STRATIFYING SUBJECTS FOR TREATMENT SELECTION WITH CENSORED EVENT TIME DATA FROM A COMPARATIVE STUDY

L. Zhao[1], T. Cai[1], L. Tian[2], H. Uno[1,3], S. D. Solomon[4], and L. J. Wei[1]

SUMMARY

The conventional approach to comparing a new treatment with a standard therapy is often based on a summary measure for the treatment difference over the entire study population. A positive trial with respect to such a global measure, however, does not mean that all individual future patients would benefit from the new treatment. On the other hand, a negative finding may not be sufficiently conclusive to claim that the new treatment is entirely futile. In this article, we propose a systematic approach to identify future patients who would benefit from the new treatment with respect to an event time outcome via a two-stage inference procedure. We first develop a scoring index to stratify study patients based on parametric or semiparametric survival models with the observed event times and covariates. We then use a nonparametric method to estimate the average treatment difference for each stratum defined by the score. Sampling variation of the resulting estimator is also provided across the entire spectrum of the score by controlling certain local and global error rates. With a numerical study, we show that the new proposal performs well under various practical settings. Our method is illustrated with the data from a recent clinical trial to evaluate whether a specific anti-hypertensive drug would prolong the lives for patients with stable coronary artery disease and normal or slightly reduced left ventricular function.

Keywords: Cox's model; Nonparametric function estimation; Personalized medicine; Perturbation-resampling method; Stratified medicine; Subgroup analysis; Survival analysis.

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

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

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

[4]Cardiovascular Division, Brigham and Women's Hospital, Boston, MA 02115, U.S.A.

1. INTRODUCTION

In a typical randomized clinical trial comparing a new treatment with a standard therapy, the study participants are generally quite heterogeneous and their responses can be drastically different. However, the design and monitoring of the study are often guided by a summary treatment difference measure for the entire study population. The determination of whether the new treatment is superior to the control is usually based on the results from statistical inferences about such a summary measure. This "one-size-fits-all" approach may not be adequate for evaluating a new drug or device. A "positive" trial, which shows a treatment benefit with respect to this global measure, does not imply that all future patients would benefit from the new treatment. On the other hand, a "negative" study does not mean that all future patients should take the standard therapy. As an example, consider a recent clinical trial "Prevention of Events with Angiotensin Converting Enzyme Inhibition (PEACE)" to study whether the ACE inhibitors (ACEi) are effective for reducing certain future cardiovascular-related events for patients with stable coronary artery disease and normal or slightly reduced left ventricular function (Braunwald et al., 2004). In this study, 4158 and 4132 patients were randomly assigned to the ACEi treatment and placebo arms, respectively. One main endpoint for the study was the patient's survival time. The median follow-up time was 4.8 years. By the end of the study, 334 and 299 deaths occurred in the control and treatment arms, respectively. As shown in Figure 1, no differences between the Kaplan-Meier curves of the two groups are apparent except for the "unstable" tail parts. The proportional hazards ratio estimate is 0.89 with a 0.95 confidence interval of (0.76, 1.04). Based on the results of this study, it is not clear whether ACEi therapy would help the patient with respect to overall mortality. However, with further analysis of the PEACE survival data, Solomon et al. (2006) reported that the ACEi might significantly prolong survival for the patient whose kidney function at the study entry time was not normal (for example, the estimated glomerular filtration rate, eGFR, < 60). This finding can be quite useful in practice. On the other hand, such a subgroup analysis

has to be executed properly and the results of such analysis have to be interpreted cautiously (Rothwell, 2005; Pfeffer and Jarcho, 2006; Wang et al., 2007). Moreover, post ad hoc subgroup analyses are often conducted by examining the interaction between the treatment indicator variable and *each* covariate. Such a procedure can be quite inefficient.

When there is a single baseline covariate, novel methods for identifying a subgroup of patients who would benefit from the new treatment have been proposed by Song and Pepe (2004), and Bonetti and Gelber (2000, 2005). In this article, we consider the case for censored survival data with multiple covariates. In theory, one may use a nonparametric function estimation procedure to make inferences about the subject-specific treatment differences. However, when there is more than one covariate involved, such a procedure performs rather poorly. In this paper, we first fit the data with a parametric or semi-parametric survival model for each treatment group. If the models are correctly specified, one may make valid inferences about the subject-specific treatment differences directly from such parametric analysis. Although these working models are likely misspecified, they can be quite useful to group subjects with similar treatment difference profiles. In this paper, we stratify subjects with a univariate scoring system constructed from these parametric models. The score is the subject-specific parametric treatment difference estimate. Subjects in each stratum would have the same score. Next, we utilize a univariate nonparametric function estimation method to make inferences about the stratum-specific treatment differences across the entire spectrum of the score, for example, via pointwise and simultaneous confidence interval estimates. Conceptually, our approach is similar to that taken by Cai et al. (2010b), which dealt with non-censored observations. The derivation of the procedure in the presence of censoring, however, is complex and technically involved. We illustrate the new proposal with the data from the PEACE study. Furthermore, via a simulation study, we show that our procedure performs well under various practical settings.

2. STRATIFYING SUBJECTS WITH A PARAMETRIC SCORING SYSTEM WITH RESPECT TO TREATMENT DIFFERENCE

In this section, we show how to construct the standard parametric or semi-parametric estimates for the treatment differences with subject level data on the response and baseline covariates. To this end, for a subject assigned to treatment k, let $\tilde{T}_k$ be its time to a specific event and U_k be the corresponding baseline covariate vector, where $k = 1, 2$. The event time $\tilde{T}_k$ may be censored by C_k, which is assumed to be independent of $\tilde{T}_k$ and U_k. Instead of observing $\tilde{T}_k$ directly, one observes $T_k = \min(\tilde{T}_k, C_k)$ and $\Delta_k = I(\tilde{T}_k \leq C_k)$, where $I(\cdot)$ is the indicator function. For subjects with a given covariate vector $U = u$, let $S_k(t; u) = \text{pr}(\tilde{T}_k > t | U_k = u)$ be its survival probability at time t if assigned to treatment k, $k = 1, 2$. To quantify the treatment contrast for these subjects, one may use the difference of two survival rates at t, $D(t; u) = S_2(t; u) - S_1(t; u)$. Alternatively, one may consider an *integrated* or *average* survival rate difference over a time interval $[t_0, t_1]$:

$$D(u) = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} D(t; u) dt \tag{2.1}$$

(Pepe and Fleming, 1989, 1991; Murray and Tsiatis, 1999; Zhao et al., 2010).

Suppose that the data from the kth treatment group consist of $\{(T_{ki}, \Delta_{ki}, U_{ki}); i = 1, \ldots, n_k\}$, which are n_k independent and identical copies of (T_k, Δ_k, U_k), for $k = 1, 2$. We assume that $\pi_k = \lim_{n \to \infty} n_k/n > 0$ for $k = 1, 2$, where $n = n_1 + n_2$. In theory, one may use a nonparametric function procedure to estimate $D(u)$ consistently. However, when U is not univariate, such a fully non-parametric approach is difficult, if not impossible, to estimate (2.1) well in practice. A more feasible approach is to utilize a parametric or semi-parametric model which approximates the relationship between the response variable and the covariate vector. Here, for each treatment group, we fit the data with a standard Cox proportional

hazards (Cox, 1972) *working* model:

$$S_k(t; U_k) = g\left\{\log \Lambda_k(t) + \beta_k' Z_k\right\}, \quad k = 1, 2, \tag{2.2}$$

where $g(x) = e^{-e^x}$, Z_k, a $p \times 1$ vector, is a known function of U_k, $\Lambda_k(\cdot)$ is the unknown baseline cumulative hazard function, and β_k is an unknown $p \times 1$ vector of regression parameters. To estimate $S_k(t; u)$, we first obtain an estimator $\hat{\beta}_k$ for β_k via the partial likelihood score equation truncated at time t_1 :

$$\sum_{i=1}^{n_k} \int_0^{t_1} \left\{ Z_{ki} - \frac{\sum_{j=1}^{n_k} Y_{kj}(t) e^{\beta_k' Z_{kj}} Z_{kj}}{\sum_{j=1}^{n_k} Y_{kj}(t) e^{\beta_k' Z_{kj}}} \right\} dN_{ki}(t) = 0, \tag{2.3}$$

where $N_{ki}(t) = I(T_{ki} \leq t)\Delta_{ki}$ and $Y_{ki}(t) = I(T_{ki} \geq t)$, for $i = 1, \ldots, n_k$. We then estimate the function $\Lambda_k(t)$ in (2.2) by the standard Breslow's estimator (Kalbfleisch and Prentice, 2002):

$$\hat{\Lambda}_k(t) = \sum_{i=1}^{n_k} \int_0^t \frac{dN_{ki}(s)}{\sum_{j=1}^{n_k} Y_{kj}(s) e^{\hat{\beta}_k' Z_{kj}}}.$$

