Harvard University
Harvard University Biostatistics Working Paper Series

Year 2014 *Paper* 169

A Predictive Enrichment Procedure to Identify Potential Responders to a New Therapy for Randomized, Comparative, Controlled Clinical Studies

Junlong Li, *Harvard University*
Lihui Zhao, *Northwestern University*
Lu Tian, *Stanford University*
Tianxi Cai, *Harvard University*
Brian Claggett, *Brigham & Women's Hospital*
Andrea Callegaro, *GlaxoSmithKline Vaccines*
Benjamin Dizier, *GlaxoSmihtKline Vaccines*
Bart Spiessens, *GlaxoSmithKline Vaccines*
Fernando Ulloa-Montoya, *GlaxoSmithKline Vaccines*
L. J. Wei, *Harvard University*

A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative controlled clinical studies

Junlong Li[1], Lihui Zhao[2], Lu Tian[3], Tianxi Cai[1], Brian Claggett[4], Andrea Callegaro[5], Benjamin Dizier[5], Bart Spiessens[5], Fernando Ulloa-Montoya[5] and Lee-Jen Wei[1]

Abstract

To evaluate a new therapy versus a control via a randomized, comparative clinical study or a series of trials, due to heterogeneity of the study patient population, a pre-specified, *predictive* enrichment procedure may be implemented to identify an "enrichable" subpopulation. For patients in this subpopulation, the therapy is expected to have a desirable overall risk-benefit profile. To develop and validate such a "therapy-diagnostic co-development" strategy, a three-step procedure may be conducted with three independent data sets from a series of similar studies or a single trial. At the first stage, we create various candidate scoring systems based on the baseline information of the patients via, for example, parametric models using the first data set. Each individual score reflects an anticipated average treatment difference for future patients who share similar baseline profiles. A large score indicates that these patients tend to benefit from the new therapy. At the second step, a potentially promising, enrichable subgroup is identified using the totality of evidence from these scoring systems. At the final stage, we validate such a selection via two-sample inference procedures for assessing the treatment effectiveness statistically and clinically with the third data set, the so-called holdout sample. When the study size is not large, one may combine the first two steps using a "cross-training-evaluation" process. The entire enrichment procedure is illustrated with the data from a cardiovascular trial to evaluate a beta-blocker versus the placebo for treating chronic heart failure patients.

Keywords: Cox model; Cross-validation; Stratified medicine; Survival analysis; Therapy-diagnostic co-development

[1]Department of Biostatistics, Harvard University, Boston, MA, 02115, USA
[2]Department of Preventive Medicine, Northwestern University, Chicago, IL 60611, USA
[3]Department of Health Research and Policy, Stanford University, Stanford, CA, 94305, USA
[4]Division of Cardiovascular Medicine, Brigham and Women's Hospital, Boston, MA, 02115, USA
[5]GlaxoSmithKline Vaccines, 89 Rue de l'Institut, 1330 Rixensart, Belgium

1 Introduction

In a typical randomized clinical trial, generally the assessment of a new therapy versus a control with respect to the risk-benefit profile is made for the entire study patient population. Due to heterogeneity of the study population, the conclusion of a positive (or negative) study on an average sense does not guarantee the new therapy benefits (or does not benefit) uniformly for all patients in the study population (Rothwell, 1995; Rothwell *et al.*, 2005; Kent and Hayward, 2007). Recently various enrichment strategies were suggested and implemented in comparative trials (Freidlin and Simon, 2005; Jiang *et al.*, 2007; Wang *et al.*, 2007; Simon, 2008; Karuri and Simon, 2012; U.S. Food and Drug Administration, 2012). For predictive enrichment, the idea is to apply a systematic, pre-specified procedure to identify and validate a subpopulation whose patients would significantly benefit from the new therapy clinically and statistically. This procedure may be utilized in a single study or a series of trials conducted under similar settings.

A properly executed predictive enrichment procedure generally consists of three stages: at the first stage using data set $\mathcal{A}$, we utilize all the relevant baseline information to fit the data with various prediction models to create several competing scoring systems for stratifying future patients. The individual score is an estimated average treatment difference for future patients with similar baseline profiles. A large score indicates that the patient has a high chance to benefit from the new therapy. At the second stage with data set $\mathcal{B}$, we evaluate and compare these scoring systems and then develop a rule for identifying an enrichable subpopulation. At the last step with data set $\mathcal{C}$, the holdout sample, we validate such a selection rule via proper two-sample inference procedures for comparing the two treatments for the enrichable subset of patients in $\mathcal{C}$. Ideally each step would be conducted via an independent data set from the same underlying study population and interventions. With the data from a single comparative study, one may split the entire data set into two independent parts and use the first part as $\mathcal{A}$ & $\mathcal{B}$ for scoring system building and evaluation, as well as the identification of an enrichable subpopulation via a conventional "cross-training-evaluation" scheme. We then use the second part of the data as $\mathcal{C}$ to examine if there is a strong evidence that patients in the enrichable subpopulation would respond favorably to the new therapy compared with the control.

As an illustrative example, consider the data from a clinical trial "Beta-Blocker Evaluation of Survival Trial (BEST)" to investigate if bucindolol, a beta-blocker, would benefit patients with advanced chronic heart failure (Eichhorn *et al.*, 2001). There were 2708 patients enrolled and

followed for an average of two years. One of the primary goals of the study was to examine if the beta-blocker could reduce the overall hospitalization and mortality rates of the patients. With the time to either death or the first hospitalization as the endpoint, the Kaplan-Meier curves are given in Figure 1. The p-value based on the standard two-sample log-rank test is 0.14, with a hazard ratio estimate of 0.93 and its corresponding 0.95 confidence interval of $(0.85, 1.02)$. For this outcome variable, the evidence that the beta-blocker was better than the control for the entire study population is not strong.

[Figure 1 about here.]

