study, it appears that our proposal is quite robust even when the censoring is dependent on the covariates.

In this article, we showed an inference procedure not only for C_τ but also for the difference between two C_τ 's for comparing two competing prediction models. Although C-statistics are commonly used for quantifying the predictability of working models, they are not sensitive for capturing overall added values from a new marker on top of conventional predictors [11, 13]. Alternatively, one might use measures such as explained variation for survival data to compare two models [7]. If the model is correct, the likelihood based measures may be more sensitive in detecting differences in prediction ability, compared to rank-based measures such as C-indexes. Recently, alternative statistical estimation procedures were proposed for comparing prediction models [23, 24], which may be utilized to quantify incremental values. Further research is needed for developing intuitively interpretable and sensitive methods for comparing prediction models.

Appendix A: Large sample properties of \hat{C}_τ

Throughout, we assume that the non-separable condition for (T, Z) given in Section 2 holds, and thus β_0 is the unique solution to the limit of the following partial likelihood score equation

$$U(\beta) = n^{-1} \sum_{i=1}^n U_i(\beta) = n^{-1} \sum_{i=1}^n \int_0^\tau \left\{ Z_i - \frac{\sum_j Y_j(t) \exp(\beta' Z_j) Z_j}{\sum_j Y_j(t) \exp(\beta' Z_j)} \right\} dN_i(t) = 0,$$

where $N_i(t) = I(X_i \leq t, \Delta_i = 1)$, $Y_i(t) = I(X_i \geq t)$. We assume that β_0 lies in a compact parameter space and the joint density of (T, Z) is continuous and bounded. To show the consistency of \hat{C}_τ , we first note

that

$$n^{1/2}(\hat{\beta} - \beta_0) = n^{-1/2} \sum_{i=1}^n A(\beta_0) U_i(\beta_0) + O_p(n^{-1}), \quad (\text{A1})$$

where $A(\beta) = -\{\partial E(U(\beta))/\partial \beta\}^{-1}$ [25].

Now, for a fixed β , let

$$C_\tau(\beta) = \text{pr}(\beta' Z_1^0 > \beta' Z_2^0 | T_1^0 < T_2^0, T_1^0 < \tau).$$

It follows from the uniform consistency of $\hat{G}(\cdot)$, the convergence of $\hat{\beta}$ to β_0 , and a uniform law of large numbers for U -processes [26], that \hat{C}_τ converges to $C_\tau(\beta_0)$ in probability as $n \rightarrow \infty$. On the other hand, it follows from the asymptotic expansion of $\hat{\beta}$ given in (A1) that $C_\tau(\beta_0) - C_\tau = O(n^{-1})$. Thus, $\hat{C}_\tau - C_\tau \rightarrow 0$ in probability.

To approximate the distribution of

$$W = n^{1/2} \{\hat{C}_\tau(\hat{\beta}) - C_\tau\},$$

we first obtain asymptotic expansions for $W(\beta) = n^{1/2} \{\hat{C}_\tau(\beta) - C_\tau(\beta)\}$, where $\hat{C}_\tau(\beta)$ is obtained by replacing $\hat{\beta}$ in \hat{C}_τ of (6) with β . To this end, we write $W(\beta) = W_a(\beta) + W_b(\beta)$, where

$$W_a(\beta) = n^{1/2} \frac{\sum_{i=1}^n \sum_{j=1}^n I_{ij}(\tau) G(X_i)^{-2} \{I(\beta' Z_i > \beta' Z_j) - C_\tau(\beta)\}}{\sum_{i=1}^n \sum_{j=1}^n \hat{G}(X_i)^{-2} I_{ij}(\tau)},$$

$$W_b(\beta) = n^{1/2} \frac{\sum_{i=1}^n \sum_{j=1}^n \{\hat{G}(X_i)^{-2} - G(X_i)^{-2}\} I_{ij}(\tau) \{I(\beta' Z_i > \beta' Z_j) - C_\tau(\beta)\}}{\sum_{i=1}^n \sum_{j=1}^n \hat{G}(X_i)^{-2} I_{ij}(\tau)}$$

and $I_{ij}(\tau) = I(X_i < X_j, X_i < \tau) \Delta_i$. Now, it follows from the standard asymptotic theory for the Kaplan–Meier estimator [27]