Note that $\hat{\beta}_k$ and $\hat{\Lambda}_k(t)$ consistently estimate their true counterparts when Model (2.2) is correctly specified. When Model (2.2) is misspecified, under a rather mild regularity condition, $\hat{\beta}_k$ converges to a finite constant β_{0k} and $\hat{\Lambda}_k(t)$ to a deterministic function $\Lambda_{0k}(t)$, as $n_k \to \infty$ (Hjort, 1992; Cai et al., 2010a). This stability property is critical for developing our inference procedures. It follows that a model based estimator for $D(u)$ is

$$\hat{D}(u) = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \left[g\left\{\log \hat{\Lambda}_2(t) + \hat{\beta}_2' z\right\} - g\left\{\log \hat{\Lambda}_1(t) + \hat{\beta}_1' z\right\} \right] dt.$$

Again, $\hat{D}(u)$ converges in probability to a deterministic function $\bar{D}(u)$, even when Model (2.2) is misspecified. When Model (2.2) is correctly specified, $\hat{D}(u)$ is consistent for $D(u)$.

Now, let U^0 be the baseline covariate vector of a future subject from a population similar to the study population. Suppose that the event time of this subject is $\tilde{T}_k^0$ if treated by

treatment k, $k = 1, 2$. For a given U^0, one may use $\hat{D}(U^0)$ to decide which treatment should be assigned to this specific subject. However, the adequacy of such a decision heavily depends on the appropriateness of Model (2.2). On the other hand, even if $\hat{D}(u)$ does not approximate $D(u)$ well, $\hat{D}(\cdot)$ can be quite useful as an index system for clustering future subjects with potentially similar treatment differences. Thus, we propose to stratify future subjects based on their values of $\hat{D}(U^0)$ and non-parametrically estimate the mean value of $D(U^0)$ for each stratum $\{U^0 : \hat{D}(U^0) = v\}$, where v is any given possible value of the estimated score.

3. NONPARAMETRIC POINT AND INTERVAL ESTIMATION FOR THE MEAN VALUE OF $D(\cdot)$ FOR FUTURE SUBJECTS WITH THE SAME ESTIMATED PARAMETRIC SCORE

The average value of $D(U^0)$ for the aforementioned subgroup of subjects, whose estimated index score is v, is

$$\mathcal{D}(v) = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \left[\text{pr} \left\{ \tilde{T}_2^0 > t | \bar{D}(U^0) = v \right\} - \text{pr} \left\{ \tilde{T}_1^0 > t | \bar{D}(U^0) = v \right\} \right] dt,$$

where the probabilities are with respect to $(\tilde{T}^0, U^0)$ and $\{(T_{ki}, \Delta_{ki}, U_{ki}); k = 1, 2, i = 1, \ldots, n_k\}$. To estimate $\mathcal{D}(v)$, let

$$\Lambda_{k,v}(t) = -\log \left[\text{pr} \left\{ \tilde{T}_k^0 > t | \hat{D}(U^0) = v \right\} \right], \quad 0 \le t \le t_1,$$

be the cumulative hazard function for future subjects with estimated score $\hat{D}(U^0) = v$. As in Cai et al. (2010a), we use a nonparametric kernel Nelson-Aalen estimator smoothed over v for estimating $\Lambda_{k,v}(t)$ based on the triplets $\{(T_{ki}, \Delta_{ki}, \hat{D}(U_{ki})), i = 1, \ldots, n_k\}$. Specifically, we consider the class of potential estimators which are step functions over t and only jump at the observed event time points with jump sizes $d\Lambda_{k,v}(t) = \Lambda_{k,v}(t) - \Lambda_{k,v}(t-)$. Then a

nonparametric functional estimator for $d\Lambda_{k,v}(t)$ can be obtained by minimizing

$$\sum_{i=1}^{n_k} K_{h_k}(\hat{Q}_{ki,v}) \left\{dN_{ki}(t) - Y_{ki}(t)d\Lambda_{k,v}(t)\right\}^2,$$

where $dN_{ki}(t) = N_{ki}(t) - N_{ki}(t-)$, $K(\cdot)$ is a smooth density function, $K_{h_k}(x) = K(x/h_k)/h_k$, h_k is a bandwidth such that $h_k \to 0$ and $nh_k^2 \to \infty$, as $n \to \infty$, $\hat{Q}_{ki,v} = \psi\{\hat{D}(U_{ki})\} - \psi(v)$, and $\psi(\cdot)$ is a known increasing transformation function. In practice, a proper choice of $\psi(\cdot)$ can be quite helpful for increasing precision of the nonparametric function estimation procedure (Wand et al., 1991; Park et al., 1997; Cai et al., 2010a). The resulting estimator for $\Lambda_{k,v}(t)$ is

$$\hat{\Lambda}_{k,v}(t) = \int_0^t \frac{\sum_{i=1}^{n_k} K_{h_k}(\hat{Q}_{ki,v})dN_{ki}(s)}{\sum_{i=1}^{n_k} K_{h_k}(\hat{Q}_{ki,v})Y_{ki}(s)}. \tag{3.1}$$

We can then estimate $\mathcal{D}(v)$ by

$$\hat{\mathcal{D}}(v) = \frac{1}{t_1 - t_0}\int_{t_0}^{t_1} \left\{e^{-\hat{\Lambda}_{2,v}(t)} - e^{-\hat{\Lambda}_{1,v}(t)}\right\} dt. \tag{3.2}$$

In Appendix A, we show that under some mild regularity conditions, $\hat{\mathcal{D}}(v)$ is uniformly consistent for $\mathcal{D}(v)$, for $v \in \mathcal{J}$, an interval properly contained in the support of the estimated score $\hat{D}(\cdot)$ when $h_k = O(n^{-\nu})$ with $1/5 < \nu < 1/2$. For any fixed $v \in \mathcal{J}$, using a similar argument in Cai et al. (2010a), we show in Appendix A that

$$(n_1h_1 + n_2h_2)^{1/2}\left\{\hat{\mathcal{D}}(v) - \mathcal{D}(v)\right\} \tag{3.3}$$

converges in distribution to a mean zero normal random variable as $n \to \infty$. To approximate the distribution of (3.3), we utilize a perturbation-resampling procedure which is similar to the so-called wild bootstrapping (Wu, 1986; Mammen, 1992), and has been successfully applied to a number of estimation problems, especially in survival analysis (Lin et al., 1993; Park and Wei, 2003; Cai et al., 2005). Specifically, let $\{V_{ki} : k = 1, 2, i = 1, \ldots, n_k\}$ be a random sample

from the distribution of a positive random variable with mean and variance of one, which is independent of the data. Let

$$\hat{\mathcal{D}}^*(v) = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \left\{ e^{-\hat{\Lambda}^*_{2,v}(t)} - e^{-\hat{\Lambda}^*_{1,v}(t)} \right\} dt, \tag{3.4}$$

where, for the standard perturbation method, $\hat{\Lambda}^*_{k,v}(t)$ is defined as

$$\hat{\Lambda}^*_{k,v}(t) = \int_0^t \frac{\sum_{i=1}^{n_k} V_{ki} K_{h_k}(\hat{Q}^*_{ki,v}) dN_{ki}(s)}{\sum_{i=1}^{n_k} V_{ki} K_{h_k}(\hat{Q}^*_{ki,v}) Y_{ki}(s)}, \tag{3.5}$$

$$\hat{Q}^*_{ki,v} = \psi\{\hat{D}^*(U_{ki})\} - \psi(v),$$

$$\hat{D}^*(U_{ki}) = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1} \left[g\left\{ \log \hat{\Lambda}^*_2(t) + \hat{\beta}^{*\prime}_2 Z_{ki} \right\} - g\left\{ \log \hat{\Lambda}^*_1(t) + \hat{\beta}^{*\prime}_1 Z_{ki} \right\} \right] dt,$$