Now, suppose that we had a pre-specified, aforementioned enrichment strategy at the beginning of the study with respect to this endpoint. We first split the entire data set into two parts, using the data from the first 900 patients on the data listing for BEST provided by US National Institutes of Health as $\mathcal{A}$ & $\mathcal{B}$ for building and evaluating scoring systems to identify a potentially enrichable subpopulation. We assume that this set of observations is a random sample from the BEST database. Then, we use the data from the remaining 1807 patients as the holdout sample $\mathcal{C}$ to examine if the beta-blocker would benefit the patients of the selected subpopulation by reducing the rate of death or the first hospitalization compared with the standard control. For this study, there are 16 clinically relevant baseline covariates: 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 versus Class IV), obesity (Body mass index > 30 versus ≤ 30), resting heart rate, smoking status (ever versus never), history of hypertension, history of diabetes, ischemic heart failure etiology, presence of atrial fibrillation and race (white versus 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. In general, the baseline covariates may consist of clinical and genetic markers of the patients.

We will use this example to illustrate the proposed procedure step by step. Although the endpoint of the study is the (possibly censored) time to a specific event, the procedure can be trivially generalized to the case with a non-censored endpoint, for example, a binary, continuous or ordinal categorical observation. In Section 2, with data set $\mathcal{A}$ of BEST described above, we use a single set of regression models, which relate the event time to its 16 baseline covariates, to estimate an average treatment difference (e.g., the hazard ratio of placebo versus beta-blocker) for patients who share similar baseline covariate profiles. This creates a continuous scoring

system. If the model is a reasonably good approximation to the truth, a future patient with a large score is expected to benefit from the new therapy. An enrichable subpopulation would consist of patients whose scores are greater than an appropriately chosen threshold value. To avoid the problem of potential over-fitting, we discuss methods to choose this threshold value with an independent data set $\mathcal{B}$ in Section 3. Since there is only one study for evaluating the beta-blocker in our illustrative example, the size of the data set at each step may not be large enough for obtaining a reliable scoring system. In Section 4, we present a "cross-training-evaluation" process by implementing the above two steps iteratively with various competing models and regularized estimation procedures to create the scoring systems. The final recommendation for an enrichable subgroup is then obtained by considering the totality of evidence from all of the candidate scoring systems and the related threshold values. In Section 5, we validate our selection by applying two-sample inference procedures to the enrichable subgroup in the holdout sample. Note that the primary criteria utilized for identifying a threshold value in Section 4 should be closely related to those for the final validation with the holdout sample. For instance, if the final primary assessment of the enrichable subgroup is based on the standard two-sample interval estimation procedure for the hazard ratio with the holdout sample, the choice of the threshold value in Section 4 would be made via the corresponding *predicted* interval estimate for the hazard ratio. That is, we choose the threshold value such that the predicted interval estimate suggests that using the data from the potential enrichable subgroup in the holdout sample, the beta-blocker can be demonstrated to have a clinically meaningful advantage over the control via the confidence interval estimate.

There are numerous novel proposals in the literature for identifying future patients to be treated by the new therapy (Qian and Murphy, 2011; Zhang *et al.*, 2012a,b; Zhao *et al.*, 2013). Generally, each proposal is based on maximizing a specific utility function. For example, using data from the BEST study, one may recommend treating future patients whose hazard ratio scores are greater than one with beta-blocker, and leaving the rest untreated. However, such a selection procedure may not be applied directly to our problem. For instance, we may not be able to validate the superiority of the beta-blocker over the control using the data from the holdout sample (especially if the enrichable subgroup includes a sizable subset of patients for whom the benefit of beta-blocker is not clinically meaningful). Furthermore, the selection of a robust, enrichable subgroup may depend on the totality of evidence, which is based on more

than one utility function and multiple competing scoring systems as proposed in this article.

2 Creating a patient-specific scoring system for quantifying the between-group differences

In this section, we present the predictive enrichment scheme under the survival analysis setting. Let T be a time-to-event outcome, U be a $p \times 1$ vector of baseline covariates and G denote the group indicator with 1 for the new therapy group and 0 for control. Furthermore, it is assumed that T may be censored by a censoring variable C, which is independent of T and U. For each patient, we observe (X, Δ, G, U), where $X = min(T, C)$ and Δ denotes the censoring indicator with $\Delta = 1$ if X is observed and 0 otherwise. The observations, $\{(X_i, \Delta_i, G_i, U_i), i = 1, \cdots, n_1\}$, from data set $\mathcal{A}$ consist of n_1 independent and identical copies of (X, Δ, G, U). Conditional on U, let the population parameter for the group contrast, i.e. the treatment difference, be denoted by $D(U)$ and a large value of $D(U)$ indicates that the new therapy is better than the control for prolonging the event time. For example, $D(U)$ may be the difference of two event time medians or the model-based "constant" hazard ratio for the control versus the new therapy (Cox, 1972).

For $U = u$, we first consider $D(u)$ being the hazard ratio for patients with baseline vector u. To estimate $D(u)$ with the data set $\mathcal{A}$, one may utilize semi-parametric or parametric regression models. For example, we may fit the data from each treatment group with a Cox model to obtain $\widehat{D}(u)$ to approximate $D(u)$. Specifically, for patients in Group $G = k$ $(k = 0, 1)$, consider the Cox model:

$$P(T > t | U = u, G = k) = g\big(log\Lambda_k(t) + \beta_k' u\big)\,, \tag{2.1}$$

where $t > 0$, $g(x) = \exp\{-e^x\}$, $\Lambda_k(\cdot)$ is the group-specific underlying cumulative baseline hazard function and β_k denotes the vector of unknown coefficients.

