Harvard University
Harvard University Biostatistics Working Paper Series

Year 2012 *Paper* 153

Treatment Selections using Risk-benefit Profiles Based on Data from Comparative Randomized Clinical Trials with Multiple Endpoints

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

*Harvard School of Public Health, bclagget@hsph.harvard.edu
†Stanford University School of Medicine, lutian@stanford.edu
‡University of Turin
**Harvard School of Public Health, wei@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper153

Treatment Selections using Risk-Benefit Profiles Based on Data from Comparative Randomized Clinical Trials with Multiple Endpoints

Brian Claggett, Lu Tian, Davide Castagno, and L. J. Wei*

Abstract

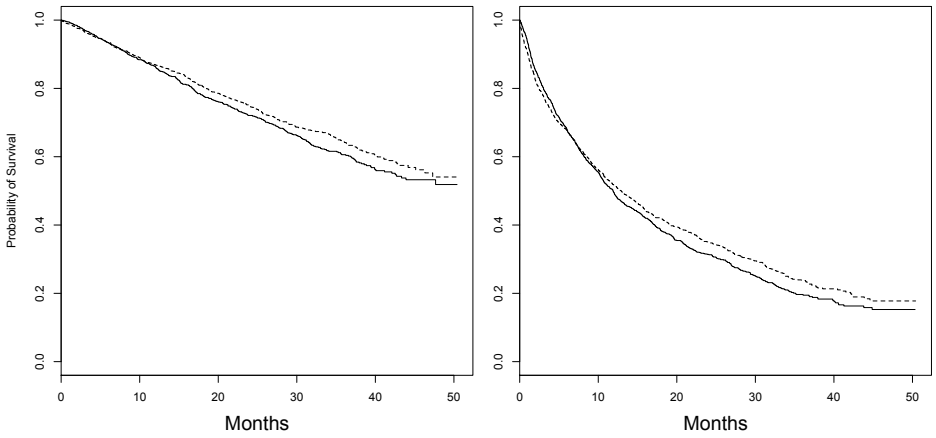
In a longitudinal, randomized clinical study to compare a new treatment with a control, oftentimes each study subject may experience any of several distinct outcomes during the study period, which collectively define the "risk-benefit" profile. To assess the treatment difference, it is desirable to utilize the entirety of such outcome information. The times to these events, however, may not be observed completely due to competing risks. The standard analyses based on the time to the first event, or individual component analyses with respect to each event time, are not ideal. In this paper, we classify each patient's risk-benefit profile, by considering all event times during follow-up, into several clinically meaningful ordinal categories. We first show how to make inferences for the treatment difference in a two-sample setting with incomplete categorical data. To bring the study results to the individual patient's bedside, we then present a systematic procedure to identify patients who would benefit from a specific treatment using baseline covariate information. Specifically, we split the data set into two independent pieces. Using the data from the first piece, we build, as a function of the baseline covariates, a scoring system for assessing treatment differences, with the final model(s) chosen via a cross-validation process. With the data from the second piece, we non-parametrically estimate the treatment differences across a range of the resulting scores. A desirable subgroup of patients can then be identified with respect to the size of the treatment difference for treatment selection. The proposal is illustrated with the data from a clinical trial to evaluate a beta-blocker for treating chronic heart failure patients, for whom it was unclear whether known risks of beta-blocking agents would outweigh anticipated benefits (Beta-Blocker Evaluation of Survival Trial Investigators, 2001).

KEY WORDS: Clinical trial; Nonparametric estimation; Ordinal regression model; Personalized medicine; Perturbation-resampling method; Stratified medicine; Subgroup analysis; Survival analysis.

*Brian Claggett is Postdoctoral Fellow, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 (E-mail: *bclagget@hsph.harvard.edu*). Lu Tian is Assistant Professor, Department of Health Research & Policy, Stanford University School of Medicine, Palo Alto, CA 94304 (E-mail: *lutian@stanford.edu*). Davide Castagno is Assistant Professor, Division of Cardiology, Department of Medical Sciences, University of Turin, Italy. L.J. Wei is Professor, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115 (E-mail: *wei@hsph.harvard.edu*).

1. INTRODUCTION

Consider a randomized, comparative clinical trial in which a treatment is assessed against a control with respect to their risk-benefit profiles. For each study patient, the outcome variables include a set of distinct event time observations reflecting such profiles during the study period. Often these event times cannot be observed completely due to the presence of competing risks. For example, to investigate if the beta-blocking drug, bucindolol, would benefit patients with advanced chronic heart failure (HF), a clinical trial, Beta-Blocker Evaluation of Survival Trial (BEST), was conducted (Beta-Blocker Evaluation of Survival Trial Investigators, 2001). There were 2708 patients enrolled and followed for an average of two years. The primary endpoint of the study was the patient's overall survival time. For patients in the two treatment arms, the Kaplan-Meier estimates for survival are given in Figure 1(a) with a p-value of 0.10 based on the standard two-sample logrank test, favoring the beta-blocker but not reaching statistical significance. Although mortality is an important endpoint, the treatment benefit evaluation should also include morbidity for chronic heart failure patients. One important morbidity measure is the time to hospitalization, especially due to worsening HF. Unfortunately such event times may be "informatively" censored, for example, by the patient's survival time. To avoid such competing-risk problems, one may consider the time to the first event among all competing events of interest as the endpoint. For example, for the BEST study, the competing events are death and HF or non-HF hospitalization. In Figure 1(b), we present the corresponding KM estimates for such an event time analysis. With this endpoint, the beta-blocker is not statistically significantly better than the control, with a p-value of 0.14. Note that this type of endpoint does not fully reflect the disease burden or progression over the entire duration of the patient's follow-up, since only one event at most is utilized per patient. In Table 1, we show the frequencies of the occurrences of these component endpoints from the study patients whose data were obtained from the National Heart, Lung and Blood Institute. Note that mortality may be classified as either cardiovascular (CV) or non-CV related, in which case it may be expected that an effective



(a) Overall survival (b) Time to first event

Figure 1: Kaplan-Meier estimates from BEST trial. Solid and dashed lines represent control and treated groups, respectively.

beta-blocker would lower the rate of events classified into the former category, but have no impact on the latter category. In general, it is not expected that a beta-blocker would have any beneficial effect on non-CV outcomes. In addition, part of the undesirable side effects of beta-blockers may be captured by non-CV outcomes (for example, non-CV related death or non-HF hospitalization). It is also important to note that a patient may have multiple events during the study follow-up reflecting the disease progression. The times to all these events contain valuable clinical information for comparing two groups with respect to overall disease burden and should not be ignored for the analysis.

