Harvard University
Harvard University Biostatistics Working Paper Series

Year 2011 *Paper* 125

Estimating Subject-Specific Treatment Differences for Risk-Benefit Assessment with Competing Risk Event-Time Data

Brian Claggett* Lihui Zhao† Lu Tian‡
Davide Castagno** L. J. Wei††

*Harvard University, bclagget@hsph.harvard.edu
†Harvard University, lhzhao@hsph.harvard.edu
‡Stanford University School of Medicine, lutian@stanford.edu
**Brigham & Women's Hospital and Universtiy of Turin
††Harvard University, wei@hsph.harvard.edu

Estimating Subject-Specific Treatment Differences for Risk-Benefit Assessment with Competing Risk Event-Time Data

B. Claggett[1], L. Zhao[1], L. Tian[2], D. Castagno[3,4], and L. J. Wei[1]

Summary

To evaluate treatment efficacy or toxicity using event-time data from a randomized comparative study, we usually make inference about a summary measure which quantifies an overall treatment difference. However, a single measure for efficacy, even when coupled with that for toxicity, is difficult to be utilized for treating a future patient at his or her bedside. A positive (negative) study result based on such a measure does not mean that every future subject should (should not) be treated by the new therapy. For clinical practice, it is desirable to identify subjects who would benefit from the new treatment from a risk-benefit perspective. In this paper, we propose a systematic approach to achieve this goal using competing risk event-time data from two similar, but independent studies. We first utilize data from a study to build a parametric score with respect to a primary event for the purpose of stratifying the patients in the second study. We then use the data from the second study to obtain a nonparametric estimate of the treatment difference, with respect to each competing risk event, for any fixed score. Furthermore, confidence interval and band estimates are constructed to quantify the uncertainty of our inferences for the treatment differences over a range of scores. To illustrate the new proposal, we use the data sets from two cardiovascular studies for evaluating specific beta-blockers in patients with heart failure. The score is based on time to death, and the competing events are myocardial infarction, hospitalization and toxicity.

Keywords: Clinical trial; Cox model; Nonparametric estimation; Personalized medicine; Perturbation-resampling method; Stratified medicine; Subgroup analysis; Survival analysis.

[1]Department of Biostatistics, Harvard University, Boston, MA
[2]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA
[3]Department of Internal Medicine, Cardiovascular Division, Brigham and Women's Hospital, Boston, MA
[4]Cardiology Unit, Department of Internal Medicine, University of Turin, Turin, Italy

1. Introduction

Consider a randomized, comparative clinical trial in which a treatment is assessed against a control with respect to their risk-benefit profiles. These profiles are quantified using event-time data under a competing risk setting. Conventionally, a single treatment contrast is utilized to assess an overall treatment difference with respect to efficacy, in addition to a global measure for toxicity, over a rather heterogeneous population. Unfortunately, the resulting inference about these two measures are rather difficult to interpret in clinical practice. A positive (negative) trial based on these two overall measures does not mean that every future patient should (should not) be treated by the new treatment. To bring the clinical trial results to the patient's bedside, we may utilize the patient's characteristics which relate to the response variable to perform so-called personalized or stratified medicine. Unfortunately, the typical ad hoc subgroup analysis of clinical studies is not credible (Wang et al., 2007). Moreover, such subgroup analysis is often conducted by investigating the effect of only a single predictor at a time and is therefore not effective in identifying patients who would benefit from the new treatment.

In this paper, we present a systematic approach to estimate subject-specific treatment differences from a risk-benefit perspective where the risk and benefit are quantified using event times. That is, for each study subject, the observations are times to events that define the efficacy and toxicity of the treatment. Since formal evaluations of new drugs or devices usually require two well-conducted studies, one may use one study to build a parametric scoring system with baseline variables and then stratify subjects in the second study and estimate the treatment differences nonparametrically with respect to the risk-benefit profiles. Note that this univariate scoring system can be constructed using a primary or composite endpoint of interest. Also note that by using the proposed "two-study approach", one may avoid the nontrivial "self-serving" bias that can result from creating the score, stratifying subjects, and estimating subject-specific treatment differences with the same data set. To

control the overall error rate, using the event-time data from the second study, we provide the simultaneous, nonparametric confidence band of the treatment differences for each competing risk event over a range of parametric scores obtained from the first study via an extensive model or variable selection process.

When there is a single baseline covariate involved, Song and Pepe (2004) and Bonetti and Gelber (2000, 2005) have proposed novel statistical procedures for identifying a subgroup of patients who would benefit from the new treatment with respect to efficacy. A recent paper by Janes et al. (2011), which is based on previous work by Pepe (2003), Huang et al. (2007), and Pepe et al. (2008), provides practical guidelines for measuring the performance of individual markers for treatment selection. By incorporating more than one baseline covariate at a time, our approach is similar in spirit to Cai et al. (2010b) and Li et al. (2010). However, they both used a single study to create a scoring system by fitting a prespecified model without model evaluation or variable selection. They then use the same data set to make inferences for either the treatment difference without competing risks or risk predictions for a single treatment group.

To illustrate our proposal, we utilize the data from a recent clinical trial, "Beta-Blocker Evaluation of Survival Trial" (BEST), which compared a beta-blocker to placebo in patients with heart failure, with a primary endpoint of all-cause mortality. Here, the competing risk events include myocardial infarction (MI), hospitalization, and adverse events (AE) (BEST, 2001). Note that in the BEST trial, the non-fatal event times were not censored by other non-fatal event times, which is different from the conventional competing risk setting. We first used the mortality data from a similar study, "Cardiac Insufficiency Bisoprolol Study II" (CIBIS-II), for evaluating a beta-blocker in patients with heart failure, to build a parametric score (CIBIS II, 1999). We then used this scoring system to stratify the patients in the BEST study and make inferences about the treatment differences with respect to the aforementioned competing risks across a range of scores via simultaneous and pointwise confidence interval estimates.