To obtain an estimate of β_k from (2.1) when the dimension of U is high, we may utilize a regularized estimation procedure, for example, via lasso, adaptive lasso, elastic net or ridge regression (Verweij and Van Houwelingen, 1994; Tibshirani *et al.*, 1997; Fan and Li, 2002; Zhang and Lu, 2007; Friedman *et al.*, 2010; Simon *et al.*, 2011) with the partial likelihood function $L(\beta_k)$. Taking the elastic net procedure as an example, $\widehat{\beta}_k$ is a minimizer of

$$\log\{L(\beta_k)\} + \lambda_1 ||\beta_k||_1 + \lambda_2 ||\beta_k||_2\,,$$

where $||\beta_k||_1$ and $||\beta_k||_2$ denote the L_1- and L_2-norm of β_k, respectively. Here, λ_1 and λ_2 are the non-negative regularization parameters, which can be selected by a standard cross-validation procedure (Tibshirani, 1996). Note that the estimation procedure becomes the standard lasso method by setting $\lambda_2 = 0$. On the other hand, when $\lambda_1 = 0$, the resulting procedure is the standard ridge regression.

Now, assuming that the underlying hazard functions for two treatment groups are proportional to each other, one may summarize the treatment difference based on $\widehat{D}(u) = \exp\{\widehat{\beta}_0' u\}/\exp\{\widehat{\beta}_1' u\}$. It is important to note that (2.1) is simply a working model, which is an approximation to the truth. The proportional hazard assumption may not be valid. Therefore, $\widehat{D}(\cdot)$ may not be able to capture the patient-specific treatment differences. In fact, even if the two Cox models are correctly specified, $\widehat{D}(\cdot)$ can only estimate $D(\cdot)$ up to a positive constant. On the other hand, this scoring system may work reasonably well for ranking the future patients with respect to the benefit from the new therapy. As an example, with the data from the first 900 patients in the BEST study discussed in Section 1, we randomly choose a subset of 450 patients and fit the corresponding data with the above procedure under the ridge regression setting. The resulting $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given in the Table 1. For each given covariate vector u, we then obtain $\widehat{D}(u)$. With such a scoring system, we can rank patients in a given data set. There is a one-to-one correspondence between the scores and their ranks. Thus, we can use relative ranks as scores to identify an enrichabe subgroup.

[Table 1 about here.]

An alternative and model-free between-group contrast measures can be considered , for example, the difference or ratio of two median survival times. However, due to censoring, we may not be able to estimate median survival time well with the observed data. Instead of the median, the mean survival time is another good summary of the survival time distribution, but again, due to censoring it cannot be estimated well either. On the other hand, unlike the median, we can modify the mean with so-called restricted mean survival time (RMST) as a summary to accommodate the study follow-up time, say, τ_0 (Irwin, 1949; Karrison, 1987; Zucker, 1998; Murray and Tsiatis, 1999; Chen and Tsiatis, 2001; Anderson *et al.*, 2004; Tsiatis *et al.*, 2008; Zhang and Schaubel, 2011; Royston and Parmar, 2011; Tian *et al.*, 2012; Zhao *et al.*, 2012). The RMST is simply the population average of the event-free times for τ_0-year follow up. This quantity can be easily estimated by the area under the Kaplan-Meier curve up to τ_0

and the corresponding one- or two-sample inference procedures can be made accordingly. For example, for the entire data from BEST, if we let $\tau_0 = 3.5$ (years), the estimated RMSTs for beta-blocker and the control are 1.51 and 1.42 (years), respectively. The p-value for testing the equality of the underlying two RMSTs based on their empirical counterparts is 0.19 and a 0.95 confidence interval for the difference of the RMSTs is $(-0.11, 0.30)$. Note that unlike the hazard ratio, this between-group difference measure is model-free and the inferential results are readily interpretable clinically.

To construct a scoring system using the difference of RMSTs, one may use the regression models studied by Anderson *et al.* (2004) and Tian *et al.* (2014). Alternatively, we may use the Cox models above to estimate $D(u)$, the difference of two RMSTs given $U = u$. The resulting estimate

$$\widehat{D}(u) = \int_0^{\tau_0} \left\{ g_1(\log \widehat{\Lambda}_1(t) + \widehat{\beta}_1' u) - g_0(\log \widehat{\Lambda}_0(t) + \widehat{\beta}_0' u) \right\} dt \,,$$

where

$$\widehat{\Lambda}_k(t) = \sum_{i=1}^{n_1} \int_0^t \frac{dN_i^{(k)}(s)}{\sum_{j=1}^{n_1} Y_j^{(k)}(s) e^{\widehat{\beta}_k' Z_j}} \,,$$

which denotes the estimated cumulative hazard function for $k = 0, 1$ with

$$N_i^{(k)}(t) = I(X_i \le t, G_i = k)\Delta_i, \;\; i = 1, \ldots, n_1$$

and

$$Y_j^{(k)}(t) = I(X_j \ge t, G_j = k), \;\; j = 1, \ldots, n_1.$$

Note that instead of fitting the data from each treatment group with a Cox model, respectively, one may use a single model including the treatment indicator, main covariate effects, and treatment and covariate interactions with those regularized estimation procedures discussed above to fit the entire data set $\mathcal{A}$ to create the scoring system. In Section 4, we will provide more details about other models with various regularized estimation procedures for building the scoring systems. Moreover, for uncensored outcome variables, one may use the generalized linear working models with various regularized estimation procedures to construct patient-specific scoring systems.

3 Choosing an enrichable subpopulation with a scoring system

From data set $\mathcal{A}$, we obtain a scoring system based on $\widehat{D}(\cdot)$ and define its corresponding rank score $\widehat{Q}(\cdot) = 1 - \widehat{F}\{\widehat{D}(\cdot)\}$, where $\widehat{F}(\cdot)$ is the empirical distribution function of $\widehat{D}(U)$. If the working models utilized for establishing the individual scores are reasonably good approximations to the truth, we expect that a patient whose rank score is large would more likely to benefit from the new therapy. Therefore, it seems natural to consider an enrichable subgroup of patients in data set $\mathcal{B}$, whose $\widehat{Q}(\cdot) \geq 1-q_0$, where q_0 is an appropriately selected threshold value representing the fraction of the enrichable subpopulation with respect to the entire study population. That is, given $0 < q_0 < 1$, the selected enrichable subgroup based on q_0 consist of patients whose rank scores exceed the $100(1 - q_0)$th percentile.