$\hat{\beta}^*_k$ is the solution to

$$\sum_{i=1}^{n_k} \int_0^{t_1} V_{ki} \left\{ Z_{ki} - \frac{\sum_{j=1}^{n_k} V_{kj} Y_{kj}(t) e^{\beta'_k Z_{kj}} Z_{kj}}{\sum_{j=1}^{n_k} V_{kj} Y_{kj}(t) e^{\beta'_k Z_{kj}}} \right\} dN_{ki}(t) = 0,$$

and

$$\hat{\Lambda}^*_k(t) = \sum_{i=1}^{n_k} \int_0^t \frac{V_{ki} dN_{ki}(s)}{\sum_{j=1}^{n_k} V_{kj} Y_{kj}(s) e^{\hat{\beta}^{*\prime}_k Z_{kj}}}.$$

We show in Appendix B that conditional on the data, the limiting distribution of

$$(n_1 h_1 + n_2 h_2)^{1/2} \left\{ \hat{\mathcal{D}}^*(v) - \hat{\mathcal{D}}(v) \right\} \tag{3.6}$$

is the same as the unconditional limiting distribution of (3.3). Note $\hat{\mathcal{D}}(v)$ converges at a rate slower than $n^{-1/2}$ and thus the variation due to $\{\hat{\beta}_k, \hat{\Lambda}_k(\cdot)\}$ is asymptotically negligible. However, we find in the literature that incorporating the variability due to $\hat{\beta}_k$ and $\hat{\Lambda}_k(\cdot)$ in the perturbation process can significantly improve the approximation to the distribution of (3.3) for settings with practical sample sizes. Moreover, through our numerical study reported in

section 5, when the study sample size is not large or the event rate is low, we find that the above resampling method and the standard bootstrapping may result in conservative interval estimates, that is, their coverage levels tend to be larger than the nominal counterparts. In Appendix C, we present a simple modified perturbation-resampling version of $\hat{\Lambda}^*_{k,v}(t)$. This modification may substantially reduce the conservativeness of the resulting interval estimation procedure for finite sample cases. Such a modification preserves all the large sample properties for (3.4). For the rest of the paper, we utilize this modified version in our presentation and analysis.

With the above large sample approximation, for any fixed $v \in \mathcal{J}$, one may obtain a variance estimate of the distribution of (3.3), denoted by $\hat{\sigma}^2(v)$, based on the empirical variance of, say, M perturbation samples. It follows that for any given $\alpha \in (0,1)$, a two-sided $1-\alpha$ confidence interval for $\mathcal{D}(v)$ is

$$\left(\hat{\mathcal{D}}(v) - z_{(1-\alpha/2)}(n_1h_1+n_2h_2)^{-1/2}\hat{\sigma}(v), \hat{\mathcal{D}}(v) + z_{(1-\alpha/2)}(n_1h_1+n_2h_2)^{-1/2}\hat{\sigma}(v)\right), \tag{3.7}$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

To make inference about the subject-specific treatment differences over a range of risk scores v's, one may construct simultaneous confidence intervals for $\{\mathcal{D}(v), v \in \mathcal{J}\}$. However, for the present case, we cannot use the conventional method based on a sup-type statistic:

$$W = \sup_{v\in\mathcal{J}} \left| \frac{(n_1h_1+n_2h_2)^{1/2}\left\{\hat{\mathcal{D}}(v) - \mathcal{D}(v)\right\}}{\hat{\sigma}(v)} \right|, \tag{3.8}$$

due to the fact that as a process in v, the limiting distribution of (3.3) does not exist (Cai et al., 2010a). On the other hand, by a strong approximation theory (Bickel and Rosenblatt, 1973), in Appendix B, we show that a standardized version of W converges in distribution to a proper random variable. Thus in practice, for large n, one may approximate the distribution of W by its empirical counterpart W^*, based on the same set of aforementioned perturbation

variables $\{V_{ki}; k = 1, 2, i = 1, \ldots, n_k\}$ simultaneously for all $v \in \mathcal{J}$. It follows that a $1 - \alpha$ simultaneous confidence interval for $\mathcal{D}(v)$ is

$$\left(\hat{\mathcal{D}}(v) - c_\alpha(n_1h_1 + n_2h_2)^{-1/2}\hat{\sigma}(v), \hat{\mathcal{D}}(v) + c_\alpha(n_1h_1 + n_2h_2)^{-1/2}\hat{\sigma}(v)\right), \tag{3.9}$$

where c_α is chosen such that $P(W^* \leq c_\alpha) \geq 1 - \alpha$.

As for any nonparametric functional estimation problem, the choice of the smoothing parameters h_1 and h_2 are crucial for making inference about $\mathcal{D}(v)$. Here, via a standard K-fold cross-validation, we obtain the smooth parameters by minimizing the sum of integrated squared martingale residuals (Tian et al., 2005; Cai et al., 2010a). Specifically, to choose h_1, we randomly split the data into K disjoint subsets of about equal sizes, denote the subjects that are assigned to treatment group k and also in the rth subset by $\mathcal{I}_{kr}$, $k = 1, 2; r = 1, \ldots, K$. For each r, we use all the data except the rth subset to build the score and estimate $\Lambda_{1,v}(t)$ with a given h_1. Let the resulting estimator be $\hat{\Lambda}^{(r)}_{1,v}(t)$. We then use the observations from $\mathcal{I}_{1r}$ to obtain the sum of integrated squared martingale residuals

$$\int_0^{t_1} \sum_{j \in \mathcal{I}_{1r}} \left\{ N_{1j}(t) - \int_0^t Y_{1j}(s) d\hat{\Lambda}^{(r)}_{1,\hat{v}_{1j}}(s) \right\}^2 d\left\{ \sum_{i \in \mathcal{I}_{1r}} N_{1i}(t) \right\}, \tag{3.10}$$

where $\hat{v}_{1j} = \hat{D}(u_{1j})$ is the score for a subject with covariate vector u_{1j} estimated using all the data except the rth subset. Lastly, we sum (3.10) over r from 1 to K, and choose $\hat{h}_1$, which minimizes the summation. The smooth parameter $\hat{h}_2$ is chosen similarly. Note that the above empirically selected bandwidths are of order $n^{-1/5}$ (Fan and Gijbels, 1995). To ensure the validity of the aforementioned large sample properties for $\hat{\mathcal{D}}(v)$, in practice we choose the smooth parameters values h's for the nonparametric function estimates by multiplying $\hat{h}$'s with $n^{-\xi}$, where ξ is a small positive number such that $\xi < 3/10$.

4. ILLUSTRATION OF THE PROPOSAL WITH THE DATA FROM PEACE STUDY

We illustrate the new proposal with data from the PEACE study discussed in the Introduction section. Here, patients received placebo in group 1 and ACEi in group 2. For illustration, we let the time interval for the integrated difference of survival rates be $[t_0, t_1] = [60, 72]$ (months) and let $U = Z$, which consists of seven baseline covariates previously identified as statistically and clinically important predictors of the overall mortality (Solomon et al., 2006). These covariates are eGFR, age, gender, left ventricular ejection fraction (lveejf), history of hypertension (yes or no), diabetes (yes or no), and history of myocardial infarction (yes or no). To construct the parametric scoring system, we fitted a Cox model to the mortality data from each treatment group with the above seven covariates. In our analysis, we included all patients ($n = 7865$) who had complete information of these seven covariates. Table 1 gives us the estimated regression coefficients and their standard error estimates. The empirical distribution function of the parametric score $\mathcal{D}(\cdot)$ is given in Figure 2(a). Note that the scores for the majority of the study subjects are between -0.02 and 0.06. If the Cox models are correctly specified, future patients whose scores are greater than zero would benefit from the new treatment.

To obtain a nonparametric estimate for $\mathcal{D}(v)$, we let $K(\cdot)$ be the Epanechnikov kernel, and $\psi(v)$ be the identity function. The smoothing parameters were chosen by minimizing (3.10) with a 10-fold cross validation procedure. We then multiplied the above minimizers by $n^{-0.05}$ as the final smoothing parameter values. Furthermore, we chose the 5th and 95th percentiles of the empirical distribution based on $\{\hat{D}(U_{ki}), k = 1, 2; i = 1, \ldots, n_k\}$ as the boundary points for interval $\mathcal{J}$. To approximate the distributions of (3.3) and W, we used the perturbation-resampling method with $M = 1000$ independent realizations of the random sample from the standard exponential distribution. In Figure 2 (b), we report the point estimate for each treatment group with respect to the group-specific integrated survival rate over [60, 72]. The

estimated integrated difference is reported in Figure 2 (c) (solid curve). Note that if the parametric models fit the data well, one expect that these point estimates would be close the 45° line. The dotted curves in the figure are the boundaries of the pointwise 0.95 confidence intervals and the shaded area is the 0.95 simultaneous confidence band. From a conservative view (using the simultaneous confidence band), patients whose scores are beyond 0.02 would benefit from ACEi with respect to overall mortality. If the drug is safe and not costly, one may recommend the treatment for patients whose scores are larger than zero, since the confidence band is relatively tight in that neighborhood and its lower bound is quite close to 0.