Table 1: Numbers of Patients Experiencing Specific Clinical Endpoints in Control and Treatment Groups in BEST

Outcome	Control	Treated
Death	448	411
CV Death	388	342
Non-CV Death	60	69
Any Hosp.	874	829
HF Hosp.	568	476
Non-HF Hosp.	634	619
Total Patients	1353	1354

**CV=Cardiovascular, *HF=Heart Failure*

For a typical cardiovascular study like BEST with multiple event time observations,

conventional secondary analyses for risk-benefit assessments are often conducted with each type of endpoint (for example, the time to HF hospitalization). The conclusions of such component analyses can be misleading due to competing risks. Moreover, because component events are analyzed separately rather than jointly, they cannot provide a global, clinically meaningful evaluation of the new treatment. To this end, there are novel procedures for handling multiple event time observations proposed, for example, by Andersen and Gill (1982), Wei et al. (1989), and Lin et al. (2000). In the presence of competing risks, however, the above procedures or their modifications are not entirely satisfactory for assessing the treatment's overall risk and benefit (Li and Lagakos, 1998; Ghosh and Lin, 2003; Pocock et al., 2012).

In this article, we create an ordinal categorical outcome variable which reflects the individual patient's morbidity, including toxicity, and mortality over the entire study period for evaluating the treatments. Such a classification system may be constructed by a panel of clinical experts who can classify various possible patient outcome patterns into categories (e.g., "improved", "stabilized", or "worsened", or finer ordinal subcategories). For example, for the BEST study, with guidance from our cardiologist co-author, we classified patient response, using five ordinal categories, based on the disease burden during 18 months of follow-up. Category 1 is assigned if the patient has experienced neither death nor any hospitalization prior to the time of evaluation. The patient is classified as Category 2 if he or she is alive and has experienced only non-HF hospitalization (reflecting potential toxicity). Category 3 denotes patients who are alive, but have experienced HF-hospitalization (reflecting lack of efficacy). Category 4 is assigned if the patient has died from non-CV related causes. Finally, Category 5 refers to those patients who suffered CV-related death. Note that often, study patients might not have their entire clinical history, until their time of death or at 18 months after randomization, available due to non-informative, or administrative, censoring.

In the paper, we first present methods for analyzing such ordinal, possibly incomplete, data in a two-sample overall comparison setting. Note that making patient-specific decisions

based on estimated population-averaged effects can lead to sub-optimal patient care (Kent and Hayward, 2007). A positive (negative or neutral) trial based on some overall assessment 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 to perform 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 therefore may not be effective, from a risk-benefit perspective, in identifying patients who would benefit from the new treatment. Here, we present a systematic approach to create a scoring system using the patient's multiple baseline covariates and utilize this system to stratify the patients for evaluation with respect to the ordinal categorical outcomes. More specifically, to avoid overly optimistic model selections, we first divide the data set into two pieces. The two pieces may be obtained by splitting the entire data set randomly or sequentially, dividing the data according to the order in which patients entered into the study. With the first piece, a cross-validation procedure is utilized to select the best scoring system among all of the competing models of interest for ordinal categorical data. We then use the second piece (the so-called holdout sample) to make inferences about the treatment differences over a range of the score selected from the first stage. All proposals are illustrated with the data from the BEST study.

When there is a single baseline covariate involved, Bonetti et al. (2000), Song and Pepe (2004), and Bonetti and Gelber (2004) have proposed novel statistical procedures for identifying a subgroup of patients who would benefit from the new treatment with respect to a single outcome. A recent paper by Janes et al. (2011), based on previous work by Pepe (2004), Huang et al. (2007), and Pepe et al. (2008), provides practical guidelines for assessing the performance of individual markers for the purposes of treatment selection. By incorporating more than one baseline covariate, our approach is similar in spirit to Cai et al. (2011) and Li et al. (2011). However, they used the data from the entire study to create a scoring

system by fitting a *prespecified* model without involving model evaluation or variable selection and then used the same data set to make inferences for either the treatment difference with respect to a single outcome or for risk predictions for a single treatment group only. Note that Chuang-Stein et al. (1991) utilized an ordinal categorical outcome with individual patients' subjective weightings of the categories for a summary measure for personalized treatment selection. Their novel approach is quite different from our proposal.

2. TWO-SAMPLE ASSESSMENT OF TREATMENT EFFECT USING INCOMPLETE CATEGORICAL OBSERVATIONS

For the jth patient in the ith treatment group, $(j = 1, \ldots, n_i;\ i = 1, 2)$, let T_{ij} be the time to the first occurrence of a *terminal* event from among the competing risks of interest. Note that T_{ij} may be infinite if there is no terminal event. Let C_{ij} be the independent censoring variable for T_{ij}. Let $X_{ij} = \min(T_{ij}, C_{ij})$ and Δ_{ij} is an indicator variable, which is one if $T_{ij} \leq C_{ij}$. Let G_i be the survival function of the censoring variable C_{ij}. For each study patient, assume that based on his/her entire morbidity and mortality endpoint information up to time t_0, where $\mathrm{pr}(C_{ij} > t_0) > 0$ $(i = 1, 2)$, one can classify the outcome ϵ as one of K ordered categories. In general, we make no assumptions about how the patient's classification may change over time. That is, the patient's classification may improve or worsen during the follow-up time. Note that we do not require traditional "competing risks" methods to account for informative censoring because we include such informative events in the definition of the patient outcome categories.

Let π_{ik} be the probability of $\epsilon = k$ with treatment i, $(i = 1, 2; k = 1, \ldots, K)$. Let $\epsilon_{ij}, j = 1, \ldots, n_i$, be the response for the jth patient in the ith group. Noting that a patient's outcome status is observable only when $\min(T_{ij}, t_0) \leq C_{ij}$, the cell probabilities π_{ik} can be consistently estimated, via inverse probability of censoring weighting, by

$$\hat{\pi}_{ik} = \sum_{j=1}^{n_i} \frac{w_{ij} I(\epsilon_{ij} = k)}{\hat{G}_i(X_{ij} \wedge t_0)} \Big/ \sum_{j=1}^{n_i} \frac{w_{ij}}{\hat{G}_i(X_{ij} \wedge t_0)}, \tag{1}$$

where $I(\cdot)$ is the indicator function, $w_{ij} = I(X_{ij} \leq t_0)\Delta_{ij} + I(X_{ij} > t_0)$, and $\hat{G}_i(\cdot)$ is the Kaplan-Meier estimator for $G_i(\cdot)$ (Li et al., 2011). In order to compare two treatment groups with such ordinal categorical outcomes, one may compare the cumulative distributions $\gamma_{ik} = \sum_{l=1}^{k} \pi_{il}$, $i = 1, 2$; $k = 1, \ldots, K$. Let $\Gamma_k = \gamma_{2k} - \gamma_{1k}$ and let $\hat{\gamma}_{ik}$ be the corresponding estimators via $\hat{\pi}_{ik}$. Note that each value Γ_k, $k = 1, \ldots, K-1$, may be interpreted as the risk difference with respect to a binomial outcome in which "success" is defined by a patient experiencing ($\epsilon \leq k$). To make inferences on the difference of these two distribution functions, we may use bootstrapping or perturbation-resampling methods (Uno et al., 2007). Details are provided in the Appendix.

