Published in final edited form as:
Biometrics. 2011 December ; 67(4): 1389–1396. doi:10.1111/j.1541-0420.2011.01600.x.

Graphical Procedures for Evaluating Overall and Subject-specific Incremental Values from New Predictors with Censored Event Time Data

Hajime Uno[1,2], **Tianxi Cai**[2], **Lu Tian**[3], and **L. J. Wei**[2,*]

[1]Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute, Boston, Massachusetts 02115, U.S.A.

[2]Department of Biostatistics, Harvard University, Boston, MA 02115, U.S.A

[3]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California 94305, U.S.A

Summary

Quantitative procedures for evaluating added values from new markers over a conventional risk scoring system for predicting event rates at specific time points have been extensively studied. However, a single summary statistic, for example, the area under the receiver operating characteristic curve or its derivatives, may not provide a clear picture about the relationship between the conventional and the new risk scoring systems. When there are no censored event time observations in the data, two simple scatterplots with individual conventional and new scores for “cases” and “controls” provide valuable information regarding the overall and the subject-specific level incremental values from the new markers. Unfortunately, in the presence of censoring, it is not clear how to construct such plots. In this paper, we propose a nonparametric estimation procedure for the distributions of the differences between two risk scores conditional on the conventional score. The resulting quantile curves of these differences over the subject-specific conventional score provide extra information about the overall added value from the new marker. They also help us to identify a subgroup of future subjects who need the new predictors, especially when there is no unified utility function available for cost-risk-benefit decision making. The procedure is illustrated with two data sets. The first is from a well-known Mayo Clinic PBC liver study. The second is from a recent breast cancer study on evaluating the added value from a gene score, which is relatively expensive to measure compared with the routinely used clinical biomarkers for predicting the patient's survival after surgery.

Keywords

Discriminant analysis; Nonparametric function estimation; Prediction; Receiver operating characteristic curve

1. Introduction

For a binary phenotypic outcome, numerical and graphical methods for evaluating an overall incremental value from a new set of markers over a conventional risk scoring system have

* wei@hsph.harvard.edu .

Supplementary Materials Web Appendix referenced in Section 2 and R code to implement the proposed method are available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

been extensively studied (Bamber, 1975; Zhou, Obuchowski and McClish, 2002; Pepe, 2003; Pepe et al., 2004; Greenland and O'Malley, 2005; Ware, 2006; Pencina et al., 2008). Novel generalizations of these procedures to handle censored event time data have also been proposed (Hanley and McNeil, 1982; Harrell, Lee and Mark, 1996; D'Agostino et al., 1997; Pencina and D'Agostino, 2004; Heagerty and Zheng, 2005; Cook, Buring and Ridker, 2006; Cai and Cheng, 2008; Uno et al., 2009; Pencina, D'Agostino and Steyerberg, 2010; Uno et al., 2011). Evaluating the added value from the new markers with an overall summary measure is an important first step for establishing a potential new prediction rule. On the other hand, even when the new markers have an impressive overall incremental value, it is not clear whether all the future subjects in the population of interest would benefit from these new markers. Moreover, when the new markers have no meaningful overall incremental value, it does not imply that we should not utilize these markers for any future subject. The next critical step for evaluation is to identify subgroup of patients who would or would not need the additional markers for better prediction via their conventional risk scores. Unfortunately, relatively little effort has been made for establishing a systematic, analytic procedure for such “subgroup analysis” in the statistical or medical literature (D'Agostino, 2006). Recently, Tian, Cai and Wei (2009) proposed a procedure for this type of subject-specific level analysis by controlling a pre-specified simultaneous inference error rate. However, their proposal does not incorporate censoring and depends heavily on the choice of the utility function, a weighted average between the false positive and negative rates.

For binary outcomes, simple scatterplots of individual conventional risk scores vs. new ones provide valuable information about an overall and also personalized-level incremental values of the new markers (Gu and Pepe, 2009). For example, in selecting patients with advanced or end-stage primary biliary cirrhosis, PBC, for orthotopic liver transplantation, five patients' baseline covariates, namely age, albumin, bilirubin, edema and prothrombin time, were identified to be important predictors for the patient's survival based on the data from a Mayo Clinic study (Dickson et al., 1989; Fleming and Harrington, 1991, pp. 160). Suppose that we would like to know the added value from the bilirubin measure over the other four variables with respect to prediction of 5-year survival based on observations from 416 patients with complete information on those predictors. To this end, we first obtain a risk score based on these four variables without bilirubin,

$$0.29\times(\text{age}/10)-3.49\times\log(\text{albumin})+1.33\times\text{edema}+3.07\times\log(\text{prothrombin time}), \quad (1.1)$$

by fitting the data with a simple additive Cox model using partial likelihood estimation procedure (Cox, 1972). Based on (1.1) and the standard Breslow estimator for the baseline cumulative hazard, we obtain individual patients' 5-year cumulative mortality risk, denoted by $\widehat{p}_{1i}, i=1,\cdots,416$. Next, we fit the data using another additive Cox model with all five covariates including bilirubin. The resulting risk score is

$$0.40\times(\text{age}/10)-2.51\times\log(\text{albumin})+0.86\times\log(\text{bilrubin})+0.90\times\text{edema}+2.4\times\log(\text{prothrombin time}). \quad (1.2)$$

Let $\widehat{p}_{2i}$ denote the *i*th individual five-year mortality rate based on (1.2).