5. A SIMULATION STUDY FOR EVALUATING THE PERFORMANCE OF THE NEW INTERVAL ESTIMATION PROCEDURE

We conducted a simulation study to examine the performance of the proposed inference procedures. We found that the proposed pointwise and simultaneous interval estimators behave well under various practical settings. That is, the empirical coverage probabilities for the interval estimators preserve their nominal levels. For example, in one of our simulation setups, we mimicked the PEACE study and generated survival data for each treatment group based on a Weibull model. The parameters of this Weibull model are obtained by fitting the PEACE data with a Weibull using the aforementioned seven covariates via the maximum likelihood method. To generate covariate vector U, first we simulated the discrete variables from their empirical distribution observed in the PEACE data set. Conditional on these discrete covariates, we generated the continuous covariates from a multivariate normal whose mean and covariance matrix were estimated empirically using PEACE data. For each treatment group, we used the above Weibull survival model to generate the survival time $\tilde{T}$ for a patient with a given realization of U. Furthermore, the censoring is generated based on the observed Kaplan-Meier curve for each treatment group.

For ease of computation, the bandwidth for constructing the nonparametric estimate was fixed and chosen as the average of the bandwidths selected based on (3.10) with $\xi = 0.05$

from the first 10 simulated datasets. We computed the empirical coverage probabilities of the pointwise and simultaneous confidence interval estimators for the integrated survival rate difference over [60,72]. The results from 1000 replicates for cases with sample sizes of $n_1 = n_2 = 4000$ and $n_1 = n_2 = 8000$ are summarized in Figure 3. The pointwise empirical coverage levels tend to be slightly higher than their nominal levels for $n_1 = n_2 = 4000$. The degree of conservativeness of our interval estimators appears to be decreasing with larger sample sizes (see, for example, $n_1 = n_2 = 8000$). The coverage levels of the 0.95 simultaneous confidence interval estimators are 0.977 with $n_1 = n_2 = 4000$ and 0.961 with $n_1 = n_2 = 8000$ for the average survival rate difference over [60, 72].

6. REMARKS

In this article, we used an integrated (or average) difference of survival rates over a specific time interval to quantify the treatment contrast. This measure is purely nonparametric and has an intuitive interpretation even when the differences of two survival rates are not constant over time. Moreover, this average quantity provides an overall difference of two survival curves when our interest is not restricted to the survival rate difference at a specific time point. Alternatively, one may use the conventional two-sample Cox's proportional hazards estimate to quantify the treatment difference. Unfortunately, it is not clear how to interpret this estimate when the proportional hazards model assumption is violated (Prentice and Kalbfleisch, 1981; Lin and Wei, 1989, Xu and O'Quigley, 2000).

In this paper, we assume that the set of baseline covariates is given and we stratify future patients with such covariates without involving a variable selection process. If the dimension of the baseline covariate vector is large, the usual variable selection procedure to identify the treatment and covariate interactions for constructing a scoring system can be rather inefficient or unstable. In his unpublished Ph.D. thesis, Signorovitch (2007) proposed a novel method for modeling a treatment contrast measure directly with covariates. Heuristically his approach is more efficient for locating important treatment and covariate interactions than the above

conventional variable selection procedure. Further research is needed along this line with censored event time data.

For evaluating different prediction models in a typical one-sample problem, there are various novel criteria available in the literature (Pepe, 2003; Tian et al., 2007). However, for the present problem with two treatment groups involved, it is rather difficult, if not impossible, to utilize these conventional methods for evaluating the scoring systems for treatment selections. The problem is that for each patient in the validation sample, she/he can only receive a single treatment, not both. Therefore, one cannot compare the observed treatment difference and its predicted counterpart at the individual patient level. On the other hand, heuristically for a good system, the distribution of its score would spread out over a large support. Moreover, this system would produce tight interval estimates like those presented in Figure 2 (c). We plan to pursue this challenging research problem in the future.

Although our method is valid when dealing with a given set of covariates, generally we must undertake a nontrivial variable selection process before considering the final parametric models to construct the scoring system. Therefore, it would be ideal to have a clearly defined proposal for implementing a systematic procedure including model building and selection for stratified medicine at the design stage of the clinical study.

APPENDIX A: ASYMPTOTIC PROPERTIES OF $\hat{\mathcal{D}}(V)$

We assume that $\pi_k = \lim_{n\to\infty} n_k/n > 0$ for $k = 1, 2$. Without loss of generality, we let ψ be the identity function. Assume that both h_1 and h_2 are of order $O(n^{-\nu})$ with $1/5 < \nu < 1/2$. For the ease of presentation and without loss of generality, we assume that $h_1 = h_2$, which is denoted by h. Let $\bar{D}(\cdot)$ be the limit of $\hat{D}(\cdot)$, $\zeta_k(\cdot)$ be the density function of $\bar{D}(U_k)$ and $H_{k,v}(s) = pr(T_k \geq s|\bar{D}(U_k) = v)$, $k = 1, 2$. Let $K(\cdot)$ be a symmetric smooth kernel function with a bounded support $[-1, 1]$. Let $m_2 = \int_{-1}^{1} K^2(x)dx$. In addition, assume that the covariate vector U_k is bounded, $k = 1, 2$.

We first show that $\hat{\mathcal{D}}(v)$ is uniformly consistent for $\mathcal{D}(v)$, for $v \in \mathcal{J} = [\rho_1, \rho_2]$, an interval

which is properly contained in the support of the estimated score $\hat{D}(\cdot)$. To this end, let

$$\tilde{\Lambda}_{k,v}(t) = \int_0^t \frac{\sum_{i=1}^{n_k} K_h(\bar{D}(U_{ki}) - v) dN_{ki}(s)}{\sum_{i=1}^{n_k} K_h(\bar{D}(U_{ki}) - v) Y_{ki}(s)},$$

Then it follows from an integration by part and similar arguments given in Cai et al. (2010a) that

$$\sup_{v\in\mathcal{J}, t\in[t_0,t_1]} (n_k h)^{\frac{1}{2}} \left|\hat{\Lambda}_{k,v}(t) - \tilde{\Lambda}_{k,v}(t)\right| = o_p(n_k^{-\frac{1}{4}} h^{-\frac{1}{2}} \log(n_k)), \quad k = 1, 2. \tag{A.1}$$

On the other hand, the arguments given in Li and Doss (1995) can be used to show that $\sup_{v\in\mathcal{J}, t\in[t_0,t_1]} \left|\tilde{\Lambda}_{k,v}(t) - \Lambda_{k,v}(t)\right| = O_p\{(n_k h_k)^{-1/2} \log(n_k)\}$. Thus it follows that

$$\sup_{v\in\mathcal{J}, t\in[t_0,t_1]} \left|\hat{\Lambda}_{k,v}(t) - \Lambda_{k,v}(t)\right| = O_p\{(n_k h_k)^{-1/2} \log(n_k) + (n_k h_k)^{-1/2} n_k^{-1/4} h^{-1/2} \log(n_k)\}$$

in probability as $n_k \to \infty$. In view of (3.2), we have $\sup_{v\in\mathcal{J}} \left|\hat{\mathcal{D}}(v) - \mathcal{D}(v)\right| \to 0$, in probability as $n \to \infty$, which concludes the uniformly consistency of $\hat{\mathcal{D}}(v)$.