For the data from BEST, let $t_0 = 18$ months. Using the five ordinal categories described in the Introduction, Table 2 displays the profiles of the estimated distribution functions for each treatment group γ_{ik} and the differences Γ_k. For each level k, the estimated distribution function for the beta-blocker group ($\hat{\gamma}_{2k}$) is larger than for the control group ($\hat{\gamma}_{1k}$), indicating that the beta-blocker group is numerically better than its control counterpart with respect to each outcome.

Table 2: Estimated distribution functions for control and treated groups with BEST data with $t_0 = 18$ months

	Control ($\hat{\gamma}_1$)		**Treated ($\hat{\gamma}_2$)**		**Contrast ($\hat{\Gamma}$)**	
Outcome Category	n	pr($\epsilon \leq k$)	n	pr($\epsilon \leq k$)	Est	SE
1	397	0.37	442	0.41	0.04	0.03
2	174	0.54	224	0.62	0.08	0.02
3	251	0.77	190	0.80	0.03	0.02
4	35	0.80	39	0.83	0.03	0.02
5	246	1.00	211	1.00	-	-
(censored)	250	-	248	-	-	-

To compare two groups with respect to ordinal categorical outcomes, a conventional way to summarize the treatment difference is to use an ordinal regression model. Let $\tau_{ij} = 1$ for patients in treated group and 0 otherwise, then this model is:

$$g(\text{pr}(\epsilon_{ij} \leq k)) = \alpha_k - \beta\tau_{ij}, \tag{2}$$

where $g(\cdot)$ is a known, increasing function and α_k and β are unknown parameters. Here β can be interpreted as an overall measure of the treatment difference even if the model is not correctly specified. Under such parameterization, for the present case, a negative value for β corresponds to an reduction in overall "risk" associated with treatment. With censored observations, the treatment difference β can be estimated by maximizing the standard weighted multinomial log-likelihood function:

$$\sum_{ij} \frac{w_{ij}}{\hat{G}_i(X_{ij} \wedge t_0)} [\sum_{k=1}^{K} I(\epsilon_{ij} = k) log\{g^{-1}(\alpha_k - \beta\tau_{ij}) - g^{-1}(\alpha_{k-1} - \beta\tau_{ij})\}], \tag{3}$$

where $\alpha_0 = -\infty, \alpha_K = \infty$, and standard error estimates can be obtained via perturbation-resampling methods. Under mild conditions, the estimator $\hat{\beta}$ from the above model converges to a finite constant β as $n \to \infty$ even when the model is not correctly specified (Zheng et al., 2006; Uno et al., 2007; Li et al., 2011). For the data from BEST, when $g(\cdot)$ is the logit function, $\hat{\beta}$ is -0.227 with a standard error estimate of 0.072. This indicates that the beta-blocker indeed reduces the disease burden. Details are given in the Appendix.

Rather than using a parametric summary of the treatment difference, an intuitively interpretable, nonparametric summary measure is the so-called general risk difference, which has been studied extensively as an extension of the simple risk difference for ordinal data (Simonoff et al., 1986; Agresti, 1990; Edwardes, 1995; Edwardes and Baltzan, 2000; Lui, 2002). In this setting, the general risk difference, which is closely related to Wilcoxon's rank-sum statistic, is $D = \text{pr}(\epsilon_1 > \epsilon_2) - \text{pr}(\epsilon_1 < \epsilon_2)$, where ϵ_i, $i = 1, 2$, is a patient response randomly chosen from treatment group i, with positive values suggesting that treated patients ($i = 2$) are generally more likely to be in a "healthier" state rather than an "unhealthier" state, compared to their control counterparts ($i = 1$). Here ϵ_1 and ϵ_2 are independent. A consistent estimator for D then is $\hat{D} = \sum_{k=2}^{K} \hat{\pi}_{1,k}\hat{\gamma}_{2,k-1} - \hat{\pi}_{2,k}\hat{\gamma}_{1,k-1}$. The standard error estimate can be obtained via bootstrap (Simonoff et al., 1986) or perturbation-resmapling methods as in Uno et al. (2007). For the data from the BEST trial, $\hat{D} = 0.069$ with standard error estimate of 0.023. Using this model-free summary of the treatment difference, the beta-blocker again

appears better than the control. Details are given in the Appendix.

3. CONSTRUCTION AND SELECTION OF A PATIENT-LEVEL STRATIFICATION SYSTEM

Suppose that U_i is the baseline covariate vector for a subject randomly chosen from the ith treatment group ($i = 1, 2$). Our goal is to make inference about the treatment difference based on ϵ_1 and ϵ_2, conditional on $U_1 = U_2 = u$, any given value in the support of the covariate vector. Ideally, one would estimate this conditional treatment difference via a nonparametric procedure. However, if the dimension of U is greater than one, it seems difficult, if not impossible, to do so. A practical alternative is to model the relationship between the treatment difference and U parametrically and then validate the selected model. To avoid an "overly optimistic" personalized prediction model, we split the data set into two pieces, say, part A and part B. With the data from part A, we build various candidate models for the conditional treatment differences and evaluate them via a cross-validation procedure. This results in a univariate scoring system with which to stratify the patients, which we refer to as a treatment selection score. In this section, we present the first step using the part A data, i.e., the construction and selection of the scoring system, and in the next section, we show how to make inferences about the treatment differences based on the selected scoring system using the part B data. It is important to note that, to validate the scoring system, we need a model-free summary measure for the treatment difference. For the present case with the ordinal categorical response discussed in Section 2, the treatment contrast,

$$D(u) = \text{pr}(\epsilon_1 > \epsilon_2 | U_1 = U_2 = u) - \text{pr}(\epsilon_1 < \epsilon_2 | U_1 = U_2 = u), \tag{4}$$

is model-free and heuristically interpretable. Note also that to obtain a coherent prediction system, it is preferable to use the same treatment contrast measure for model building, selection and validation.

3.1 Creating Treatment Difference Scoring Systems

There are numerous ways to estimate (4) parametrically. For instance, one can model the ordinal categorical response via two separate ordinal regression models, that is, for each treatment i and conditional on U_{ij}:

$$g_i(\gamma_{ik}(U_{ij})) = \alpha_{ik} - \beta_i' Z_{ij}, \qquad i = 1, 2; j = 1, \ldots, n_i \tag{5}$$

where $\gamma_{ik}(U_{ij}) = \text{pr}(\epsilon_i \leq k|U_{ij})$, Z_{ij} is a function of U_{ij}, $g_i(\cdot)$ is a known monotone increasing function, and α_{ik} and β_i are unknown parameters. It follows that a parametric estimate $\hat{D}(u)$ for $D(u)$ is given by

$$\hat{D}(u) = \sum_{k=1}^{K} \hat{\pi}_{1,k}(u)\hat{\gamma}_{2,k-1}(u) - \hat{\pi}_{2,k}(u)\hat{\gamma}_{1,k-1}(u) \tag{6}$$

where estimated probabilities $\hat{\gamma}_{ik}(u)$ are obtained from the fitted models (5) and $\hat{\pi}_{i,k}(u) = \hat{\gamma}_{i,k}(u) - \hat{\gamma}_{i,k-1}(u)$, with $\hat{\gamma}_{i,0} = 0$, $i = 1, 2$.