2. Building a Scoring System with Respect to a Specific Event Time via a Training Data Set

To begin, we use the first study, say, the training data set, to build a scoring system using the patients' baseline characteristics with respect to a particular or composite event time which is of primary interest for the target population. This event time may be subject to censoring from competing event times. Specifically, for this training set, each subject was assigned to a particular treatment j, where j=1,2. For the jth treatment group, let T_j be an event time, representing the minimum of the time to this primary event and other competing event times. Let the indicator function $\epsilon_j = 1$, if T_j is not censored by a competing event time. Also, let U_j be the vector of baseline covariates. In addition to those competing risks, let C_j be the censoring variable, which is assumed to be independent of U_j and all the underlying competing event times. Furthermore, let $X_j = \min(T_j, C_j)$ and Δ_j be the indicator function, which is one if $T_j \leq C_j$. The data consist of $\{(X_{ij}, \Delta_{ij}, \Delta_{ij}\epsilon_{ij}, U_{ij})', j = 1, \cdots, n_j\}$, n_j independent copies of $(X_j, \Delta_j, \Delta_j\epsilon_j, U_j)'$, $j = 1, 2$. Note that ϵ_{ij} is observable if $\Delta_{ij} = 1$.

Now, suppose that we are interested in estimating the t_0-year event rates $\pi_j(U), j = 1, 2$, with respect to the primary event, where

$$\pi_j(U) = \text{pr}(T_j \leq t_0, \epsilon_j = 1|U) \tag{2.1}$$

for a pre-specified time point t_0. To obtain an estimate for $\pi_j(U)$, one may use the following working models

$$\pi_j(U_{ij}) = g_j(\beta_j' Z_{ij}), j = 1, 2, \tag{2.2}$$

where Z_{ij} is a function of U_{ij}, $g_j(\cdot)$ is a given monotone function, and β_j is an unknown vector of parameters. An estimating function for β_j with the above censored competing risks data is

$$R(\beta_j) = n_j^{-1} \sum_{i=1}^{n_j} \frac{w_{ij}}{\hat{G}_j(X_{ij} \wedge t_0)} Z_{ij} \{ I(X_{ij} \leq t_0, \epsilon_{ij} = 1) - g_j(\beta_j' Z_{ij}) \}, \tag{2.3}$$

where $w_{ij} = I(X_{ij} \leq t_0)\Delta_{ij} + I(X_{ij} > t_0)$, $I(\cdot)$ is the indicator function, and $\hat{G}_j(\cdot)$ is the Kaplan-Meier estimator for $G_j(\cdot)$, the survival function of the censoring variable for the jth group and this primary event time. (Uno et al., 2007; Li et al., 2010 and Zheng et al., 2006). The point estimator $\hat{\beta}_j$ for β_j can be obtained by solving the equation $R(\beta) = 0$. Under some mild conditions, the resulting estimator $\hat{\beta}_j$ converges to a finite constant vector as $n \to \infty$ even when the model (2.2) is not correctly specified (Uno et al., 2007).

Note that one may repeatedly utilize (2.2) and (2.3) with various Z and $g_j(\cdot)$ via, for instance, the standard stepwise regression with U, to obtain a final estimate $\hat{\pi}_j(U)$. For a given U, let the score for the treatment contrast be denoted by $\hat{D}(U) = \hat{\pi}_2(U) - \hat{\pi}_1(U)$, which intends to estimate $D(U) = \pi_2(U) - \pi_1(U)$. Now, suppose that the control group corresponds to j=1. Then one may use $\hat{\pi}_1(\cdot)$ as the risk score for grouping the patients in the second study for estimating the subject-specific treatment differences. On the other hand, because the scoring system $\hat{D}(\cdot)$ is constructed using the interactions between the treatment and covariates, intuitively, it may more effectively stratify patients with similar treatment difference profiles.

Models other than (2.3) may also be used to build the scoring systems $\hat{\pi}_j(\cdot)$ and $\hat{D}(\cdot)$. For example, we may use a generalized Cox proportional hazards model, which deals with competing event time data, to estimate the survival rate at t_0 (Fine and Gray, 1999). Variable selection procedures can also be utilized with such a global survival model. However, it is possible that certain covariates may be highly predictive for short-term survival, but not for long-term survival (or vice versa), in which case it may be more appropriate to use a logistic regression model as described above to build a scoring system.

Since many explanatory models can be considered as candidates for estimating $\pi_j(\cdot), j = 1, 2$, it is important to formally evaluate their relative merits. To this end, we first note that

the adequacy of a survival model for t_0-year survival can be quantified by the area under the receiver operating characteristic curve (AUC). Specifically, for treatment j, the AUC_j for the t_0-year survival rate is

$$\text{pr}(\hat{\pi}_j(U_{ij}) > \hat{\pi}_j(U_{lj}) | I(T_{ij} \leq t_0) > I(T_{lj} \leq t_0)),$$

where $I(T_{ij} \leq t_0) \neq I(T_{lj} \leq t_0)$. A large value of AUC indicates that the model fits the event-time data well. With censored data, we may use a similar procedure proposed by Uno et al. (2007) to consistently estimate the AUC nonparametrically.

Next we use an M-fold cross validation procedure to evaluate all the candidate models. Specifically, we split the training data set into M disjoint subsets of approximately equal size, denoted by $\{\mathcal{I}_m, m = 1, ..., M\}$. For each m, we use all observations not in $\mathcal{I}_m$ to build a working model and apply the estimates $\hat{\pi}_j(U_{ij})$ to the data in $\mathcal{I}_m$, with the resulting AUC estimate denoted by $\widehat{AUC}_j^{(m)}$. Lastly, we average these AUC estimates over $m = 1, ..., M$ to obtain a final estimate $\widehat{AUC}_j$. The model which yields the largest cross-validated AUC_j value among all candidate models will be chosen as the final model for treatment group j. We then refit the entire training data set with this model in order to construct the final score.