We next derive the asymptotic distribution of $(nh)^{1/2} \left\{\hat{\mathcal{D}}(v) - \mathcal{D}(v)\right\}$. From (A.1), we have

$$(n_k h)^{\frac{1}{2}} \left\{\hat{\Lambda}_{k,v}(t) - \Lambda_{k,v}(t)\right\} = (n_k h)^{\frac{1}{2}} \left\{\tilde{\Lambda}_{k,v}(t) - \Lambda_{k,v}(t)\right\} + o_p(1).$$

On the other hand, by decomposition and a Taylor series expansion,

$$(n_k h)^{\frac{1}{2}} \left\{\tilde{\Lambda}_{k,v}(t) - \Lambda_{k,v}(t)\right\} = (n_k h)^{\frac{1}{2}} \int_0^t \frac{\sum_{i=1}^{n_k} K_h(\bar{D}(U_{ki}) - v) dM_{ki}(s)}{\sum_{i=1}^{n_k} K_h(\bar{D}(U_{ki}) - v) Y_{ki}(s)} + O_p(n_k^{\frac{1}{2}} h^{\frac{5}{2}}),$$

where $M_{ki}(t) = N_{ki}(t) - \int_0^t Y_{ki}(s) d\Lambda_{k,\bar{D}(U_{ki})}(s)$. Then, by a martingale central limit theorem,

$$\mathrm{Var}\left[(n_k h)^{\frac{1}{2}} \left\{\tilde{\Lambda}_{k,v}(t) - \Lambda_{k,v}(t)\right\}\right] = n_k h \int_0^t \frac{\sum_{i=1}^{n_k} K_h(\bar{D}(U_{ki}) - v)^2 Y_{ki}(s) d\Lambda_{k,\bar{D}(U_{ki})}(s)}{\left\{\sum_{i=1}^{n_k} K_h(\bar{D}(U_{ki}) - v) Y_{ki}(s)\right\}^2} + o_p(1),$$

which, by change of variable and the uniform law of large numbers (Pollard, 1990), converges

in probability to

$$m_2 \int_0^t \frac{d\Lambda_{k,v}(s)}{\zeta_k(v) H_{k,v}(s)}. \tag{A.2}$$

Furthermore, by the functional central limit theorem (Pollard, 1990), it can be shown that for each fixed v, $\left\{(n_k h)^{\frac{1}{2}} \{\hat{\Lambda}_{k,v}(t) - \Lambda_{k,v}(t)\} : t \in [t_0, t_1]\right\}$ converges weakly to a Gaussian process with independent increment and variance function given in (A.2). Let $S_{k,v}(t) = e^{-\Lambda_{k,v}(t)}$. By the functional delta-method followed with integration by parts and Gill (1983),

$$\begin{aligned} &\frac{(n_k h)^{\frac{1}{2}}}{t_1 - t_0} \int_{t_0}^{t_1} \left\{\hat{S}_{k,v}(t) - S_{k,v}(t)\right\} dt = \frac{(n_k h)^{\frac{1}{2}}}{t_1 - t_0} \int_{t_0}^{t_1} S_{k,v}(t) \left\{\hat{\Lambda}_{k,v}(t) - \Lambda_{k,v}(t)\right\} dt + o_p(1) \\ &= \frac{(n_k h)^{\frac{1}{2}}}{t_1 - t_0} \int_0^{t_1} \left\{\int_{s \vee t_0}^{t_1} S_{k,v}(u) du\right\} \frac{\sum_{i=1}^{n_k} K_h(\bar{D}(U_{ki}) - v) dM_{ki}(s)}{\zeta_k(v) H_{k,v}(s)} + o_p(1), \end{aligned} \tag{A.3}$$

for any fixed $v \in \mathcal{J}$. It then follows from a martingale central limit theorem that (A.3) converges in distribution to a mean zero normal random variable with variance

$$\frac{m_2}{(t_1 - t_0)^2} \int_0^{t_1} \left\{\int_{s \vee t_0}^{t_1} S_{k,v}(u) du\right\}^2 \frac{d\Lambda_{k,v}(s)}{\zeta_k(v) H_{k,v}(s)}.$$

It follows that for any fixed $v \in \mathcal{J}$, $(nh)^{1/2} \left\{\hat{\mathcal{D}}(v) - \mathcal{D}(v)\right\}$ converges in distribution to a mean zero normal random variable with variance

$$\sum_{k=1}^{2} \frac{m_2}{\pi_k (t_1 - t_0)^2 \zeta_k(v)} \int_0^{t_1} \left\{\int_{s \vee t_0}^{t_1} S_{k,v}(u) du\right\}^2 \frac{d\Lambda_{k,v}(s)}{H_{k,v}(s)},$$

which we denote by $\sigma^2(v)$.

APPENDIX B: JUSTIFICATION FOR THE PERTURBATION-RESAMPLING METHODS

In view of the resampling procedure, we first note that $|\hat{\beta}_k^* - \hat{\beta}_k| + \sup_t |\hat{\Lambda}_k^*(t) - \hat{\Lambda}_k(t)| = O_p(n_k^{-1/2})$. Let

$$\tilde{\Lambda}_{k,v}^*(t) = \int_0^t \frac{\sum_{i=1}^{n_k} V_{ki} K_h(\bar{D}(U_{ki}) - v) dN_{ki}(s)}{\sum_{i=1}^{n_k} V_{ki} K_h(\bar{D}(U_{ki}) - v) Y_{ki}(s)}.$$

It follows from the arguments given in Cai et al. (2010a) that,

$$(n_k h)^{\frac{1}{2}} \left\{ \hat{\Lambda}^*_{k,v}(t) - \hat{\Lambda}_{k,v}(t) \right\} = (n_k h)^{\frac{1}{2}} \left\{ \tilde{\Lambda}^*_{k,v}(t) - \hat{\Lambda}_{k,v}(t) \right\} + \mathcal{E}_{k1}(t,v),$$

where $\text{pr}(\sup_{t,v} n^{\delta} |\mathcal{E}_{k1}(t,v)| \geq \epsilon | \text{data}) \to 0$ in probability as $n_k \to \infty$ for some $\delta > 0$. Noting that

$$(n_k h)^{\frac{1}{2}} \left\{ \tilde{\Lambda}^*_{k,v}(t) - \hat{\Lambda}_{k,v}(t) \right\} = (n_k h)^{\frac{1}{2}} \int_0^t \frac{\sum_{i=1}^{n_k} V_{ki} K_h(\bar{D}(U_{ki}) - v) \left\{ dN_{ki}(s) - Y_{ki}(s) d\hat{\Lambda}_{k,v}(s) \right\}}{\sum_{i=1}^{n_k} V_{ki} K_h(\bar{D}(U_{ki}) - v) Y_{ki}(s)},$$

it follows from the similar arguments for deriving (A.3) and the convergence rate of $\hat{\Lambda}_{k,v}(s)$ give in Appendix A that $(nh)^{1/2} \left\{ \hat{\mathcal{D}}^*(v) - \hat{\mathcal{D}}(v) \right\}$ can be written as

$$\frac{(n_k h)^{\frac{1}{2}}}{t_1 - t_0} \int_0^{t_1} \left\{ \int_{s \vee t_0}^{t_1} \hat{S}_{k,v}(u) du \right\} \frac{\sum_{i=1}^{n_k} (V_{ki} - 1) K_h(\bar{D}(U_{ki}) - v) dM_{ki}(s)}{\sum_{i=1}^{n_k} K_h(\bar{D}(U_{ki}) - v) Y_{ki}(s)} + \mathcal{E}_{k2}(v),$$

where $\text{pr}(\sup_v n^{\delta} |\mathcal{E}_{k2}(v)| \geq \epsilon | \text{data}) \to 0$ in probability for some $\delta > 0$. Thus by a Lindeberg central limit theorem, conditional on the data, $(nh)^{1/2} \{ \hat{\mathcal{D}}^*(v) - \hat{\mathcal{D}}(v) \}$ is approximately a normal random variable with mean zero and variance

$$\frac{h \sum_{i=1}^{n_k}}{n_k (t_1 - t_0)^2} \int_0^{t_1} \left\{ \int_{s \vee t_0}^{t_1} \hat{S}_{k,v}(u) du \right\}^2 \frac{\left[K_h(\bar{D}(U_{ki}) - v) dM_{ki}(s) \right]^2}{\{ \zeta_k(v) H_{k,v}(s) \}^2} + o_p(1),$$

which converges to the limiting variance of $(nh)^{1/2} \left\{ \hat{\mathcal{D}}(v) - \mathcal{D}(v) \right\}$.

