Harvard University
Harvard University Biostatistics Working Paper Series

Year 2009 *Paper* 107

Comparing Risk Scoring Systems Beyond the ROC Paradigm in Survival Analysis

Hajime Uno* Lu Tian† Tianxi Cai‡
Isaac S. Kohane** L. J. Wei††

*Dana Farber Cancer Institute and Harvard University, huno@jimmy.harvard.edu

†Stanford University School of Medicine, lutian@stanford.edu

‡Harvard University, tcai@hsph.harvard.edu

**Harvard University and Massachusetts Institute of Technology, isaac.kohane@tch.harvard.edu

††Harvard University, wei@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper107

COMPARING RISK SCORING SYSTEMS BEYOND THE ROC PARADIGM IN SURVIVAL ANALYSIS

Hajime Uno[1,2], Lu Tian[3], Tianxi Cai[2], Isaac S. Kohane[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 Health Research and Policy, Stanford University School of Medicine, Stanford, CA, USA*

[4]*Division of Health Sciences and Technology, Harvard University and Massachusetts Institute of Technology, Cambridge, MA, USA*

SUMMARY

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 on top of the conventional markers. The question is how to quantify the improvement from the new markers for prediction of the patient's risk for cost-benefit decisions. The standard method using the area under the receiver operating characteristic curve (AUROC) to measure the added value may not be sensitive enough to capture incremental improvements from the new markers. In this article, we address this issue for the case that the response variable is the time, possibly censored, to a specific event of interest. We present graphical and numerical methods for evaluating the predictive ability of the new markers. Our proposal includes most of the recent procedures in the literature as special cases for alternatives to the AUROC-based methods. The new inference procedures are theoretically justified and illustrated with data from a cancer study to evaluate a new gene score for the prediction of patient's survival.

Keywords: Area under the receiver operating characteristic curve; C-statistic; Cox's regression; Gaussian process; Integrated discrimination improvement; Improvement in the area under the curve; 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 $p(Z) = \text{pr}(T \leq t_0 \mid Z)$, where t_0 is a pre-specified time point. Let $Z_{(1)}$ be a function of Z, which consists of the "conventional" predictor values and $Z_{(2)}$ be a function of Z, which contains $Z_{(1)}$, but also new 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 decisions.

A commonly used statistical method to answer the first question is to fit the data with a "working" survival model, for example, 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 from new markers. To answer the second question, a popular procedure is to use the improvement in the area under the receiver operating characteristic curve (AUROC), that is, compute the difference between two AUROC's based on $Z_{(1)}$ and $Z_{(2)}$ (D'Agostino et al., 1997; Bamber, 1975; Hanley & McNeil, 1982). Recently the time-specific AUROC methods have been modified to deal with the censored event time data (Heagerty and Zheng, 2005; Cai and Cheng, 2008; Uno et al, 2009). The resulting summary measures are called C-statistics (Harrell et al., 1996; Pencina & D'Agostino, 2004). 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 (Pepe et al., 2004; Greenland & O'Malley, 2005; Ware, 2006). One possible reason is that the difference of two AUROC's does not leverage the pairing information

of two risk scores within each study subject.

Recently, a number of new measures to quantify incremental values have been proposed (Cook et al., 2006; Pencina et al., 2008). For the case in which there are no prespecified or well-defined risk categories, Pencina et al. (2008) introduced the "integrated discrimination improvement" (IDI) index and an "improvement in the AUROC" (IAUC) as new criteria for evaluating the added value of new markers. Specifically, for a random independent subject, which 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. With the censored event time data, let $\hat{p}_2(Z^0_{(2)})$ and $\hat{p}_1(Z^0_{(1)})$ be two estimates for $p(Z^0)$, for example, via two survival working models. Define $\hat{D}(Z^0) = \hat{p}_2(Z^0_{(2)}) - \hat{p}_1(Z^0_{(1)})$. Then, the IDI index is the limit of

$$E\{\hat{D}(Z^0) \mid T^0 \le t_0\} - E\{\hat{D}(Z^0) \mid T^0 > t_0\}, \tag{1.1}$$

as the sample size goes to infinity. The IAUC is the limit of

$$\mathrm{pr}(\hat{D}(Z^0) \ge 0 \mid T^0 \le t_0) - \mathrm{pr}(\hat{D}(Z^0) \ge 0 \mid T^0 > t_0). \tag{1.2}$$

Pepe et al. (2008) discussed the IDI extensively and connected it to other interesting discrimination measures between cases and controls. Note that (1.1) is composed of two marginal differences. The first one is the difference of two marginal means, $\hat{p}_2(\cdot)$ and $\hat{p}_1(\cdot)$, for those who will develop the event. The second part of (1.1) is the difference of these two marginal means for those subjects who will not have the event. The pairing information between $\hat{p}_2(\cdot)$ and $\hat{p}_1(\cdot)$ from the same subject is not utilized to compute the IDI index. On the other hand, (1.2) does take advantage of such pairing information. Under the usual random censorship model in survival analysis, the large sample properties of the existing estimators for the IDI index and IAUC have

not been studied in the literature for inferences. The standard bootstrapping method may not work due to the fact that the estimation procedures are not smooth.

