Published in final edited form as:
Stat Med. 2013 June 30; 32(14): 2430–2442. doi:10.1002/sim.5647.

A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data

Hajime Uno[a,b,*,†], **Lu Tian**[c], **Tianxi Cai**[b], **Isaac S. Kohane**[d], and **L. J. Wei**[b]

[a]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Boston, MA, U.S.A.

[b]Department of Biostatistics, Harvard University, Boston, MA, U.S.A.

[c]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA, U.S.A.

[d]Division of Health Sciences and Technology, Harvard University and Massachusetts Institute of Technology, Cambridge, MA, U.S.A.

Abstract

Risk prediction procedures can be quite useful for the patient's treatment selection, prevention strategy, or disease management in evidence-based medicine. Often, potentially important new predictors are available in addition to the conventional markers. The question is how to quantify the improvement from the new markers for prediction of the patient's risk in order to aid cost–benefit decisions. The standard method, using the area under the receiver operating characteristic curve, to measure the added value may not be sensitive enough to capture incremental improvements from the new markers. Recently, some novel alternatives to area under the receiver operating characteristic curve, such as integrated discrimination improvement and net reclassification improvement, were proposed. In this paper, we consider a class of measures for evaluating the incremental values of new markers, which includes the preceding two as special cases. We present a unified procedure for making inferences about measures in the class with censored event time data. The large sample properties of our procedures are theoretically justified. We illustrate the new proposal with data from a cancer study to evaluate a new gene score for prediction of the patient's survival.

Keywords

area under the receiver operating characteristic curve; *C*-statistic; Cox's regression; integrated discrimination improvement; net reclassification improvement; risk prediction

1. Introduction

Consider the case that the response variable T is the time to a specific event of interest, which is possibly censored. Also let Z be its corresponding vector of baseline covariates or predictors. Suppose that we are interested in predicting the risk $p(Z; t_0) = \mathrm{pr}(T \leq t_0 \mid Z)$, where t_0 is a pre-specified time point. Let $Z_{(1)}$ be a function of Z that consists of the 'conventional' predictor values and $Z_{(2)}$ be a function of Z that contains $Z_{(1)}$ and also new

[*] Correspondence to: Hajime Uno, Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02115, U.S.A.
[†]huno@jimmy.harvard.edu

predictor values. The question is whether a prediction model with $Z_{(2)}$ can improve the predictive ability over a model with $Z_{(1)}$. The next question would be how to quantify the added value from the new markers for cost–benefit purposes over the entire study population.

A commonly used statistical method to answer the first question is to fit the data with a working survival model (e.g., the Cox proportional hazards model) with $Z_{(2)}$ and then utilize statistical significance tests for association of the new markers with the risk to identify important new predictors. Unfortunately, this approach sheds little light on the degree of improvement resulting from new markers. To answer the second question, a popular procedure is to use the improvement in the area under the receiver operating characteristic (ROC) curve (i.e., compute the difference between two areas under the ROC curves based on $Z_{(1)}$ and $Z_{(2)}$) [1–3]. Recently, the time-specific ROC curve methods have been modified to deal with censored event time data [4, 5]. Various C-statistics type summary measures have also been proposed to quantify the overall adequacy of risk prediction [6–8]. However, it has been shown that these metrics are not sensitive enough to capture a meaningful improvement from the new markers over the conventional counterparts [9–11].

Recently, a number of new measures for quantifying the added value from the new markers were proposed [12, 13]. For example, *the integrated discrimination improvement index* (IDI) and *the net classification improvement* (NRI) studied by Pencina *et al.* [13]. These two measures have drawn much attention in medical research, especially in evaluation of markers for cardiovascular disease progression or burden. For example, as of August 2011, more than 200 publications in clinical studies had utilized or cited these new measures (see PubMed.gov: http://www.ncbi.nlm.nih.gov/pubmed), indicating that there is a real need for alternatives to the ROC-related measures. Further insightful comments about these two metrics can be found in [14, 15].

The original estimate for the IDI could not handle cases in the presence of censoring. More recently, Chambless *et al.* [16] and Pencina *et al.* [17] provided modified estimators to account for censoring in survival data. To obtain the corresponding standard error estimates, the standard bootstrap method was utilized. It is not clear, however, that such a resampling method can be justified under a general random censorship model. In this paper, we present a class of measures that quantify the added values from the new markers. This class includes the popular IDI and NRI [17] metrics as special cases. We then apply a novel resampling-perturbation inference procedure to obtain the confidence interval estimates with theoretical justification.

For a random independent subject that is not in the study sample, let $Z = Z^0$, $Z_{(1)}=Z^0_{(1)}$, and $Z_{(2)}=Z^0_{(2)}$ denote its covariate vectors, and let $T = T^0$ denote event time. Assume that we are interested in estimated risk probabilities at a particular time point, t_0. We define subjects who have events by t_0 as *cases* (i.e., $T^0 \leq t_0$) and those who are event-free as *controls* (i.e., $T^0 > t_0$). Let $\hat{p}_2\left(Z^0_{(2)};t_0\right)$ and $\hat{p}_1\left(Z^0_{(1)};t_0\right)$ be two approximations to $p(Z^0; t_0)$ via two *working* survival models. Define $\hat{D}(Z^0;t_0)=\hat{p}_2\left(Z^0_{(2)};t_0\right)-\hat{p}_1\left(Z^0_{(1)};t_0\right)$, which denotes the change in estimated risk score. Heuristically, if $Z^0_{(2)}$ gives a better prediction than $Z^0_{(1)}$, it can be expected that $\hat{D}(Z^0; t_0)$ tends to be positive for a case and negative for a control. The class of measures we consider here is a set of global measures for the 'distance' between these two distributions of $\hat{D}(Z^0; t_0)$. The aforementioned IDI [13], for example, belongs to this class. The limit $M^{(1)}(t_0)$ for this index, as n goes to ∞, is simply the difference of the mean values of $\hat{D}(Z^0; t_0)$ for the cases and controls. That is,

$$M_n^{(1)}(t_0)=E\left\{\hat{\mathrm{D}}(Z^0;t_0)|T^0\le t_0\right\}-E\left\{\hat{\mathrm{D}}(Z^0;t_0)|T^0>t_0\right\}, \quad (1)$$

where the expectation is with respect to (Z^0, T^0). Note that the individual average values for the cases and controls in (1) are also quite informative. Pencina *et al.* [13] and Pepe *et al.* [14] discussed $M_n^{(1)}(t_0)$ extensively and connected it to other quantitative discrimination measures between the cases and controls. Another measure in this class is $M^{(2)}(t_0)$, the limit of

$$M_n^{(2)}(t_0)=\mathrm{pr}\left(\hat{\mathrm{D}}(Z^0;t_0)>0|T^0\le t_0\right)-\mathrm{pr}\left(\hat{\mathrm{D}}(Z^0;t_0)>0|T^0>t_0\right) \quad (2)$$

proposed by Pencina *et al.* [17] as an extension of NRI discussed in [13].

Estimators for $M^{(1)}(t_0)$ and $M^{(2)}(t_0)$ to account for censoring have been proposed by Chambless *et al.* [16] and Pencina *et al.* [17], respectively. In particular, Chambless *et al.* [16] utilized the fact that $M_n^{(1)}(t_0)$ can be written as

$$\frac{\mathrm{Var}\left\{\mathrm{pr}\left(T^0\le t_0|Z^0_{(2)}\right)\right\}-\mathrm{Var}\left\{\mathrm{pr}\left(T^0\le t_0|Z^0_{(1)}\right)\right\}}{\mathrm{pr}(T^0\le t_0)\{1-\mathrm{pr}(T^0\le t_0)\}}, \quad (3)$$

when $\hat{\mathrm{p}}_1\left(Z^0_{(1)};t_0\right)$ and $\hat{\mathrm{p}}_2\left(Z^0_{(2)};t_0\right)$ are well calibrated [14], that is, they are consistent with respect to $\mathrm{pr}\left(T^0\le t_0|Z^0_{(1)}\right)$ and $\mathrm{pr}\left(T^0\le t_0|Z^0_{(2)}\right)$, respectively. Chambless *et al.* [16] then used