We now show that after proper standardization, the supermum type statistics

$$W = \sup_{v \in \mathcal{J}} \left| \frac{(nh)^{1/2} \left\{ \hat{\mathcal{D}}(v) - \mathcal{D}(v) \right\}}{\hat{\sigma}(v)} \right|,$$

defined in (3.8), converges weakly. It follows from (A.3) and the uniform consistency of $\hat{\sigma}(v)$

for $\sigma(v)$ that

$$W = \sup_{v\in\mathcal{J}} \left| \sum_{k=1}^{2} (-1)^k \frac{(nh)^{\frac{1}{2}}}{t_1 - t_0} \int_0^{t_1} \left\{ \int_{s\vee t_0}^{t_1} S_{k,v}(u)du \right\} \frac{\sum_{i=1}^{n_k} K_h(\bar{D}(U_{ki}) - v) dM_{ki}(s)}{\zeta_k(v) H_{k,v}(s)\sigma(v)} \right| + o_p(n^{-\delta}),$$

for some $\delta > 0$. To apply the strong approximation arguments and extreme value limit theorem given in Bickel and Rosenblatt (1973), we represent the observed data $\{(T_{ki}, \Delta_{ki}, U_{ki}), k = 1, 2, i = 1, \ldots, n_k\}$ as $\{(T_j, \Delta_j, U_j, G_j), j = 1, \ldots, n\}$, where G_j is the treatment group indicator for subject j ($G_j = 1$ if subject j is in group 2, and $G_j = 0$ otherwise). We can rewrite W as

$$\sup_{v\in\mathcal{J}} \left| \sum_{j=1}^{n} (nh)^{\frac{1}{2}} K_h(\bar{D}(U_j) - v)\xi_j \right| + o_p(n^{-\delta}),$$

where

$$\xi_j = \frac{1}{t_1 - t_0} \int_0^{t_1} \left\{ \frac{\int_{s\vee t_0}^{t_1} \{G_j S_{2,v}(u) - (1 - G_j) S_{1,v}(u)\} du}{\{G_j \zeta_2(v) H_{2,v}(s) + (1 - G_j)\zeta_1(v) H_{1,v}(s)\}\sigma(v)} \right\} dM_j(s).$$

Using similar arguments as in Bickel and Rosenblatt (1973) and Cai et al. (2010a), we have

$$pr\{a_n(W - d_n\} < x\} \to e^{-2e^{-x}},$$

where

$$a_n = \left[2\log\left\{ \frac{\psi(\rho_2) - \psi(\rho_1)}{h} \right\} \right]^{\frac{1}{2}} \text{ and } d_n = a_n + a_n^{-1} \log\left\{ \frac{1}{4m_2\pi} \int K'(t)^2 dt \right\},$$

where $K'(\cdot)$ is the derivative of $K(\cdot)$.

To justify the resampling procedure for constructing the simultaneous confidence intervals, we note that

$$\frac{(nh)^{1/2}\left\{\mathcal{D}^*(v) - \hat{\mathcal{D}}(v)\right\}}{\hat{\sigma}(v)} = \sum_{j=1}^{n} (nh)^{\frac{1}{2}} K_h(\bar{D}(U_j) - v)\hat{\xi}_j V_j + \mathcal{E}_3(v)$$

where $\hat{\xi}_j$ is obtained by replacing all the unknown quantities in ξ_j by their empirical counterparts, $\{V_{ki}, k=1,2; i=1,\cdots,n_k\} = \{V_j, j=1,\cdots,n\}$ and $\text{pr}(\sup_v n^\delta|\mathcal{E}_3(v)| \geq \epsilon \mid \text{data}) \to 0$ in probability for some $\delta > 0$. Therefore,

$$W^* = \sup_{v\in\mathcal{J}} \left| \sum_{j=1}^{n} (nh)^{\frac{1}{2}} K_h(\bar{D}(U_j) - v)\hat{\xi}_j V_j \right| + \mathcal{E}_4,$$

where $\text{pr}(|n^\delta \mathcal{E}_4| \geq \epsilon|\text{data}) \to 0$ in probability. It follows from similar arguments in Tian et al. (2005) and Li et al. (2010) that

$$\sup_x \left| pr\{a_n(W^* - d_n\} < x|(T_{ki}, \Delta_{ki}, U_{ki}), k=1,2, i=1,\ldots,n_k\} - e^{-2e^{-x}} \right| \to 0,$$

in probability as $n \to \infty$. Thus the conditional distribution of $a_n(W^* - d_n\}$ can be used to approximate the unconditional distribution of $a_n(W - d_n)$. When $h_1 \neq h_2$, in general, the standardized W does not converge to the extreme value distribution. However, when $h_1/h_2 = k \in (0,\infty)$, the distribution of the suitable standardized version of W still can be approximated by that of the standardized W^* conditional on the data (Gilbert et al. 2002).

APPENDIX C: A MODIFIED PERTURBATION PROCEDURE

When the study sample size is not large or the event rate is low, the resulting interval estimates tend to be conservative. Here we propose a modified perturbation-resampling version for $\hat{\Lambda}^*_{k,v}(t)$ in (3.5), which may substantially improve the precision of the resulting inference procedure for finite sample cases. Specifically, we replace $\hat{\Lambda}^*_{k,v}(t)$ in (3.5) by

$$\int_0^t \frac{\sum_{i=1}^{n_k} V_{ki} K_{h_k}(\hat{Q}_{ki,v}) dN_{ki}(s)}{\sum_{i=1}^{n_k} V_{ki} K_{h_k}(\hat{Q}_{ki,v}) Y_{ki}(s)} + \int_0^t \left\{ \frac{\sum_{i=1}^{n_k} K_{\hbar_k}(\hat{Q}^*_{ki,v}) dN_{ki}(s)}{\sum_{i=1}^{n_k} K_{\hbar_k}(\hat{Q}^*_{ki,v}) Y_{ki}(s)} - \frac{\sum_{i=1}^{n_k} K_{\hbar_k}(\hat{Q}_{ki,v}) dN_{ki}(s)}{\sum_{i=1}^{n_k} K_{\hbar_k}(\hat{Q}_{ki,v}) Y_{ki}(s)} \right\}. \tag{A.4}$$

Note that we use two potentially different sets of smoothing parameters in (A.4). When $\hbar_k = h_k$, (A.4) reduces to (3.4). Also note that the second term is a difference function with

respect to $\hat{Q}^*$ and $\hat{Q}$, which can be approximated by a product of a derivative-like function and a function of differences $\hat{\beta}^* - \hat{\beta}$ and $\hat{\Lambda}^*(\cdot) - \hat{\Lambda}(\cdot)$. To make this term more stable for finite sample cases, one may use a larger bandwidth $\hbar_k$. Since this resembles estimating a derivative function in the nonparametric function estimation literature, we recommend choosing smooth parameters $\hbar_k$'s in the second term of (A.4) with order of $O(n_k^{-1/7})$, which is an optimal choice in estimating a derivative function (Fan et al., 1997). It follows that we let $\hbar_k = h_k \times n_k^{1/5-1/7}$ in our analysis.

Since $|\hat{\beta}_k^* - \hat{\beta}_k| + \sup_t |\hat{\Lambda}_k^*(t) - \hat{\Lambda}_k(t)| = O_p(n_k^{-1/2})$, it is straightforward to show that the standardized (A.4) is asymptotically equivalent to the standardized (3.5).

REFERENCES

Bickel, P. J. and Rosenblatt, M. (**1973**), "On some global measures of the deviations of density function estimates (Corr: V3 p1370)," The Annals of Statistics **1**, 1071–1095.

Bonetti, M. and Gelber, R. D. (**2000**), "A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data," Statistics in Medicine **19**, 2595–609.

—— (**2005**), "Patterns of treatment effects in subsets of patients in clinical trials," Biostatistics **5**, 465–81.

Braunwald, E., Domanski, M. J., Fowler, S. E., and et al., The PEACE Trial Investigators (**2004**), "Angiotensin-converting-enzyme inhibition in stable coronary artery disease," The New England Journal of Medicine **351**, 2058–2068.

Cai, T., Tian, L., Uno, H., Solomon, S. D., and Wei, L. J. (**2010**a), "Calibrating parametric subject-specific risk estimation," Biometrika **97**(2), 389–404.

Cai, T., Tian, L., and Wei, L. J. (**2005**), "Semiparametric Box-Cox power transformation models for censored survival observations," Biometrika **92**, 619–632.

Cai, T., Tian, L., Wong, P. H., and Wei, L. J. (**2010**b), "Analysis of Randomized Comparative Clinical Trial Data for Personalized Treatment Selections," Biostatistics **to appear**.

Cox, D. R. (**1972**), "Regression models and life-tables (with discussion)," Journal of the Royal Statistical Society, Series B **34**, 187–220.

Fan, J. and Gijbels, I. (**1995**), "Data-driven bandwidth selection in local polynomial regression: variable bandwidth selection and spatial adaptation," Journal of the Royal Statistical Society, Series B **57**, 371–394.

Fan, J., Gijbels, I., and King, M. (**1997**), "Local likelihood and local partial likelihood in hazard regression," The Annals of Statistics **25**, 1661–1690.