In this article, we generalize the above discrimination measures by considering two distribution functions based on the *paired* difference $\hat{D}(\cdot)$. The first one is

$$F_n(s) = \text{pr}(\hat{D}(Z^0) \leq s|\ T^0 \leq t_0\}, \tag{1.3}$$

and the second one is

$$G_n(u) = \text{pr}(\hat{D}(Z^0) \leq u|\ T^0 > t_0\}, \tag{1.4}$$

where $(s, u) \in [-1, 1] \times [-1, 1]$ and the probabilities are with respect to the data and (T^0, Z^0). If we know (1.3) and (1.4), a plot of these two functions jointly can be quite informative as shown in Figure 1 as an example. If there is no difference between two competing working models, $F_n(\cdot) \approx G_n(\cdot)$, and thus we expect that $F_n(\cdot) - G_n(\cdot)$ 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 older ones. Any metric which quantifies the distance between these two curves would be a reasonable measure of the added value. The IDI index is simply the area between these two curves and the IAUC is the vertical distance between these two functions evaluated at $s = u = 0$ (the distance between two gray dots in Figure 1). In this paper, in the presence of censoring, we proposed consistent estimators for the limits of (1.3) and (1.4). Furthermore, we show that as a process of (s, u), the joint distribution of the standardized estimators for (1.3) and (1.4) 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 (Wu, 1986). With this approximation, one can then make inferences about any "smooth" function of (1.3) and (1.4), for example, confidence interval estimates for the IDI

and IAUC. We also derive inference procedures for other distance metrics between the above two curves, for example, the difference of two medians from (1.3) and (1.4) (the horizontal distance between two black dots in Figure 1). Lastly, 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 metastasis or mortality.

2. ESTIMATING THE DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO COMPETING RISK SCORES

Consider a general case that the event time T may not be observed completely. Let C be the censoring variable, which is independent of T and Z. One can observe $X = \min(T, C)$ and a binary indicator function Δ, which is one if T is observed. Let $\{(T_i, C_i, Z_i)\}, i = 1, \cdots, n$, be n independent copies of (T, C, Z). Let $(X_i, \Delta_i, Z_{(1i)}, Z_{(2i)})$ be the ith counterpart of $(X, \Delta, Z_{(1)}, Z_{(2)})$, in the sample. Also, let $\hat{p}_k(Z_{(k)})$ be an estimator for $p(Z)$ with the data $\{(X_i, \Delta_i, Z_{(ki)}), i = 1, \cdots, n\}, k = 1, 2$.

To obtain estimates $\hat{p}_k(Z_{(k)}), k = 1, 2$, one may use the conventional Cox regression models (Cox, 1972). 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 most likely these models are not correctly specified. On the other hand, under a mild regularity condition, the standard maximum partial likelihood estimator $\hat{\beta}_k$ for β_k converges to a constant vector, as $n \to \infty$ (Hjort, 1992). 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 $\Lambda_{k0}(t)$ (Kalbfleish

& Prentice, 2002), one may estimate the risk $p(Z^0)$ by

$$\hat{p}_k(Z^0_{(k)}) = 1 - \exp\{\hat{\Lambda}_{k0}(t_0)\exp(\hat{\beta}'_k Z^0_{(k)})\}, k = 1, 2, \tag{2.7}$$

where $\hat{\Lambda}_{k0}(t) = \sum_{i=1}^n \int_0^t \left\{\sum_{j=1}^n Y_j(s) e^{\hat{\beta}'_k Z_{(kj)}}\right\}^{-1} dN_i(s)$, $N_i(t) = I(X_i \le t)\Delta_i$, $I(\cdot)$ is the indicator function and $Y_i(t) = I(X_i \ge t)$. The difference $\hat{D}(Z^0)$ can then be defined accordingly. From the large sample stability property of $\hat{\beta}_k$, it follows that $\hat{D}(\cdot)$ converges to a finite deterministic function $D(\cdot)$, as $n \to \infty$. Also, let the limits of F_n and G_n be denoted by F and G, respectively.

To estimate F and G in the presence of censoring, one may use the technique employed by Chen et al. (1995). Specifically, let

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

and

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

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 above estimators is given in the Appendix. Heuristically, the expected value of $n^{-1}\times$ numerator of $\hat{F}(\cdot)$ is approximately equal to

$$E[\Delta_1 \{H(X_1)\}^{-1} I\{D(Z_1) \le s, X_1 \le t_0\}]$$

$$= E[\Delta_1 \{H(T_1)\}^{-1} I\{D(Z_1) \le s, T_1 \le t_0 \mid T_1, Z_1\}] \approx \mathrm{pr}\{D(Z_1) \le s, T_1 \le t_0\}.$$

Similarly, the expected value of the standardized denominator of $\hat{F}$ is approximately equal to $\mathrm{pr}(T_1 \le t_0)$.