Now, the question is how to choose this threshold value q_0 at this stage. To answer this question, we need to know how we will validate the potentially promising subgroup using a threshold value q_0, with the data from the holdout sample $\mathcal{C}$. As an example, let us consider the hazard ratio estimate based $\widehat{Q}(\cdot)$ as the score created by two Cox models with the ridge regression method discussed in Section 2. We apply the above binary rule to identify a set $\mathcal{E}$ of patients in the holdout sample-whose ranks of the hazard ratios are greater than $n_3 q_0$, where n_3 is the sample size of the holdout sample. With the observations from $\mathcal{E}$, we test the null hypothesis that there is no difference between the new therapy and control via the logarithm of the two-sample hazard ratio estimator $\widetilde{HR}_{q_0}$. If the resulting standardized test statistic Z is significantly large, then one claims that this selected subgroup is successfully validated.

Now for each given q, let HR_q be the limit of $\widetilde{HR}_q$ and σ_q^2 be the asymptotic variance of $\sqrt{n_3 q} log(\widetilde{HR}_q)$. Then the effective size of the above Z-test is $\sqrt{q} log(HR_q)/\sigma_q$. With the data from data set $\mathcal{B}$, the Z-test statistic for the holdout sample may be predicted via

$$\sqrt{n_3 q}\, \log(\widehat{HR}_q)/\hat{\sigma}_q\,, \tag{3.1}$$

where $\widehat{HR}_q$ is the estimate for the hazard ratio from observations whose $\widehat{Q}(\cdot) \geq 1 - q$ in data set $\mathcal{B}$ and $\hat{\sigma}_q$ is the corresponding estimate for σ_q.

Utilizing the data from the first 900 patients in the BEST study, we considered using the hazard ratio scores created from Model (2.1) with ridge regression procedure and the data from 450 patients excluded from those utilized for fitting the models discussed in the previous section to obtain the predicted Z-test scores. Figure 2 gives the curve of such predicted Z-test scores

over q, where $0 < q < 1$. Although the right tail of this curve may not be stable due to the small sample size, this predicted Z-test score curve is quite informative for choosing an appropriate threshold value q_0. For example, the Z-test score achieves its maximal value of 2.98 when $q = 0.6$. This means that if we apply $q_0 = 0.6$ as the threshold value to the holdout sample, this would be a good approximation to the observed two-sample standardized test statistic resulted in the final validation stage. If one would like to choose a larger enrichable subgroup, we may select a larger threshold value q_0 but with a compromise of a smaller Z-test score.

[Figure 2 about here.]

Alternatively, we may utilize a 0.95 confidence interval estimate for the two-sample hazard ratio as the inferential tool for decision making with the holdout sample. As in the aforementioned hypothesis testing case, one can estimate the expected bounds of the confidence interval based on the data in $\mathcal{B}$. For example, the expected value for the lower bound of a one-sided 0.95 confidence interval can be predicted as

$$\mathcal{L}_q = \exp\{\log(\widehat{HR}_q) - 1.645\hat{\sigma}_q/\sqrt{n_3 q}\}. \tag{3.2}$$

Figure 3 gives the corresponding curve based on (3.2), which provides information about the size of the treatment effect over q. If we choose $q_0 = 0.6$, the predicted lower bound would be about 1.1. Thus, we expect that for the holdout sample, it is likely that the observed lower bound of 0.95 one-sided interval for HR_q would be around 1.1.

[Figure 3 about here.]

4 Using a "cross-training-evaluation" procedure iteratively to identify an enrichable subgroup

For each given scoring system, with two independent data sets $\mathcal{A}$ and $\mathcal{B}$, the predicted Z-test score or the confidence interval lower bound curve as in Figure 2 or Figure 3 may not be stable due to small sample sizes. An alternative is to use the conventional "cross-training-evaluation" procedure iteratively for model fitting and evaluation. Specifically, we randomly divide the combined data set from the first and second stages into two independent pieces, $\mathcal{A}$ and $\mathcal{B}$ evenly,

as the training and the evaluation sets. As before, we use the data from $\mathcal{A}$ to build the scoring system and the data from $\mathcal{B}$ to obtain the predicted Z-test score curve as a function of q. By repeating such random cross-validation process M times, the final curve $\bar{Z}_q$ can be obtained by averaging those predicted Z-test score curves. Note that similar procedure can be performed based on K-fold cross-validation.

With the data from the first 900 patients in BEST, using this random cross-validation procedure with two Cox working models and ridge regression setting as discussed in the previous sections, the resulting $\bar{Z}_q$ curve with $M = 100$ is given in Figure 4. Note that this curve is much more smoother than that in Figure 2 using a single training and an evaluation data set. The $\bar{Z}_q$ has its highest value of 2.3 at $q = 0.6$.

[Figure 4 about here.]

Now, we may consider various working models and regularized estimation procedures to create candidate scoring systems using data set $\mathcal{A}$ and then construct, for example, the $\bar{Z}_q$ curves using data set $\mathcal{B}$ iteratively via the above random cross-validation scheme. We use the data from BEST to illustrate how to implement this process. In Table 2, we list several working, Cox-type models with various regularized estimation criteria for analyzing survival data. For instance, a procedure labeled as "Two.Cox_Ridge" means that we fit the data from each treatment group via a Cox model additively with baseline covariates using the partial likelihood function and the L_2 penalty as described in the previous section. On the other hand, a procedure labeled as "One.Cox_Lasso" means that we fit the data from two treatment groups together via a single Cox model with main covariate effects and the first order treatment and covariate interactions and the L_1 penalty. Note that "ALasso" stands for adaptive lasso, a new version of lasso, where adaptive weights are used for penalizing different coefficients based on the L_1 penalty (Zou, 2006).

