Harvard University
Harvard University Biostatistics Working Paper Series

Year 2009 *Paper* 101

On The C-Statistics For Evaluating Overall Adequacy Of Risk Prediction Procedures With Censored Survival Data

Hajime Uno* Tianxi Cai† Michael J. Pencina‡
Ralph B. D'Agostino** L. J. Wei††

*Harvard University, huno@hsph.harvard.edu
†Harvard University, tcai@hsph.harvard.edu
‡Boston University, mpencina@bu.edu
**Boston University, ralph@bu.edu
††Harvard University, wei@hsph.harvard.edu

ON THE C-STATISTICS FOR EVALUATING OVERALL ADEQUACY OF RISK PREDICTION PROCEDURES WITH CENSORED SURVIVAL DATA

Hajime Uno[1,2], Tianxi Cai[2], Michael J. Pencina[3], Ralph B. D'Agostino[4] and L.J. Wei[2]

[1]*Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Boston, MA, USA*

[2]*Department of Biostatistics, Harvard University, Boston, MA, USA*

[3]*Department of Biostatistics, Boston University, Boston, MA, USA*

[4]*Department of Mathematics and Statistics, Boston University, Boston, MA, USA*

SUMMARY

For modern evidence-based medicine, a well thought-out risk scoring system for predicting the timing of 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. When the event time is possibly censored, however, the commonly used C-statistics estimate parameters which 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 this measure. Numerical studies are also conducted to investigate the performance of the new procedure.

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 diseases (CVD), cancer or many other conditions within a certain time period (Anderson et al., 1991; D'Agostino et al., 2008; Shariat et al., 2008; Parikh et al., 2008). 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 using, for example, the estimated AUC, the area under the observed Receiver Operating Characteristics (ROC) curve, which is also referred as a "C-statistic" (Bamber, 1975). 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" (Hanley & McNeil, 1982).

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 the theoretical counterpart of the estimated risk score for the subject with Z. Consider two independent copies $\{(T_1, Z_1, g(Z_1))', (T_2, Z_2, g(Z_2))'\}$ of $(T, Z, g(Z))'$. A commonly used concordance measure is

$$C = \text{pr}(g(Z_1) > g(Z_2) \mid \ T_2 > T_1) \tag{1.1}$$

(Heagerty & Zheng, 2005). When T is subject to right censoring, as discussed in Heagerty &

Zheng (2005) one would typically consider a modified C_τ with a fixed, prespecified follow-up period $(0, \tau)$, where

$$C_\tau = \text{pr}(g(Z_1) > g(Z_2) \mid \ T_2 > T_1, T_1 < \tau). \tag{1.2}$$

Estimation of (1.1) or (1.2) when the event time may be censored, however, is not straightforward (Harrell et al., 1996; Pencina and D'Agostino, 2004; Gönen & Heller, 2005; Chambless & Diao, 2006).

The estimator for C or C_τ proposed by Heagerty & Zheng (2005) is derived under a proportional hazards model. If this parametric working model is not correctly specified, the resulting estimator may be biased. A popular nonparametric C-statistic for estimating C was proposed by Harrell et al. (1996), which was extensively studied by Pencina & D'Agostino (2004). Note that this generalization is a weighted average of the time-dependent AUCs, which corresponds to the "*incident/dynamic*" ROC curve (Heagerty & Zheng, 2005; Cai et al., 2006) with weights depending on the study-specific censoring distribution.

When the study individuals had differential follow-up times, the C-statistic studied by Harrell et al. (1996) converges to an association measure, which depends on the study-specific censoring distribution. In this article, under the general random censorship assumption, we provide a simple, purely non-parametric estimator for (1.2), which is free of the censoring distribution. 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.