$$\frac{\mathrm{Var}\left\{\hat{\mathrm{p}}_2\left(Z^0_{(2)};t_0\right)\right\}}{E\left\{\hat{\mathrm{p}}_2\left(Z^0_{(2)};t_0\right)\right\}\left[1-E\left\{\hat{\mathrm{p}}_2\left(Z^0_{(2)};t_0\right)\right\}\right]}-\frac{\mathrm{Var}\left\{\hat{\mathrm{p}}_1\left(Z^0_{(1)};t_0\right)\right\}}{E\left\{\hat{\mathrm{p}}_1\left(Z^0_{(1)};t_0\right)\right\}\left[1-E\left\{\hat{\mathrm{p}}_1\left(Z^0_{(1)};t_0\right)\right\}\right]}$$

to estimate (3). It is not clear whether this would estimate $M^{(1)}(t_0)$ well when $\hat{\mathrm{p}}_1\left(Z^0_{(1)};t_0\right)$ and $\hat{\mathrm{p}}_2\left(Z^0_{(2)};t_0\right)$ are not well calibrated. Recently, Pencina *et al.* [17] showed that $M_n^{(2)}(t_0)$ can be rewritten as

$$\frac{\mathrm{pr}\{T^0\le t_0)|\hat{\mathrm{D}}(Z^0;t_0)>0\}\mathrm{pr}\left\{\hat{\mathrm{D}}(Z^0;t_0)>0)\right\}}{\mathrm{pr}(T^0\le t_0)}-\frac{\mathrm{pr}\{T^0>t_0)|\hat{\mathrm{D}}(Z^0;t_0)>0\}\mathrm{pr}\left\{\hat{\mathrm{D}}(Z^0;t_0)>0)\right\}}{\mathrm{pr}(T^0>t_0)}. \quad (4)$$

Note that, because of the association between T^0 and Z^0, the probabilities in the preceding Equation (4) should be estimated by estimating conditional probabilities given Z^0 firstly and integrating them with respect to the distribution of Z^0. Unless the covariate vector can be discretized, the implementation may not be easy. Note that Chambless *et al.* [16] and Pencina *et al.* [17] utilized the standard bootstrap method to construct interval estimates for $M^{(1)}(t_0)$ and $M^{(2)}$. It is not clear, however, if such a resampling method can be justified for the current case.

In this article, we consider a class of measures for evaluating the added values of the new markers, which includes the preceding two as special cases. Moreover, we utilize a theoretically justifiable perturbation-resampling method to make inferences about those measures. Specifically, we consider two distribution functions based on the paired difference $\hat{D}(\cdot;\, t_0)$:

$$F_n(s;t_0)=\mathrm{pr}\left\{\hat{\mathrm{D}}(Z^0;t_0) \le s|T^0 \le t_0\right\} \quad (5)$$

and

$$G_n(u;t_0)=\mathrm{pr}\left\{\hat{\mathrm{D}}(Z^0;t_0) \le u|T^0 \le t_0\right\}, \quad (6)$$

where $(s, u) \in [-1, 1] \times [-1, 1]$ and the probabilities are with respect to the data and (T^0, Z^0). A plot of these two functions jointly can be quite informative, with an example shown in Figure 1. If there is no difference between two competing working models, $F_n(\cdot; t_0) \approx G_n(\cdot; t_0)$, and we expect that $F_n(\cdot; t_0) - G_n(\cdot; t_0)$ would be symmetric around 0. The larger the separation between these two curves, the larger the improvement in performance of the new markers with respect to the conventional ones. Any metric that quantifies the distance between these two curves would be a reasonable measure of the added value. Note that $M_n^{(1)}(t_0)$ is the difference between two areas under the curves, and $M_n^{(2)}(t_0)$ is the vertical distance between these two functions evaluated at $s = u = 0$, the distance between the two black dots in Figure 1. Another measure of improvement from the new markers is the difference of two medians of two distributions, $M_n^{(3)}(t_0)$, which is the distance between the two gray dots in Figure 1. In this paper, in the presence of censoring, we propose consistent estimators for the limits of (5) and (6). Furthermore, we show that, under mild regularity conditions, as a process of (s, u), the joint distribution of the standardized estimators for (5) and (6) is asymptotically Gaussian. We then show that this limiting distribution can be approximated easily via a perturbation-resampling method, which is similar to wild bootstrapping [18]. With this approximation, one can then make inferences about any smooth functionals of (5) and (6), for example, constructing confidence interval estimates for $M^{(l)}(t_0)$, $l = 1, 2, 3$. We illustrate the new proposal with the data from a breast cancer study to evaluate the degree of improvement from a new gene-expression score over the conventional clinical markers for predicting the event rates for metastasis or mortality. We examine the performance of the new proposal extensively via a simulation study under various practical settings.

2. Estimating the distribution of the difference between two competing risk scores

Consider the case that the event time T may be censored by a random variable C, which is independent of T and Z. One can observe $X = \min(T, C)$ and a binary indicator function Δ, which is 1 if T is observed. Let $\{(T_i, C_i, Z_i), i = 1, \ldots, n\}$, be n independent copies of (T, C, Z). Let $(X_i, \Delta_i, Z_{(1i)}, Z_{(2i)})$ be the i th counterpart of $(X, \Delta, Z_{(1)}, Z_{(2)})$ in the sample. Also, let $\hat{p}_k(Z_{(k)}; t_0)$ be an estimator for $p(Z; t_0)$ with the data $\{(X_i, \Delta_i, Z_{(ki)}), i = 1, \ldots, n\}$, $k = 1, 2$.

To obtain estimates $\hat{p}_k(Z_{(k)}; t_0)$, $k = 1, 2$, one may use the conventional Cox regression models [19]. Specifically, at time point t, we model the cumulative hazard function $\Lambda(t; Z_{(k)})$ of T given $Z_{(k)}$ as $\Lambda_{k0}(t)\exp(\beta_k' Z_{(k)})$, where $\Lambda_{k0}(\cdot)$ is the underlying cumulative hazard function, and β_k is an unknown vector of parameters, for $k = 1, 2$. It is important to note that these models are unlikely to be correctly specified. Even so, under a mild regularity condition, the standard maximum partial likelihood estimator $\hat{\beta}_k$ for β_k converges to a constant vector, as $n \to \infty$ [20]. This stability feature is essential for developing the large sample properties of estimators for F_n and G_n. Using the standard Breslow estimator $\hat{\Lambda}_{k0}(t)$ for $\Lambda_{k0}(t)$ [21], one may estimate the risk $p(Z^0; t_0)$ by

$$\hat{\mathrm{p}}_k\left(Z^0_{(k)};t_0\right)=1-\exp\left\{-\hat{\Lambda}_{k0}(t_0)\exp\left(\hat{\beta}_k^{\prime}Z^0_{(k)}\right)\right\},k=1,2, \quad (7)$$

where $\hat{\Lambda}_{k0}(t)=\sum_{i=1}^{n}\int_0^t\left\{\sum_{j=1}^{n}Y_j(s)e^{\hat{\beta}_k^{\prime}Z_{(kj)}}\right\}^{-1}dN_i(s)$, $N_i(t) = I(X_i \le t)\Delta_i$, $I(\cdot)$ is the indicator function and $Y_i(t) = I(X_i \ge t)$. The difference $\hat{D}(Z^0; t_0)$ can then be defined accordingly. From the large sample stability property of $\hat{\beta}_k$, it follows that $\hat{D}(\cdot; t_0)$ converges to a finite deterministic function $D(\cdot; t_0)$ as $n \to \infty$ [22].

Now, let F and G denote the limits of F_n and G_n, respectively. To estimate F and G in the presence of censoring, one may use the technique employed by Cheng *et al.* [23]. Specifically, let

$$\hat{\mathrm{F}}(s;t_0)=\frac{\sum_{i=1}^{n}\Delta_i\left\{\hat{H}(X_i)\right\}^{-1}I\left\{\hat{\mathrm{D}}(Z_i;t_0)\le s,X_i\le t_0\right\}}{\sum_{i=1}^{n}\Delta_i\left\{\hat{H}(X_i)\right\}^{-1}I(X_i\le t_0)} \quad (8)$$

and

$$\hat{G}(s;t_0)=\frac{\sum_{i=1}^{n}I\left\{\hat{\mathrm{D}}(Z_i;t_0)\le s,X_i>t_0\right\}}{\sum_{i=1}^{n}I(X_i>t_0)},$$

where $\hat{H}(\cdot)$ is the Kaplan–Meier estimator for the censoring distribution, $H(t) = \mathrm{pr}(C > t)$. The proof of uniform consistency of the preceding estimators is given in the Appendix. Heuristically, the expected value of $n^{-1}\times$ numerator of $\hat{F}(\cdot; t_0)$ is approximately equal to

$$E[\Delta\{H(X)\}^{-1}I\{D(Z;t_0)\le s,X\le t_0\}]=E(E[\Delta\{H(T)\}^{-1}|T,Z]I\{D(Z;t_0)\le s,T\le t_0\}])\approx\mathrm{pr}\{D(Z;t_0)\le s,T\le t_0\}.$$

Similarly, the expected value of the standardized denominator of $\hat{F}$ is approximately equal to $\mathrm{pr}(T \le t_0)$.