To make further inferences about $F(\cdot)$ and $G(\cdot)$ or functions thereof, in the Appendix we show that as $n \to \infty$, the joint distribution of $W_F(s) = \sqrt{n}\{\hat{F}(s) - F(s)\}$ and $W_G(u) =$

$\sqrt{n}\{\hat{G}(u)-G(u)\}$ converges to a mean-zero Gaussian process indexed by $(s,u)\in[-1,1]\times[-1,1]$. However, with the conventional method, the covariance functions of these limiting processes, which involves the unknown density functions, cannot be estimated well. On the other hand, a perturbation-resampling method, which is similar to a "wild" bootstrapping procedure, can be utilized to generate independent realizations of a process which has the same distribution of the above limiting Gaussian process. Specifically, let (x,δ,z), $\tilde{F}(\cdot)$ and $\tilde{G}(\cdot)$ be the observed value of (X,Δ,Z), $\hat{F}(\cdot)$ and $\hat{G}(\cdot)$. Let $\{V_i,\ i=1,\ldots,n\}$, be a random sample from the standard exponential distribution. let $W_F^*(s)=n^{1/2}\{F^*(s)-\tilde{F}(s)\}$ and $W_G^*(u)=n^{1/2}\{G^*(u)-\tilde{G}(u)\}$ where

$$F^*(s)=\frac{\sum_{i=1}^n \delta_i\{H^*(x_i)\}^{-1}I\{D^*(z_i)\le s,\ x_i<t_0\}V_i}{\sum_{i=1}^n \delta_i\{H^*(x_i)\}^{-1}I(x_i<t_0)V_i}, \tag{2.8}$$

$$G^*(u)=\frac{\sum_{i=1}^n I\{D^*(z_i)\le u,\ x_i\ge t_0\}V_i}{\sum_{i=1}^n I(x_i\ge t_0)V_i}, \tag{2.9}$$

where $H^*(\cdot)$ and $D^*(\cdot)$ are perturbed counterparts of $\hat{H}(\cdot)$ and $\hat{D}(\cdot)$ by the same set of $\{V_i\}$, respectively. The details are given in the Appendix. It can be shown that when n is large, the joint unconditional distribution of the process $\{W_F(\cdot),W_G(\cdot)\}$ can be approximately well with that of the process $\{W_F^*(\cdot),W_G^*(\cdot)\}$. In practice, the distribution of $\{W_F(\cdot),W_G(\cdot)\}$ can be approximated by a large number of realizations from $\{W_F^*(\cdot),W_G^*(\cdot)\}$ via realized $\{V_i,i=1,\cdots,n\}$. It is interesting to note that $F^*(\cdot)$ and $G^*(\cdot)$ are non-decreasing functions.

Now, to make inferences about a "differentiable" function (van der Vaart, 1998, Chapter 20) $\mathcal{H}\{F(\cdot),G(\cdot)\}$ of $\{W_F(\cdot),W_G(\cdot)\}$, the distribution of $n^{1/2}[\mathcal{H}\{\hat{F}(\cdot),\hat{G}(\cdot)\}-\mathcal{H}\{F(\cdot),G(\cdot)\}]$ can be approximated by the conditional (on the data) distribution of $n^{1/2}[\mathcal{H}\{F^*(\cdot),G^*(\cdot)\}-\mathcal{H}\{\tilde{F}(\cdot),\tilde{G}(\cdot)\}]$. Note that under the sup-norm metric or topology, one can use this approximation to construct confidence intervals for the IDI and IAUC. Moreover, for making inference about

the difference of two medians, we let $\mathcal{H}(F, G) = F^{-1}(1/2) - G^{-1}(1/2)$.

3. EXAMPLE

We illustrate the proposed method with the data from a breast cancer study to evaluate the predictive value of a new biomarker, "wound-response gene expression signature", for patient's survival (Chang et al., 2005). For each patient, this gene score was derived from her microarray gene expression data. The 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-node, tumor grade, vascular invasion, estrogen receptor status, chemo/hormonal therapy or not, and mastectomy or breast conserving surgery. The data are available at http://microarray-pubs.stanford.edu/wound_NKI/explore.html. The gene expression data and the conventional biomarker values were collected at the Netherlands Cancer Institute by van't Veer et al. (2002) and van de Vijver et al. (2002) to investigate the predictive ability of a gene score based on 70 specific gene expression data. The scoring system created by Chang et al. (2005) is different from the so-called Dutch 70 scoring system. For this data set, the median follow-up duration among the 295 patients was 6.7 years and the range was from 0.05 to 18.3 years (van de Vijver, 2002). Here, we are interested in quantifying the added value from the gene score by Chang et al. over the above conventional predictors.

For illustration, we let T be the time to either the first metastasis or death. The Kaplan-Meier curve based on the entire event times is given in Figure 1. The ten year event-free survival rate is 61.5%. Now, to evaluate the added value of the gene score over the conventional markers, we 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. Let $t_0 = 10$ (years). We fit the data with two Cox proportional hazards models described in Section 2 with $Z_{(1)}$ and $Z_{(2)}$, respectively. The regression coefficient estimates with the corresponding standard error estimates are reported in Table 1. 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 in the sample, we then obtain a pair of risk scores $\{\hat{p}_1(Z_{(1i)}), \hat{p}_2(Z_{(2i)})\}$ for approximating $p(Z_i)$. In Figure 2, we present a scatter diagram whose x-axis and y-axis are $\hat{p}_1(\cdot)$ and $\hat{p}_2(\cdot)$, respectively. The black dots represent the subjects who had events, the gray ones are those who were event-free, and the open circles are for the censored observations before ten years. If there are relatively few censored observations, visually this type of plot can be quite informative to examine the added value of the gene score. If the gene score is "useless", one would expect that the black and gray dots are symmetrically distributed around the 45-degree line. For the present case, the black dots tend to scatter above the 45-degree line and the gray ones are under the diagonal line, indicating that the gene score indeed improves prediction. Moreover, this diagram provides the "conventional" risk score value and the contrast between two scores within each patient qualitatively and also quantitatively. Note that the observed standard C-statistics with and without the gene score are 0.71 and 0.69, respectively. The improvement from the gene score in C-statistic is only 0.02, and the corresponding 0.95 confidence interval is (-0.01, 0.05), which covers null value zero.