[Table 2 about here.]

For the data from the first 900 patients in BEST, we utilize the same setup to implement the scoring system building and evaluation iteratively as the case "Two.Cox_ridge". The resulting predicted $\bar{Z}_q$ curves for all these candidate scoring systems are given in Figure 5. Note that the largest $\bar{Z}_q$ score is with the blue curve evaluated at $q = 0.6$. If the size of the enrichable subgroup and the significance level are acceptable, we then choose the scoring system which generated the blue curve to identify the enriched subgroup. For this case, it is based on two

Cox models with the ridge regression. It is interesting to observe that scores built with two separate models appear to perform uniformly better compared to those constructed by fitting an overall interaction model in this data set. This suggests that fitting a more flexible model could potentially lead to better performance. We then use the two Cox models with ridge regression procedure on the entire data set of the first 900 patients in BEST to obtain the final scoring system for ranking the future patients with respect to the benefit of beta-blocker compared with the control. The enrichable subgroup would consist of patients whose scores are among the top 60% of the entire study population. This final selected enriched subgroup will then be formally validated using the data from the holdout sample.

[Figure 5 about here.]

Now, if we are interested in using the size of the treatment difference for the final assessment of the treatment benefit, Figure 6 provides the corresponding average curves of the lower bounds of the 0.95 one-sided confidence interval estimates under the same setting. It appears that the above choice based on the blue curve is appropriate.

[Figure 6 about here.]

If we are interested in utilizing the difference of RMSTs (up to 3.5 years) for quantifying the treatment contrast, with the data from BEST and under the same setting discussed above, Figure 7 gives the average predicted $\bar{Z}_q$ curves with various scoring systems listed in Table 2 for testing the hypothesis that there is no difference between two treatment groups using the difference of two estimated RMSTs. It is interesting to note that the choice of the blue curve with $q_0 = 0.6$ would give us a reasonably sized of the enrichable subgroup and largest predicted Z-test score. Similar conclusions can be made via the expected lower bound of the 0.95 one-sided confidence interval estimates, which is provided in Figure 8.

[Figure 7 about here.]

[Figure 8 about here.]

It is important to note that the selection rule for an enrichable subgroup at this stage may depend on the totality of evidence based on various criteria rather than a single one. For instance, if the primary analysis for the holdout sample is based on the test for the hazard ratio being one, one would use Figure 5 first to choose a set of subgroups with different sizes, whose predicted

Z-test scores ranging from 2.0 to 2.4. Then among those subgroups, we can then examine how these subgroups behave with respect to other features (say, with respect to the difference of the RMSTs) and make a final recommendation on the choice of the enrichable subpouplation.

5 Using the holdout sample for the final validation of the selected enrichable subgroup

In Section 4, we choose a specific scoring system and a threshold value q_0 for selecting a potentially promising subgroup. We apply this rule to the data in the holdout sample. For this selected subgroup of patients in data set $\mathcal{C}$, one can then make inferences about the relative merits between two treatment groups.

As an example, for the data from BEST, the holdout sample is obtained from the remaining 1807 patients. If we are interested in making inference about the "constant" hazard ratio, we use the scoring system generated from two independent Cox models with ridge regression and the threshold value q_0 of 0.6. With the data from corresponding enrichable subgroup in the holdout sample, the right panel of Figure 9 provides the Kaplan-Meier curves for two treatment groups, p-value of the two sample test based on the hazard ratio estimates and the lower bound of the one-sided 0.95 confidence interval for the hazard ratio. The beta-blocker for patients in this enriched subgroup appears to prolong the event time compared with the control. On the other hand, the left panel shows the non-significant inferential results for the non-enrichable subgroup.

[Figure 9 about here.]

Now, if we are interested in quantifying the treatment contrast using the difference of RMSTs up to $\tau_0 = 3.5$ years, the right panel of Figure 10 provides the comparison results for the enrichable subgroup and the left panel gives the results for non-enrichable subset. For the enrichable subgroup, the one-sided p-value is 0.05 with the lower bound of the confidence interval of 0. Note that for the entire study population, the corresponding one-sided p-value via the test statistic based on the difference of two estimated RMSTs is 0.10 and one-sided 0.95 confidence interval is $(-0.08, +\infty)$.

[Figure 10 about here.]

With the event time data from the BEST study, we created two sets of patient-specific scores using the hazard ratio and the RMST difference as the treatment effect summary measures. If

we choose the "optimal" threshold values based on their average predicted Z-test scores from Figures 5 and 7, respectively, to define the enrichable subgroups, these two resulting subgroups in the holdout sample are almost identical to each other. This may suggest that the selection of the enrichable subgroup is quite robust to the choice of the summary measure.

6 Remarks

If there are two clinical trials conducted under the same setting, the data from one study can be used for building and evaluating the scoring systems via the "cross-training-evaluation" procedure described in the paper. Then we use the data from the second study for the assessment of the treatment effectiveness with pre-specified two-sample inference procedures to the sample of the selected enrichable population. It is important to note that the criteria for evaluating scoring systems and selection of the potentially promising subgroup should be consistent with those utilized for the holdout sample. Like the conventional cross-validation procedure for model building and selection, it is not clear how to choose the size of the data set at each stage of the process. Empirically we find that the final validation set size should be relatively large compared to that for the training and evaluation to obtain greater precision of the inference procedures conducted for the holdout sample (Shao, 1993).