3. Making inferences About the Treatment Differences over a Range of Scores with Respect to all Competing Risk Event-Time Data from the Target Data Set

Let the final parametric score for a patient with the covariate vector U in the target study be denoted by S(U), which may be the risk score $\hat{\pi}_1$(U) based on the control group or the treatment selection score $\hat{D}$(U) discussed in the previous section. Now, with the event-time data from the second study, in order to obtain the personalized risk-benefit assessment, we construct the confidence interval or band estimates for the treatment differences with respect to each competing risk event over the score. To this end, let T_{ijk} be the minimum

of the time to the kth event and its competing event times, where $i = 1, ..., n_j; j = 1, 2; k = 1, ..., K$. Moreover, let ϵ_{ijk} be a binary indicator, which is one if we observe the time to the kth event. The data consist of n_j independent and identically distributed observations $\{(X_{ijk}, \Delta_{ijk}, \Delta_{ijk}\epsilon_{ijk}, U_{ij}), k = 1, \cdots, K; i = 1, \cdots, n_j\}, j = 1, 2$, where $X_{ijk} = \min(T_{ijk}, C_{ij})$, $\Delta_{ijk} = I(T_{ijk} \leq C_{ij})$ and C_{ij} is the censoring time, which is independent of $\{(T_{ijk}, U_{ij}), i = 1, \cdots, n_j; j = 1, 2; k = 1, \cdots, K\}$. Furthermore, we let $Y_{ijk} = I(X_{ijk} \leq t_0, \epsilon_{ijk} = 1)$. For the kth event, we are interested in estimating the treatment difference conditional on $S(U) = s$, that is,

$$E_k(s) = \text{pr}(T_{i2k} \leq t_0, \epsilon_{i2k} = 1 | S(U) = s) - \text{pr}(T_{i1k} \leq t_0, \epsilon_{i1k} = 1 | S(U) = s). \tag{3.1}$$

To estimate $E_k(s)$ nonparametrically, we use a kernel estimator for each term on the right hand side of (3.1). Specifically, we estimate $p_{jk}(s) = \text{pr}(T_{ijk} \leq t_0, \epsilon_{ijk} = 1 | S(U) = s)$ with $\hat{p}_{jk}(s)$

$$= \left\{ \sum_i \frac{w_{ijk}}{\hat{G}_{jk}(X_{ijk} \wedge t_0)} K_{h_{jk}}(V_{ij} - s) Y_{ijk} \right\} / \left\{ \sum_i \frac{w_{ijk}}{\hat{G}_{jk}(X_{ijk} \wedge t_0)} K_{h_{jk}}(V_{ij} - s) \right\}, \tag{3.2}$$

where $V_{ij} = S(U_{ij})$, $w_{ijk} = I(X_{ijk} \leq t_0)\Delta_{ijk} + I(X_{ijk} > t_0)$, $\hat{G}_{jk}(\cdot)$ is the Kaplan-Meier estimator of $G_j(\cdot)$, the survival distribution of the censoring variable C_j, estimated using observations $\{(X_{ijk}, \Delta_{ijk}), i = 1, \cdots, n_j; j = 1, 2\}$, $K_{h_{jk}}(s) = K(s/h_{jk})/h_{jk}$, $K(\cdot)$ is a smooth symmetric kernel with finite support and h_{jk} is a smoothing parameter. When $h_{jk} = O(n^{-v}), 1/5 < v < 1/2$, it follows from a similar argument by Li et al. (2010) that $\hat{p}_{jk}(s)$ converges to $p_{jk}(s)$ uniformly over the interval $s \in \mathcal{S}$, where $\mathcal{S}$ is an interval contained properly in the support of S(U). Let $\mathbf{E}(s) = \{E_1(s), ..., E_K(s)\}' = \mathbf{p}_2(s) - \mathbf{p}_1(s)$ and its empirical counterpart $\hat{\mathbf{E}}(s) = \{\hat{E}_1(s), ..., \hat{E}_K(s)\}' = \hat{\mathbf{p}}_2(s) - \hat{\mathbf{p}}_1(s)$, where $\hat{E}_k(s) = \hat{p}_{2k}(s) - \hat{p}_{1k}(s)$, $\mathbf{p}_j(s) = \{p_{j1}(s), \cdots, p_{jK}(s)\}'$ and $\hat{\mathbf{p}}_j(s) = \{\hat{p}_{j1}(s), \cdots, \{\hat{p}_{jK}(s)\}'$

It follows from a similar argument by Li et al. (2010) that when h_{jk} is of the same order

as above, for a fixed s, the joint distribution

$$\text{diag}\{(n_1h_{11}+n_2h_{21})^{1/2},...,(n_1h_{1K}+n_2h_{2K})^{1/2}\}\{\hat{\mathbf{E}}(s)-\mathbf{E}(s)\} \tag{3.3}$$

converges in distribution to a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}(s)$ as $n\to\infty$, where diag$\{\cdots\}$ is a $K\times K$ diagonal matrix.

To approximate the distribution in (3.3), we may use a perturbation-resampling method, which is similar to 'wild bootstrapping' (Wu, 1986; Mammen, 1993) and has been successfully implemented in many estimation problems (Lin et al., 1993; Park and Wei, 2003; Cai et al. 2010). Specifically, let $\{B_{ij}: i=1,...,n_j; j=1,2\}$ be independent random samples from a strictly positive distribution with mean and variance equal to one. Let $p^*_{jk}(s)$ be the perturbed version of $\hat{p}_{jk}(s)$ with $p^*_{jk}(s)$

$$=\left\{\sum_i \frac{B_{ij}w_{ijk}}{\hat{G}^*_{jk}(X_{ijk}\wedge t_0)}K_{h_{jk}}(V_{ij}-s)Y_{ijk}\right\}/\left\{\sum_i \frac{B_{ij}w_{ijk}}{\hat{G}^*_{jk}(X_{ijk}\wedge t_0)}K_{h_{jk}}(V_{ij}-s)\right\}. \tag{3.4}$$

Here, $\hat{G}^*_{jk}(\cdot)$ is the perturbed estimator for the survival function $G_j(\cdot)$

$$\hat{G}^*_{jk}(t)=\exp\left[-\sum_{i=1}^{n_j}\int_0^t \frac{B_{ij}d\{I(C_{ij}\le u\wedge X_{ijk})\}}{\sum_{l=1}^{n_j}B_{lj}I(X_{ljk}\ge u)}\right]. \tag{3.5}$$

Denote $\mathbf{E}^*(s)=\mathbf{p}^*_2(s)-\mathbf{p}^*_1(s)$, where $\mathbf{p}^*_j(s)=\{p^*_{j1}(s),\cdots,p^*_{jK}(s)\}'$. Using the arguments by Cai et al. (2010), the limiting distribution, conditional on the target data set, of

$$\text{diag}\{(n_1h_{11}+n_2h_{21})^{1/2},...,(n_1h_{1K}+n_2h_{2K})^{1/2}\}\{\mathbf{E}^*(s)-\hat{\mathbf{E}}(s)\}, \tag{3.6}$$

is also multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}(s)$.