Next we plot the estimated distribution functions $\hat{F}(\cdot)$ and $\hat{G}(\cdot)$ in Figure 3. Graphically the gene score appears to provide extra information regarding the prediction of the ten year event rates. The area between two curves is an estimated IDI index, which is 0.05 with a 0.95 confidence interval of (0.02, 0.09). The vertical distance between two gray dots is an estimated IAUC, which

is 0.27 with a 0.95 confidence interval of (0.09, 0.45). The horizontal distance between two black dots is an estimate of the median difference, which is 0.06 with a 0.95 confidence interval of (0.02, 0.10). Note that to obtain the standard error estimates, we utilized the perturbation-resampling method discussed in Section 2 with 1000 realized independent samples of the unit exponential.

4. REMARKS

If there are very few censored observations before t_0, the scatter diagram like Figure 3 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 3, 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, for the cancer example, the censoring proportion at year 10 is about 40%. Figure 3 by itself is not particularly useful. The distribution function plot in Figure 4 is informative for the contrast of two scoring systems. However, 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.

If we have pre-specified risk categories, for example, 0-10, 10-20, $>$ 20 per cent ten-year risk, one may use the net reclassification improvement (NRI) suggested by Cook et al. (2006) and Pencina et al. (2008). With the perturbation-resampling method, it would be straightforward to obtain an approximation to the distribution of the estimator of NRI. The confidence intervals for such a metric can be obtained accordingly.

For the analysis of the data from the cancer study presented in Section 3, we discretized the event time using ten-year cutoff time point to define "cases" and "controls". Such a binary outcome variable may not be able to capture differences between long and short term survivors.

It would be interesting to generalize the inference procedures for the binary to ordinal categorical outcomes or continuous responses.

5. APPENDIX

Let $\theta_k = (\log\{\Lambda_{0k}(t_0)\}, \beta_k')'$ be a vector of parameters for k=1,2, and let $p_k(Z_{(k)};\theta_k) = 1 - e^{-\exp\{(1,Z_{(k)}')\theta_k\}}$ 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 $n \to \infty$, 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 $\mathrm{B}_k, k = 1, 2$. To derive the asymptotic properties, we assume that B_k is 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 density function continuous in $\theta_1 \in \mathrm{B}_1$ and $\theta_2 \in \mathrm{B}_2$.

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

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

It follows, from the uniform consistency of $\hat{H}(\cdot)$ (Kalbfleish & Prentice, 2002) and a uniform law of large numbers (Pollard, 1990), that

$$\sup_{(s,\theta_1,\theta_2)\in[-1,1]\times \mathrm{B}_1\times\mathrm{B}_2} \left|\hat{F}(s,\theta_1,\theta_2) - F(s,\theta_1,\theta_2)\right| \to 0,$$

where

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

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 $F(s,\theta_{10},\theta_{20}) = F(s)$. The uniform consistency of $\hat{G}(\cdot)$ is shown with the same argument.

Secondary, to derive the limiting distribution of $W_F(s) = \sqrt{n}\left\{\hat{F}(s) - F(s)\right\}$ let

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

and

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

Note that

$$W_F(s) = W_{Fa}(s,\hat{\theta}_1,\hat{\theta}_2) + W_{Fb}(s), \tag{5.1}$$

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

$$n^{-1/2}\sum_{i=1}^{n}\left[\frac{\Delta_i}{\hat{H}(X_i)}I\{D(Z_i;\theta_1,\theta_2)\le s, X_i\le t_0\} - \mathrm{pr}\{D(Z_i;\theta_1,\theta_2)\le s, T_i\le t_0\}\right] \tag{5.2}$$

is tight. From the standard asymptotic theory for the Kaplan-Meier estimator (Kalbfleish and Prentice, 2002),

$$\frac{\Delta_i}{H(X_i)} = 1 - \int_0^{\tau}\frac{dM_i(u)}{H(u)} \text{ and } 1 - \frac{\hat{H}(X_i)}{H(X_i)} = \int_0^{X_i}\frac{dM(u)}{\pi_X(u)} + o_p(n^{-1/2}),$$

where $\pi_X(t) = \mathrm{pr}(X_i \ge t)$, $M_i(t) = I(X_i \le t, \Delta_i = 0) - \int_0^t I(X_i \ge u)d\Lambda_C(u)$, $M(t) = \sum_{i=1}^n M_i(t)/n$, and $\Lambda_C(\cdot)$ is the cumulative hazard function for the common censoring variable. Using the aforementioned relationship (Bang & Tsiatis, 2000), (5.2) can be rewritten as