In the PBC dataset, there are 196 censored survival observations by Year 5 and 114 patients died during this time period. Figure 1(a) shows the scatterplot of $\widehat{p}_{1i}$ vs. the difference $(\widehat{p}_{2i} - \widehat{p}_{1i})$ for those 114 observable "cases." The majority of those black dots in the figure are above the horizontal line, indicating that globally the bilirubin provides extra information about the 5-year mortality rate for those "cases." Moreover, for a subject with $\widehat{p}_1$ between 0.2 and 0.6, the corresponding $\widehat{p}_2$ tends to be substantially higher. Figure 1(b) shows the scatterplot for the observable "controls," who survived and were still under follow-up by Year 5. Here, most of $\widehat{p}_2$ tend to be smaller than their $\widehat{p}_1$, indicating that bilirubin has an overall incremental value. At the personalized level, it appears that for the survived patients whose conventional risk scores are between 0.15 and 0.35, bilirubin provides nontrivial improvement for predicting survival beyond 5 years. If there were no censored observations in the data, the scatterplots in Figure 1, coupled with the standard lowess curves for the scatter diagram (dark curves), would provide a valuable tool for quantifying global and subject-specific level performance using Model (1.2) with bilirubin. Unfortunately, for the present example, the number of censored observations is substantial and it is not clear how to construct valid plots like Figure 1.

In this paper, with censored survival data we propose a nonparametric procedure to consistently estimate quantiles of the distributions of the difference $(\widehat{p}_2 - \widehat{p}_1)$ given $\widehat{p}_1$ for cases and controls. The resulting quantile curves are then presented using a similar configuration to Figure 1. The new method is derived under a more general setting. Here, a case is defined as the survival time being in a time interval I_1, while a control is defined as the survival time in an interval I_0, where I_1 is entirely on the left hand side of I_0. By repeating the analysis with various pairs of I_0 and I_1, one may find, for example, that the new predictors are not useful when these two intervals are widely separated (for instance, short- vs. long-term survival), but may have substantial incremental values when these two intervals are relatively close. This type of finding can be quite informative for cost-benefit decision making. The graphical procedure proposed here can be utilized to evaluate an overall and also subject-specific incremental values of a new set of markers over the conventional risk scoring system. That is, by first checking the conventional score for an individual patient, one may decide whether the additional markers are needed from cost-benefit perspectives. For example, if a subject's old and new estimated risks are 0.20 and 0.22, respectively, the new markers would add rather little value to our decision making on the patient's treatment and prevention selections. The new procedure is illustrated with the above Mayo Clinic data and also with the data set from a breast cancer study to evaluate the additional prediction ability based on a new gene risk score on top of conventional clinical markers. The second example is particularly interesting due to the fact that it is relatively expensive to measure the gene score compared with clinical markers, which are routinely obtained after patients' surgery for breast cancer.

2. Estimating the distribution of the new risk score conditional on the old risk score

Let T be the time to an event of interest and Z be its corresponding vector of baseline covariates. For the two specific time intervals $I_1 \varepsilon [t_1, t_2)$ and $I_0 \varepsilon [t_3, t_4)$ discussed in Section 1, suppose that for a given Z we are interested in estimating the risk of a case:

$$\mathrm{pr}(T \in I_1|Z) / \{\mathrm{pr}(T \in I_0|Z) + \mathrm{pr}(T \in I_1|Z)\}. \tag{2.1}$$

Let U and V be two vectors, which are functions of Z. Here, U is a function of conventional markers only, but V is a function of both conventional and new predictors. One of the

questions is how to identify patients with U, who may need V for better prediction of (2.1). This is a particularly important question when it is costly or invasive to measure the new markers. Often, the event time T may be censored by a censoring variable C. Assume that C is independent of T and Z. Let $G(\cdot)$ be the survival function of C. Moreover, let the binary variable $E = 0$, if $T \varepsilon I_0$; $= 1$, if $T \varepsilon I_1$. Note that one can assign an arbitrary value (other than 0 or 1) for E when T is outside of these two time intervals. Now, let $\{(T_i, C_i, E_i, Z_i, U_i, V_i), i = 1, \ldots, n\}$ be n independent copies of (T, C, E, Z, U, V). For T_i, we observe $\{X_i, \Delta_i\}$, where $X_i = \min(T_i, C_i)$, and $\Delta_i = 1$, if $X_i = T_i$, and 0, otherwise, $i = 1, \cdots, n$. Due to potential censoring, the binary variable E may not be observable.

To construct a risk scoring system with U, let us consider the standard Cox proportional hazards model with the risk score $\beta' U$, where β is an unknown vector of regression parameters. With the above observed data, let $\widehat{\beta}$ be the maximum partial likelihood estimator for β. In practice, this semi-parametric model is simply an approximation of the "true" model. Under a mild condition, $\widehat{\beta}$ converges to a constant, as $n \to \infty$ (Hjort,1992), regardless of the adequacy of the Cox model. This property is critical for developing our new procedure. Similarly, for V, we fit the data with another additive Cox's model with the risk score $\gamma' V$. Let $\widehat{\gamma}$ be the corresponding estimator for γ.