In order to obtain an approximation to $\mathbf{\Sigma}(s)$, we generate a large number of realizations of $\{B_{i1},B_{i2}\}$, and compute $\mathbf{E}^*(s)$ for each perturbation sample. The resulting sample covariance matrix based on those perturbed estimates $\mathbf{E}^*$, say, $\tilde{\mathbf{\Sigma}}(s)$, is a consistent estimator of $\mathbf{\Sigma}(s)$.

A two-sided confidence interval for an individual risk difference $E_k(s)$ is then given by

$$\hat{E}_k(s) \pm z_{(1-\alpha/2)}(n_1 h_{1k} + n_2 h_{2k})^{-1/2}\tilde{\sigma}_k(s), \tag{3.7}$$

where $\tilde{\sigma}_k(s)$ is the kth diagonal element of $\tilde{\mathbf{\Sigma}}(s)$.

To construct a $(1-\alpha)$ simultaneous confidence band for $E_k(s)$ over the pre-specified interval $\mathcal{S}$, we cannot use the conventional method based on the sup-statistic,

$$\sup_{s\in\mathcal{S}} \tilde{\sigma}_k^{-1}(s)|(n_1 h_{1k} + n_2 h_{2k})^{1/2}\{\hat{E}_k(s) - E_k(s)\}|$$

due to the fact that as a process in s, $(n_1 h_{1k} + n_2 h_{2k})^{1/2}\{\hat{E}_k(s) - E_k(s)\}$ does not converge to a process. On the other hand, one may utilize the strong approximation argument given in Bickel and Rosenblatt (1973) to show that an appropriately transformed sup of $\hat{E}_k(s) - E_k(s)$ converges to a proper random variable. In practice, to construct a confidence band, we can first find a critical value b_α such that

$$\mathrm{pr}(\sup_{s\in\mathcal{S}} |E_k^*(s) - \hat{E}_k(s)|/\{(n_1 h_{1k} + n_2 h_{2k})^{-1/2}\tilde{\sigma}_k(s)\} > b_\alpha) \approx \alpha.$$

Then the confidence band for $E_k(s) : s \in \mathcal{S}$ is given by

$$\hat{E}_k(s) \pm b_\alpha(n_1 h_{1k} + n_2 h_{2k})^{-1/2}\tilde{\sigma}_k(s). \tag{3.8}$$

Here $\mathcal{S}$ can be chosen as an interval whose lower and upper bounds are the 5th and 95th percentile of the empirical distribution of S(U).

As with any nonparametric estimation problem, it is important that we choose appropriate smoothing parameters in to make inference about $\mathbf{E}(s)$. Here, we use a 'leave-one-out' cross-validation procedure to choose the smoothing parameter $\hat{h}_{jk}$ which minimizes a weighted cross-validated mean squared error, as discussed in Altman and MacGibbon (1998).

Specifically, for any fixed values of h and (j,k), we can estimate $p_{jk}(s)$ using all observations except for the i^{th} subject, which yields estimator $\hat{p}_{(-i)jk}(s)$. The weighted leave-one-out mean squared error is

$$\sum_{i=1}^{n_j} \frac{w_{ijk}}{\hat{G}_{jk}(X_{ijk} \wedge t_0)} \{Y_{ijk} - \hat{p}_{(-i)jk}(V_{ij})\}^2. \tag{3.9}$$

Let $\hat{h}_{jk}$ be a minimizer of (3.9). In the Appendix, we show that $\hat{h}_{jk}$ is of the order $n^{-1/5}$. To ensure the validity of the above large-sample approximation, however, we let the final smoothing parameter be $\tilde{h}_{jk} = \hat{h}_{jk} \times n^{-\xi}$ where ξ is a small positive number less than 0.3.

4. Example

We illustrate the new proposal using the data from the CIBIS-II and BEST trials, as discussed in the Introduction. First we build the scoring system using the data from CIBIS-II. In the CIBIS-II trial, 2647 patients were assigned to either placebo or beta-blocker, with an average followup time of 1.3 years. By the end of the study, 156 patients in the beta-blocker group and 228 patients in the placebo group had died. For each group, we used 15 clinically relevant covariates from Table 1 of Castagno et al. (2010) to fit the mortality data with various working models to estimate the probability of death at $t_0 = 18$ months. These baseline variables are: age, sex, left ventricular ejection fraction (LVEF), estimated glomerular filtration rate adjusted for body surface area (eGFR), systolic blood pressure (SBP), class of heart failure (Class III vs. Class IV), obesity (Body mass index (BMI) > 30 vs. BMI ≤ 30), resting heart rate, smoking status (ever vs. never), history of hypertension, history of diabetes, ischemic heart failure, and atrial fibrillation. As in Castagno et al. (2010), we used 3 indicator variables to discretize eGFR values into 4 categories, with cut-points of 45, 60, and 75. Since the primary endpoint was the time to death, there were no competing risks involved in CIBIS-II. We used two classes of models to fit survival data: the standard Cox model and the t_0-year logistic regression model proposed in Section 2. For each of these

two models, we used four different methods of variable selection. The first one used all 15 variables additively. The second one used a simple backwards stepwise regression procedure. It started from the full model including all 15 covariates and successively eliminated the least significant covariate until all p-values of remaining covariates were less than 0.15. The third and fourth ones used the same backwards elimination procedure. However, the third one stopped when all remaining p-values were less than 0.05 and the fourth one stopped when no more covariates could be removed without subsequently increasing the Akaike information criterion (AIC). Therefore, a total of eight models were considered for each treatment group. Although there were 16 candidate models in total, for each treatment group, there were six distinct candidate models left after variable selection. To evaluate these models, we used a 10-fold cross validation procedure. In Table 1, we present those 12 models with their corresponding average AUC values. It is interesting to note that all these models have very similar AUC values. Our final models for the present example are the two marked with ** in Table 1.