To make further inferences about $F(\cdot; t_0)$ and $G(\cdot; t_0)$, or functions thereof, in the Appendix we show that under mild regularity conditions as $n \to \infty$, the joint distribution of $W_F(s; t_0) = n^{1/2}\{\hat{F}(s; t_0) - F(s; t_0)\}$ and $W_G(u; t_0) = n^{1/2}\{\hat{G}(u; t_0) - G(u; t_0)\}$ converges to a mean-zero Gaussian process indexed by $(s, u) \in [-1, 1] \times [-1, 1]$. However, with the conventional method, we cannot estimate well the covariance functions of these limiting processes, which involve the unknown density functions. On the other hand, we can utilize a perturbation-resampling method, which is similar to a wild bootstrapping procedure, to generate independent realizations of a process that has the same distribution as the preceding limiting Gaussian process. Specifically, let (x, δ, z), $\tilde{F}(\cdot; t_0)$, and $\tilde{G}(\cdot\ t_0)$ be the observed values of (X, Δ, Z), $\hat{F}(\cdot; t_0)$, and $\hat{G}(\cdot; t_0)$, respectively. Let $\{V_i, i = 1, \ldots, n\}$ be a random sample from the standard exponential distribution. We can approximate the joint distribution of $W_F(s; t_0)$ and $W_G(u; t_0)$ by using $W_F^*(s;t_0)=n^{1/2}\{F^*(s;t_0)-\tilde{\mathrm{F}}(s;t_0)\}$ and $W_G^*(u;t_0)=n^{1/2}\{G^*(u;t_0)-\tilde{\mathrm{G}}(u;t_0)\}$ where

$$F^*(s;t_0)=\frac{\sum_{i=1}^{n}\Delta_i\{H^*(x_i)\}^{-1}I\{D^*(z_i;t_0)\le s,x_i<t_0\}V_i}{\sum_{i=1}^{n}\delta_i\{H^*(x_i)\}^{-1}I(x_i<t_0)V_i}, \quad (9)$$

$$G^*(u;t_0)=\frac{\sum_{i=1}^{n} I\{D^*(z_i;t_0) \le u, x_i \ge t_0\}V_i}{\sum_{i=1}^{n} I(x_i \ge t_0)V_i}, \quad (10)$$

where $H^*(\cdot)$ and $D^*(\cdot;\ t_0)$ are perturbed counterparts of $\hat{H}(\cdot)$ and $\hat{D}(\cdot;\ t_0)$ by the same set of $\{V_i\}$, respectively. We give the details in the Appendix. We can show that when n is large, we can approximate well the joint unconditional distribution of the process $\{W_F(\cdot;\ t_0),\ W_G(\cdot;\ t_0)\}$ by the conditional distribution of the process $\{W_F^*(\cdot;t_0), W_G^*(\cdot;t_0)\}$ given the data. In practice, we can approximate the distribution of $\{W_F(\cdot;\ t_0),\ W_G(\cdot;\ t_0)\}$ by a large number of realizations from $\{W_F^*(\cdot;t_0), W_G^*(\cdot;t_0)\}$ via realized $\{V_i,\ i = 1, \ldots, n\}$. It is interesting to note that $F^*(\cdot;\ t_0)$ and $G^*(\cdot;\ t_0)$ are non-decreasing functions.

Now, to make inference about a differentiable functional [24], $\mathcal{H}\{F(\cdot;\ t_0),\ G(\cdot;\ t_0)\}$ of $\{W_F(\cdot;\ t_0),\ W_G(\cdot;\ t_0)\}$, we can approximate the distribution of $n^{1/2}[\mathcal{H}\ \{\hat{F}(\cdot;\ t_0),\ \hat{G}(\cdot;\ t_0)\} - \mathcal{H}\{F(\cdot;\ t_0),\ G(\cdot;\ t_0)\}]$ by the distribution, conditional on the data, of $n^{1/2}[\mathcal{H}\ \{F^*(\cdot;\ t_0),\ G^*(\cdot;\ t_0)\} - \mathcal{H}\{\tilde{F}(\cdot;\ t_0),\ \tilde{G}(\cdot;\ t_0)\}]$. One can use this approximation to construct confidence intervals for $M^{(l)}(t_0)$, $l = 1, 2, 3$, via the standard percentile method using the preceding approximation.

3. Numerical studies

3.1. Example

We illustrate the proposed method with the data from a breast cancer study [25–27]. This retrospective study consists of gene-expression data and various conventional clinical markers from 295 women who had fresh-frozen tissues collected at the Netherlands Cancer Institute. The study patients were relatively young (≤52 years) and were diagnosed between 1984 and 1995. These patients were treated by either modified radical mastectomy or breast-conserving surgery, including dissection of the axillary lymph nodes, followed by radiotherapy, if indicated. The patients' follow-up information was extracted from the medical registry of the Netherlands Cancer Institute. For illustrating our new procedure, each patient's gene-expression data are summarized with a single gene signature [27]. Our data set consists of 295 breast cancer patient files. Each file is composed of a patient's clinical outcomes (metastasis/death or censoring time), the gene score, and conventional markers collected at time of surgery, including age, tumor diameter, number of positive lymph nodes, tumor grade, vascular invasion, estrogen receptor status, chemo/hormonal therapy or not, and mastectomy or breast-conserving surgery. The median follow-up time for study patients is 6.7 years, and the range is from 0.05 to 18.3 years. The data are available at http://microarray-pubs.stanford.edu/wound_NKI/explore.html. Note that the gene score proposed by Chang *et al.* [27] is different from the Dutch 70 scoring system [25, 26].

Because the gene score is expensive to obtain, it is important to quantify the incremental value of the gene score over the conventional clinical markers for cost–benefit decisions. Here, T is the time from surgery to either the first metastasis or death. For illustration, we choose 5 and 10 years as t_0 to evaluate the added value of the gene score over the conventional clinical markers for the short-term and long-term predictions. The 5-year and 10-year event-free rates are 72.6% and 61.5%, respectively. Let Z be the vector of all the aforementioned baseline covariate values. Furthermore, let $Z_{(2)} = Z$, and $Z_{(1)}$ be the vector without the gene score. We fit the data with two additive Cox proportional hazards models described in Section 2 with $Z_{(1)}$ and $Z_{(2)}$, respectively. We report the regression coefficient estimates, with the corresponding standard error estimates, in Table I. Although some regression parameters are not statistically significantly different from 0, we include all the covariates in our analysis. For the ith patient with covariate vector Z_i, we obtain a pair of risk scores $\{\hat{p}_1(Z_{(1i)}; t_0), \hat{p}_2(Z_{(2i)}; t_0)\}$ for estimating $p(Z_i; t_0)$ under our two models. A conventional way to evaluate the added value from $Z_{(2)}$ over $Z_{(1)}$ is to compare the corresponding time-specific area under the ROC curve at t_0 [4]. For t_0 = 10 years, the area under the ROC curve is improved from 0.71 to 0.74 by adding the gene score. With the standard bootstrapping method, a 0.95 confidence interval for the difference between these two areas under the curve is (−0.01, 0.06).

Before presenting the results of our new procedure, we show scatter diagrams of $\hat{p}_1(\cdot; t_0)$ versus $\hat{p}_2(\cdot; t_0)$ for t_0 = 5 and 10 years in Figure 2. The dots in (a) and (c) represent the subjects who had events by t_0, and the open circles denote the subjects who were censored before t_0. The dots in (b) and (d) are those who were event-free at t_0. If there are no censored observations, these plots are quite informative to examine the added value of the gene score. If the gene score has very little added value, one would expect that the dots are symmetrically distributed around the 45° line. The plots in Figure 2 show that the dots tend to be above the 45° line for cases but below the line for controls. These suggest that the gene signature may have some nontrivial incremental value.