Now, consider an independent future subject from the same study population whose $(T, E, Z, U, V) = (T^0, E^0, Z^0, U^0, V^0)$. To estimate (2.1) with U^0, let $\widehat{p}_1\left(U^0\right)$ be the estimator for (2.1) constructed from the Breslow estimator for the underlying cumulative hazard function of the above Cox's model and $\widehat{\beta}' U^0$. Explicitly, letting $\widehat{\Lambda}_1(\cdot)$ denote the Breslow estimator, then $\widehat{p}_1\left(U^0\right)$ is

$$\frac{\exp\left\{-\widehat{\Lambda}_1(t_1)\, e^{\widehat{\beta}' U^0}\right\} - \exp\left\{-\widehat{\Lambda}_1(t_2)\, e^{\widehat{\beta}' U^0}\right\}}{\exp\left\{-\widehat{\Lambda}_1(t_1)\, e^{\widehat{\beta}' U^0}\right\} - \exp\left\{-\widehat{\Lambda}_1(t_2)\, e^{\widehat{\beta}' U^0}\right\} + \exp\left\{-\widehat{\Lambda}_1(t_3)\, e^{\widehat{\beta}' U^0}\right\} - \exp\left\{-\widehat{\Lambda}_1(t_4)\, e^{\widehat{\beta}' U^0}\right\}}.$$

Similarly, let $\widehat{p}_2\left(V^0\right)$ be the corresponding estimator via the covariate vector V^0. To compare these two predictors, let $\widehat{D}\left(Z^0\right) = \widehat{p}_2\left(V^0\right) - \widehat{p}_1\left(U^0\right)$. Note that to make *overall* comparisons between models with U and V, one may estimate the distribution of $\widehat{D}\left(Z^0\right)$ given $E^0 = e$, where e is either 0 or 1. If V has an overall added value over U, one would expect that for $e = 1$, that is, for those future subjects with $T^0 \in I_1$, $\widehat{D}\left(Z^0\right)$ has more positive mass, and if $e=0$, $\widehat{D}\left(Z^0\right)$ has more negative mass. Recently various analytic methods based on the distributions of $\widehat{D}\left(Z^0\right)$ for cases and controls were proposed, for example, by Pencina et al. (2008), Gu and Pepe (2009) and Uno et al. (2009) to summarize the overall incremental value of the new markers.

In this paper, we are also interested in the subject-specific level evaluation for the incremental values, that is, estimating the distribution of $\widehat{D}\left(Z^0\right)$ conditional on $E^0 = e$ and $\widehat{p}_1\left(U^0\right) = p$, where e is either 0 or 1, and p belongs to $\mathcal{I} = [jl, jr]$ is a strictly inner subset of the support of $\widehat{p}_1\left(U^0\right)$. Let $q_{\tau e}(p)$ be the τth conditional quantile of the above distribution, for $0 < \tau < 1$. To estimate $q_{\tau e}(p)$, we utilize a nonparametric quantile regression estimation technique by letting the quantile of $\widehat{p}_2\left(V^0\right)$ be *locally* linear in $\widehat{p}_1\left(U^0\right)$ (Yu and Jones, 1998).

Specifically, without censored observations, for any given p, we minimize the following objective function with respect to a and b,

$$\sum_{i=1}^{n} I(E_i=e)\, K_h\{\psi(\widehat{p}_1(U_i)) - \psi(p)\}\, \rho_\tau(\psi(\widehat{p}_2(V_i)) - a - b[\psi\{\widehat{p}_1(U_i)\} - \psi(p)]), \tag{2.2}$$

where $K_h(x) = K(x/h)/h$, $K(\cdot)$ is a symmetric probability density function, h is a bandwidth such that $h = O(n^{-\nu})$ with $\nu \in (1/5, 1/2)$ and $\rho_\tau(x)$ is the check function, which is τx if $x \geq 0$, and is $(\tau - 1)x$ if $x < 0$. Here, we choose a proper transformation $\psi\{\widehat{p}_1(U)\}$ of $\widehat{p}_1(U)$ to improve smoothing, where $\psi(\cdot)$: $(0, 1) \to (-\infty, \infty)$ is a known, non-decreasing function (Wand, Marron and Ruppert, 1991; Park et al., 1997). For example, one may let $\psi(p) = \log\{-\log(1 - p)\}$. Let the minimizer of (2.2) be $\widehat{a}$ and $\widehat{b}$. Then let

$$\widehat{q}_{\tau e}(p) = \psi^{-1}(\widehat{a}) - p \tag{2.3}$$

be an estimator for $q_{\tau e}(p)$.

Since E may not be observable in (2.2), we replace $I(E = e)$ by $I(E^\dagger = e)$, with an inverse probability weighting technique, where $E^\dagger$ is 1 if $X \in I_1$; 0 if $X \in I_0$. Specifically, let $\widehat{G}(\cdot)$ denote the Kaplan Meier estimator of $G(\cdot)$ and let η be a pre-specified time point such that $G(\eta) > 0$. The choice of the weight depends of the choice of I_1 and I_0. For the case where I_0 is an interval such that $t_4 < \eta$, the weight $\widehat{w}_i = \Delta_i / \widehat{G}(X_i)$ for both $I(E_i^\dagger = 1)$ and $I(E_i^\dagger = 0)$. This may be justified heuristically using the argument that $E\{\widehat{w}_i I(E_i^\dagger = e) | Z_i, T_i\} \approx I(E = e)$. For the case when $t_3 < \eta$, and $t_4 = \infty$ for the interval I_0, the weights $\widehat{w}_i = \Delta_i / \widehat{G}(X_i)$ and $\widehat{w}_i = 1/\widehat{G}(t_3)$ are used for $I(E_i^\dagger = 1)$ and $I(E_i^\dagger = 0)$, respectively. Heuristically this can be justified with the argument that $E\{\widehat{w}_i I(E_i^\dagger = 0) | Z_i, T_i\} \approx I(E = 0)$. This inverse probability weighting adjustment, coupled with (2.2), results in the following minimand:

$$\sum_{i=1}^{n} \widehat{w}_i I(E_i^\dagger = e)\, K_h\{\psi(\widehat{p}_1(U_i)) - \psi(p)\}\, \rho_\tau(\psi(\widehat{p}_2(V_i)) - a - b[\psi\{\widehat{p}_1(U_i)\} - \psi(p)]). \tag{2.4}$$

Then, the corresponding estimator $\widehat{q}_{\tau e}(p)$ for $q_{\tau e}(p)$ is given by (2.3), but with $\widehat{a}$ being a minimizer of (2.4) with respect to a and b. In Web Appendix, we show that for each fixed τ, $\sup_{p \in \mathcal{I}} |\widehat{q}_{\tau e}(p) - q_{\tau e}(p)| \to 0$, in probability as $n \to \infty$.

In practice, it is important to know how to choose the smooth parameter h in the above nonparametric estimation. To this end, we consider a commonly used K-fold cross-validation procedure. Specifically, we randomly partition the data into K disjoint parts, $\mathcal{I}_1, \cdots, \mathcal{I}_K$. For each k, we use the data not in $\mathcal{I}_K$ to obtain the regression parameter estimators in the above two Cox's models, denoted by $\widehat{\beta}_{(-k)}$ and $\widehat{\gamma}_{(-k)}$. Moreover, let $\{\widehat{p}_{1(-k)}(\cdot), \widehat{p}_{2(-k)}(\cdot), \widehat{q}_{\tau e(-k)}(\cdot)\}$ denote the respective estimators corresponding to $\{\widehat{p}_1(\cdot), \widehat{p}_2(\cdot), \widehat{q}_{\tau e}(\cdot)\}$ based on data not in $\mathcal{I}_k$. We propose to choose h by minimizing

$$\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\widehat{w}_i I\left(E_i^{\dagger}=e\right) I\{\widehat{p}_{1(-k)}(U_i)\in\mathcal{I}\}_{\rho\tau}\left(\psi\{\widehat{p}_{2(-k)}(V_i)\}-\psi\left[\widehat{q}_{\tau e(-k)}\{\widehat{p}_{1(-k)}(U_i)\}+\widehat{p}_{1(-k)}(U_i)\right]\right). \tag{2.5}$$

In practice, the lower and upper bounds of $\mathcal{I}$ may be chosen as, for example, the 3rd and 97th percentiles of the empirical distribution of $\widehat{p}_1\left(U^0\right)$.

3. Examples

First, let us revisit the PBC example discussed in the Introduction Section. Assume that we are interested in two time intervals, $I_0 = (5, \infty)$(years) and $I_1 = [0, 5]$ (years). On average the 5-year cumulative mortality rate is 0.30. Here, $\widehat{p}_1(U_i)$ is obtained without using bilirubin and $\widehat{p}_2(V_i)$ is with bilirubin, $i = 1, \cdots, n$, via the risk scores (1.1) and (1.2) and two working Cox's models. Before applying the proposed analysis, we present the results from conventional methods for evaluating the overall comparisons between the above two models. First, the difference in the area under the ROC curve between two models is 0.10 (Uno et al., 2007). The summary measures, "integrated discrimination improvement (IDI)" and "category-less net reclassification improvement (NRI)," are 0.18 and 1.08, respectively. Note that with our notations, the IDI and category-less NRI are $E\left\{\widehat{D}\left(Z^0\right)\middle|E^0=1\right\}-E\left\{\widehat{D}\left(Z^0\right)\middle|E^0=0\right\}$ and $2\left[\mathrm{pr}\left\{\widehat{D}\left(Z^0\right)\geq 0\middle|E^0=1\right\}-\mathrm{pr}\left\{\widehat{D}\left(Z^0\right)\geq 0\middle|E^0=0\right\}\right]$, respectively (Pencina et al., 2008, 2010; Uno et al., 2009). These results indicate a significant, overall improvement from bilirubin. However, it is not clear this marker would be valuable for all subjects. Note that from a practical point view, this would not be an issue since measuring bilirubin is neither costly nor invasive. We use this example to illustrate the point that even when a marker has a very impressive overall performance, it does not mean that we need the marker for all future subjects.