Table 1: Candidate models with average cross-validated AUC values

Placebo Group			**Treatment Group**		
Model	Covariates	AUC	Model	Covariates	AUC
logistic-full	(1-15)	0.651	logistic-full	(1-15)	0.663
logistic-stepwise(p=0.15)	(1,2,3,6,7,8,12,13)	0.666	logistic-stepwise(p=0.15)	(2,3,5,6,8,13,15)	0.695**
logistic-stepwise(p=0.05)	(1,2,3,6,7,8,12)	0.669**	logistic-stepwise(p=0.05)	(2,3,5,6,8,13)	0.690
logistic-stepwise(AIC)	(1,2,3,6,7,8,12,13)	0.666	logistic-stepwise(AIC)	(2,3,5,6,8,12,13,15)	0.689
Cox-full	(1-15)	0.657	Cox-full	(1-15)	0.670
Cox-stepwise(p=0.15)	(1,2,3,4,6,7,8,12)	0.667	Cox-stepwise(p=0.15)	(2,3,5,6,8,12,13,15)	0.689
Cox-stepwise(p=0.05)	(2,3,4,6,7,8)	0.657	Cox-stepwise(p=0.05)	(2,3,5,6,8,12,13,15)	0.689
Cox-stepwise(AIC)	(1,2,3,4,6,7,8,12)	0.667	Cox-stepwise(AIC)	(2,3,5,6,8,12,13,15)	0.689

Covariates: 1. age 2. sex: male 3. LVEF 4. I(eGFR> 75) 5. I(eGFR> 60) 6. I(eGFR> 45) 7. SBP 8. Class IV heart failure 9. I(BMI> 30) 10. never-smoker 11. heart rate 12. history of hypertension 13. history of diabetes 14. ischemic heart failure 15. atrial fibrillation

Now, we apply the final scoring systems to the patients in the BEST trial. Since the patients in CIBIS-II were predominantly white, for illustration, we only considered the data from white patients in BEST (CIBIS II, 1999; BEST, 2001). There were 1895 white patients and the average follow-up time was about 2 years. Two potential scoring systems from Table

Table 2: Regression coefficients for the final logistic models with CIBIS-II data with respect to all-cause mortality

	Placebo		Beta-Blocker
Covariates		Covariates	
(Intercept)	-0.175	(Intercept)	-1.285
Sex: male*	0.682	Sex: male*	1.167
LVEF	-0.034	LVEF	-0.041
I(eGFR>45)*	-0.712	I(eGFR>45)*	-0.822
Class IV Heart Failure*	0.775	Class IV Heart Failure*	1.060
age	0.025	I(eGFR>60)*	-0.543
SBP	-0.018	Hist. Diabetes*	0.686
Hist. Hypertension*	0.474	Atrial Fibrillation*	0.416

* Binary risk factor: 1 if factor is present, 0 otherwise.

2 are considered, $\hat{\pi}_1(\cdot)$ and $\hat{D}(\cdot)$, for the patients in BEST. The four competing risk events are all-cause mortality, MI, any hospitalization, and treatment discontinuation due to AE. Again, only the time to death is a potential competing risk for the times to the non-fatal events. We present the event counts at $t_0 = 18$ months for each treatment arm in Table 3 below, and the cumulative incidence function estimates in Figure 1.

Table 3: BEST 1.5-year Event Totals

	Control Group	**Treated Group**
Death	201	167
MI	30	15
Hospitalization	504	467
AE Discontinuation	234	198
n	950	945

To estimate $\mathbf{p}_1(s)$, $\mathbf{p}_2(s)$ and $\mathbf{E}(s)$ in our examples, we let $K(\cdot)$ be the standard Epanechnikov kernel. The smoothing parameters were chosen as the minimizers of (3.9) using "leave-one-out" cross-validation, then multiplied by $n_j^{-0.05}$.

First, we used the risk score $\hat{\pi}_1(\cdot)$ from the CIBIS-II control group to construct estimates for the 1.5-year event rates for the white patients in the control group of BEST. In the left panel of Figure 2, the point and 0.95 interval estimates are presented. We find that our

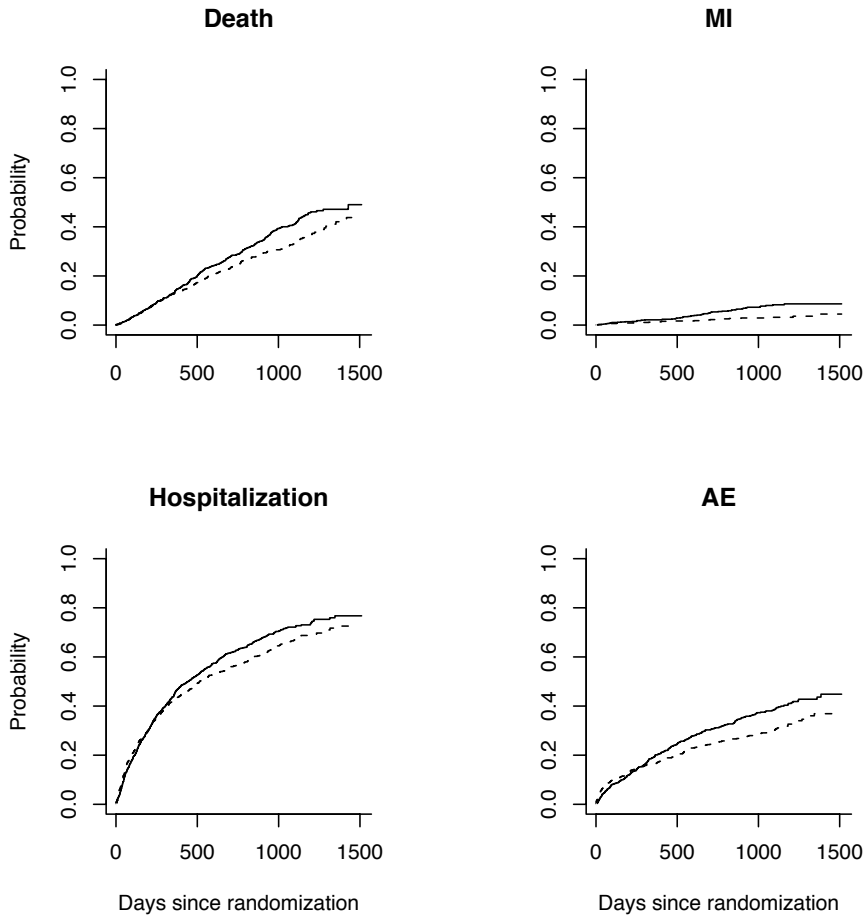

Figure 1: Cumulative incidence function estimates for each endpoint (solid: control group, dashed: treated group).

risk scoring model $\hat{\pi}_1(\cdot)$ matches the t_0-year survival profile well. Furthermore, the 1.5-year event rate for each non-fatal event increases with this risk score derived from the CIBIS-II mortality data. In Figure 2, the 0.95 pointwise interval and band are denoted by dashed and solid lines, respectively.