Alternatively, we may use a single model

$$g(\gamma_{ik}(U_{ij})) = \alpha_k - \beta' Z_{ij} - \tau_{ij}(\theta' Z^*_{ij}), \tag{7}$$

where $Z^*_{ij} = (1, Z'_{ij})'$, and α, β, and θ are unknown parameters. The resulting probability estimates $\hat{\gamma}_{ik}(u)$ may similarly be used to estimate $\hat{D}(u)$ via (6).

Models (5) may be fitted to the data by applying inverse probability of censoring weights and maximizing the group-specific weighted multinomial log-likelihood functions

$$\sum_{j=1}^{n_i} \frac{w_{ij}}{\hat{G}_i(X_{ij} \wedge t_0)} [\sum_{k=1}^{K} I(\epsilon_{ij} = k) log\{g^{-1}(\alpha_{ik} - \beta_i' Z_{ij}) - g^{-1}(\alpha_{i,k-1} - \beta_i' Z_{ij})\}], \tag{8}$$

where $\alpha_0 = -\infty, \alpha_K = \infty$, $i = 1, 2$. For model (7), the log-likelihood function is

$$\sum_{ij} \frac{w_{ij}}{\hat{G}_i(X_{ij} \wedge t_0)} [\sum_{k=1}^{K} I(\epsilon_{ij} = k) log\{g^{-1}(\alpha_k - \beta' Z_{ij} - \tau_{ij}(\theta' Z^*_{ij})) - g^{-1}(\alpha_{k-1} - \beta' Z_{ij} - \tau_{ij}(\theta' Z^*_{ij})\}], \tag{9}$$

Under some mild conditions, the resulting estimators $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ from the above models converge to a finite constant vector as $n \to \infty$ even when the model (5) or (7) is not correctly specified (Uno et al., 2007). Note that one may repeatedly utilize (5) or (7) along with (8) or (9) using various Z and $g(\cdot)$ via, for instance, a stepwise regression procedure, to obtain final estimates $\hat{\gamma}_{ik}(U)$ and $\hat{D}(U)$.

3.2 Evaluation and Selection of a Final Model for Stratification

To choose the "best" stratification system from among many possible candidates obtained via the process described in Section 3.1, we use a cross-validation procedure. Specifically, we split the data into two parts randomly. We fit the data from the first part with each of the models of interest, then use the data from the second part to evaluate them via an intuitively interpretable, model-free criterion. Note that unlike the one-sample risk prediction problem, most standard evaluation criteria based on individual prediction errors (e.g., with respect to the L_1 or L_2 norm) are not applicable here because each study patient was only assigned to either the treatment or control, not both. However, a "goodness of fit" measure using the concordance between the true treatment difference $D(u)$ in (4) and the rank of the parametric score $\hat{D}(u)$, say, $C = \text{Cov}\{H(\hat{D}(U)), D(U)\}$, can be estimated consistently under the current setting, where $H(\cdot)$ is the distribution function of $\hat{D}(U)$ and the covariance is with respect to the random covariate vector U. Here, C can be estimated by

$$\widehat{C} = \int_0^1 (1-q) \left[\frac{\sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} \frac{[I(\epsilon_{1j} > \epsilon_{2j'}) - I(\epsilon_{1j} < \epsilon_{2j'})] I\{\hat{H}(\hat{D}_{1j}) > q, \hat{H}(\hat{D}_{2j'}) > q\}}{\{\hat{G}_1(X_{1j} \wedge t_0)/w_{1j}\}\{\hat{G}_2(X_{2j'} \wedge t_0)/w_{2j'}\}}}{\sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} \frac{I\{\hat{H}(\hat{D}_{1j}) > q, \hat{H}(\hat{D}_{2j'}) > q\}}{\{\hat{G}_1(X_{1j} \wedge t_0)/w_{1j}\}\{\hat{G}_2(X_{2j'} \wedge t_0)/w_{2j'}\}}} - \hat{D} \right] \text{d}q, \tag{10}$$

where $\hat{H}(\cdot)$ is the empirical cumulative distribution function of $\hat{D}(U)$ and $\hat{D}_{ij} = \hat{D}(U_{ij})$. The justification of the consistency of (10) can be derived using similar arguments to those given by Zhao et al. (2012). Now, since the variances of $D(U)$ and $H(\hat{D}(U))$ are independent of the fitted model, the correlation ρ corresponding to C can be estimated up to a common constant across all candidate models. Therefore, to quantify the improvement of, say, Model I relative to Model II, we may take the ratio of the resulting covariance estimates $\widehat{C}_1/\widehat{C}_2$ to estimate the ratio of the two corresponding correlation coefficients $\hat{\rho}_1/\hat{\rho}_2$, to guide model selection.

We use a repeated random cross-validation procedure, in each iteration randomly dividing this part A data set into two mutually exclusive subsets, $\mathcal{B}$ and $\mathcal{E}$, the "model building set" and "evaluation set", respectively. For each model building set and for a given link function and variable selection procedure, we can construct a model, using only data in $\mathcal{B}$ to obtain $\hat{D}(\cdot)$ via (6), then compute all $\hat{D}(U_{ij})$, for all U_{ij} in $\mathcal{E}$. We repeatedly split the training data set M times. For each m, and for each modeling procedure, we obtain an estimate of the concordance $\widehat{C}^{(m)}$. Lastly, we average these estimates over $m = 1, ..., M$ to obtain final estimates $\widehat{C}$. The modeling procedure which yields the largest cross-validated C values will be used for the construction of our final working model. We then refit the entire part A data set with this specific modeling procedure in order to construct the final score.

3.3 Construction and Selection of Scoring Systems Using the BEST Data Set

We first split the data set into parts A and B, using the first 900 (33%) patients entering the study as part A and using the remaining patients as part B. In this sense, we mimic the traditional prediction process, using current data to predict the future outcomes of patients. Note that Shao (1993) presents theoretical justifications for the preference of a relatively large holdout sample, and a comparatively smaller sample size devoted to "model construction". Within part A data, 123, 60, 86, 9, and 84 patients in the control group were classified into categories 1 through 5, respectively, after 18 months of followup. The corresponding counts for the treatment group were 148, 74, 52, 13, and 68 patients, respectively. The numbers of

censored patients in part A were 94 and 89 in the control and treatment groups, respectively.