It is interesting to note that for our analysis of the data from BEST, the statistical significance of the two sample test (either for the hazard ratio or difference of two RMSTs) with the holdout sample for the enrichable subgroup is not as impressive as its predicted counterpart obtained at the training-evaluation stage. This may be partially due to the phenomenon of over-optimistic even with cross-validation which does not correct for the overfitting due to the selection of an optimal q or an optimal model among all candidate models for building the scoring system. It is also important to keep in mind that the Z score in the holdout sample is a random variable by nature and would be difficult to predict with great precision, especially when the covariates are marginally predictive of the treatment effects.

The choice of a summary measure for the between-group difference is crucial. The hazard ratio is a heavily model-based treatment contrast measure. Although the proportional hazards assumption seems to be reasonable in the BEST study example, it is likely to be violated in general. On the other hand, a model-free and clinically meaningful two-sample contrast measure such as the difference of two medians or RMSTs seems more appropriate when there is no strong

evidence to support a model-based summary at the beginning of the study.

The selection of an enrichable subpopulation should be made from a risk-benefit perspective. Therefore, the outcome measure of the treatment effectiveness needs to reflect the disease burden or progression over the entire study follow-up period (Claggett *et al.*, 2012). For analyzing the data from BEST, one may choose an endpoint which reflects the patient's entire profile of morbidity and mortality, not only the event time for the first hospitalization or death.

References

Anderson, P., Hansen, M., and Klein, J. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* **10**, 4, 335–350.

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**, 6, 607–616.

Chen, P.-Y. and Tsiatis, A. A. (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* **57**, 4, 1030–1038.

Claggett, B., Tian, L., Castagno, D., and Wei, L.-J. (2012). Treatment selections using risk-benefit profiles based on data from comparative randomized clinical trials with multiple endpoints. *Harvard University Biostatistics Working Paper Series* .

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

Eichhorn, E., Domanski, M., Krause-Steinrauf, H., and Anderson, J. (2001). A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure. *ACC Current Journal Review* **10**, 6, 49.

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *The Annals of Statistics* **30**, 1, 74–99.

Freidlin, B. and Simon, R. (2005). Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* **11**, 21, 7872–7878.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**, 1, 1–22.

Irwin, J. (1949). The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiment with mice. *Journal of Hygiene* **47**, 02, 188–189.

Jiang, W., Freidlin, B., and Simon, R. (2007). Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute* **99**, 13, 1036–1043.

Karrison, T. (1987). Restricted mean life with adjustment for covariates. *Journal of the American Statistical Association* **82**, 400, 1169–1176.

Karuri, S. W. and Simon, R. (2012). A two-stage bayesian design for co-development of new drugs and companion diagnostics. *Statistics in medicine* **31**, 10, 901–914.

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**, 10, 1209.

Murray, S. and Tsiatis, A. A. (1999). Sequential methods for comparing years of life saved in the two-sample censored data problem. *Biometrics* **55**, 4, 1085–1092.

Qian, M. and Murphy, S. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics* **39**, 2, 1180–1210.

Rothwell, P. (1995). Can overall results of clinical trials be applied to all patients? *The lancet* **345**, 8965, 1616–1619.

Rothwell, P., Mehta, Z., Howard, S., Gutnikov, S., and Warlow, C. (2005). From subgroups to individuals: general principles and the example of carotid endarterectomy. *The Lancet* **365**, 9455, 256–265.

Royston, P. and Parmar, M. K. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine* **30**, 19, 2409–2421.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association* **88**, 422, 486–494.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* **39**, 5, 1–13.

Simon, R. (2008). Designs and adaptive analysis plans for pivotal clinical trials of therapeutics and companion diagnostics. *Expert Opinion on Medical Diagnostics* **2**, 6, 721–729.

Tian, L., Cai, T., Zhao, L., and Wei, L.-J. (2012). On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial. *Biostatistics* **13**, 2, 256–273.

Tian, L., Zhao, L., and Wei, L.-J. (2014). Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics* **15**, 2, 222–233.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Tibshirani, R. *et al.* (1997). The lasso method for variable selection in the cox model. *Statistics in medicine* **16**, 4, 385–395.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in medicine* **27**, 23, 4658–4677.

U.S. Food and Drug Administration (2012). Guidance for industry: Enrichment strategies for clinical trials to support approval of human drugs and biological products. `http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.htm`.

Verweij, P. J. and Van Houwelingen, H. C. (1994). Penalized likelihood in cox regression. *Statistics in Medicine* **13**, 23-24, 2427–2436.

Wang, S.-J., O'Neill, R. T., and Hung, H. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics* **6**, 3, 227–244.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012a). Estimating optimal treatment regimes from a classification perspective. *Stat* **1**, 1, 103–114.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012b). A robust method for estimating optimal treatment regimes. *Biometrics* **68**, 4, 1010–1018.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94**, 3, 691–703.

Zhang, M. and Schaubel, D. E. (2011). Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics* **67**, 3, 740–749.

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L.-J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association* **108**, 502, 527–539.

Zhao, L., Tian, L., Uno, H., Solomon, S. D., Pfeffer, M. A., Schindler, J. S., and Wei, L. J. (2012). Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials* **9**, 5, 570–577.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101**, 476, 1418–1429.

Zucker, D. M. (1998). Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association* **93**, 442, 702–709.