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\left[I\{D(Z_i;\theta_1,\theta_2)\le s, T_i\le t_0\} - \mathrm{pr}\{D(Z_i;\theta_1,\theta_2)\le s, T_i\le t_0\}\right]\\
&- n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}\frac{dM_i(u)}{H(u)}\left[I\{D(Z_i;\theta_1,\theta_2)\le s, T_i\le t_0\} - m(\theta_1,\theta_2,s,u)\right]\\
=&n^{-1/2}\sum_{i=1}^{n}\left[I\{D(Z_i;\theta_1,\theta_2)\le s, T_i\le t_0\}\left\{1 - \int_0^{\tau}\frac{dM_i(u)}{H(u)}\right\} - \mathrm{pr}\{D(Z_i;\theta_1,\theta_2)\le s, T_i\le t_0\}\right]\\
&+ n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau} m(\theta_1,\theta_2,s,u)\frac{dM_i(u)}{H(u)},
\end{aligned} \tag{5.3}$$

where

$$m(\theta_1,\theta_2,s,u) = \mathrm{pr}\{D(Z;\theta_1,\theta_2)\le s, T < t_0 \mid T \ge u\}.$$

To prove that (5.2) 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 since the last term in (5.3) only involves a smooth deterministic function in (θ_1, θ_2, s). Since B_k is bounded, it can be covered by $N_k = O(\epsilon^{-d_k})$ balls centered at $\theta_{k[j]} \in \mathrm{B}_k$ with a radius of ϵ, where $j = 1, \cdots, N_k$ and d_k is the dimension of $\theta_k, k = 1, 2$. Coupled with the fact that Z has a bounded support, it implies that for any $\theta_k \in \mathrm{B}_k$, one can find $1 \leq j_k \leq N_k$ such that $|\theta'_{k[j_k]}\tilde{z}_k - \theta'_k\tilde{z}_k| \leq C_{1k}\epsilon$ 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 $N_3 = O(\epsilon^{-1})$ points in the interval $[-1, 1]$ such that $-1 = s_1 < s_2 < \cdots < s_{N_3} = 1$ and $s_i - s_{i-1} \leq \epsilon$. Therefore for any θ_1, θ_2 and s, we can find j_1, j_2 and j_3, such that $|\{D(z; \theta_1, \theta_2) - s\} - \{e^{-\exp(\theta'_{1[j_1]}\tilde{z}_1)} - e^{-\exp(\theta'_{2[j_2]}\tilde{z}_2)} - s_{j_3}\}| \leq C_2\epsilon$. In the following, we will estimate the bracketing number of $\mathcal{F}$. Let

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

and

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

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

$$\mathrm{E}[\{u_{ijk}(Z) - l_{ijk}(Z)\}^2] = \mathrm{pr}(|e^{-\exp\{(1, Z'_{(1)})\theta_{1[i]}\}} - e^{-\exp\{(1, Z'_{(2)})\theta_{2[i]}\}} - s_k| < C_2\epsilon)$$

$$\leq \sup_{\theta_1, \theta_2, s} \mathrm{pr}(|D(Z, \theta_1, \theta_2) - s| \leq C_2\epsilon) \leq C_3\epsilon,$$

since the density function of $D(Z, \theta_1, \theta_2)$ is uniformly bounded. Therefore, the bracketing number of $\mathcal{F}$ is $O(\epsilon^{-2(d_1+d_2+1)})$ 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,

$$W_{Fb}(s) = \dot{F}_{\theta_1}(s, \theta_{10}, \theta_{20})n^{1/2}\left(\hat{\theta}_1 - \theta_{10}\right) + \dot{F}_{\theta_2}(s, \theta_{10}, \theta_{20})n^{1/2}\left(\hat{\theta}_2 - \theta_{20}\right) + o_p(1) \qquad (5.4).$$

where $\dot{F}_{\theta_k} = \frac{\partial F}{\partial \theta_k}$. Since regardless of model adequacy, the maximum partial likelihood estimator $\hat{\theta}_k$ is a regular estimator, i.e.,

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

where $\psi_{k1}, \cdots, \psi_{kn}$ are n i.i.d mean zero random variables. Coupled with (5.1), (5.3) and (5.4),

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

where

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

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 above, 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 $[-1, 1] \times [-1, 1]$.

The perturbed version of $\hat{H}$ in (2.7) is given by

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

where $\tilde{H}(\cdot)$ is the observed $\hat{H}(\cdot)$, $\tilde{M}_i(t) = I(x_i \leq t, \delta_i = 0) - \int_0^t I(x_i > u) d\tilde{\Lambda}_C(u)$, and $\tilde{\Lambda_C}(\cdot)$ is the observed Nelson-Aalan estimator of the cumulative hazard function for the censoring variable C. $D^*(\cdot)$ in (2.8) and (2.9) is given by

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

where β_k^* and $\Lambda_k^*(t_0)$ are given as Cai et al. (2009), i.e., $\beta_k^* - \tilde{\beta}_k$ and $\log\{\Lambda_{k0}^*(t)\} - \log\left\{\tilde{\Lambda}_{k0}(t)\right\}$ are