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. Also, let $\{(T_i, Z_i, D_i), i = 1, \ldots, n\}$ be n independent copies of $\{(T, Z, D)\}$. For the ith 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 (Cox, 1972):

$$\Lambda_Z(t) = \Lambda_0(t) \exp(\beta' Z), \tag{2.1}$$

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 (2.1) is not correctly specified, under a rather mild non-separable condition that there does not exist vector ζ such that $\text{pr}(T_1 > T_2 \mid \zeta' Z_1 < \zeta' Z_2) = 1$, $\hat{\beta}$ converges to a constant vector, say, β_0, as $n \to \infty$. 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:

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

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

$$C = \text{pr}(\beta_0' Z_1^0 > \beta_0' Z_2^0 \mid \ T_1^0 < T_2^0). \tag{2.2}$$

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 (2.2), that is,

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

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

It follows from an "inverse probability weighting" technique proposed by Cheng et al. (1995) for dealing with a completely different problem in survival analysis that C_τ can be consistently, nonparametrically estimated by

$$\hat{C}_\tau = \frac{\sum_{i=1}^n \sum_{j=1}^n \Delta_i \{\hat{G}(X_i)\}^{-2} I(X_i < X_j, X_i < \tau) I(\hat{\beta} Z_i > \hat{\beta} Z_j)}{\sum_{i=1}^n \sum_{j=1}^n \Delta_i \{\hat{G}(X_i)\}^{-2} I(X_i < X_j, X_i < \tau)}, \tag{2.3}$$

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 $n \to \infty$, the denominator of (2.3) divided by n^2 converges to

$$E\left\{\frac{\Delta_1 I(X_1 < X_2, X_1 < \tau)}{G^2(X_1)}\right\} = E\left[E\left\{\frac{I(T_1 < T_2, T_1 < \tau) I(D_1 \wedge D_2 > T_1)}{G^2(T_1)} \middle| T_1\right\}\right]$$

$$= \text{pr}(T_1 < T_2, T_1 < \tau),$$

and the numerator of (2.3) divided by n^2 converges to $\text{pr}(\beta_0' Z_1^0 > \beta_0' Z_2^0, \ T_1^0 < T_2^0, T_1^0 < \tau)$.

In the Appendix, 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 (5.2) 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 $\hat{C}_\tau$ 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. (1996) is

$$\frac{\sum_{i\neq j} \Delta_i I(X_i < X_j) I(\hat{\beta}Z_i > \hat{\beta}Z_j)}{\sum_{i\neq j} \Delta_i I(X_i < X_j)}, \tag{2.4}$$

which converges to a quantity which involves the study-specific censoring distribution:

$$\mathrm{pr}(\beta_0{}'Z_1^0 > \beta_0{}'Z_2^0 \ \mid \ T_1^0 < T_2^0, T_1^0 \leq D_1^0).$$

Pencina & D'Agostino (2004) formulated C-statistic by allowing various τ. Their C-statistic is

$$\frac{\sum_{i\neq j} \Delta_i I(X_i < X_j, X_i < \tau) I(\hat{\beta}Z_i > \hat{\beta}Z_j)}{\sum_{i\neq j} \Delta_i I(X_i < X_j, X_i < \tau)}, \tag{2.5}$$

which converges to

$$\mathrm{pr}(\beta_0{}'Z_1^0 > \beta_0{}'Z_2^0 \ \mid \ T_1^0 < T_2^0, T_1^0 \leq D_1^0, T_1^0 < \tau).$$

3. NUMERICAL STUDIES

First, we illustrate the new estimation procedure 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 (HCD), systolic blood pressure (SBP) and use of medication for high blood pressure (TxBP). These individuals were then followed until Year 2006. Here, the event time T is the first time that the subject had a cardiovascular disease event (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 the event time T and the censoring D are given in Figure 1. Note that most study subjects were followed more than ten years, but less than 13 years.

We fitted the data with a Cox proportional hazards model (2.1). The resulting risk score $\hat{\beta}' Z^0$ is

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

In Table 1, for various τ, we present point estimates for our C-statistics and their corresponding 0.95 confidence intervals for C_τ. When $\tau = 8, 10, 12$ (years), our results are very similar to those

based on the conventional C-index (2.4) 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 C-statistics (2.5) 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. For 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 (Pencina & D'Agostino, 2004). Note that all the results reported in Table 1 were based on $M = 500$ independent realizations of a random sample with $n = 3087$ from the unit exponential for (5.2).