Here the covariate vector U consists of 16 clinically relevant covariates from Table 1 of Castagno et al. (2010). 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 etiology, presence of atrial fibrillation at baseline, and race (white vs. non-white). 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.

Models (5) and (7) were utilized with the logit and complementary log-log links, $g(p) = log(\frac{p}{1-p})$, and $g(p) = log(-log(1-p))$, respectively. For each of type of model, first we let Z be the vector of the above 16 covariates. We then consider a stepwise regression procedure with the weighted likelihood function as the objective function and an Akaike information criterion (AIC) as the criterion for covariate inclusion/exclusion. For each type of model (5) or (7), we started with all covariates (and first order interactions with the treatment indicator for (7)), and successively added/eliminated terms until no more covariates could be added/removed without subsequently increasing the AIC. For illustration, a total of eight modeling procedures were considered in our analysis. To evaluate these models, we used a repeated random cross validation procedure with 80% of the part A data used for model building and 20% for evaluation for each iteration of the procedure with $M = 25$ iterations.

In Table 3, we present these modeling procedures along with their relative concordance value, based on $\hat{C}$ with the modeling approach of separate logistic regression models with no variable selection as the reference model.

The model building procedure found to provide the most overall discriminatory ability was the one which models each treatment group separately, using the complementary log-log link, and performs AIC-based variable selection. This procedure is marked with * in Table

3. We then used this model building procedure to fit the entire part A data. The resulting model is given in Table 4. We note that six variables were eliminated from both models, six variables were retained in one model only, and four variables were retained in both models.

Table 3: Model building procedures with average cross-validated concordance values

Separate/Single Models	Link	Var. Selection	$\widehat{C}$ Ratio
Separate	logit	None	*(ref)*
Separate	logit	AIC	2.01
Separate	c-log-log	None	1.62
Separate	c-log-log	AIC	2.17*
Single	logit	None	0.99
Single	logit	AIC	1.91
Single	c-log-log	None	1.61
Single	c-log-log	AIC	0.90

Table 4: Regression coefficients for the final working models using BEST training data with $log(-log)$ link function

	Control Group	Treated Group
Covariate	β_1	β_2
LVEF	-0.018	-0.034
I(eGFR>75)	-0.237	-0.489
I(eGFR>45)	-0.673	-0.753
SBP	-0.012	-
Class IV Heart Failure	-	0.843
I(BMI>30)	0.218	0.212
Heart Rate	-	-0.008
History of hypertension	0.213	-
History of diabetes	0.359	-
Atrial Fibrillation	0.263	-

4. MAKING INFERENCES ABOUT THE TREATMENT DIFFERENCES OVER A RANGE OF SCORES USING THE HOLDOUT SAMPLE

Let $\hat{d}(u)$ be the observed score, obtained from the part A data set, for a patient in the part B data set with covariates u. In this section, using the data from Part B, we make inferences about the general risk difference $E(s) = \text{pr}(\epsilon_1 > \epsilon_2 | \hat{d}(u) = s) - \text{pr}(\epsilon_1 < \epsilon_2 | \hat{d}(u) = s)$ and the cumulative risk differences $\Gamma_k(s) = \text{pr}(\epsilon_2 \leq k | \hat{d}(u) = s) - \text{pr}(\epsilon_1 \leq k | \hat{d}(u) = s), k = 1, \ldots, K$,

where ϵ_i is outcome of a random patient in treatment group i from a future population identical to the part B data. Rather than using a parametric estimate for these contrast measures, we use a nonparametric kernel functional estimation procedure conditional on the treatment selection score. To this end, let the conditional cell probabilities for the ordinal response ϵ_{ij} be denoted by $\pi_{ik}(s)$ and cumulative probabilities by $\gamma_{ik}(s), i = 1, 2; j = 1, \ldots, n_i^*; k = 1, \ldots, K$. Here n_i^* is the sample size in the ith group in the part B data set. Let $\hat{\pi}_{ik}(s)$ and $\hat{\gamma}_{ik}(s)$ be their corresponding nonparametric kernel estimators. Let $Y_{ijk} = I(\epsilon_{ij} = k), k = 1, \ldots, K$. The kernel estimators for $\pi_{ik}(s)$ and $\gamma_{ik}(s)$ are

$$\hat{\pi}_{ik}(s) = \left\{ \sum_j^{n_i^*} \frac{w_{ij} Y_{ijk}}{\hat{G}_i(X_{ij} \wedge t_0)} K_{h_i}(V_{ij} - s) \right\} / \left\{ \sum_j^{n_i^*} \frac{w_{ij}}{\hat{G}_i(X_{ij} \wedge t_0)} K_{h_i}(V_{ij} - s) \right\}, \tag{11}$$

and $\hat{\gamma}_{ik}(s) = \sum_{l=1}^{k} \hat{\pi}_{il}(s)$, $i = 1, 2$; $k = 1, \ldots, K$, where $V_{ij} = \hat{d}(U_{ij})$, $w_{ij} = I(X_{ij} \leq t_0)\Delta_{ij} + I(X_{ij} > t_0)$, $\hat{G}_i(\cdot)$ is the Kaplan-Meier estimator of $G_i(\cdot)$ from the part B data, $K_{h_i}(s) = K(s/h_i)/h_i$, $K(\cdot)$ is a smooth symmetric kernel with finite support and h_i is a smoothing parameter. The resulting estimator for $E(s)$ is $\hat{E}(s) = \sum_{k=1}^{K} \hat{\pi}_{1,k}(s)\hat{\gamma}_{2,k-1}(s) - \hat{\pi}_{2,k}(s)\hat{\gamma}_{1,k-1}(s)$. When $h_i = O(n_i^{*-v}), 1/5 < v < 1/2$, it follows from a similar argument by Li et al. (2011) that $\hat{\pi}_{ik}(s)$ converges to $\pi_{ik}(s)$ uniformly over the interval $s \in \mathcal{S}$, where $\mathcal{S}$ is an interval contained properly in the support of $\hat{d}(U)$. Consequently, when h_i is of the same order as above, for a fixed s, the distribution $(n_1^* h_1 + n_2^* h_2)^{1/2}\{\hat{\Gamma}_k(s) - \Gamma_k(s)\}, k = 1, \ldots, K$ converges in distribution to a normal with mean 0 and covariance $\sigma_k(s)$ as $n_i^* \to \infty, i = 1, 2$. Similarly, the distribution $(n_1^* h_1 + n_2^* h_2)^{1/2}\{\hat{E}(s) - E(s)\}$ converges in distribution to a normal with mean 0 and variance $\sigma(s)$ as $n_i^* \to \infty, i = 1, 2$. To approximate the distributions above, we 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). In addition, $(1 - \alpha)$ simultaneous confidence bands for $E(s)$ and $\Gamma_k(s)$ over the pre-specified interval $\mathcal{S}$ can be obtained accordingly. Details are provided in the Appendix.