Now, for the new proposal, to estimate $q_{\tau e}(\cdot)$, we let ψ in (2.4) be the log(− log) function and the kernel function be the standard normal density function. To choose the smooth parameter h, we used 10-fold cross validation scheme with (2.5). For instance, to estimate the median $q_{0.5e}(\cdot)$ for patients who would die by Year 5, the resulting "optimal" h with respect to ψ-scale is 1.6. The interval $\mathcal{I}$ over which we construct the median curves is (0.10, 0.995). Figure 2(a) gives the estimated median curve of $\widehat{D}$ over the risk score $\widehat{p}_1$ (solid curve). The lower and upper boundaries of the shaded area are the corresponding 25*th* and 75*th* percentile curves. Figure 2(b) gives the plots which are the counterparts for subjects who would survive more than 5 years. In Figure 2(c), we provide the density function estimate of $\widehat{p}_1$ score. There are about 80% of patients in this specific population, whose risk scores $\widehat{p}_1$ are between 0.1 and 0.6. Based on Figure 2, the distributions of $\widehat{D}$ for "cases" over the interval (0.1, 0.995) have more positive mass, especially for $\widehat{p}_1$ between 0.2 and 0.6. The bilirubin helps greatly for the "controls" when $\widehat{p}_1$ is between 0.2 and 0.6, that is, the false positive rate can be drastically reduced with bilirubin. We have also examined extensively the added values of bilirubin for various sets of time intervals I_0 and I_1. In Figure 3, we present the plots of estimated median curves for cases and controls with respect to four different sets of time intervals I_0 and I_1. If bilirubin were not routinely measured for evaluating liver function clinically, one would recommend its usage for future subjects

whose “conventional” scores were between 0.2 and 0.6. Note that we cannot estimate the medians well for controls beyond 0.6 with this set of data.

Next, we use a more interesting example to illustrate a scenario in which a non-trivial cost is associated with measuring a new marker. The data set used for our illustration is from a breast cancer study to evaluate a new genetic marker, “wound-response gene expression signature,” for predicting patients' survival (Chang et al., 2005). For each study patient, this gene score was derived from the microarray gene expression data. Here, 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 baseline variables 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 data are available at http://microarray-pubs.stanford.edu/wound NKI/explore.html, which were collected at the Netherlands Cancer Institute by van't Veer et al. (2002) and van de Vijver et al. (2002). The median follow-up time for those 295 patients is 6.7 years and the range is 0.05 to 18.3 years; the cumulative event rate at Year 10 is 0.39.

The gene score (the so-called Dutch 70) created by the aforementioned Dutch scientists is different from that proposed by Chang et al. (2005). Here, we are interested in quantifying the added value from the gene score by Chang et al. over the above conventional clinical predictors. To this end, we fit the data with two working Cox models, one with gene score and the other without. The regression coefficient estimates are given in Table 1. Note that the gene score is statistically significant. Suppose we are interested in evaluating the incremental value from the gene score with respect to predicting events that occur by Year 10, which corresponds to set $I_0 = (10,\infty)$ (years) and $I_1 = (0, 10]$ (years) in our proposed analysis. Before applying our analysis, we present the results from conventional methods for evaluating the overall comparisons between the models with and without this marker. The difference in the area under the ROC curve is 0.05. The IDI and category-less NRI are about 0.05 and 0.40, respectively. These results indicate the gene score is not that impressive globally. Since measuring this marker is costly, it is not clear which set of future patients would benefit from it.

Now, for our analysis, we used the standard normal kernel and the log(−log) as the ψ function in the nonparametric estimation of the quantiles. Moreover, we used 10-fold cross validation procedure to choose h. For example, for estimating the medians, the optimal h for “cases” is 3.25 with respect to the ψ-scale and $\mathcal{I}$ is (0.15, 0.85). Figure 4(a) gives the median curve (solid curve) and the bands whose boundaries are the 0.25 and 0.75 quantile curves for “cases.” The x-axis is the score without using gene expression data. Figure 4(b) gives the counterparts for “controls,” those subjects who would survive beyond 10 years. The density function estimate of $\widehat{p_1}$ is given in Figure 4(c). The “conventional scores” of the majority of patients in this population are between 0.2 and 0.75. Note that the median curve is in the positive (negative) side for cases (controls). The improvement from the gene score, however, is quite modest uniformly over the conventional score. Since it is relatively expensive to measure the gene score compared with the routinely obtained clinical marker values, it is not clear from cost-benefit view if we should measure the gene score for any future patient. In Figure 5, we present plots of estimated median curves for cases and controls with respect to various sets of time intervals I_0 and I_1. Again, there seems no obvious gain from measuring gene score for predicting survival.

4. Remarks

If a study is designed for evaluating the incremental value from new predictors with respect to a specific set of time intervals I_0 and I_1, a global Cox model may not be appropriate for establishing the risk scores due to the fact that the resulting regression coefficient estimates reflect an average covariate effect over the entire study time. For this case, we may use, for example, the logistic regression for modeling the probability of a binary variable with two events $\{T \varepsilon I_1\}$ and $\{T \varepsilon I_0\}$ with predictors and use the technique developed by Uno et al. (2007) to obtain the risk scores. Then with the same argument in the present paper, nonparametric function estimates for conditional quantiles can be obtained accordingly. When there is no pre-specified set of time intervals of interest, one may use the Cox models to obtain unified scores $\widehat{\beta}' U$ and $\widehat{\gamma}' U$ first. However, it is important to note that these two scoring systems may not be comparable since we fit the data with two different models. Therefore, in this paper we convert the Cox scores to their risk counterparts with respect to a given paired I_0 and I_1 to evaluate the incremental values. When there is no pre-specified set of time intervals, by considering various sets of I_0 and I_1 in our analysis, one may identify when the new markers have practically meaningful added values for prediction. On the other hand, it is not clear how to utilize the Cox scores directly to perform such subject-level analysis without discretizing the continuous study follow-up time.