Figure 1: Kaplan-Meier curves for the time to death or the first hospitalization (Beta-blocker versus placebo) with the data from BEST

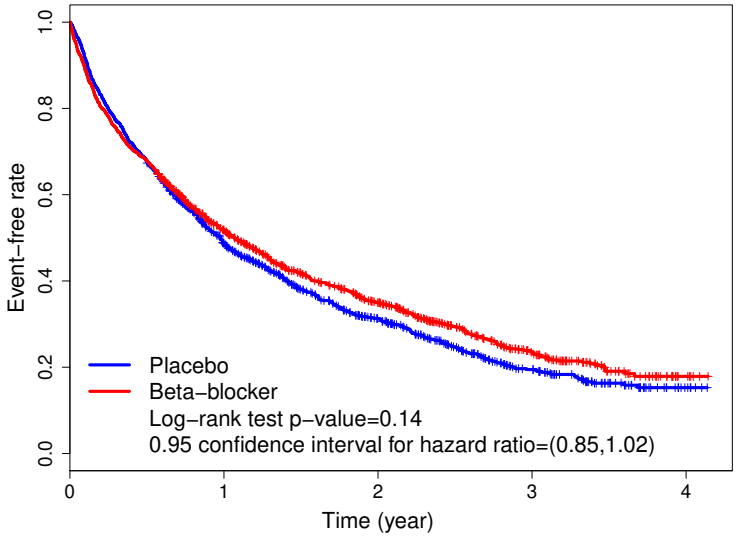

Figure 2: Predicted Z-test score curve based on hazard ratio estimates with the data from 450 patients randomly chosen from the first 900 patients in BEST study

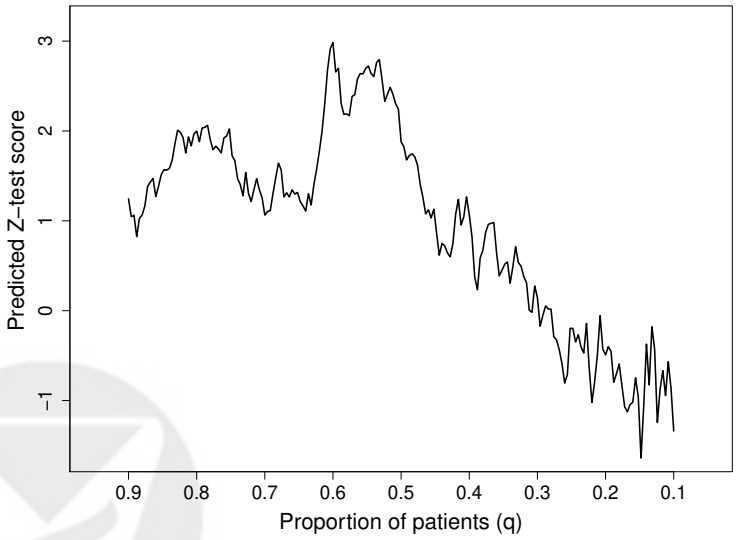

Figure 3: Predicted lower bound of one-sided 0.95 confidence interval curve based on hazard ratio estimates with the data from 450 patients randomly chosen from the first 900 patients in BEST study

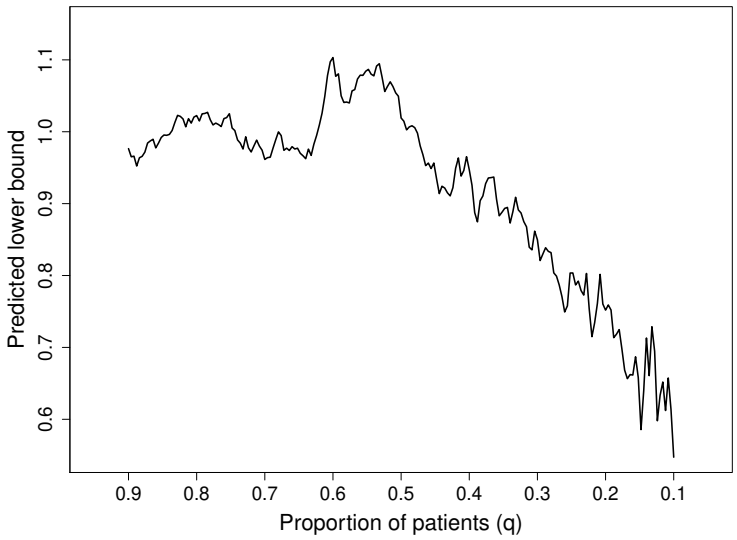

Figure 4: The predicted $\bar{Z}_q$ curve using the data from the first 900 patients entered BEST study based on 100 random cross-validation by fitting two Cox models with ridge regression

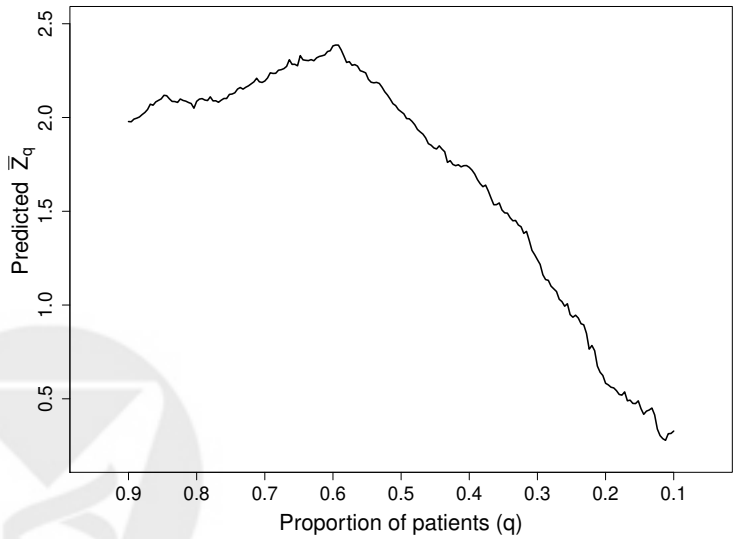

Figure 5: The predicted $\bar{Z}_q$ curves for testing treatment difference with two-sample hazard ratio estimates based on 100 random cross-validation with various candidate scoring systems listed in Table 2.