Because there are quite a few censored observations in the study, one cannot make valid inference on the basis of Figure 2. In Figure 1, we plot the estimated distribution functions $\hat{F}(\cdot; t_0)$, the thin curve, and $\hat{G}(\cdot; t_0)$, the thick curve. Graphically, the gene score appears to provide extra information regarding the prediction of both 5-year and 10-year event rates. The estimate for the integrated discrimination improvement, $\mathcal{M}^{(1)}(t_0)$, is the difference between the areas under the thin and thick curves. With 1000 perturbation samples, as discussed in Section 2, the resulting 0.95 confidence interval for the integrated discrimination improvement, $\mathcal{M}^{(1)}(t_0)$, is (0.01, 0.10) with t_0 = 10 years. Note that we used the unit exponential as the resampler for the perturbation method in the analysis. Moreover, we obtained all the confidence interval estimates via the standard percentile method.

The point estimate and 0.95 confidence interval for $\mathcal{M}^{(2)}(t_0)$, with t_0 = 10 years, are 0.27 and (0.09, 0.43). That is, on average, the improvement from $Z^{(2)}$ over $Z^{(1)}$ is about 27%. Note that these estimates were constructed using the pairing information of $\hat{p}_2\left(Z^0_{(2)};t_0\right)$ and $\hat{p}_1\left(Z^0_{(1)};t_0\right)$, which may be more sensitive than the conventional area under the ROC curve-based method.

For the difference of medians for the two curves in Figure 1(b) (the distance between two gray dots), the point and 0.95 confidence interval for $\mathcal{M}^{(3)}(t_0)$, with t_0 = 10 years, are 0.07 and (0.01, 0.12), respectively. As in other cases, generally the difference of two medians is difficult to estimate well, and the corresponding confidence interval tends to be rather large.

The point estimates for $\mathcal{M}^{(l)}(t_0)$, l = 1, 2, 3, with t_0 = 5 years, are 0.03, 0.27, and 0.03, and the corresponding 0.95 confidence intervals are (0.00, 0.08), (0.03, 0.38), and (0.00, 0.08), respectively. As Figure 1 also shows graphically, the magnitude of improvement with t_0 = 5 years is similar to that with t_0 = 10 years for this breast cancer example.

It is interesting to note that the estimates for $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(3)}$ provide us the average magnitudes of the incremental values of the gene score, which are modest. On the other hand, the improvement based on the estimate for $\mathcal{M}^{(2)}$ is quite impressive. In practice, we recommend to utilize all three measures with Figure 1 to obtain a global assessment of the added value from the new markers.

3.2. Simulation studies

To examine the performance of the new inference procedure in practice, we conducted a simulation study under practical settings. Mimicking the gene-expression breast cancer study, we generated the event and censoring times with the aforementioned clinical markers and gene score. Specifically, we first fitted the breast cancer data with a parametric survival model, for instance, a Weibull or log-normal regression model, to create the true model for each simulation scenario. We repeatedly drew the clinical markers and gene score, Z, from the empirical distribution constructed from the observed Z of the breast cancer study. Then, we generated T from the true model with each Z being drawn. For each regression model for T, we considered three different types of censoring: (a) type I censoring, that is, every subject has the same follow-up time; (b) random censoring that is independent of survival time and covariates; and (c) random censoring that is dependent on covariates but independent of survival time conditional on the covariates. Specifically, for (a), we chose a constant, 15 years, as C; that is, we truncated all subjects at 15 years. For (b), we generated C, for each subject, as the minimum of 15 years and a realization from a Weibull distribution, the parameters of which were derived by fitting the breast cancer data. For (c), using the same set of covariates for generating T, we derived a Weibull regression model for censoring time by fitting the breast cancer data. Then, C was generated, for each subject, as the minimum of 15 years and a realization from the resulting Weibull regression model with a given Z. Note that our inference proposals are valid under (a) or (b) but not generally true under (c). Investigating the performance of the new inference procedure under (c) will thus shed light on the robustness of the proposed methods with respect to the violation of independent censoring assumption.

We numerically obtained the true values for $M^{(1)}$, $M^{(2)}$, and $M^{(3)}$ by aMonte Carlo method with one million data points of (T, C, Z). We used two Cox working models, one with $Z_{(1)}$ and the other with $Z_{(2)}$ regardless of whether the true model was Weibull or log-normal. Note that, for each configuration, we used C only for estimating parameters in the working Cox models and the difference of the risk scores associated with each realization of Z.

Now, for each realized sample (T_i, Z_i, C_i), we fitted these simulated data with two Cox working models. For each simulation data set, we used the resampling method described in Section 2 with 1000 perturbation samples to construct 0.95 confidence intervals for $M^{(l)}$, $l = 1, 2, 3$, via the percentile method. We computed the empirical coverage levels for each case with 1000 realized data sets. In Table II, we present such empirical coverage probabilities with various censoring assumptions, sample sizes, and survival models. The interval estimation procedures are satisfactory, even when the censoring distribution depends on the covariates. For $M^{(2)}$, the coverage levels tend to be higher than their nominal values. This may be due to the fact that the estimator for $M^{(2)}$ involves the indicator functions and is not smooth. We may utilize some smoothing techniques for this case [28], but the choice of a proper smoothing parameter needs further research.

4. Remarks

If there are very few censored observations before t_0, the scatter diagram like Figure 2 is quite informative to evaluate the added value of the new markers. For each subject, one can easily see the incremental value of the risk score with the new markers as well as the corresponding conventional score. For example, in Figure 2, for the subjects who had events, it appears that the addition of the gene score does help when the conventional score is, say, more than 0.4. Unfortunately, however, for the cancer example, the censoring proportion at year 10 is about 40%. Figure 2 by itself is not particularly useful. The distribution function plot in Figure 1 is informative for the contrast of two scoring systems, where censoring can be handled as we propose. On the other hand, it is not clear how to add the information of the conventional score to such plots to explore where the gain would be from the new markers. A specific procedure recently proposed by Uno *et al.* [29] may provide a partial solution to this problem. Further research is needed along this line in the presence of censoring.

For the analysis of the data from the cancer study presented in Section 3, we discretized the event time using 5-year and 10-year cutoff time points to define cases and controls. In practice, the choice of t_0 will depend on the disease of interest.

We use an inverse probability weighting scheme to estimate the two distribution functions (5) and (6). Note that we can also rewrite (5) and (6) in the form similar to (4) so that one can utilize the Kaplan–Meier estimates for the event time to estimate (5) and (6) as Pencina *et al.* [17] did. However, it is not clear whether the resulting estimates for the distribution functions would be monotone non-decreasing.

In this paper, our proposed procedure assumes that the censoring distribution is independent of both the event time and covariates. This assumption on censoring is not unreasonable in a well-executed clinical study, especially when the censoring is primarily due to administrative reasons. When the covariate vector can be discretized, we can modify easily our proposed procedure using stratified Kaplan–Meier estimates for the censoring by replacing $\hat{H}(X_i)$ in (8) with $\hat{H}(X_i \mid Z_i)$. In general, it is difficult, if not impossible, to construct a conditional estimate $\hat{H}(X_i \mid Z_i)$ nonparametrically when the dimension of Z_i is more than 1. Therefore, one needs to assume a parametric model for censoring, with covariates, to estimate $M^{(l)}$, $l = 1, 2, 3$. Any model assumption for censoring is subject to the potential inadequacy of the final evaluation. We may justify the validity of the resulting estimation procedure via a simulation study. On the basis of our numerical study, it appears that our assumption about censoring is not that crucial in practice.

It is important to note that measures in the class we have discussed in this article share the issue on the null behavior of nested models recently investigated by Kerr *et al.* [30] and Demler *et al.* [31]. Heuristically, when nested models are fitted and the regression coefficients associated with all new predictors are zero under the bigger model, $D(Z, \theta_1, \theta_2) \equiv 0$ for all Z and $D(Z; \hat{\theta}_1, \hat{\theta}_2)$ converges to a degenerated distribution with a point mass at 0. The large sample theories given in the Appendix will not be valid for this degenerated case, and tools such as Edgeworth expansion [32] might be needed to derive higher order inference for this case. As Kerr *et al.* [30] or Demler *et al.* [31] suggested, for nested models, the interval estimation should be performed only when the regression coefficients for the added predictors are significantly different from 0, and those measures will not be used for testing against the null.

An R package (survIDINRI) for implementing the new inference procedure is available from the R website (http://cran.r-project.org/web/packages/survIDINRI/index.html).