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

and

$$\frac{n^{-1}}{\tilde{\Lambda}_{k0}(t)} \sum_{i=1}^{n} I(x_i \leq t)\delta_i \left\{ \frac{(V_i - 1)}{\tilde{S}_k^{(0)}(x_i, \tilde{\beta}_k)} - \frac{n^{-1}\sum_{j=1}^{n}(V_j - 1) I(x_j \geq x_i) e^{\tilde{\beta}_k' z_{(kj)}} + \tilde{S}_k^{(1)}(x_i, \tilde{\beta}_K)'(\beta_k^* - \tilde{\beta}_k)}{\tilde{S}_k^{(0)}(x_i, \tilde{\beta}_k)^2} \right\},$$

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

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

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

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 $n^{1/2}\left(\hat{\theta_k} - \theta_{k0}\right)$ converges to the same limiting normal distribution (Cai et al., 2009). Furthermore, using similar expressions given as (5.1), (5.3), and (5.4), it is also straightforward to show that $\{W_F^*(s), W_G^*(u)\}'$ can be approximated by $n^{-1/2}\sum_{i=1}^n\{\tilde{\pi}_F(s; z_i, x_i, \delta_i), \tilde{\pi}_G(u; z_i, x_i, \delta_i)\}'(V_i - 1)$, 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, the distribution of $W_H = n^{1/2}[\mathcal{H}\{\hat{F}(\cdot), \hat{G}(\cdot)\} - \mathcal{H}\{F(\cdot), G(\cdot)\}]$ can be approximated by that of $W_H^* = n^{1/2}[\mathcal{H}\{F^*(\cdot), G^*(\cdot)\} - \mathcal{H}\{\tilde{F}(\cdot), \tilde{G}(\cdot)\}]$ conditional on the observed data in the sense that $\text{pr}\{|W_H^* - W_H| \geq \epsilon | (Z_i, X_i, \Delta_i), i = 1, \cdots, n\}$ converges to 0 in probability for any $\epsilon > 0$.

References

Bang, H. & Tsiatis, A. A. (2000), "Estimating medical cost with censored data," *Biometrika,* 87, 329 – 343.

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. & Cheng, S. (2008), "Robust combination of multiple diagnostic tests for classifying censored event times," *Biostatistics,* 9, 216–233.

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.

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

Cook, N. R., Buring, J. E., & Ridker, P. M. (2006), "The effect of including C-reactive protein in cardiovascular risk prediction models for women," *Annals of Internal Medicine,* 145, 21 – 29 .

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., Griffith, J. L., Schmidt, C. H., & Terrin, N. (1997), "Measures for evaluating model performance," *Proceedings of the Biometrics Section, Alexandria, VA, U.S.A. American Statistical Association, Biometrics Section: Alexandria, VA.,* 253 – 258

Greenland, P. & O'Malley, P. G. (2005), "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,* 165(21), 2454 – 2456.

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.

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.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W. & Newcomb P. (2004), "Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker," *American Journal of Epidemiology,* 159, 882 – 890.

Pollard, D. (1990), *Empirical Processes: Theory and Applications,* Hayward, CA: Institute of Mathematical Statistics.

Uno, H., Cai, Pencina, M. J., D'Agostino, R. B. & Wei, L. J. (2009), "On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," Harvard University Biostatistics Working Paper Series. Working Paper 101. http://www.bepress.com/harvardbiostat/paper101,

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.

van der Vaart, A. W. (1998), *Asymptotic statistics,* Cambridge University Press, Cambridge.

Ware, J. H. (2006), "The limitations of risk factors as prognostic tools," *The New England Journal of Medicine,* 355, 2615 – 2617.

Wu, C.F.J. (1986), "Jackknife, bootstrap, and other resampling methods in regression analysis (with discussion)," *The Annals of Statistics,* 14, 1261 – 1295.

Table 1. *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 [yrs]	-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 error estimate

(3) p-value

Figure 1. A hypothetical example for distribution functions of $\hat{D}$ for subjects with events and for those without