Next, we estimate the treatment difference for each competing risk over the score $\hat{\pi}_1(\cdot)$. We present the estimates in the right panel of Figure 2. A common clinical practice is to use such a risk scoring system to guide us for the patient's treatment management. The results of our analysis suggest that patients with a low risk score would experience a nontrivial benefit

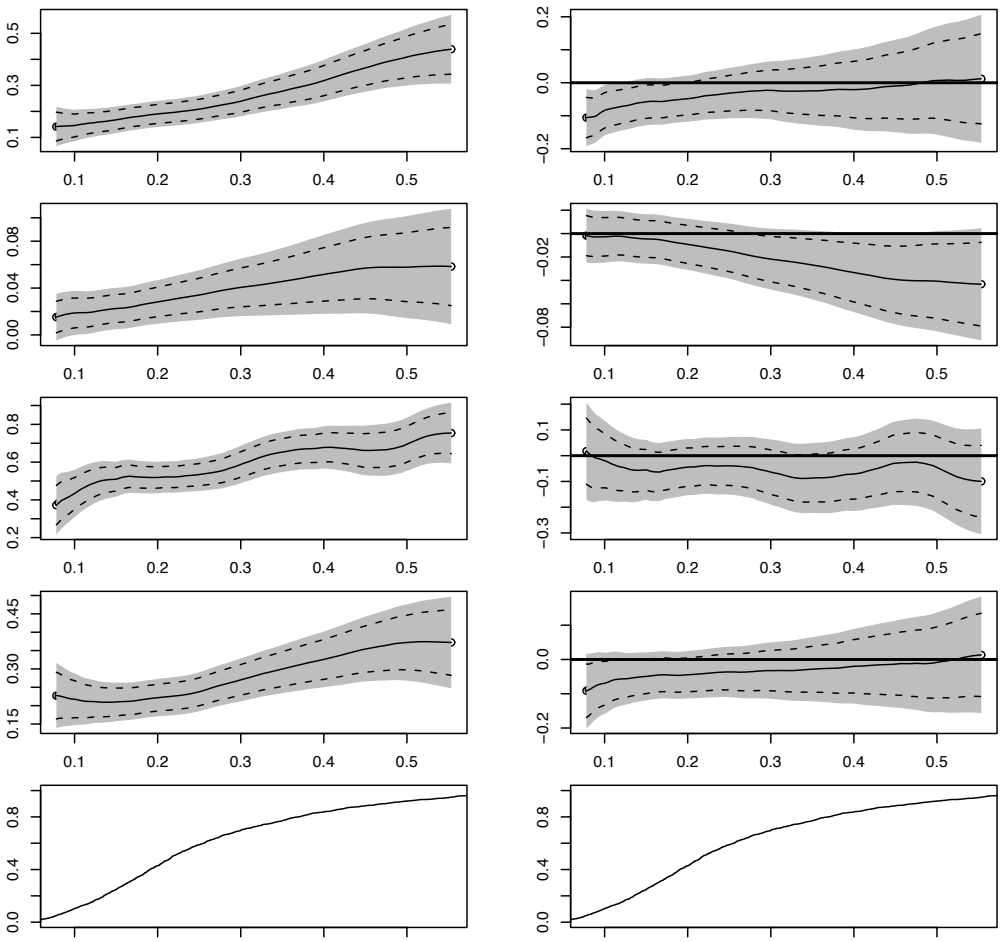

Figure 2: Use of $\hat{\pi}_1(U)$ as a scoring system (left panel: 1.5-year event probabilities in control group; right panel: 1.5-year treatment effect; from top: death, MI, hospitalization, AE discontinuation; bottom row: empirical CDF of risk scores).

from beta-blockers with respect to overall mortality. The pointwise 0.95 confidence interval estimates indicate that patients with risk scores below 0.2 show a significant benefit in terms of reduced risk of death, and those with risk scores below 0.12 show significant effects using the simultaneous confidence band estimate. These two subsets represent approximately 36% and 7%, respectively, of the BEST patient population. The risk of discontinuation due to adverse events also appears to be decreased for patients with low risk scores, showing (pointwise) significant effects for those with scores below 0.14. On the other hand, the subset of patients with scores greater than 0.28 showed significant reductions in MI. These same pa-

tients show no evidence of reduction with respect to any of the other risks, with considerable uncertainty surrounding the estimates for these patients. These high-risk patients make up approximately 30% of the BEST patient population. There is no strong evidence to claim any increase or decrease in risk of hospitalization over the score. Note that high risk scores were most strongly associated with increased age and low values of LVEF.