As with any nonparametric estimation problem, it is important that we choose appropriate smoothing parameters in order to make inference about the treatment differences. Here, we use a L-fold cross-validation procedure to choose the smoothing parameter $\hat{h}_i$ which maximizes a weighted cross-validated multinomial log-likelihood, as in Li et al. (2011). Specifically, we may randomly divide the entire data set into L mutually exclusive, approximately equally sized subsets. For any fixed values of h_i and (i,k), we can estimate $\pi_{ik}(s)$ using all observations except for those contained in the same subset as the j^{th} subject, which yields the estimator $\hat{\pi}_{i(-j)k}(s)$. The cross-validated log-likelihood, adjusted for censoring, is

$$\sum_{V_{ij}\in\mathcal{S}} \frac{w_{ij}}{\hat{G}_i(X_{ij}\wedge t_0)}\{\sum_{k=1}^{K} Y_{ijk} log(\hat{\pi}_{i(-j)k}(V_{ij}))\}. \tag{12}$$

Let $\hat{h}_i$ be a maximizer of (12). As in Li et al. (2011), $\hat{h}_i$ is of the order $n^{*-1/5}$. To ensure the bias of the estimator is asymptotically negligible and that the above large-sample approximation is valid, however, we slightly undersmooth the data and let the final smoothing parameter be $\tilde{h}_i = \hat{h}_i \times n_i^{*-\xi}$ where ξ is a small positive number less than 0.3.

4.1 Making Inference About Treatment Differences using the BEST Data Set

Next, we apply the final scoring system derived from the part A data set to the patients in the part B data set mentioned in Section 3.3. In Figure 2 below, we show the empirical cumulative distribution function of the scores in the part B data set. The vertical line indicates $\hat{d}(u) = 0$, and we note that 75% of the scores fall to the right of this line, indicating an anticipated treatment benefit for a majority of patients. For all kernel estimators, we let $K(\cdot)$ be the standard Epanechnikov kernel, with the smoothing parameters chosen as the maximizers of (12), then multiplied by $n_i^{*-0.05}$.

The resulting estimates of the patient-specific treatment differences $\hat{E}(s)$, with 0.95 pointwise and simultaneous confidence interval estimates, are displayed in Figure 3. Using the final score derived from the model in Table 4 over the range $s \in (-0.26, 0.40)$, we find $\hat{E}(s) > 0$ for $s > -0.11$ and $\hat{E}(s) < 0$ for $s < -0.11$. The point and interval estimates

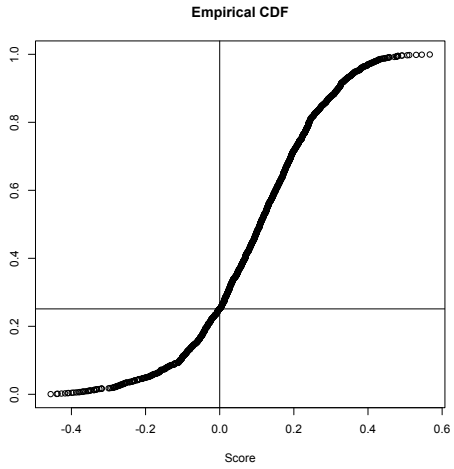


Figure 2: Distribution of treatment selection scores $\hat{d}(u)$ for BEST patients in the holdout sample.

displayed in Figure 3 are quite informative for identifying subgroups of patients who would benefit from the beta-blocker with various desired levels of treatment differences.

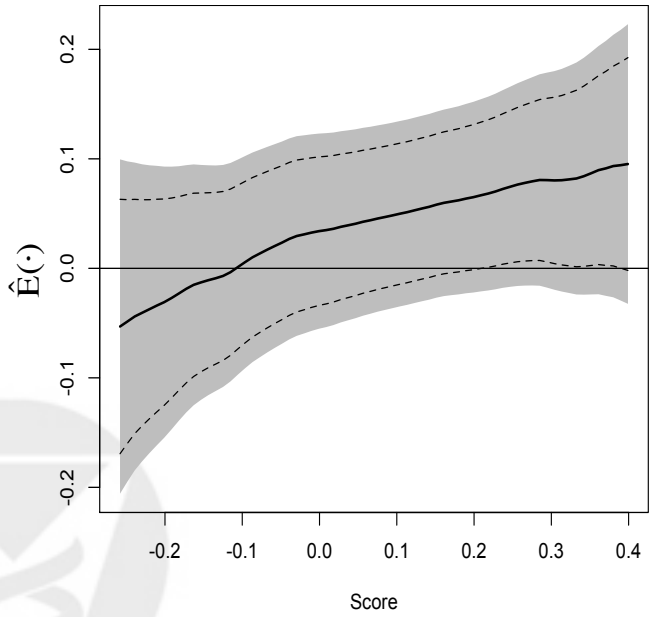


Figure 3: Estimated BEST treatment effect $\hat{E}(s)$ using treatment selection score presented in Table 4. Solid curve represents point estimates, with 0.95 pointwise and simultaneous confidence intervals denoted by dashed lines and shaded region, respectively.

In Figure 4, we show the corresponding treatment differences with respect to the cumulative outcome probabilities $\gamma_{ik}(\cdot)$. Note that each value $\Gamma_k(s)$ allows for the estimation of

the treatment contrast with respect to a different composite outcome. For example, $\Gamma_1(s)$ refers to the effect of treatment on the composite outcome "any hospitalization or death". It can be seen that $\hat{\Gamma}_1(s) > 0$ for $s > 0.02$ and $\hat{\Gamma}_1(s) < 0$ for $s < 0.02$, indicating that our score is also informative for identifying patients would experience "treatment success" with respect to this outcome as well. Furthermore, using $\hat{\Gamma}_2$, patients with scores > 0.05 and > 0.24 are found to experience significant treatment benefits (via the 95% confidence intervals and bands, respectively) with respect to the desirable outcome ($\epsilon \leq 2$) (alive with no HF hospitalization). Finally, we note that the estimated effect of treatment with respect to death, $\hat{\Gamma}_3(s)$, is relatively constant with a (non-significant) risk reduction of approximately 2% across the scores.

5. CONCLUSIONS

The proposed procedures can be applied to any study with multiple endpoints which reflect a patient's risk-benefit profile. For example, a longitudinal trial may collect repeated measurements for an endpoint over time. The standard analysis, for example, via GEE techniques (Liang and Zeger, 1986) provides a treatment comparison using an average mean difference of a response variable during the study follow-up. Such a contrast may not provide a clinically interpretable summary, particularly when the temporal profile of such repeated measures should be considered for the outcome. One may instead classify the repeated measure profile for each individual patient into several clinically meaningful categories, such as those presented in this paper for evaluating the treatments risk(s) and benefit(s) together.

For ordinal categorical outcomes, one may give each stratum a weight and create a single univariate outcome as the final endpoint. However, the estimates for the treatment difference for each category are quite informative for treatment selections. To avoid post-hoc subgroup analysis, we highly recommend pre-specifying a systematic procedure for identifying patients who would benefit from the new treatment in the study protocol or its statistical analysis plan.