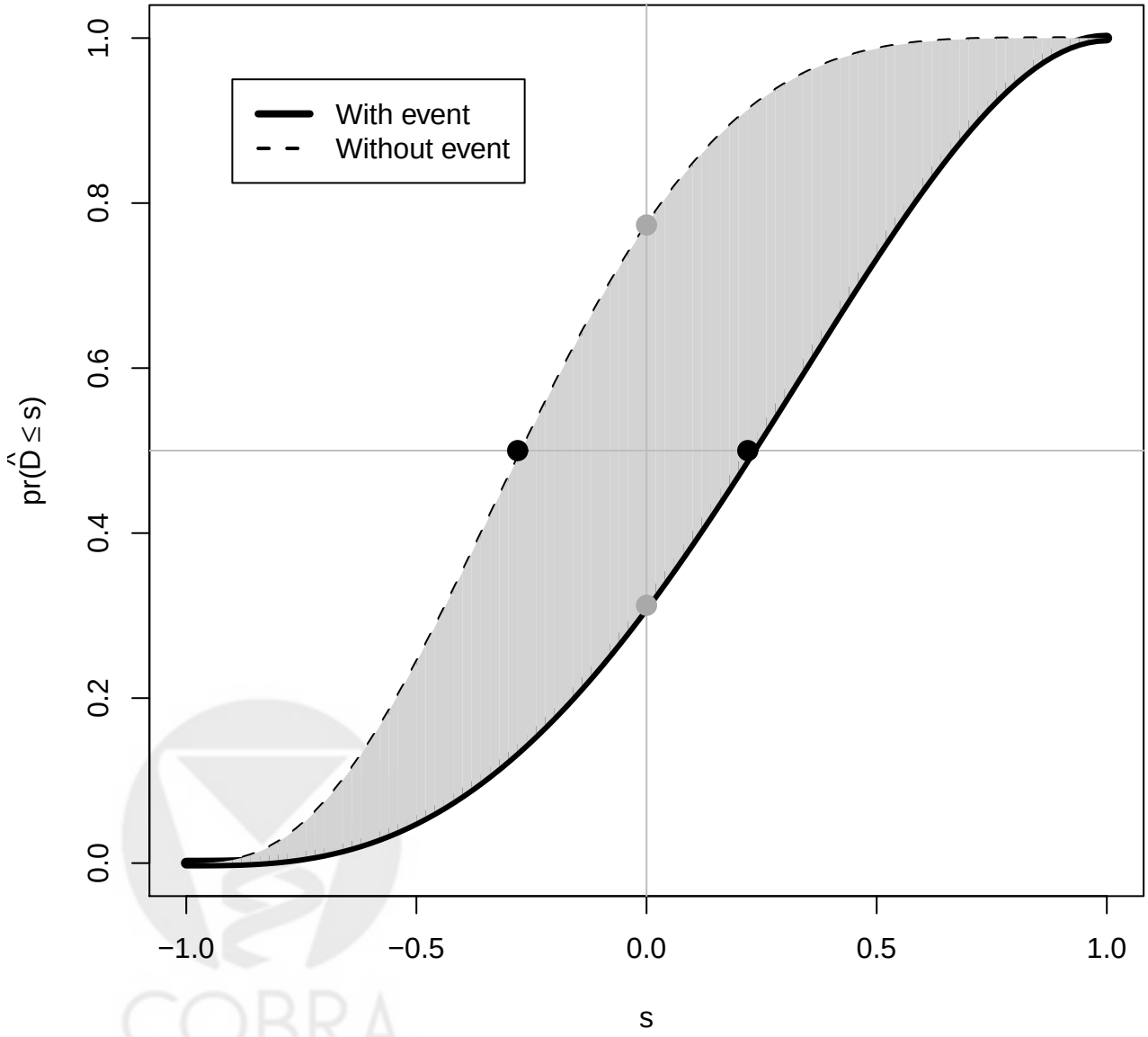

Figure 2. Kaplan-Meier estimate for metastasis/death with breast cancer data

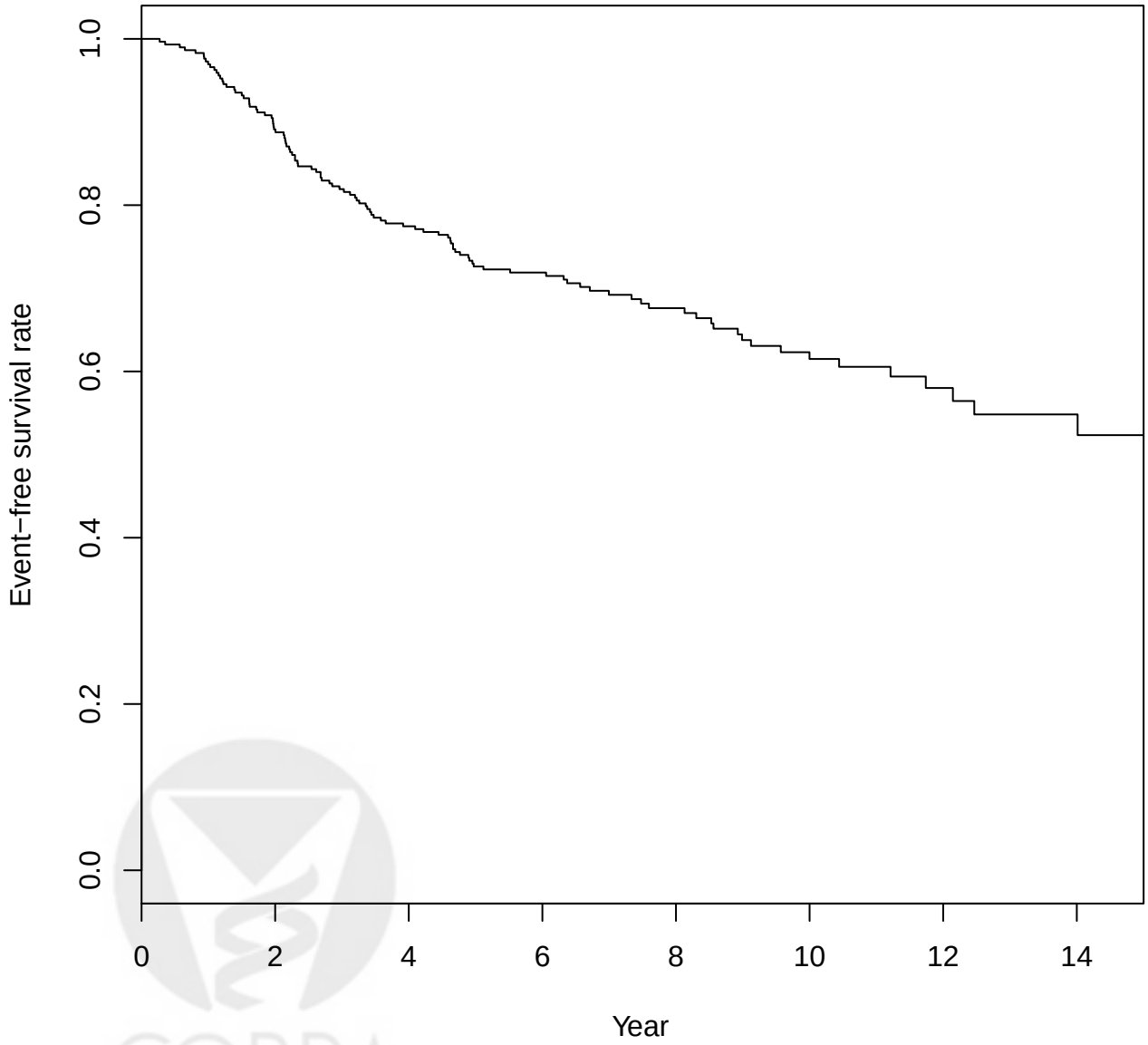

Figure 3. Scatter diagram of $\hat{p}_1$ vs. $\hat{p}_2$ for "with event" and "without event" with breast cancer data

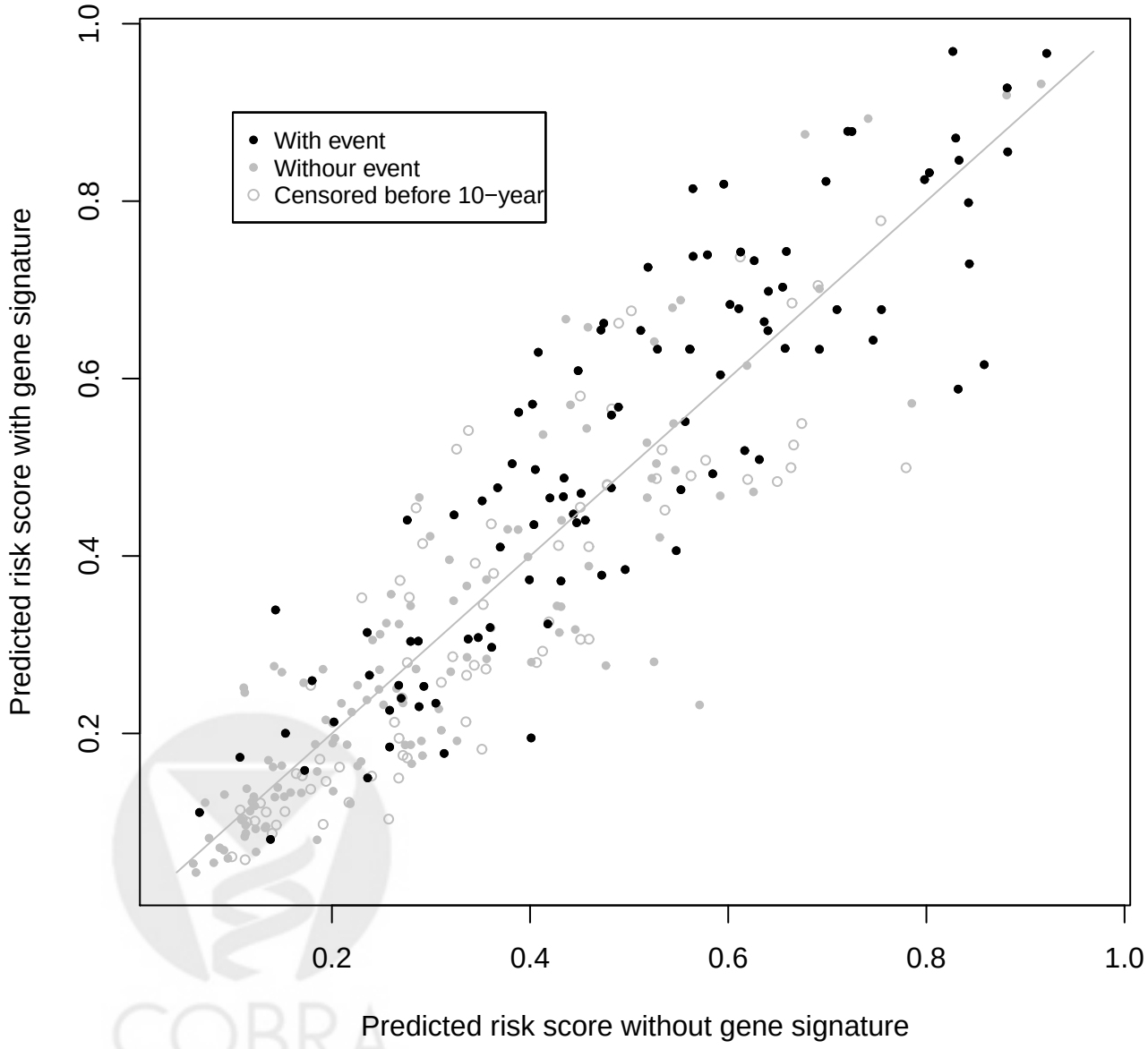

Figure 4. Empirical distribution function with $\hat{D}(s)$ for "with event" and "without event" with breast cancer data