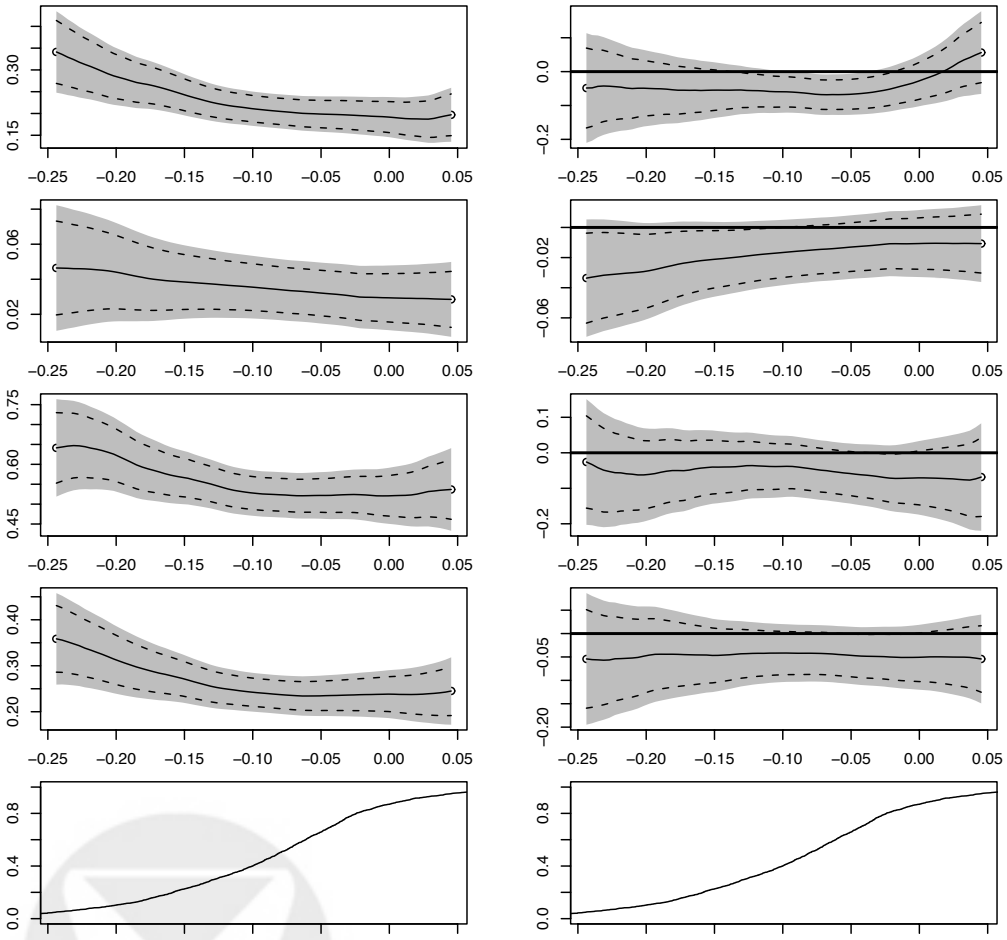

Figure 3: Use of $\hat{\pi}_2(U) - \hat{\pi}_1(U)$ as a scoring system (left panel: 1.5-year event probabilities in control group; right panel: 1.5-year treatment effect; from top: death, MI, hospitalization, AE discontinuation; bottom row: empirical CDF of risk scores).

Next, we used the treatment selection score $S(U) = g(\hat{\beta}_2' Z_2) - g(\hat{\beta}_1' Z_1)$ from CIBIS-II to estimate the subject-specific treatment difference with respect to each competing risk. A patient who has a negative score would be anticipated to experience a decrease in risk of death

by taking the treatment, while a patient who receives a positive score would be expected to experience an increase in risk of death with the beta-blocker. The resulting scores range from -0.38 to +0.25, with empirical 5th and 95th percentile values of approximately -0.24 and +0.05, respectively. The median score in the BEST population was -0.08. It is important to note that due to the highly significant treatment benefit found in CIBIS-II, the score has a large mass on the negative side. On the other hand, the overall treatment benefit from beta-blocker in BEST was rather modest. Therefore, this score may not reflect the true treatment difference well for mortality in the BEST population.

The results from our analysis suggest that patients with negative scores show reductions in mortality and MI. In particular, the 0.95 confidence intervals indicate that a score below -0.02 (81% of the patient population) is associated with a significant reduction in either death or MI. On the other hand, patients with scores greater than 0.02 showed an estimated (non-significant) increased risk of mortality.

Note that for the treatment difference with respect to overall mortality, the treatment selection score system cannot differentiate well between patients with negative scores. For example, the curve is relatively flat for scores between -0.25 and -0.02. For the present example, it appears that the risk score $\pi_1(\cdot)$ works well. Note that the treatment selection score was most strongly associated with increased SBP values and history of diabetes.

5. Remarks

In this paper, we use two independent data sets to construct a systematic, subject-specific treatment selection procedure. The final scoring system may be chosen via a complex, exploratory model and variable selection process using the training data set. We then apply this system to stratify the patients in the second study and make inferences about the treatment differences with respect to various competing risks for each stratum. If two similar studies are unavailable, one may instead split a single data set randomly into two

pieces and implement our proposal accordingly. However, if different random splits result in markedly different profiles for the treatment differences, this indicates that our data may not have enough information for making inference for risk prediction or personalized treatment selection.

Although using a treatment selection score $\hat{D}(\cdot)$ may be a more effective procedure to group subjects with similar treatment difference profiles within each stratum, the resulting score may not represent the patients in the second study well, due to nontrivial differences in treatment efficacy between the two studies. On the other hand, the risk score built using the control group patients in the training data set may be validated with the data from the second study, as shown in our example. Moreover, clinical practitioners seem more comfortable with using a risk score from the control arm to make treatment decisions, especially when there is more than one treatment involved in the comparison.

In this paper, we used the t_0-year event rates as the parameters of interest, where t_0 may be chosen from a clinical perspective. In practice, one may repeat our procedure with various time points. It would be interesting to choose a global measure to quantify the treatment contrast, for example, the difference, between two treatment groups, of the areas under the cumulative incidence function, truncated at a specific time point. Further research is warranted along this line.

Our model and variable selection procedure is intended to select the "best" model for each treatment group among all candidate models, where the two models are built and evaluated independently of one another. When the endpoint is the treatment difference, it is not clear that our approach of using two independent models would produce the "best" score. In his unpublished thesis, Signorovitch (2007) proposed a novel method for modeling the treatment contrast directly with covariates. Intuitively, his approach would more effectively select the treatment and covariate interactions for creating the score. Further research is needed to evaluate scoring systems with respect to the subject-specific treatment differences.

Lastly, the choice of treatment based on a risk-benefit perspective is quite individual-

ized. A global summary of the treatment *effectiveness*, for example, the risk-benefit ratio, may not provide enough information for personalized medicine. Separate summaries for the treatment's toxicity and efficacy at the subject level, as we have proposed in this article, can be quite useful for the patient's bedside management.

Acknowledgments

This manuscript was prepared using BEST and CIBIS-II Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and from the CIBIS-II Scientific Committee and does not necessarily reflect the opinions or views of the BEST and CIBIS-II investigators or the NHLBI.