$$\hat{W}_G(t) = \frac{n^{1/2} \{G(t) - \hat{G}(t)\}}{G(t)} \approx n^{-1/2} \sum_{i=1}^n \psi_i(t)$$

and $\hat{W}_G(t)$ converges weakly to a zero-mean Gaussian process indexed by t for $t \leq \tau$, where $\psi_i(t) = \int_0^t dM_i(u)/\pi_X(u)$, $\pi_X(t) = \text{pr}(X_i \geq t)$ and $M_i(t) = I(X_i \leq t, \Delta_i = 0) - \int_0^t I(X_i \geq u) d\Lambda_D(u)$, $\Lambda_D(\cdot)$ is the cumulative hazard function for the common censoring variable. Also, it follows from the uniform consistency of $\hat{G}(\cdot)$ and a functional central limit theorem for U -processes [28] that

$$W_a(\beta) = n^{-3/2} p(\tau)^{-1} \sum_{i=1}^n \sum_{j=1}^n G(X_i)^{-2} I_{ij}(\tau) \{I(\beta' Z_i > \beta' Z_j) - C_\tau(\beta)\} + o_p(1),$$

where $p(\tau) = P(T_1 < T_2, T_1 < \tau)$. Furthermore,

$$W_b(\beta) = \int_0^\tau n^{1/2} \left\{ \frac{G(t)^2}{\hat{G}(t)^2} - 1 \right\} d\hat{\gamma}(t, \beta),$$

where

$$\hat{\gamma}(t, \beta) = \frac{\sum_{i=1}^n \sum_{j=1}^n G(X_i)^{-2} I(X_i \leq t) I_{ij}(\tau) \{I(\beta' Z_i > \beta' Z_j) - C_\tau(\beta)\}}{\sum_{i=1}^n \sum_{j=1}^n \hat{G}(X_i)^{-2} I_{ij}(\tau)}.$$

By a uniform law of large numbers for U -processes [26] and the uniform consistency of $\hat{G}(\cdot)$, we have

$$\sup_{t \in [0, \tau], \beta} |\hat{\gamma}(t, \beta) - \gamma(t, \beta)| \rightarrow 0 \quad \text{in probability,}$$

where

$$\gamma(t, \beta) = \frac{p(t) \{\text{pr}(\beta' Z_i > \beta' Z_j | T_i < T_j, T_i < t) - C_\tau(\beta)\}}{p(\tau)}.$$

This, coupled with the weak convergence of $\hat{W}_G(t)$ implies that

$$W_b(\beta) = 2 \int \hat{W}_G(t) d\gamma(t, \beta) + o_p(1) = n^{-1/2} \sum_{i=1}^n 2 \int \psi_i(t) d\gamma(t, \beta).$$

Therefore,

$$W(\beta) = \binom{n}{2}^{-1} \sum_{i < j} \{\mathcal{V}_{ij}(\beta) + \phi_{ij}(\beta)\} + o_p(1),$$

where $\mathcal{V}_{ij}(\beta) = (V_{ij}(\beta) + V_{ji}(\beta))/2$,

$$V_{ij}(\beta) = p(\tau)^{-1} G(X_i)^{-2} I_{ij}(\tau) \{I(\beta' Z_i > \beta' Z_j) - C_\tau(\beta)\},$$

and $\phi_{ij}(\beta) = \int \{\psi_i(t) + \psi_j(t)\} d\gamma(t, \beta)$. It then follows from a functional central limit theorem for U -processes that $W(\beta)$ converges weakly to a zero-mean Gaussian process. This, together with the continuity of $C_\tau(\beta)$ and the asymptotic expansion of $n^{1/2}(\hat{\beta} - \beta_0)$ implies that

$$W = W(\beta_0) + \dot{C}_\tau(\beta_0) n^{1/2}(\hat{\beta} - \beta_0) + o_p(1) = \binom{n}{2}^{-1} \sum_{i < j} \mathcal{W}_{ij} + o_p(1) \quad (\text{A2})$$

where $\dot{C}_\tau(\beta) = \partial C_\tau(\beta) / \partial \beta$ and

$$\mathcal{W}_{ij} = \mathcal{V}_{ij}(\beta_0) + \phi_{ij}(\beta_0) + \dot{C}_\tau(\beta_0) A(\beta_0) \{U_i(\beta_0) + U_j(\beta_0)\} / 2.$$

This, together with the standard asymptotic theory of U -statistics, W is asymptotically normal with mean 0 and variance $E(\mathcal{W}_{12} \mathcal{W}_{13})$.