The second data set for illustration is from a recent cancer study (Chang et al., 2005). This study was designed to evaluate the prognostic value of a new gene signature constructed from the patient's microarray gene expression data to predict the time of the future patient's death or metastasis. The data set consists of 295 breast cancer patient records from the Netherlands Cancer Institute. The details of participants of the study are given in van't Veer et al. (2002) and van de Vijver et al. (2002). One of the clinical implications for establishing a risk score system is to identify future patients who may benefit from adjuvant systemic, but potentially toxic, therapies. For illustrating our proposal, we let T be the patient's survival time. The Kaplan-Meier estimates for the survival functions of the event and the censoring 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. The patient's baseline covariates consist of the new gene score and other conventional markers such as age, number of positive lymph-node, estrogen receptor status, diameter of tumor, histologic grade, vascular invasion, chemotherapy, hormonal therapy, mastectomy or breast

conserving surgery et al. (http://microarray-pubs.stanford.edu/wound_NKI/explore.html).

We first fitted the data with the clinical variables only via the standard stepwise procedure with Cox's proportional hazards model. The covariate vector Z for the final model consists of age, estrogen receptor (ER: positive or negative) and tumor grade (GR: 1, 2, 3). The resulting risk score $\hat{\beta}'Z$ is

$$-0.42 \times (\text{AGE}/10) - 0.74 \times I(\text{ER=positive}) + 1.53 \times I(\text{GR=2}) + 2.01 \times I(\text{GR=3}).$$

In Table 2, with various τ, based on our C-statistic, we report the point estimates and the corresponding 0.95 confidence intervals for C_τ. Our standard error estimates tend to be larger than that from the conventional C-statistic. When $\tau = 15$, our interval is much wider than the conventional one by taking care of the unstable inverse weighting probability estimate $\hat{G}(\cdot)$ at the tail part of the curve. The conventional C-statistic gives a biased estimator for C_τ and its interval estimator may not have the adequate coverage level.

Now, if we fit the data with a Cox model, which includes the gene score variable (GS) and the above three conventional variables, the risk score is

$$2.43 \times (\text{GS}) - 0.56 \times (\text{AGE}/10) - 0.55 \times I(\text{ER=positive}) + 1.25 \times I(\text{GR=2}) + 1.52 \times I(\text{GR=3}).$$

We also report the results in Table 2, which are similar to those without using gene score.

To examine the performance of the new proposal, for example, the interval estimation procedure, we conducted an extensive numerical study under various practical settings. To be specific, we first created a true survival regression model to relate T to Z and also a censoring distribution for D. For example, for one case in our study, we generated the model by fitting the above cancer

data with a Weibull regression survival model that has four baseline covariates: age, estrogen receptor status, tumor grade and gene score. We then fitted the data with a one-sample Weibull distribution (with two unknown parameters) for the censoring distribution $G(\cdot)$. For this case, the resulting censoring is heavy (about 70% censoring by Year 15). To simulate a data set similar to the cancer data, the covariate vector Z with the aforementioned four components was randomly generated from its empirical distribution based on the above observed gene-expression cancer data. For each selected Z, we generated T via the above Weibull regression model and also an independent censoring D from the one-sample Weibull. The resulting data point is $(X, \Delta, Z)'$. With one million of such simulated data points, we obtain the "true" C_{15}, which is 0.75. Next, to examine the adequacy of the coverage probabilities of our 0.95 confidence interval estimator for C_{15}, for a given sample size n, we generated 1000 independent realizations of $\{(X_i, \Delta_i, Z_i)', i = 1, \cdots, n\}$ under the above simulation setting. We then fitted each simulated data set with a *working* Cox regression model (not necessarily the correct model) and obtained estimated risk scores. With this scoring system, we constructed a 0.95 confidence interval for C_{15} in (2.3). We repeated this process and obtained 1000 realized confidence intervals with the nominal level of 0.95. With various working models and $n = 300$, the empirical coverage levels of our interval estimators are around 0.97. With $n = 500$ and various working Cox's models, the empirical levels are about 0.95. We also considered a case when the censoring is relatively light (about 60% censoring at Year 15) by multiplying the above censoring variable by 2.7, the empirical coverage levels are from 0.94 to 0.95. Based on the results of the extensive simulation study, we find that the new interval estimation procedure performs well, that is, the empirical

coverage levels are practically identical to their nominal counterparts.

4. REMARKS

In this article, we show that an over-all performance of predicting the subject-level survival over the entire interval $(0, \tau)$ based on a parametric or semi-parametric model can be evaluated via a simple, unbiased estimation procedure for C_τ. Based on our extensive numerical study, generally we find that the new estimation procedure is robust with respect to the choice of τ. However, if the pre-specified 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 $\hat{C}_\tau$ 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).