Appendix

With slight abuse of notation, we only consider the smoothing parameter selection in one treatment group and one type of event. For convenience, we drop the subscripts for treatment arms as well as type of event. Specifically, we let $Y_i = I(X_i \leq t_0, \epsilon_i = 1), i = 1, \cdots, n$, $p(v) = \text{pr}(T \leq t_0, \epsilon = 1 | V = v)$, where $V = S(U)$, $\hat{W}_i = w_i / \hat{G}(X_i \wedge t_0)$ and $W_i = w_i / G(X_i \wedge t_0)$, where $G(\cdot)$ is the survival function for the censoring distribution. $\hat{h}_{CV}$, the bandwidth selected via cross-validation, is the minimizer of the weighted least square loss function

$$\hat{CV}(h) = \frac{1}{n} \sum_{j=1}^{n} \hat{W}_j \left[Y_j - \frac{\sum_{i \neq j} K_h(V_i - V_j) \hat{W}_i Y_i}{\sum_{i \neq j} K_h(V_i - V_j) \hat{W}_i} \right]^2 .$$

Firstly, one can show that $\hat{CV}(h)$ can be approximated by

$$CV(h) = \frac{1}{n} \sum_{j=1}^{n} W_j \left[Y_j - \frac{(n-1)^{-1} \sum_{i \neq j} K_h(V_i - V_j) W_i Y_i}{f(V_j)} \right]^2$$

in that $|\hat{CV}(h) - CV(h)|/\hat{CV}(h) = o_p(1)$ uniformly in $h \in H_n = [n^{-1+\delta}, n]$, where $\delta > 0$, and $f(\cdot)$ is the density function of V_i. Let

$$Q(h) = \frac{1}{n}\sum_{j=1}^{n} W_j \left[Y_j - \frac{n^{-1}\sum_{i=1}^{n} K_h(V_i - V_j) W_i Y_i}{f(V_j)} \right]^2 .$$

$CV(h)$ can be expressed as $Q(h) + A + B$ where

$$A = \frac{2K(0)}{n^2 h}\sum_{j=1}^{n} W_j \left[Y_j - \frac{n^{-1}\sum_{i=1}^{n} K_h(V_i - V_j) W_i Y_i}{f(V_j)} \right] \frac{W_j Y_j}{f(V_j)} + O_p(n^{-1})$$

and

$$B = n^{-1}\sum_{j=1}^{n} W_j \left[\frac{(n-1)^{-1}\sum_{i\neq j} K_h(V_i - V_j) W_i Y_i}{f(V_j)} - \frac{n^{-1}\sum_{i=1}^{n} K_h(V_i - V_j) W_i Y_i}{f(V_j)} \right]^2 .$$

Furthermore, one can verify that

$$Q(h) = E[p(V)\{1 - p(V)\}] + O_p(n^{-1}h^{-1})$$

$$A = \frac{2K(0)}{nh} E\left[\frac{\{Y - p(V)\}^2}{G(X \wedge t_0) f(V)} \right] + O_p(n^{-1}h^{-1})$$

and $B = O_p(n^{-2}h^{-2})$ uniformly in $h \in H_n$. Therefore

$$\hat{CV}(h) = Q(h)\left(1 + \frac{2K(0)}{nh} E\left[\frac{\{Y - p(V)\}^2}{G(X \wedge t_0) f(V)} \right] \frac{1}{E\left[p(V)\{1 - p(V)\}\right]} + O_p(n^{-2}h^{-2}) \right)$$

It follows from Theorems 1 and 2 of Hardle et al. (1988) that $\hat{h}_{CV}$ is consistent to the optimal bandwidth and in the order of $O_p(n^{-\frac{1}{5}})$.

References

Altman, N. and MacGibbon, B. (1998). Consistent bandwidth selection for kernel binary regression. *Journal of Statistical Planning and Inference*, (**70**), 121.

The Beta-Blocker Evaluation of Survival Trial Investigators (**2001**). A Trial of the Beta-Blocker Bucindolol in Patients with Advanced Chronic Heart Failure. *New England Journal of Medicine*, **344**(22), 1659.

Bickel, P. and Rosenblatt, M. (**1973**). On Some Global Measures of the Deviations of Density Function Estimates. *Annals of Statistics*, **1**, 1071.

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.

Bonetti, M. and Gelber, R. D. (**2005**). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, **5**, 465.

Cai, T., Tian, L., Uno, H., Solomon, S. and Wei, L.J. (**2010**) Calibrating Parametric Subject-specific Risk Estimation. *Biometrika*, (**97**), 389.

Cai, T., Tian, L., Wong ,P.H., and Wei, L.J. (**2010b**) Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, Ahead of print.

Castagno, D., Jhund, P., McMurray, J.J.V., Lewsey, J., Erdmann, E., Zannad, F., Remme, W., Lopez-Sendon, J.L., Lechat, P., Follath, F., Höglund, C., Mareev, V., Sadowski, Z., Seabra-Gomes, R.J., and Dargie, H.J. (**2010**). Improved survival with bisoprolol in patients with heart failure and renal impairment: an analysis of the cardiac insufficiency bisoprolol study II (CIBIS-II) trial. *European Journal of Heart Failure*, **12**, 607.

CIBIS II Investigators and Committees (**1999**). The Cardiac Insufficiency Bisoprolol Study II (CIBIS II): a randomised trial. *Lancet*, **353**, 9.

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.

Fine, J. Gray, R (**1999**). A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*, **94**, 496.

Hardle, W., Hall, P. and Marron, J. S. (**1988**). How far are automatically chosen regression smoothing parameters from their optimum? with discussion. *J. Amer. Statist. Assoc.*, 83, 86-101.

Huang, Y., Pepe, M.S. and Feng, Z. (**2007**). Evaluating the predictiveness of a continuous marker. *Biometrics*, **63**, 1181.

Janes, H., Pepe, M.S., Bossuyt, P., and Barlow, W. E. (**2011**). Measuring the Performance of Markers for Guiding Treatment Decisions. *Annals of Internal Medicine*, **154**, 253.

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.

Mammen, E. (**1993**) Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *Annals of Statistics*, (**21**), 255-285.

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

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

Pepe, M.S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I.M. and Zheng, Y. (**2008**). Integrating the predictiveness of a marker with its performance as a classifier. *Am J Epidemiol.*, **167**, 362.

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

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

Uno, H., Cai, T., Tian, T. and Wei, L.J. (**2007**). Evaluating Prediction Rules for t-Year Survivors With Censored Regression Models. *Journal of the American Statistical Association*, **102**, 527-537.

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.

Zhao, L., Cai, T., Tian, L., Uno, H., Solomon, S.D., and Wei, L.J. (**2010**), Stratifying Subjects for Treatment Selection With Censored Event Time Data From a Comparative Study. *Harvard University Biostatistics Working Paper Series*, Working Paper 122.

Zheng, Y., Cai, T., Feng, Z. (**2006**)Application of the Time-Dependent ROC Curves for Prognostic Accuracy with Multiple Biomarkers. *Biometrics*, (**62**), 279.