If the conventional scoring system is well-established, one may not need to fit the current data with the conventional markers. However, for this situation we recommend examining closely whether the present study population is similar to that from which the conventional score was constructed.

The graphical method presented here can also be utilized as a quantitative way to assess relative merits of two proposed models for fitting survival data. Unlike the lack of fit tests for model checking or a single summary statistic such as the likelihood ratio, the plots in Figures 3 and 5 with different sets of I_0 and I_1 provide much more information to help us to choose an appropriate model.

It is important to note that the parametric or semi-parametric models used for constructing the risk scores are simply approximations for the true models. If the “old” model does not fit the data well, it is difficult if not impossible to determine whether the improvement from the “new” model is the incremental value from the new predictors or a better model fitting.

The new proposal can be quite useful when the cost functions for obtaining new markers vary from region to region. The graphical displays at a subject-level can be utilized to identify subgroups of patients who would benefit from new markers under, for example, certain financial constrains in a specific region.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to the editor, associate editor and the reviewers for insightful comments on the article. This work was supported in part by grants R01-AI052817, U54-LM008748, RC4-CA155940 and R01-GM079330 from the National Institutes of Health. H. Uno thanks Department of Biostatistics and Computational Biology, Dana Farber Cancer Institute for the support from the Research Scientist Developmental Funds.

References

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.

Cai T, Cheng S. Robust combination of multiple diagnostic tests for classifying censored event times. Biostatistics. 2008; 9:216–233. [PubMed: 18056687]

Chang HY, Nuyten DSA, Sneddon JB, Hastie T, Tibshirani R, Sørlieb T, Dai H, He YD, van't Veer LJ, Bartelinke H, van de Rijnj M, Brownb PO, van de Vijver MJ. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. PNAS. 2005; 102:3738–43. [PubMed: 15701700]

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]

Cox DR. Regression Models and Life Tables" (with Discussion). Journal of the Royal Statistical Society. 1972; 34:187–220.Series B

D'Agostino RB. Risk prediction and finding new independent prognostic factors. Journal of Hypertension. 2006; 24:643–645. [PubMed: 16531791]

D'Agostino, RB.; Griffith, JL.; Schmidt, CH.; Terrin, N. Measures for evaluating model performance. Biometrics Section, Alexandria, VA, U.S.A.; Alexandria, VA: American Statistical Association, Biometrics Section; 1997. p. 253-258.

Dickson E, Fleming T, Grambsch P, Fisher L, Langworthy A. Prognosis in Primary Biliary Cirrhosis: Model for Decision Making. Hepatology. 1989; 10:1–7. [PubMed: 2737595]

Fleming, TR.; Harrington, DP. Counting Processes and Survival Analysis. John Wiley & Sons, Inc.; New York: 1991.

Gu W, Pepe M. Measures to summarize and compare the predictive capacity of markers. The International Journal of Biostatistics. 2009; 5:2454–2456.

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; 165:2454–2456. [PubMed: 16314539]

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]

Harrell FE, Lee KL, Mark DB. Tutorial in Biostatistics: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996; 15:361–87. [PubMed: 8668867]

Heagerty PJ, Zheng Y. Survival Model Predictive Accuracy and ROC Curves. Biometrics. 2005; 61:92–105. [PubMed: 15737082]

Hjort N. On inference in parametric survival data models. International Statistical Review. 1992; 60:355–87.

Park B, Kim W, Ruppert D, Jones M, Signorini D, Kohn R. Simple transformation techniques for improved non-parametric regression. Scandinavian Journal of Statistics. 1997; 24:145–163.

Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Statistics in Medicine. 2004; 23:2109–23. [PubMed: 15211606]

Pencina MJ, D'Agostino RB Sr. D'Agostino RB Jr. Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond (with Commentaries & Rejoinder). Statistics in Medicine. 2008; 27:157–212. [PubMed: 17569110]

Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statistics in Medicine. 2010 online ahead print.

Pepe, MS. The statistical evaluation of medical tests for classification and prediction. Oxford University Press; New York: 2003.

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; 159:882–890. [PubMed: 15105181]

Tian L, Cai T, Wei LJ. Identifying subjects who benefit from additional information for better prediction of the outcome variables. Biometrics. 2009; 65:894–902. [PubMed: 18945268]

Uno H, Cai T, Tian L, Wei LJ. Evaluating prediction rules for t-year survivors with censored regression models. Journal of the American Statistical Association. 2007; 102:527–537.

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 online ahead print.

Uno, H.; Tian, L.; Cai, T.; Kohane, IS.; Wei, LJ. Comparing Risk Scoring Systems Beyond the ROC Paradigm in Survival Analysis. Harvard University Biostatistics. 2009. Working Paper SeriesWorking Paper 107. http://www.bepress.com/harvardbiostat/paper107.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415:530–6. [PubMed: 11823860]

van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. The New England Journal of Medicine. 2002; 347:1999–2009. [PubMed: 12490681]

Wand M, Marron J, Ruppert D. Transformation in density estimation (with comments). Journal of the American Statistical Association. 1991; 36:343–361.

Ware JH. The limitations of risk factors as prognostic tools. The New England Journal of Medicine. 2006; 355:2615–2617. [PubMed: 17182986]

Yu K, Jones MC. Local linear quantile regression. Journal of the American Statistical Association. 1998; 93:228–237.