When the dimension of the covariate vector Z is greater than one, the estimated risk scoring system may not perform well based on a purely nonparametric function estimation procedure. A feasible alternative is to construct a risk index system via a parametric or semi-parametric model as we did in this article. One may then use this scoring system to group future subjects, but calibrate the corresponding stratum-specific risk estimates via a univariate nonparametric function estimation procedure (Cai et al., 2009). In any event, to obtain an efficient and reliable final product, it is crucial to examine the adequacy of the parametric risk index system. Recently

alternative quantification methods other than using C-statistics for evaluating the performance of risk prediction procedures have been proposed (Pencina et al., 2008; Cai et al., 2008). Further research along this line is warranted.

5. APPENDIX: LARGE SAMPLE PROPERTIES OF $\hat{C}_\tau$

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 limiting 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 $N_i(t) = I(X_i \le t, \Delta_i = 1)$, $Y_i(t) = I(X_i \ge t)$. 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 $\hat{C}_\tau$, we first note that

$$n^{1/2}(\hat{\beta} - \beta_0) = n^{-1/2}\sum_{i=1}^{n} A(\beta_0)U_i(\beta_0) + O_p(n^{-1}) \tag{5.1}$$

where $A(\beta) = -\left\{\partial E(U(\beta))/\partial\beta\right\}^{-1}$ (Hjort, 1992).

Now, for a fixed β, let

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

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 (Nolan & Pollard, 1987), that $\hat{C}_\tau$ converges to $C_\tau(\beta_0)$ in probability as $n \to \infty$. On the other hand, it follows from the asymptotic expansion of $\hat{\beta}$ given in (5.1) that $C_\tau(\beta_0) - C_\tau = O(n^{-1})$. Thus, $\hat{C}_\tau - C_\tau \to 0$ in probability.

To approximate the distribution of

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

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 $\hat{C}_\tau$ of (2.3) 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 (Kalbfleish & Prentice, 2002),

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

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 (Nolan & Pollard, 1988) 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

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

By a uniform law of large numbers for U-processes (Nolan & Pollard, 1987) and the uniform consistency of $\hat{G}(\cdot)$, we have

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

where

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

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

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

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 (Lin et al., 1993; Lin et al., 1994). To be specific, let $\{\Xi_i, i = 1, \ldots n\}$ be a sets of n iid random variable from a known distribution with mean 1 and variance 1. For large n, we can approximate W with the conditional distribution (conditional on the data) of

$$\tilde{W} = \binom{n}{2}^{-1} \sum_{i<j} \hat{\mathcal{V}}_{ij}(\hat{\beta})\Xi_i\Xi_j + n^{1/2}\{\hat{K}(G^*) - \hat{K}(\hat{G})\} + n^{1/2}\{\hat{C}_\tau(\beta^*) - \hat{C}_\tau(\hat{\beta})\}, \qquad (5.2)$$

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 version of $\hat{G}(\cdot)$ and $\hat{\beta}$. Specifically, $G^*(t)$ is generated by

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

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

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

Note that only the random quantity in $\tilde{W}$ is $\{\Xi_i, i = 1, \ldots, n\}$. 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 $\{\Xi_i, i = 1, \ldots n\}$.

References

Anderson, K. M., Odell, P. M., Wilson, P. W. F., & Kannel, W. B. (1991), "General cardiovascular risk profile for use in primary care," *American Heart Journal,* 121, 293–8.

Bamber D. (1975), "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology,* 12, 387-415.

Cai, T., Pepe, M. S., Zheng, Y., Lumley, T. & Jenny, N. S. (2006), "The sensitivity and specificity of markers for event times," *Biostatistics,* 72, 182-97.

Cai, T., Tian, L., Lloyd-Jones, D. M. & Wei, L. J. (2008), "Evaluating subject-level incremental values of new markers for risk classification rule," Harvard University Biostatistics Working Paper Series. Working Paper 91. http://www.bepress.com/harvardbiostat/paper91

Cai, T., Tian, L., Uno, H. Solomon, S. D. & Wei, L. J. (2009), "Calibrating parametric subject-specific risk estimation," Harvard University Biostatistics Working Paper Series. Working Paper 92. http://www.bepress.com/harvardbiostat/paper92,

Chang, H. Y., Nuyten, D. S. A., Sneddon, J. B., Hastie, T., Tibshirani, R., Sørlieb, T., Dai, H., He, Y. D., van't Veer, L. J., Bartelinke, H., van de Rijnj, M., Brownb, P. O. & van de Vijver, M.J. (2005), "Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival," *PNAS,* 102, 3738–43.