Appendix A

Let $\theta_k = \left(\log\{\Lambda_{0k}(t_0)\}, \beta_k'\right)'$ be a vector of parameters for $k = 1, 2$. Note that throughout the Appendix, all arguments involving θ_k contain a fixed time point t_0, although t_0 does not explicitly appear in the arguments. Let $p_k(Z_{(k)}; \theta_k) = 1 - e^{-\exp\left\{\left(1, Z_{(k)}'\right)\theta_k\right\}}$ and $D(Z; \theta_1, \theta_2) = p_2(Z_{(2)}; \theta_2) - p_1(Z_{(1)}; \theta_1)$. Suppose that the estimator $\hat{\theta}_k = \left(\log\left\{\hat{\Lambda}_{0k}(t_0)\right\}, \hat{\beta}_k'\right)'$ converges to θ_{k0}, as $n \rightarrow \infty$, and then $\hat{p}_k(Z_{(k)}) = p_k(Z_{(k)}; \hat{\theta}_k)$ and $\hat{D}(Z) = p_2(Z_{(2)}; \hat{\theta}_2) - p_1(Z_{(1)}; \hat{\theta}_1)$. Furthermore, we denote the parameter space for θ_k by B_k, $k = 1, 2$. To derive the asymptotic properties, we assume that B_k is a compact set containing θ_{k0}, and $Z_{(k)}$ has bounded support. We also assume that $D(Z; \theta_1, \theta_2)$ is a continuous random variable with a non-degenerated density function continuous in $\theta_1 \in B_1$ and $\theta_2 \in B_2$.

Firstly, we will show the uniform consistency of $\hat{F}(s)$ and $\hat{G}(u)$. To this end, let

$$\hat{\mathrm{F}}(s,\theta_1,\theta_2)=\frac{\sum_{i=1}^{n}\Delta_i\hat{H}(X_i)^{-1}I\{D(Z_i;\theta_1,\theta_2)\le s, X_i\le t_0\}}{\sum_{i=1}^{n}\Delta_i\hat{H}(X_i)^{-1}I(X_i\le t_0)}.$$

It follows, from the uniform consistency of $\hat{H}(\cdot)$ [21] and a uniform law of large numbers [33], that

$$\sup_{(s,\theta_1,\theta_2)\in[-1,1]\times B_1\times B_2}\left|\hat{\mathrm{F}}(s,\theta_1,\theta_2)-F(s,\theta_1,\theta_2)\right|\to 0,$$

where

$$F(s,\theta_1,\theta_2)=\mathrm{pr}\{D(Z;\theta_1,\theta_2)\le s|T\le t_0\}.$$

Coupled with the convergence of $\hat{\theta}_k \to \theta_{k0}$, this implies that $\hat{F}(s, \hat{\theta}_1, \hat{\theta}_2)$ is uniformly consistent for $F(s, \theta_{10}, \theta_{20}) = F(s)$. We show with the same argument the uniform consistency of $\hat{G}(\cdot)$.

Secondly, to derive the limiting distribution of $W_F(s) = n^{1/2}\{\hat{F}(s) - F(s)\}$, let

$$W_{F_a}(s,\theta_1,\theta_2)=n^{1/2}\left\{\hat{\mathrm{F}}(s,\theta_1,\theta_2)-F(s,\theta_1,\theta_2)\right\}$$

and

$$W_{F_b}(s)=n^{1/2}\left\{F(s,\hat{\theta}_1,\hat{\theta}_2)-F(s)\right\}.$$

Note that

$$W_F(s)=W_{F_a}(s,\hat{\theta}_1,\hat{\theta}_2)+W_{F_b}(s); \quad \text{(A1)}$$

we will first show the stochastic equicontinuity of the process $W_{F_a}(s, \theta_1, \theta_2)$ indexed by s, θ_1, and θ_2. To this end, it is adequate to show that

$$n^{-1/2}\sum_{i=1}^{n}\left[\frac{\Delta_i}{\hat{H}(X_i)}I\{D(Z_i;\theta_1,\theta_2)\le s, X_i\le t_0\}-\mathrm{pr}\{D(Z_i;\theta_1,\theta_2)\le s, T_i\le t_0\}\right] \quad \text{(A2)}$$

is tight. From the standard asymptotic theory for the Kaplan–Meier estimator [21],

$$\frac{\Delta_i}{H(X_i)}=1-\int_0^{\tau}\frac{\mathrm{d}M_i(u)}{H(u)}\mathrm{and}1-\frac{\hat{H}(X_i)}{H(X_i)}=\int_0^{X_i}\frac{\mathrm{d}M(u)}{\pi_X(u)}+o_p(n^{-1/2}),$$

where $\pi_X(t) = \mathrm{pr}(X_i \geq t)$, $M_i(t)=I(X_i \leq t, \Delta_i=0) - \int_0^t I(X_i \geq u)\mathrm{d}\Lambda_C(u)$, $M(t)=\sum_{i=1}^n M_i(t)/n$, and $\Lambda_C(\cdot)$ is the cumulative hazard function for the common censoring variable. Using the aforementioned relationship, we can rewrite (A2) as

$$\begin{aligned} &\mathrm{n}^{-1/2}\sum_{i=1}^n [I\{D(Z_i;\theta_1,\theta_2) \leq s, T_i \leq t_0\} - \mathrm{pr}\{D(Z_i;\theta_1,\theta_2) \leq s, T_i \leq t_0\}] \\ &- n^{-1/2}\sum_{i=1}^n \int_0^\tau \frac{\mathrm{d}M_i(u)}{H(u)}[I\{D(Z_i;\theta_1,\theta_2) \leq s, T_i \leq t_0\} - m(\theta_1,\theta_2,s,u)] \\ &=n^{-1/2}\sum_{i=1}^n \left[I\{D(Z_i;\theta_1,\theta_2) \leq s, T_i \leq t_0\}\left\{1 - \int_0^\tau \frac{\mathrm{d}M_i(u)}{H(u)}\right\} - \mathrm{pr}\,\{D(Z_i;\theta_1,\theta_2) \leq s, T_i \leq t_0\}\right] + n^{-1/2}\sum_{i=1}^n \int_0^\tau m(\theta_1,\theta_2,s,u)\frac{\mathrm{d}M_i(u)}{H(u)}, \end{aligned} \tag{A3}$$

where

$$m(\theta_1,\theta_2,s,u)=\mathrm{pr}\{D(Z;\theta_1,\theta_2) \leq s, T<t_0|T \geq u\}.$$

To prove that (A2) is tight in θ_1, θ_2, and s, one only needs to show that $\mathcal{F} = \{D(z, \theta_1, \theta_2) - s : \theta_1, \theta_2, s\}$ is Donsker as the last term in (A3) only involves a smooth deterministic function in (θ_1, θ_2, s). Because B_k is bounded, it can be covered by $N_k = \mathcal{O}\varepsilon^{-d_k})$ balls centered at $\theta_{k[j]} \in B_k$ with a radius of ε, where $j = 1, \ldots, N_k$, and d_k is the dimension of θ_k, $k = 1, 2$. Coupled with the fact that Z has a bounded support, it implies that for any $\theta_k \in B_k$, one can find $1 \leq j_k \leq N_k$ such that $|\theta'_{k[j_k]}\tilde{z}_k - \theta'_k\tilde{z}_k| \leq C_{1k}\varepsilon$ for a positive constant C_{1k}, where $\tilde{z}_k=(1,z'_k)'$ and $z_k \in$ support of $Z_{(k)}$. Furthermore, we can select $N_3 = \mathcal{O}\varepsilon^{-1})$ points in the interval $[-1, 1]$ such that $-1 = s_1 < s_2 < \cdots < s_{N_3} = 1$ and $s_i - s_{i-1} \leq \varepsilon$. Therefore, for any θ_1, θ_2, and s, we can find j_1, j_2, and j_3, such that $\left|\{D(z;\theta_1,\theta_2) - s\} - \left\{e^{-\exp\left(\theta'_{1[j_1]}\tilde{z}_1\right)} - e^{-\exp\left(\theta'_{2[j_2]}\tilde{z}_2\right)} - s_{j_3}\right\}\right| \leq C_2\varepsilon$. In the following, we will estimate the bracketing number of $\mathcal{F}$. Let

$$l_{ijk}(z)=I\left(e^{-\exp\left(\theta'_{1[i]}\tilde{z}_1\right)} - e^{-\exp\left(\theta'_{2[j]}\tilde{z}_2\right)} - s_k+C_2\varepsilon \leq 0\right)$$

and

$$u_{ijk}(z)=I\left(e^{-\exp\left(\theta'_{1[i]}\tilde{z}_1\right)} - e^{-\exp\left(\theta'_{2[j]}\tilde{z}_2\right)} - s_k - C_2\varepsilon \leq 0\right),$$