Zhou, XH.; Obuchowski, NA.; McClish, DK. Statistical methods in diagnostic medicine. Wiley Interscience; New York: 2002.

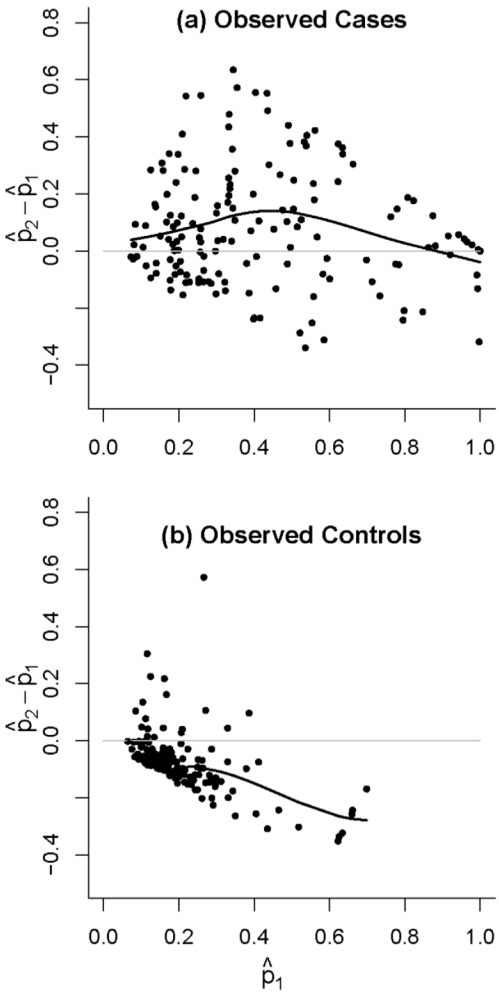

Figure 1.

Scatterplots of the risk score without bilirubin, $\widehat{p}_1$, (x-axis) versus the difference between the risk scores with and without bilirubin, $\widehat{p}_2 - \widehat{p}_1$, (y-axis) and estimated lowess curves (solid line): (a) subjects who died by Year 5; (b) subjects who survived beyond Year 5.

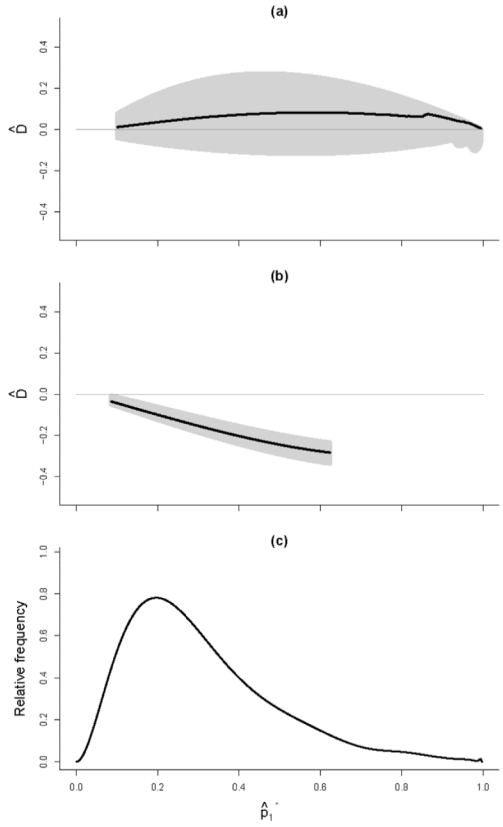

Figure 2.
Estimated quartiles — median (solid line); 25th and 75th percentiles (upper and lower boundaries of the shaded area) — for the conditional distributions for the diffrences between the risk scores with and without bilirubin at Year 5. (a) For subjects who would die by Year 5; (b) For subjects who would survive beyond Year 5; (c) Estimated density function of the score without using bilirubin.

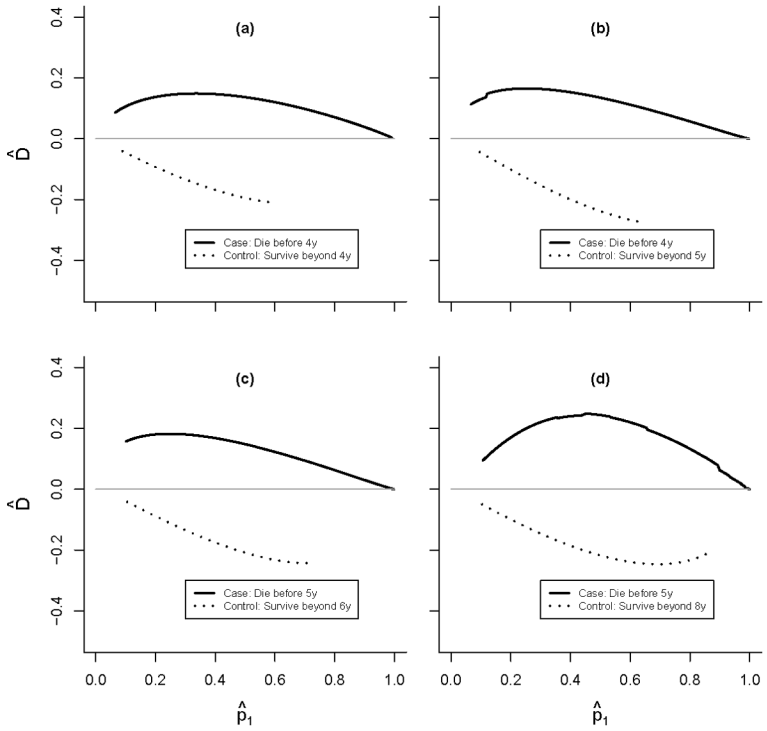

Figure 3.