Gilbert, P. B., Wei, L. J., Kosorok, M. R., and Clemens, J. D. (**2002**), "Simultaneous Inferences on the Contrast of Two Hazard Functions with Censored Observations," Biometrics **58**, 773–780.

Gill, R. D. (**1983**), "Large Sample Behaviour of the Product-Limit Estimator on the Whole Line," The Annals of Statistics **11**, 49–58.

Hjort, N. (**1992**), "On inference in parametric survival data models," Int. Statist. Rev. **60**, 355–387.

Kalbfleisch, J. D. and Prentice, R. L. (**1981**), "Estimation of the average hazard ratio," Biometrika **68**, 105–112.

—— (**2002**), *The Statistical Analysis of Failure Time Data* (New York: JohnWiley & Sons).

Li, G. and Doss, H. (**1995**), "An approach to nonparametric regression for life history data using local linear fitting," The Annals of Statistics **23**, 787–823.

Li, Y., Tian, L., and Wei, L. J. (**2010**), "Estimating Subject-Specific Dependent Competing Risk Profile with Censored Event Time Observations," Biometrics **In press**.

Lin, D. Y., Wei, L. J., and Ying, Z. (**1993**), "Checking the Cox model with cumulative sums of martingale-based residuals," Biometrika **80**, 557–572.

Mammen, E. (**1992**), "Bootstrap, wild bootstrap, and asymptotic normality," Prob. Theory Rel. Fields **93**, 439–455.

Murray, S. and Tsiatis, A. A. (**1999**), "Sequential Methods for Comparing Years of Life Saved in the Two-Sample Censored Data Problem," Biometrics **55**, 1085–1092.

Park, B., Kim, W., Ruppert, D., Jones, M., Signorini, D., and Kohn, R. (**1997**), "Simple transformation techniques for improved non-parametric regression," Scand. J. Statist. **24**, 145–163.

Park, Y. and Wei, L. J. (**2003**), "Estimating subject-specific survival functions under the accelerated failure time model," Biometrika **90**, 717–723.

Pepe, M. S. (**2003**), *The statistical evaluation of medical tests for classification and prediction* (Oxford University Press).

Pepe, M. S. and Fleming, T. R. (**1989**), "Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data," Biometrics **45**, 497–507.

—— (**1991**), "Weighted Kaplan-Meier statistics: Large sample and optimality considerations," Journal of the Royal Statistical Society, Series B **53**, 341–352.

Pfeffer, M. and Jarcho, J. (**2006**), "The Charisma of Subgroups and the Subgroups of CHARISMA," New England Journal of Medicine **354**(16), 1744.

Pollard, D. (**1990**), *Empirical Processes: Theory and Applications, Regional Conference Series in Probability and Statistics 2* (Institute of Mathematical Statistics, Hayward, CA).

Rothwell, P. (**2005**), "External validity of randomised controlled trials: "to whom do the results of this trial apply?"," The Lancet **365**(9453), 82–93.

Signorovitch, J. E. (**2007**), "Identifying Informative Biological Markers in High-Dimensional Genomic Data and Clinical Trials," Ph.D. thesis, Harvard University.

Solomon, S. D., M., R. M., Jablonski, K. A., and et al., for the Prevention of Events with ACE inhibition (PEACE) Investigators (**2006**), "Renal Function and Effectiveness of Angiotensin-Converting Enzyme Inhibitor Therapy in Patients With Chronic Stable Coronary Disease in the Prevention of Events with ACE inhibition (PEACE) Trial," Circulation **114**, 26–31.

Song, X. and Pepe, M. S. (**2004**), "Evaluating markers for selecting a patient's treatment," Biometrics **60**, 874–83.

Tian, L., Cai, T., Goetghebeur, E., and Wei, L. (**2007**), "Model evaluation based on the sampling distribution of estimated absolute prediction error," Biometrika **94**(2), 297.

Tian, L., Zucker, D., and Wei, L. J. (**2005**), "On the Cox model with time-varying regression coefficients," Journal of American Statistical Association **100**, 172–183.

Wand, M., Marron, J., and Ruppert, D. (**1991**), "Transformation in density estimation (with comments)," Journal of American Statistical Association **86**, 343–361.

Wang, R., Lagakos, S., Ware, J., Hunter, D., and Drazen, J. (**2007**), "Statistics in Medicine–Reporting of Subgroup Analyses in Clinical Trials," New England Journal of Medicine **357**(21), 2189.

Wu, C. (**1986**), "Jackknife, bootstrap and other resampling methods in regression analysis," Ann. Statist. **14**, 1261–1295.

Xu, R. and O'Quigley, J. (**2000**), "Estimating Average Regression Effect under Non-Proportional Hazards," Biostatistics **1**, 423–439.

Zhao, L., Tian, L., Uno, H., Solomon, S. D., Pfeffer, M. A., Schindler, J. S., and Wei, L. J. (**2010**), "Utilizing the Integrated Difference of Two Survival Functions to Quantify the Treatment Contrast for Designing, Monitoring and Analyzing a Comparative Clinical Study," Harvard University Biostatistics Working Paper Series, working Paper 115.

Table 1: Estimated (Est) regression coefficients, their standard errors (SE) and p-values by fitting the Cox model to the PEACE data based on all the mortality information up to study month 72

Covariates	Placebo			ACEi		
	Est	SE	p-value	Est	SE	p-value
eGFR	-0.006	0.003	0.05	0.000	0.003	0.96
Age	0.072	0.008	<0.01	0.063	0.008	<0.01
Gender[1]	-0.179	0.155	0.25	-0.577	0.178	<0.01
lveejf	-0.026	0.007	<0.01	-0.009	0.007	0.17
Medical histories (0: no, 1: yes)						
Hypertension	0.330	0.117	<0.01	0.245	0.120	0.04
Diabetes	0.515	0.135	<0.01	0.647	0.133	<0.01
Myocardial infarction	0.016	0.119	0.89	0.244	0.124	0.05

[1] 0: Male, 1: Female

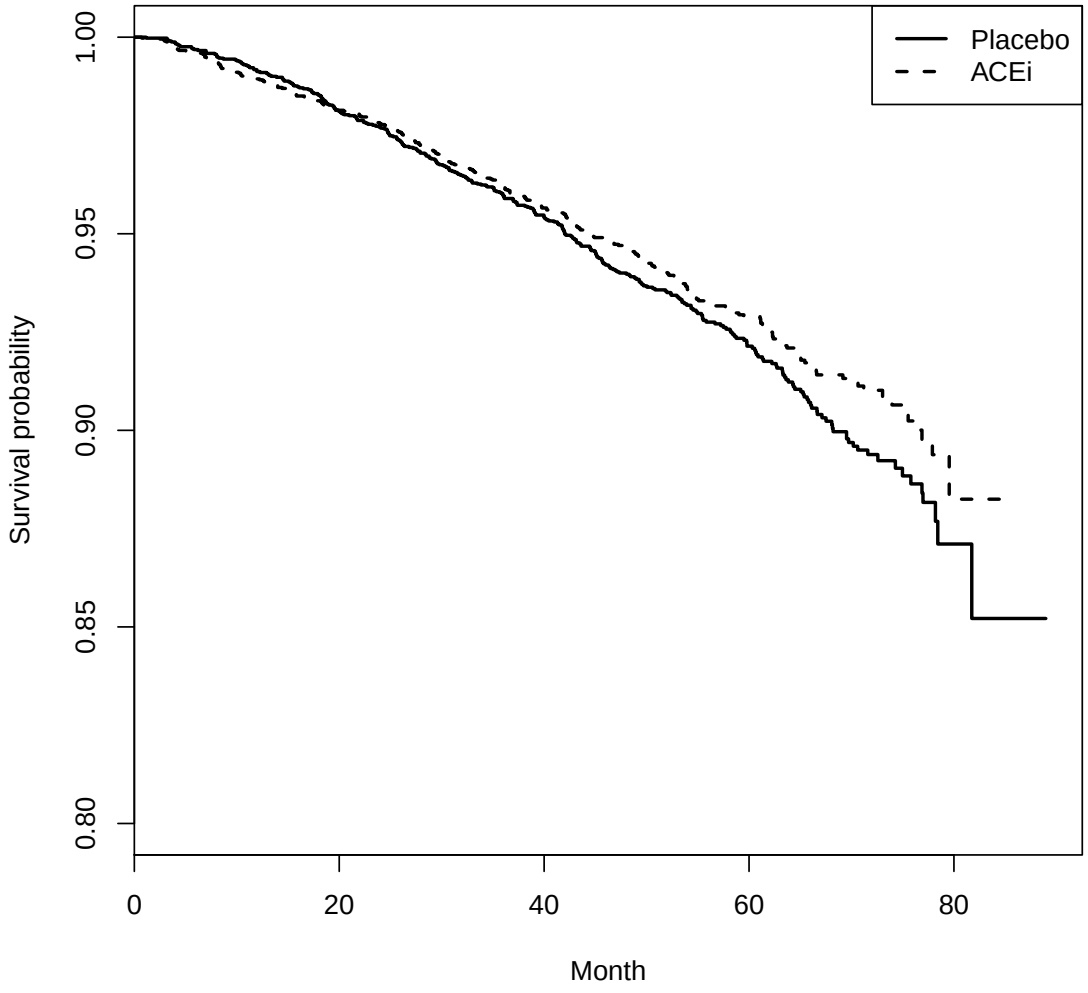

Figure 1: The Kaplan-Meier estimates for the survival functions of patients in the PEACE study