where $1 \leq i \leq N_1$, $1 \leq j \leq N_2$, $1 \leq k \leq N_3$. The brackets $[l_{ijk}(z), u_{ijk}(z)]$, $1 \leq i \leq N_1$, $1 \leq j \leq N_2$, $1 \leq k \leq N_3$ covers $\mathcal{F}$ and

$$\mathrm{E}[\{u_{ijk}(Z) - l_{ijk}(Z)\}^2]=\mathrm{pr}\left(\left|e^{-\exp\left\{\left(1,Z'_{(1)}\right)\theta_{1[i]}\right\}} - e^{-\exp\left\{\left(1,Z'_{(2)}\right)\theta_{2[i]}\right\}} - s_k\right|<C_2\varepsilon\right) \leq \sup_{\theta_1,\theta_2,s} \mathrm{pr}(|D(Z,\theta_1,\theta_2) - s| \leq C_2\varepsilon) \leq C_3\varepsilon,$$

as the density function of $D(Z, \theta_1, \theta_2)$ is uniformly bounded. Therefore, the bracketing number of $\mathcal{F}$ is $\mathcal{O}(\varepsilon^{-2(d_1+d_2+1)})$, and thus $\mathcal{F}$ is Donsker. Thus, $W_{Fa}(\cdot, \theta_1, \theta_2)$ is tight, and asymptotically, $W_{Fa}(\cdot, \hat{\theta}_1, \hat{\theta}_2)$ is equivalent to $W_{Fa}(\cdot, \theta_{10}, \theta_{20})$, uniformly in s.

Next, by a Taylor series expansion,

$$W_{Fb}(s)=\dot{F}_{\theta_1}(s,\theta_{10},\theta_{20})n^{1/2}\left(\hat{\theta}_1-\theta_{10}\right)+\dot{F}_{\theta_2}(s,\theta_{10},\theta_{20})n^{1/2}(\hat{\theta}_2-\theta_{20})+o_p(1). \quad \text{(A4)}$$

where $\dot{F}_{\theta_k}=\dfrac{\partial F}{\partial\theta_k}$. Because regardless of model adequacy, the maximum partial likelihood estimator $\hat{\theta}_k$ is a regular estimator, that is,

$$n^{1/2}\left(\hat{\theta}_k-\theta_{k0}\right)=n^{-1/2}\sum_{i=1}^{n}\psi_{ki}+o_p(1)$$

where $\psi_{k1}, \cdots, \psi_{kn}$ are n i.i.d mean-zero random variables. Coupled with (A1), (A3), and (A4),

$$W_F(s)=n^{-1/2}\sum_{i=1}^{n}\pi_F(s,Z_i,X_i,\Delta_i)+o_p(1)$$

where

$$\begin{aligned}&\pi_F(s,Z_i,X_i,\Delta_i)\\&\quad=\dot{F}_{\theta_1}(s,\theta_{10},\theta_{20})\psi_{1i}\\&\quad+\dot{F}_{\theta_2}(s,\theta_{10},\theta_{20})\psi_{2i}\\&\quad+\frac{I\{D(Z_i;\theta_{10},\theta_{20})\le s,T_i\le t_0\}}{\{1-S_T(t_0)\}}\\&\quad-F(s)\\&\quad-\int_0^\tau\frac{\mathrm{d}M_i(u)}{\{1-S_T(t_0)\}H(u)}[I\{D(Z_i;\theta_{10},\theta_{20})\le s,T_i\le t_0\}-m(\theta_{10},\theta_{20},s,u)]-F(s)\Bigg[\frac{I(T_i\le t_0)}{1-S_T(t_0)}-1-\int_0^\tau\frac{\mathrm{d}M_i(u)}{\{1-S_T(t_0)\}H(u)}\{I(T_i\le t_0)\\&\qquad-\mathrm{pr}(T_i\le t_0|T\ge u)\}],\end{aligned}$$

where $S_T(t_0) = \mathrm{pr}(T > t_0)$. Similarly, one may show that

$$W_G(u)=n^{-1/2}\sum_{i=1}^{n}\pi_G(u,Z_i,X_i,\Delta_i)+o_p(1)$$

uniformly in u. Therefore,

$$\begin{pmatrix}W_F(s)\\W_G(u)\end{pmatrix}=n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix}\pi_F(s;Z_i,X_i,\Delta_i)\\\pi_G(u;Z_i,X_i,\Delta_i)\end{pmatrix}+o_p(1).$$

Following the similar arguments as earlier, one may show that the class of functions $\{\pi_F(s;z,x,\delta),\pi_G(u;z,x,\delta)\}'$ indexed by s and u is Donsker, and thus $\{W_F(s),W_G(u)\}'$ converges to a mean-zero- two-dimensional Gaussian process on [−1, 1] × [−1, 1].

The perturbed version of $\hat{H}$ in (7) is given by

$$H^*(t)=\tilde{\mathrm{H}}(t)-\tilde{\mathrm{H}}(t)\sum_{i=1}^{n}V_i\int_0^t\left\{\sum_{j=1}^{n}I(x_j>u)\right\}^{-1}\mathrm{d}\tilde{\mathrm{M}}_i(u),$$

where $\tilde{H}(\cdot)$ is the observed $\hat{H}(\cdot)$, $\tilde{\mathrm{M}}_i(t)=I(x_i\le t,\delta_i=0)-\int_0^t I(x_i>u)\mathrm{d}\tilde{\Lambda}_C(u)$ and $\tilde{\Lambda}_C(\cdot)$ is the observed Nelson–Aalan estimator of the cumulative hazard function for the censoring variable C. $D^*(\cdot)$ in (9) and (10) is given by

$$D^*(z)=p_2^*(z_{(2)})-p_1^*(z_{(1)})=\exp\{\Lambda_1^*(t_0)\exp(\beta_1^{*\prime}z_{(1)})\}-\exp\{\Lambda_2^*(t_0)\exp(\beta_2^{*\prime}z_{(2)})\},$$

where β_k^* and $\Lambda_k^*(t_0)$ are the same ones as given in [22]; that is, $\beta_k^*-\tilde{\beta}_k$ and $\log\{\Lambda_{k0}^*(t)\}-\log\{\tilde{\Lambda}_{k0}(t)\}$ are

$$\tilde{A}_k^{-1}\sum_{i=1}^{n}\delta_i\left[(V_i-1)\left\{z_{(ki)}-\frac{\tilde{\mathrm{S}}_k^{(1)}(x_i,\tilde{\beta}_k)}{\tilde{\mathrm{S}}_k^{(0)}(x_i,\tilde{\beta}_k)}\right\}-\frac{n^{-1}\sum_{j=1}^{n}(V_j-1)I(x_j\ge x_i)e^{\hat{\beta}_k^{\prime}z_{(kj)}}\left\{\tilde{\mathrm{S}}_k^{(0)}(x_i,\tilde{\beta}_k)z_{(kj)}-\tilde{\mathrm{S}}_k^{(1)}(x_i,\hat{\beta}_k)\right\}}{\tilde{\mathrm{S}}_k^{(0)}\left(x_i,\tilde{\beta}_k\right)^2}\right],$$

and

$$\frac{n^{-1}}{\tilde{\Lambda}_{k0}(t)}\sum_{i=1}^{n}I(x_i\le t)\delta_i\left\{\frac{(V_i-1)}{\tilde{\mathrm{S}}_k^{(0)}(x_i,\tilde{\beta}_k)}-\frac{n^{-1}\sum_{j=1}^{n}(V_j-1)I(x_j\ge x_i)e^{\tilde{\beta}_k^{\prime}z_{(kj)}}+\tilde{\mathrm{S}}_k^{(1)}(x_i,\tilde{\beta}_K)^{\prime}\left(\beta_k^*-\tilde{\beta}_k\right)}{\tilde{\mathrm{S}}_k^{(0)}\left(x_i,\tilde{\beta}_k\right)^2}\right\},$$

respectively, where $\tilde{\beta}_k$ is the observed $\hat{\beta}_k$, $\tilde{\Lambda}_{k0}(t)$ is the observed $\hat{\Lambda}_{k0}(t)$, $\tilde{\mathrm{S}}_k^{(m)}(t,\beta_k)=n^{-1}\sum_{i=1}^{n}I(x_i\ge t)e^{\beta_k z_{(ki)}}z_{(ki)}^{\otimes m}$,

$$\tilde{A}_k=\int\left[\frac{\tilde{\mathrm{S}}_k^{(2)}(t,\tilde{\beta}_k)}{\tilde{\mathrm{S}}_k^{(0)}(t,\tilde{\beta}_k)}-\left\{\frac{\tilde{\mathrm{S}}_k^{(1)}(t,\tilde{\beta}_k)}{\tilde{\mathrm{S}}_k^{(0)}(t,\tilde{\beta}_k)}\right\}^{\otimes 2}\right]\tilde{\mathrm{S}}_k^{(0)}(t,\tilde{\beta}_k)\mathrm{d}\tilde{\Lambda}_{k0}(t)$$

and for any vector x, $x^{\otimes 0}=1$, $x^{\otimes 1}=x$, $x^{\otimes 2}=x^{\prime}x$.