Estimated median curves for the conditional distributions for the differences between the risk scores with and without bilirubin for various sets of time intervals I_0 and I_1. (a) $I_1 = (0, 4]$, $I_0 = (4, \infty)$; (b) $I_1 = (0, 4]$, $I_0 = (5, \infty)$; (c) $I_1 = (0, 5]$, $I_0 = (6, \infty)$; and (d) $I_1 = (0, 5]$, $I_0 = (8, \infty)$.

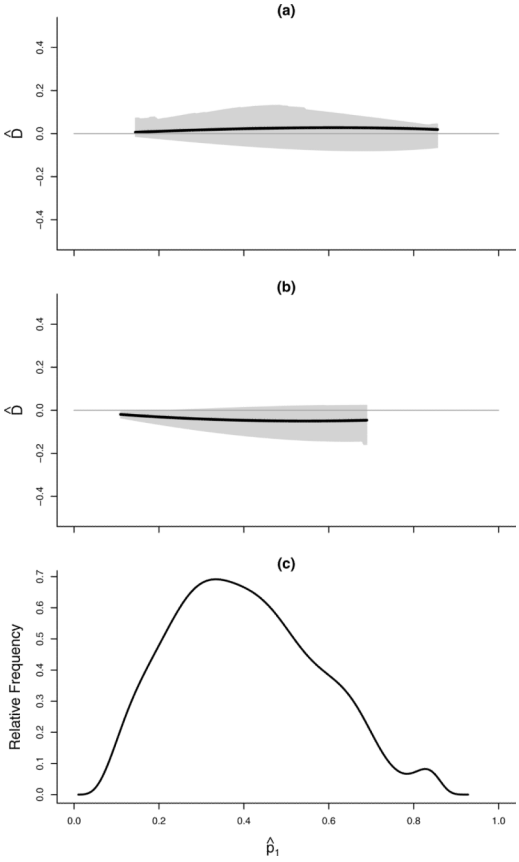

Figure 4.
Estimated quartiles — median (solid line); 25th and 75th percentiles (upper and lower boundaries of the shaded area) — for the conditional distributions for the differences between the risk scores with and without gene score at Year 10. (a) For subjects who would die by Year 10; (b) For subjects who would survive beyond Year 10; (c) Estimated density function of the score without using gene score.

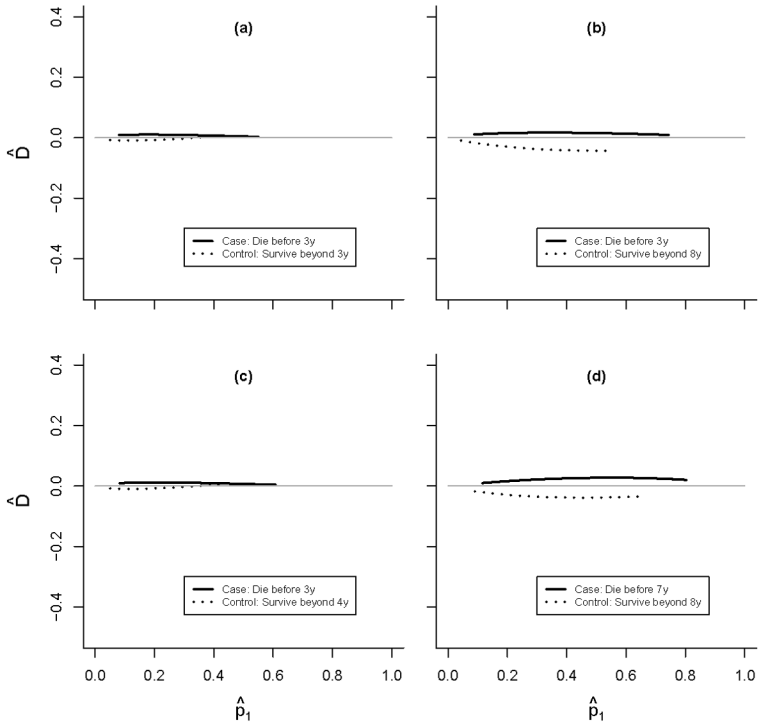

Figure 5.

Estimated median curves for the conditional distributions for the differences between the risk scores with and without gene score for various sets of time intervals I_0 and I_1. (a) I_1 = (0, 3], I_0 = (3, ∞); (b) I_1 = (0, 3], I_0 = (8, ∞); (c) I_1 = (0, 3], I_0 = (4, ∞); and (d) I_1 = (0, 7], I_0 = (8, ∞).

Table 1

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

	Without gene score		With gene score	
	Est (s.e.)[(1)]	**p**[(2)]	**Est (s.e.)**	**p**
Age/10 [yrs]	−0.47 (0.17)	¡0.01	−0.57 (0.18)	¡0.01
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)	¡0.01	0.74 (0.35)	0.04
Grade = 3 vs 1	1.11 (0.35)	¡0.01	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)	¡0.01

[(1)] Estimate (Standard error estimate)

[(2)] p-value