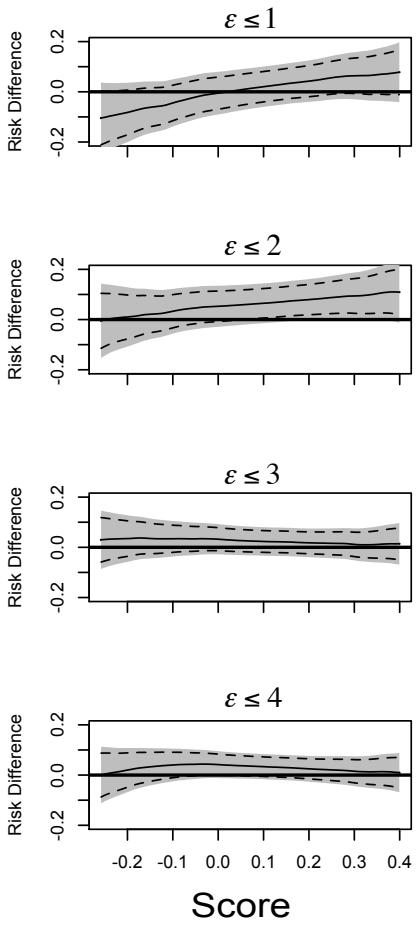


Figure 4: BEST target treatment differences (treated minus untreated) using treatment selection score presented in Table 4. Solid curve represents point estimates, with 0.95 pointwise and simultaneous confidence intervals denoted by dashed lines and shaded region, respectively.

When there are more than two treatments available for selection, one may create a scoring system such as a "risk score", for example, based on the data from the control or standard care arm. Then we may use the holdout sample to estimate the treatment effectiveness nonparametrically over the selected score. For comparing two treatment groups only, we recommend using the treatment difference score rather than the risk score from the control only.

If the disease progression is not reversible, i.e., a patient's classification cannot improve over time, one may utilize more information from a censored observation rather than using

the inverse probability of censoring weighting scheme. For example, to estimate $\{\pi_{ik}, i = 1, 2; k = 1, \ldots, K\}$, if the ordinal response of the patient is l at the censored time, then the contribution to the weighted likelihood from this patient is $\sum_{k \geq l} \pi_{ik}$.

For comparing scoring systems constructed for the treatment difference, we use a concordance measure between the observed and expected treatment differences. More research is needed to explore if other measures, which may be more intuitively interpretable, can be used for model evaluation and selection. Moreover, it is important to consider a parsimonious model as the final candidate even if it is not the optimal one based on the selection criteria. The application of a parsimonious scoring system can have more clinical utility than an optimal, but complex, system. For ordinal categorical outcomes, we use the general risk difference D, the net treatment improvement rate, to estimate the treatment contrast. It would be interesting to consider other measures for quantifying the contrast, reflecting the size of the treatment difference.

APPENDIX

For all inference using large sample approximations, we employ perturbation-resampling procedures using 1000 realizations from the standard exponential distribution. Details are provided below.

Construction of Confidence Intervals for Two-Sample Inference

Let $\{B_{ij} : i = 1, 2; j = 1, \ldots, n_i\}$ be independent random samples from a strictly positive distribution with mean and variance equal to one. Let π^*_{ik} be the perturbed version of $\hat{\pi}_{ik}$ with

$$\pi^*_{ik} = \left\{ \sum_j \frac{B_{ij} w_{ij} Y_{ijk}}{\hat{G}^*_j(X_{ij} \wedge t_0)} \right\} / \left\{ \sum_j \frac{B_{ij} w_{ij}}{\hat{G}^*_i(X_{ij} \wedge t_0)} \right\}, \tag{A.1}$$

and $\gamma^*_{ik} = \sum_{l=1}^{k} \pi^*_{il}$. Here, $\hat{G}^*_i(\cdot)$ is the perturbed estimator for the survival function $G_i(\cdot)$

$$\hat{G}^*_i(t) = \exp\left[- \sum_{j=1}^{n_i} \int_0^t \frac{B_{ij} d\{I(C_{ij} \leq u \wedge X_{ij})\}}{\sum_{l=1}^{n_i} B_{il} I(X_{il} \geq u)} \right]. \tag{A.2}$$

Let β^* be the maximizer of the perturbed version of the weighted log-likelihood function in (3):

$$\sum_{ij} \frac{B_{ij} w_{ij}}{\hat{G}_i^*(X_{ij} \wedge t_0)} [\sum_{k=1}^{K} Y_{ijk} log\{g^{-1}(\alpha_k - \beta\tau_{ij}) - g^{-1}(\alpha_{k-1} - \beta\tau_{ij})\}]. \tag{A.3}$$

The limiting distribution, conditional on the data, of

$$(n_1 + n_2)^{1/2}\{\beta^* - \hat{\beta}\}, \tag{A.4}$$

is normal with mean 0 and variance $\hat{\sigma}_b^2$, which is a consistent estimator of σ_b^2, the variance associated with the distribution $(n_1 + n_2)^{1/2}\{\hat{\beta} - \beta\}$. Thus, the empirical variance of the perturbed estimates β^* can be used to estimate the standard error associated with $\hat{\beta}$ (Zheng et al., 2006; Uno et al., 2007; Li et al., 2011).

Denote $\mathbf{\Gamma}^* = \boldsymbol{\gamma}_2^* - \boldsymbol{\gamma}_1^*$, where $\boldsymbol{\gamma}_i^* = \{\gamma_{i1}^*, \cdots, \gamma_{iK}^*\}'$. Using the arguments by Cai et al. (2010), the limiting distribution, conditional on the target data set, of

$$(n_1^* + n_2^*)^{1/2}\{\mathbf{\Gamma}^* - \hat{\mathbf{\Gamma}}\}, \tag{A.5}$$

is multivariate normal with mean zero and covariance matrix $\hat{\mathbf{\Sigma}}$ which is a consistent estimator of $\mathbf{\Sigma}$, the covariance matrix associated with the distribution $(n_1^* + n_2^*)^{1/2}\{\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\}$. The resulting sample covariance matrix based on those perturbed estimates $\mathbf{\Gamma}^*$, say, $\tilde{\mathbf{\Sigma}}$, is a consistent estimator of $\mathbf{\Sigma}$. A two-sided confidence interval for the two-sample risk difference Γ_k is then given by

$$\hat{\Gamma}_k \pm z_{(1-\alpha/2)}(n_1^* h_1 + n_2^* h_2)^{-1/2}\tilde{\sigma}_k, \tag{A.6}$$

where $\tilde{\sigma}_k^2$ is the kth diagonal element of $\tilde{\mathbf{\Sigma}}$. Furthermore, one may use a similar approach for making inference on $\hat{D}$ by perturbed $D^* = \sum_{k=2}^{K} \pi_{1,k}^* \gamma_{2,k-1}^* - \pi_{2,k}^* \gamma_{1,k-1}^*$.