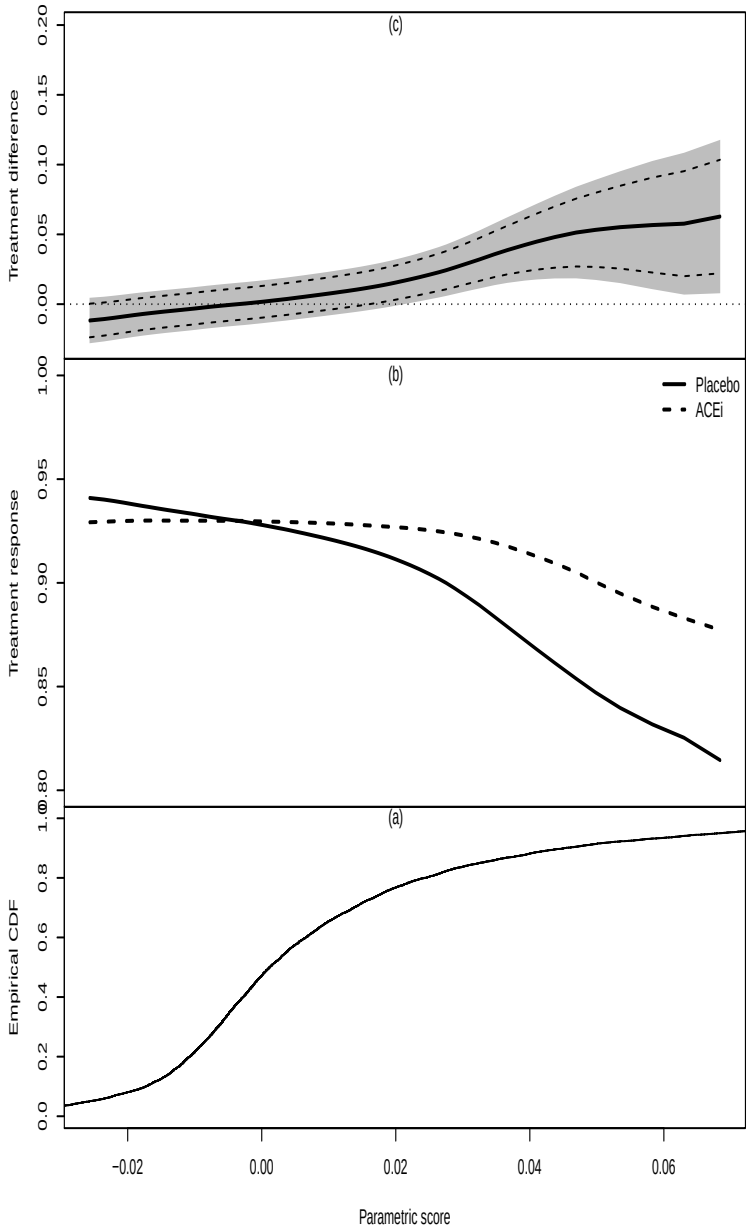

Figure 2: Calibrated parametric estimates for the integrated difference of survival rates for the time interval [60, 72] with the data from PEACE study; (a). The empirical distribution function of the parametric score; (b). The calibrated estimates for the average of survival rates for the time interval [60, 72]; (c). The calibrated estimates for the integrated difference of survival rates (solid curve), 0.95 pointwise confidence interval (dashed lines) and 0.95 simultaneous confidence region (shaded area)

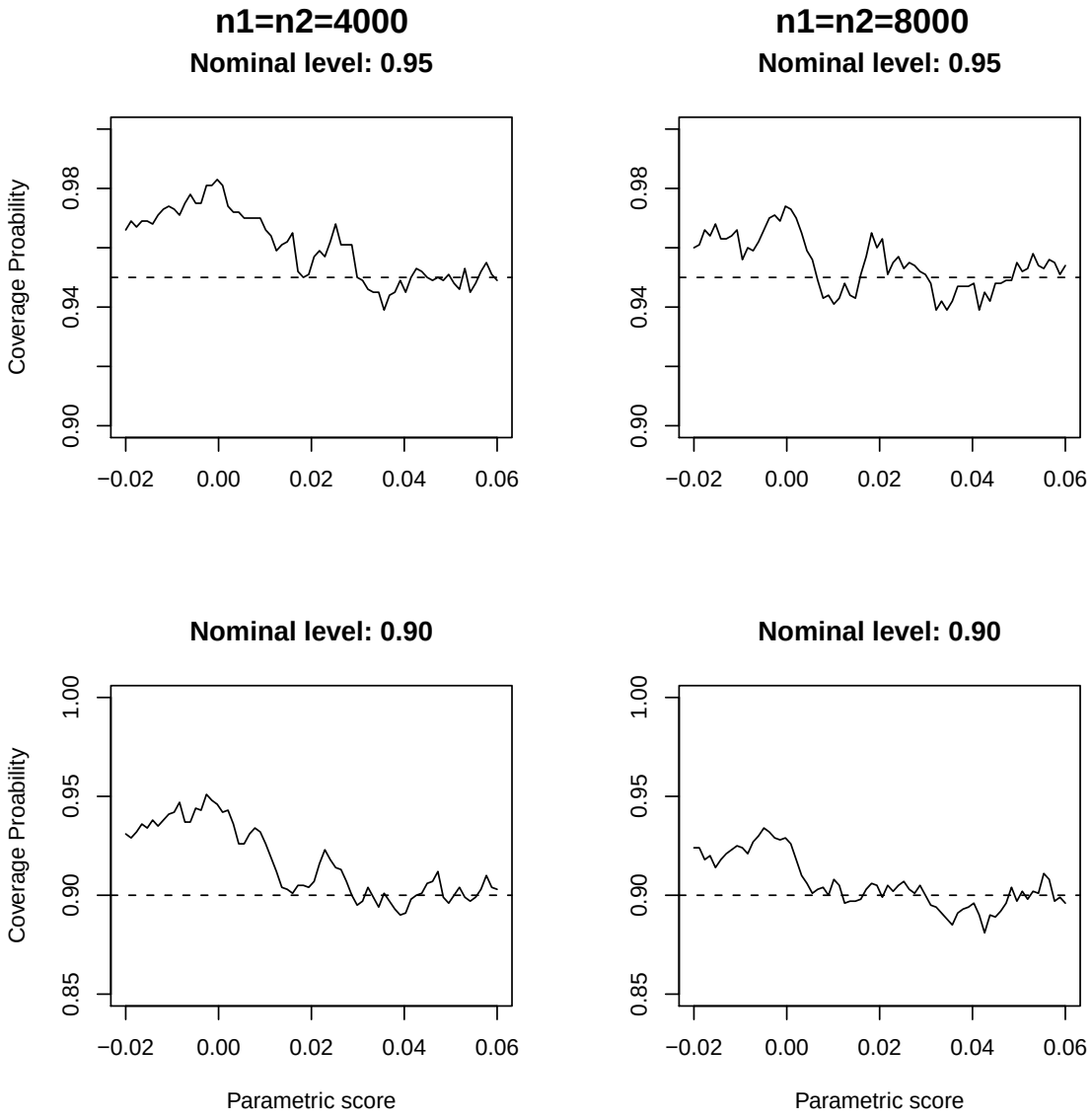

Figure 3: Empirical coverage probabilities of pointwise confidence interval estimators. Left panel: $n_1 = n_2 = 4000$, Right panel: $n_1 = n_2 = 8000$