Chambless LE & Diao G. (2006), "Estimation of time-dependent area under the ROC curve for long-term risk prediction," *Statistics in Medicine,* 25, 3474-3486.

Cheng, S. C., Wei, L. J. & Ying, Z. (1995), "Analysis of Transformation Models with Censored Data," *Biometrika,* 82, 835–845.

Cox, D. R. (1972), "Regression Models and Life Tables" (with Discussion), *Journal of the Royal Statistical Society,* Ser. B, 34, 187–220.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M. R., Massaro, J. M., & Kannel, W. B. (2008), "General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study," *Circulation,* 117, 743–53.

Gönen, M. & Heller, G. (2005), " Concordance probability and discriminatory power in proportional hazards regression," *Biometrika,* 92, 965-970.

Hanley, J. A. & McNeil, B. J. (1982), "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology,* 143, 29-36.

Harrell, F. E., Lee, K. L., & Mark, D.B. (1996), "Tutorial in Biostatistics: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in Medicine,* 15, 361–87.

Heagerty, P. J. & Zheng, Y. (2005), "Survival Model Predictive Accuracy and ROC Curves," *Biometrics,* 61, 92–105.

Hjort, N. (1992), "On inference in parametric survival data models," *International Statistical Review,* 60, 355–87.

Kalbfleish, J. D. & Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data* (2nd ed.), New York: John Wiley & Sons, Inc.

Lin, D. Y., Wei, L. J. & Ying, Z. (1993), "Checking the Cox model with cumulative sums of martingale-based residuals," *Biometrika,* 80, 557–72.

Lin, D. Y., Fleming, T. R. & Wei, L. J. (1994), "Confidence bands for survival curves under the

proportional hazards model," *Biometrika,* 81, 73–81.

Nolan, D. & Pollard, D. (1987), "U-processes: Rates of convergence," *The Annals of Statistics,* 15, 780–99.

Nolan, D. & Pollard, D. (1988), "Functional Limit Theorems for U-processes," *The Annals of Statistics,* 16, 1291–98.

Parikh, N. I., Pencina, M. J., Wang, T. J., Benjamin, E. J., Lanier, K. J., Levy, D., D'Agostino, R. B. Sr., Kannel, W. B., & Vasan, R. S. (2008), "A risk score for predicting near-term incidence of hypertension: the Framingham Heart Study," *Ann Intern Med,* 148, 102-10.

Pencina, M. J. & D'Agostino, R. B. (2004), "Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation," *Statistics in Medicine,* 23, 2109–23.

Pencina, M. J., D'Agostino, R. B. Sr., D'Agostino, R. B. Jr., & Vasan, R. S. (2008), "Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond," *Statistics in Medicine,* 27, 157–72.

Shariat, S. F., Karakiewicz, P. I., Roehrborn, C. G., & Kattan, M. W. (2008), "An updated catalog of prostate cancer predictive tools," *Cancer,* 113(11), 3075-99.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002), "Gene expression profiling predicts clinical outcome of breast cancer," *Nature,* 415, 530–6.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen,

A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. & Bernards, R. (2002), "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer," *The New England Journal of Medicine,* 347, 1999–2009.