To estimate the variance of W , we utilize a perturbation-resampling method which has been successfully used for handling numerous inference problems in survival analysis [29, 30]. To be specific, let $\{\Xi_i, i = 1, \dots, n\}$ be a set of n iid random variables from a known distribution with mean 1 and variance 1. For a large n , we can approximate W with the conditional distribution (conditional on the data) of

$$\tilde{W} = \binom{n}{2}^{-1} \sum_{i < j} \hat{\mathcal{V}}_{ij}(\hat{\beta}) \Xi_i \Xi_j + n^{1/2} \{\hat{K}(G^*) - \hat{K}(\hat{G})\} + n^{1/2} \{\hat{C}_\tau(\beta^*) - \hat{C}_\tau(\hat{\beta})\}, \quad (\text{A3})$$

where

$$\hat{K}(G) = \hat{p}(\tau)^{-1} \binom{n}{2}^{-1} \sum_{i < j} G(X_i)^{-2} I_{ij}(\tau) \{I(\hat{\beta}' Z_i > \hat{\beta}' Z_j) - \hat{C}_\tau(\hat{\beta})\},$$

$\hat{p}(\tau) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \hat{G}(X_i)^{-2} I_{ij}(\tau)$, and $G^*(\cdot)$ and β^* are the corresponding perturbed versions of $\hat{G}(\cdot)$ and $\hat{\beta}$. Specifically, $G^*(t)$ is generated by

$$G^*(t) = \hat{G}(t) - \hat{G}(t) \binom{n}{2}^{-1} \sum_{i < j} \int_0^t \hat{\pi}_X^{-1}(u) \{d\hat{M}_i(u) + d\hat{M}_j(u)\} \Xi_i \Xi_j / 2$$

where $\hat{\pi}_X(u) = n^{-1} \sum_{i=1}^n I(X_i \geq u)$, $\hat{M}_i(t) = I(X_i \leq t, \Delta_i = 0) - \int_0^t I(X_i \geq u) d\hat{\Lambda}_D(u)$ and $\hat{\Lambda}_D(\cdot)$ is a consistent estimator of the cumulative hazard function for the censoring variable. We generate β^* as

$$\beta^* = \hat{\beta} + \binom{n}{2}^{-1} \sum_{i < j} [\hat{A}(\hat{\beta}) \{U_i(\hat{\beta}) + U_j(\hat{\beta})\} / 2] \Xi_i \Xi_j.$$

Note that only the random quantity in \tilde{W} is $\{\Xi_i, i = 1, \dots, n\}$. The unknown quantities are replaced with their empirical counterparts. The distribution of \tilde{W} (and therefore the distribution of W) can be approximated by generating a large number of realized random samples from $\{\Xi_i, i = 1, \dots, n\}$.

Appendix B: Large sample approximation to W_ξ

From (A2), W_ξ is denoted by $\binom{n}{2}^{-1} \sum_{i < j} (\mathcal{W}_{ij}^{(A)} - \mathcal{W}_{ij}^{(B)}) + o_p(1)$, where $\mathcal{W}_{ij}^{(A)}$ and $\mathcal{W}_{ij}^{(B)}$ correspond to \mathcal{W}_{ij} in (A2) for Model A and Model B, respectively. Thus, W_ξ is asymptotically normal with mean 0 and variance $E\{(\mathcal{W}_{12}^{(A)} - \mathcal{W}_{12}^{(B)})(\mathcal{W}_{13}^{(A)} - \mathcal{W}_{13}^{(B)})\}$. To estimate the variance of W_ξ , let $\tilde{W}^{(A)}$ and $\tilde{W}^{(B)}$ be the approximation to $W^{(A)} = n^{\frac{1}{2}}(\hat{C}_\tau^{(A)} - C_\tau^{(A)})$ and $W^{(B)} = n^{\frac{1}{2}}(\hat{C}_\tau^{(B)} - C_\tau^{(B)})$ by the perturbation described above, respectively. Let $\tilde{W}_k^{(A)}$ and $\tilde{W}_k^{(B)}$ be a pair of realizations obtained with the k th set of random variables $\{\Xi_i, i = 1, \dots, n\}$. Then, the distribution of W_ξ can be numerically approximated by a large number of M realizations of $\tilde{W}_k^{(A)} - \tilde{W}_k^{(B)}$, $k = 1, \dots, M$.