Construction of Confidence Intervals and Bands for Stratified Inference

For personalized medicine, we let $\pi^*_{ik}(s)$ be the perturbed version of $\hat{\pi}_{ik}(s)$ with $\pi^*_{ik}(s)$

$$= \left\{ \sum_j \frac{B_{ij} w_{ij}}{\hat{G}^*_j(X_{ij} \wedge t_0)} K_{h_i}(V_{ij} - s) Y_{ijk} \right\} / \left\{ \sum_j \frac{B_{ij} w_{ij}}{\hat{G}^*_i(X_{ij} \wedge t_0)} K_{h_i}(V_{ij} - s) \right\}. \tag{A.7}$$

and $\gamma^*_{ik}(s) = \sum_{l=1}^k \pi^*_{il}(s)$. Using identical arguments to those above, we denote $\mathbf{\Gamma}^*(s) = \boldsymbol{\gamma}^*_2(s) - \boldsymbol{\gamma}^*_1(s)$, where $\boldsymbol{\gamma}^*_i(s) = \{\gamma^*_{i1}(s), \cdots, \gamma^*_{iK}(s)\}'$, and can show that the distribution for

$$(n^*_1 h_1 + n^*_2 h_2)^{1/2} \{\mathbf{\Gamma}^*(s) - \hat{\mathbf{\Gamma}}(s)\}, \tag{A.8}$$

conditional on the observed data is multivariate normal and asymptotically equivalent to that of $(n^*_1 h_1 + n^*_2 h_2)^{1/2}(\hat{\mathbf{\Gamma}}(s) - \mathbf{\Gamma}(s))$. Therefore, the point-wise confidence interval for $\Gamma(s)$ can be constructed using generated $\Gamma^*(s)$ as in (A.4).

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

$$\sup_{s \in \mathcal{S}} \tilde{\sigma}_k^{-1}(s) |(n^*_1 h_1 + n^*_2 h_2)^{1/2} \{\hat{\Gamma}_k(s) - \Gamma_k(s)\}| \tag{A.9}$$

due to the fact that as a process in s, $(n^*_1 h_1 + n^*_2 h_2)^{1/2}\{\hat{\Gamma}_k(s) - \Gamma_k(s)\}$ does not converge weakly to a tight 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 $\hat{\Gamma}_k(s) - \Gamma_k(s)$ converges to a proper random variable. In practice, to construct a confidence band, we can first find a critical value b_α such that

$$\text{pr}(\sup_{s \in \mathcal{S}} |\Gamma^*_k(s) - \hat{\Gamma}_k(s)| / \{(n^*_1 h_1 + n^*_2 h_2)^{-1/2} \tilde{\sigma}_k(s)\} > b_\alpha) \approx \alpha. \tag{A.10}$$

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

$$\hat{\Gamma}_k(s) \pm b_\alpha (n_1^* h_1 + n_2^* h_2)^{-1/2} \tilde{\sigma}_k(s). \tag{A.11}$$

Similar arguments are used for the construction of the confidence band for $E(s) : s \in \mathcal{S}$.

ACKNOWLEDGEMENTS

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

REFERENCES

Agresti, A. (1990). *Categorical data analysis*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics* **10,** 1100–1120.

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,** 1659–1667.

Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics* pages 1071–1095.

Bonetti, M. and Gelber, R. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* **5,** 465–481.

Bonetti, M., Gelber, R., et al. (2000). A graphical method to assess treatment-covariate interactions using the cox model on subsets of the data. *Statistics in medicine* **19,** 2595–2609.

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

Cai, T., Tian, L., Wong, P., and Wei, L. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* **12,** 270–282.

Castagno, D., Jhund, P., McMurray, J., Lewsey, J., Erdmann, E., Zannad, F., Remme, W., Lopez-Sendon, J., Lechat, P., Follath, F., et al. (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,** 607–616.

Chuang-Stein, C., Mohberg, N., and Sinkula, M. (1991). Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. *Statistics in medicine* **10,** 1349–1359.

Edwardes, M. and Baltzan, M. (2000). The generalization of the odds ratio, risk ratio and risk difference to r x k tables. *Statistics in medicine* **19,** 1901–1914.

Edwardes, M. D. d. (1995). A confidence interval for pr(x < y) - pr(x > y) estimated from simple cluster samples. *Biometrics* **51,** pp. 571–578.

Ghosh, D. and Lin, D. (2003). Semiparametric analysis of recurrent events data in the presence of dependent censoring. *Biometrics* **59,** 877–885.

Huang, Y., Sullivan Pepe, M., and Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. *Biometrics* **63,** 1181–1188.

Janes, H., Pepe, M., Bossuyt, P., and Barlow, W. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of internal medicine* **154,** 253.

Kent, D. and Hayward, R. (2007). Limitations of applying summary results of clinical trials to individual patients. *JAMA: the journal of the American Medical Association* **298,** 1209–1212.

Li, Q. and Lagakos, S. (1998). Use of the wei–lin–weissfeld method for the analysis of a recurring and a terminating event. *Statistics in Medicine* **16,** 925–940.

Li, Y., Tian, L., and Wei, L. (2011). Estimating subject-specific dependent competing risk profile with censored event time observations. *Biometrics* **67,** 427–435.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Lin, D., Wei, L., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62,** 711–730.

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

Lui, K.-J. (2002). Notes on estimation of the general odds ratio and the general risk difference for paired-sample data. *Biometrical Journal* **44,** 957–968.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics* **21,** 255–285.

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

Pepe, M. (2004). *The statistical evaluation of medical tests for classification and prediction.* Oxford University Press, USA.

Pepe, M., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I., and Zheng, Y. (2008). Integrating the predictiveness of a marker with its performance as a classifier. *American journal of epidemiology* **167,** 362–368.

Pocock, S., Ariti, C., Collier, T., and Wang, D. (2012). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European heart journal* **33,** 176–182.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88,** 486–494.

Simonoff, J. S., Hochberg, Y., and Reiser, B. (1986). Alternative estimation procedures for pr(x < y) in categorized data. *Biometrics* **42,** pp. 895–907.

Song, X. and Pepe, M. (2004). Evaluating markers for selecting a patient's treatment. *Biometrics* **60,** 874–883.

Uno, H., Cai, T., Tian, L., 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 medicinereporting of subgroup analyses in clinical trials. *New England Journal of Medicine* **357,** 2189–2194.

Wei, L., Lin, D., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84,** 1065–1073.

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

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L. (2012). Effectively selecting a target population for a future comparative study. *JASA* page under review.

Zheng, Y., Cai, T., and Feng, Z. (2006). Application of the time-dependent roc curves for prognostic accuracy with multiple biomarkers. *Biometrics* **62,** 279–287.