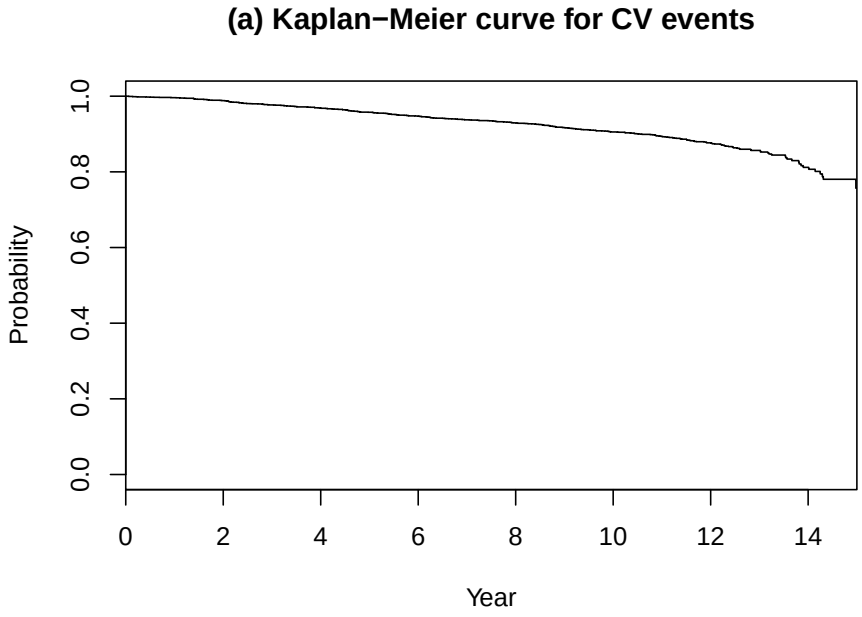
Table 1. *Point estimates (Est), standard error estimates (SE) and 0.95 confidence intervals (CI) for C_τ with Framingham study Data*

C-index	New Method			Conventional		
τ	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 2. *Point estimates (Est), standard error estimates (SE) and 0.95 confidence intervals (CI) for C_τ with breast cancer data*

C-index	New Method			Conventional		
τ	Est	SE	CI	Est	SE	CI
Without Gene Score Model						
6	0.75	0.04	(0.68, 0.82)	0.76	0.03	(0.70, 0.82)
8	0.74	0.03	(0.67, 0.81)	0.76	0.03	(0.70, 0.81)
10	0.74	0.03	(0.68, 0.81)	0.75	0.03	(0.70, 0.81)
15	0.68	0.05	(0.58, 0.78)	0.75	0.03	(0.70, 0.81)
∞	NA	NA	NA	0.75	0.03	(0.70, 0.81)
With Gene Score Model						
6	0.77	0.03	(0.71, 0.84)	0.78	0.03	(0.72, 0.83)
8	0.77	0.03	(0.71, 0.83)	0.78	0.03	(0.72, 0.83)
10	0.77	0.03	(0.71, 0.83)	0.77	0.03	(0.72, 0.82)
15	0.71	0.05	(0.61, 0.81)	0.77	0.03	(0.72, 0.82)
∞	NA	NA	NA	0.77	0.03	(0.72, 0.82)

Figure 1. Estimates for survival functions for CV events and censoring variables with Framingham study data.



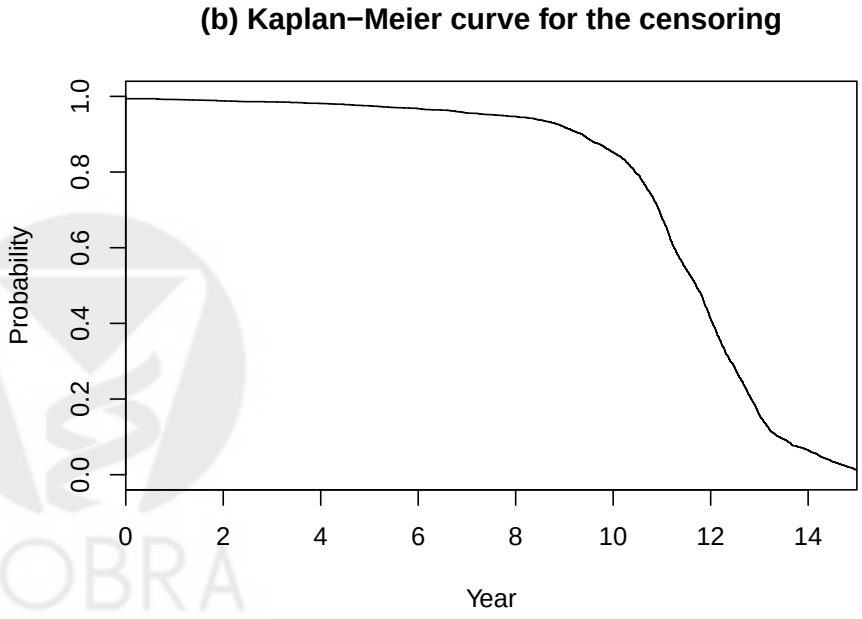


Figure 2. Estimates for survival functions for CV events and censoring variables with breast cancer data.