Acknowledgements

The authors are grateful to the associate editor, three referees, and the editor for insightful comments on the article. This work was supported in part by grants R01-GM079330, R01-AI052817, and N01-HC-25195 from the NIH.

References

- Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular risk profiles. *American Heart Journal* 1991; **121**:293–298.
- D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain MR, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham heart study. *Circulation* 2008; **117**(6):743–753. DOI: 10.1161/CIRCULATIONAHA.107.699579.
- Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. *Cancer* 2008; **113**(11):3075–3099. DOI: 10.1002/cncr.23908.
- Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, D'Agostino Sr RB, Kannel WB, Vasan RS. A risk score for predicting near-term incidence of hypertension: the Framingham heart study. *Annals of Internal Medicine* 2008; **148**(2):102–110.
- Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**:387–415.
- Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
- Korn EL, Simon R. Measures of explained variation for survival data. *Statistics in Medicine* 1990; **9**:487–503.
- Hielscher T, Zucknick M, Werft W, Benner A. On the prognostic value of survival models with application to gene expression signatures. *Statistics in Medicine* 2010; **29**:818–829. DOI: 10.1002/sim.3768.
- Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *Journal of the American Medical Association* 1982; **247**:2543–2546.
- Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine* 1984; **3**:143–152. DOI: 10.1002/sim.4780090503.
- Harrell FE, Lee KL, Mark DB. Tutorial in Biostatistics: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387. DOI: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
- Brown BW, Hollander M, Korwar RM. Nonparametric tests of independence for censored data, with applications to heart transplant studies. In *Reliability and Biometry*, Proschan F, Serfling RJ (eds). SIAM: Philadelphia, 1974.
- 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**:2109–2123. DOI: 10.1002/sim.1802.
- Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics* 2005; **61**:92–105. DOI: 10.1111/j.0006-341X.2005.030814.x.
- Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Statistics in Medicine* 2006; **25**:3474–3486. DOI: 10.1002/sim.2299.
- Cai T, Pepe MS, Zheng Y, Lumley T, Jenny NS. The sensitivity and specificity of markers for event times. *Biostatistics* 2006; **72**:182–197. DOI: 10.1093/biostatistics/kxi047.
- Gönen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005; **92**:965–970. DOI: 10.1093/biomet/92.4.965.
- Cox DR. Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**:187–220.
- Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. *Biometrika* 1995; **82**:835–845. DOI: 10.1093/biomet/82.4.835.
- Chang HY, Nuyten DSA, Sneddon JB, Hastie T, Tibshirani R, Sørlied T, Dai H, He YD, van't Veer LJ, Bartelink H, van de Rijn M, Brown PO, van de Vijver MJ. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *PNAS* 2005; **102**:3738–3743.
- van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**:530–536. DOI: 10.1038/415530a.

22. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine* 2002; **347**:1999–2009.
23. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* 2008; **27**:157–172. DOI: 10.1002/sim.2929.
24. Cai T, Tian L, Lloyd-Jones DM, Wei LJ. Evaluating subject-level incremental values of new markers for risk classification rule. *Harvard University Biostatistics Working Paper Series*, 2008. Working Paper 91. <http://www.bepress.com/harvardbiostat/paper91>.
25. Hjort N. On inference in parametric survival data models. *International Statistical Review* 1992; **60**:355–873.
26. Nolan D, Pollard D. *U*-processes: Rates of convergence. *The Annals of Statistics* 1987; **15**:780–799.
27. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data* (2nd edn). Wiley: New York, 2002.
28. Nolan D, Pollard D. Functional limit theorems for *U*-processes. *The Annals of Statistics* 1988; **16**:1291–1298.
29. Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 1993; **80**:557–572. DOI: 10.1093/biomet/80.3.557.
30. Lin DY, Fleming TR, Wei LJ. Confidence bands for survival curves under the proportional hazards model. *Biometrika* 1994; **81**:73–81. DOI: 10.1093/biomet/81.1.73.