Now, let $\theta_k^*=(\log\{\Lambda_{0k}^*(t_0)\},\beta_k^{*\prime}$ and $\tilde{\theta}_{k0}$ be the observed $\hat{\theta}_{k0}$; it can be shown that $n^{1/2}\left(\theta_k^*-\tilde{\theta}_k\right)$ conditional on data and $n^{1/2}(\hat{\theta}_k-\theta_{k0})$ converges to the same limiting normal distribution [22]. Furthermore, using similar expressions given as (A1), (A3), and (A4), it is also straightforward to show that $\{W_F^*(s),W_G^*(u)\}^{\prime}$ can be approximated by $n^{-1/2}\sum_{i=1}^{n}\{\tilde{\pi}_F(s;z_i,x_i,\delta_i),\tilde{\pi}_G(u;z_i,x_i,\delta_i)\}^{\prime}(V_i-1)$, where $\tilde{\pi}_F(s;z,x,\delta)$ and $\tilde{\pi}_G(u;z,x,\delta)$ are observed counterparts of $\pi_F(s;z,d,\delta)$ and $\pi_G(u;z,d,\delta)$, respectively. Therefore, by functional delta method, we can approximate the distribution of $W_H=n^{1/2}[\mathscr{H}\{\hat{F}(\cdot),\hat{G}(\cdot)\}-\mathscr{H}\{F(\cdot),G(\cdot)\}]$ by that of $W_H^*=n^{1/2}[\mathscr{H}\{F^*(\cdot),G^*(\cdot)\}-\mathscr{H}\{\tilde{\mathrm{F}}(\cdot),\tilde{\mathrm{G}}(\cdot)\}]$ conditional on the observed data in the sense that $\mathrm{pr}\left\{\left|W_H^*-W_H\right|\ge\varepsilon|(Z_i,X_i,\Delta_i),i=1,\ldots,n\right\}$ converges to 0 in probability for any $\varepsilon>0$.

References

1. 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.
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143(1):29–36. [PubMed: 7063747]
3. D'Agostino, RB.; Griffith, JL.; Schmidt, CH.; Terrin, N. Proceedings of the Biometrics Section. Alexandria, VA, U.S.A: American Statistical Association; 1997. Measures for evaluating model performance; p. 253-258.
4. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics. 2005; 61(1):92–105. [PubMed: 15737082]
5. Cai T, Cheng S. Robust combination of multiple diagnostic tests for classifying censored event times. Biostatistics. 2008; 9(2):216–233. [PubMed: 18056687]
6. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996; 15(4):361–387. [PubMed: 8668867]
7. Pencina MJ, D'Agostino RB. Overall *C* as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in Medicine. 2004; 23(13): 2109–2123. [PubMed: 15211606]
8. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the *C*-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine. 2011; 30:1105–1116. [PubMed: 21484848]
9. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. American Journal of Epidemiology. 2004; 159(9):882–890. [PubMed: 15105181]
10. Greenland P, O'Malley PG. When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. Archives of Internal Medicine. 2005; 165(21):2454–2456. [PubMed: 16314539]
11. Ware JH. The limitations of risk factors as prognostic tools. The New England Journal of Medicine. 2006; 355(25):2615–2617. [PubMed: 17182986]
12. Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. Annals of Internal Medicine. 2006; 145:21–29. [PubMed: 16818925]
13. Pencina M, D'Agostino RB Sr, D'Agostino RB Jr, Vasan R. Comments on 'integrated discrimination and net reclassification improvements—practical advice'. Statistics in Medicine. 2008; 27(2):207–212.
14. Pepe MS, Feng Z, Gu JW. Comments on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond' by Pencina et al. Statistics in Medicine. 2008; 27:173–181. [PubMed: 17671958]
15. Cook N, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Annals of Internal Medicine. 2009; 150:795–802. [PubMed: 19487714]
16. Chambless LE, Cummiskey CP, Cui G. Several methods to assess improvement in risk prediction models: extension to survival analysis. Statistics in Medicine. 2011; 30(1):22–38. [PubMed: 20827726]
17. Pencina M, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statistics in Medicine. 2011; 30(1):11–21. [PubMed: 21204120]
18. Wu C. Jackknife, bootstrap and other resampling methods in regression analysis. The Annals of Statistics. 1986; 14(4):1261–1295.
19. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological). 1972; 34(2):187–220.
20. Hjort N. On inference in parametric survival data models. International Statistical Review / Revue Internationale de Statistique. 1992; 60(3):355–387.

21. Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. 2nd ed.. New York: John Wiley & Sons, Inc.; 2002.

22. Cai T, Tian L, Uno H, Solomon SD, Wei LJ. Calibrating parametric subject-specific risk estimation. Biometrika. 2010; 97(2):389–404. [PubMed: 23049123]

23. Cheng S, Wei LJ, Ying Z. Analysis of transformation models with censored data. Biometrika. 1995; 82(4):835–845.

24. van der Vaart, A. Asymptotic Statistics. Cambridge: Cambridge University Press; 2000.

25. 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(25): 1999–2009. [PubMed: 12490681]

26. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, 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(6871):530–536. [PubMed: 11823860]

27. Chang HY, Nuyten DSA, Sneddon JB, Hastie T, Tibshirani R, Sørlie 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. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(10): 3738–3743. [PubMed: 15701700]

28. van der Vaart A. Weak convergence of smoothed empirical processes. Scandinavian Journal of Statistics. 1994; 21(4):501–504.

29. Uno H, Cai T, Tian L, Wei LJ. Graphical procedure for evaluating overall and subject-specific incremental values from new predictors with censored event time data. Biometrics. 2011; 67:1389–1395. [PubMed: 21504421]

30. Kerr KF, McClelladm RL, Brown ER, Lumley T. Evaluating the incremental value of new biomarkers with integrated discrimination improvement. American Journal of Epidemiology. 2011; 174(3):364–374. [PubMed: 21673124]

31. Demler OV, Pencina MJ, D'gostino RB. Misuse of DeLong test to compare AUCs for nested models. Statistics in Medicine. 2012; 31(23):2577–2587. [PubMed: 22415937]

32. Hall, P. The Bootstrap and Edgeworth Expansion. New York: Springer-Verlag; 1997.

33. Pollard, D. Empirical Processes: Theory and Applications. Hayward, CA: Institute of Mathematical Statistics; 1990.

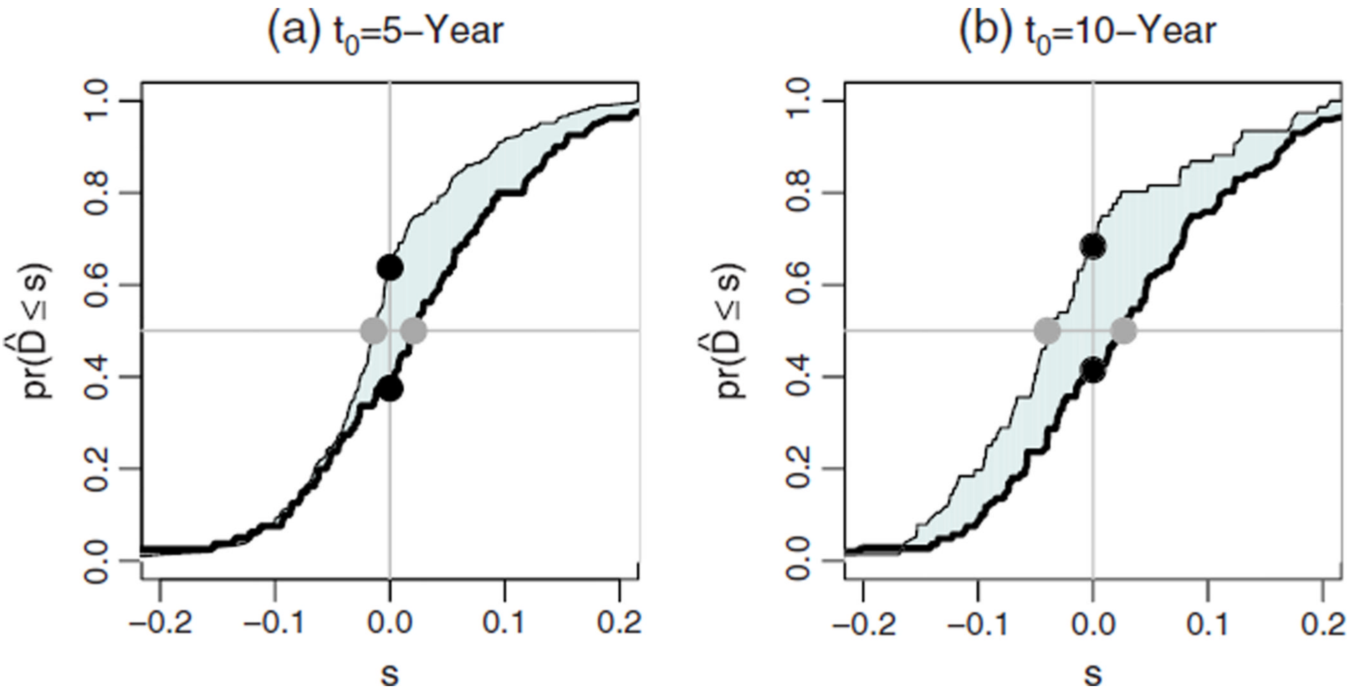

Figure 1.

Empirical distribution function of $\hat{D}$ for $T^0 \le t_0$ (thick solid line) and $T^0 > t_0$ (thin solid line). The difference between areas under two curves is $M_n^{(1)}(t_0)$, and the distances between two black dots and between two gray dots are $M_n^{(2)}(t_0)$ and $M_n^{(3)}(t_0)$, respectively.

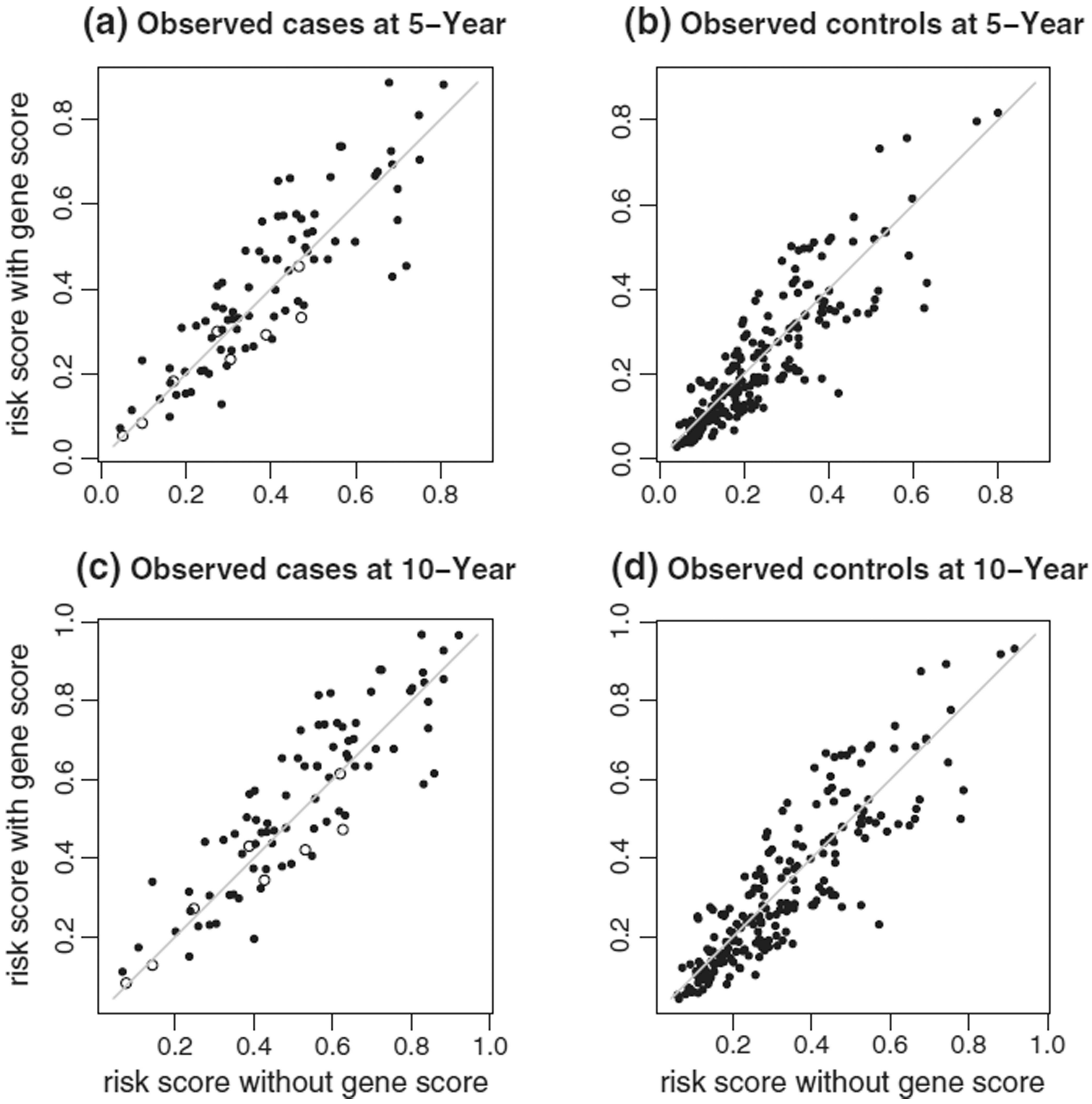

Figure 2.

Scatter diagram of $\hat{p}_1$ (*x*-axis) versus $\hat{p}_2$ (*y*-axis) with the breast cancer data: (a) subjects with event (black dot) and censored (white circle) by 5 years; (b) subjects without event by 5 years; (c) subjects with event (black dot) and censored (white circle) by 10 years; and (d) subjects without event by10 years.

Table I

Estimates of regression parameters for Cox's models with breast cancer data.

	Model without gene score			Model with gene score		
	Est.[1]	SE[2]	*p*[3]	Est.	SE	*p*
Age/10 (years)	−0.47	0.17	0.01	−0.57	0.18	0.00
Diameter of tumor (cm)	0.19	0.11	0.10	0.18	0.12	0.12
Lymph nodes	0.00	0.08	0.98	−0.01	0.08	0.90
Grade = 2 vs. 1	1.00	0.35	0.00	0.74	0.35	0.04
Grade = 3 vs. 1	1.11	0.35	0.00	0.66	0.37	0.08
Vascular invasion 1–3 vs. 0	0.08	0.37	0.83	−0.10	0.37	0.78
Vascular invasion > 3 vs. 0	0.81	0.62	0.19	0.64	0.63	0.30
Estrogen status=positive	−0.39	0.23	0.09	−0.16	0.24	0.51
Chemo or hormonal=Yes	−0.54	0.33	0.11	−0.49	0.33	0.14
Mastectomy=Yes	0.13	0.21	0.54	0.21	0.22	0.34
Gene score	–	–	–	2.43	0.67	0.00

[1] Estimate

[2] Standard

[3] *p*-value

Table II

Empirical coverage probabilities of 0.95 confidence intervals for $M^{(1)}(10)$, $M^{(2)}(10)$ and $M^{(3)}(10)$ based on 1000 iterations.

True model	Censoring	N	$M^{(1)}(10)$	$M^{(2)}(10)$	$M^{(3)}(10)$
Weibull	(a) Type I	200	0.935	0.975	0.934
		300	0.941	0.976	0.949
	(b) Independent	200	0.938	0.988	0.935
		300	0.943	0.985	0.945
	(c) Cond. independent	200	0.945	0.988	0.955
		300	0.941	0.990	0.971
Log-normal	(a) Type I	200	0.939	0.988	0.954
		300	0.934	0.973	0.950
	(b) Independent	200	0.938	0.996	0.946
		300	0.924	0.983	0.950
	(c) Cond. independent	200	0.943	0.992	0.969
		300	0.920	0.982	0.955