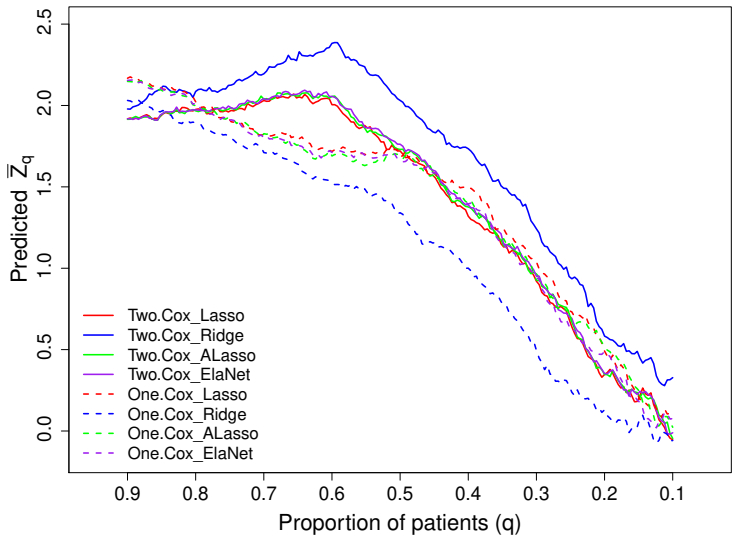

Figure 6: The average predicted lower bound curves of 0.95 one-sided confidence interval estimates for two-sample hazard ratios based on 100 random cross-validation with various candidate scoring systems listed in Table 2.

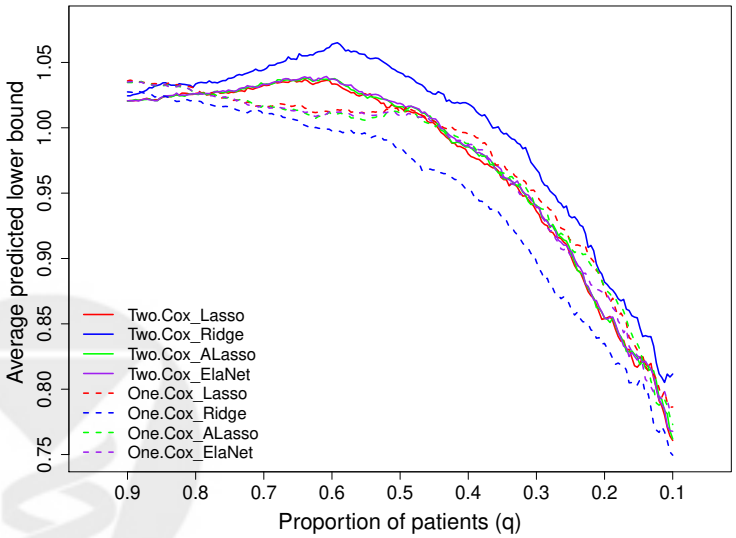

Figure 7: The predicted $\bar{Z}_q$ curves for testing treatment difference with the difference of two estimated RMSTs based on 100 random cross-validation with candidate scoring systems listed in Table 2

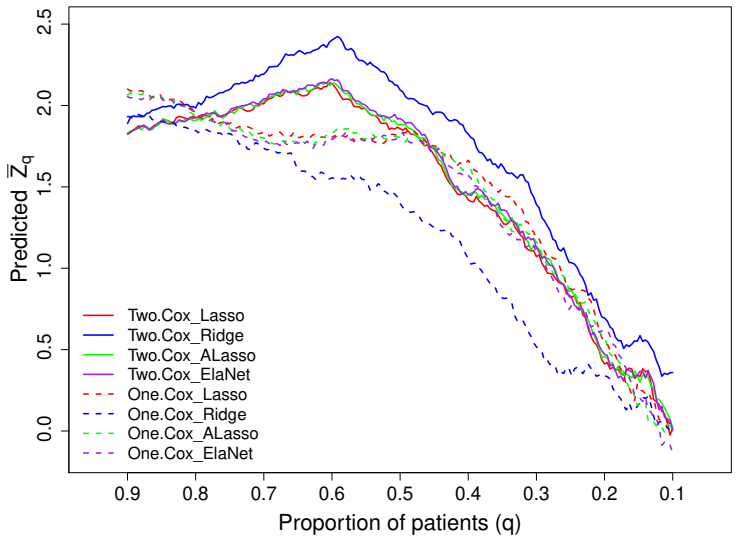

Figure 8: The average predicted lower bound curves of 0.95 one-sided confidence interval estimates for two estimated RMSTs based on 100 random cross-validation with candidate scoring systems in Table 2.

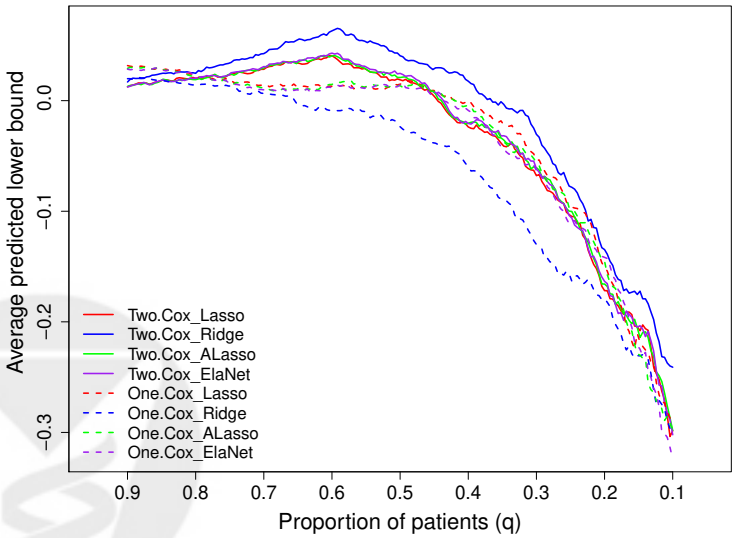

Figure 9: Two-sample inferences for the hazard ratio

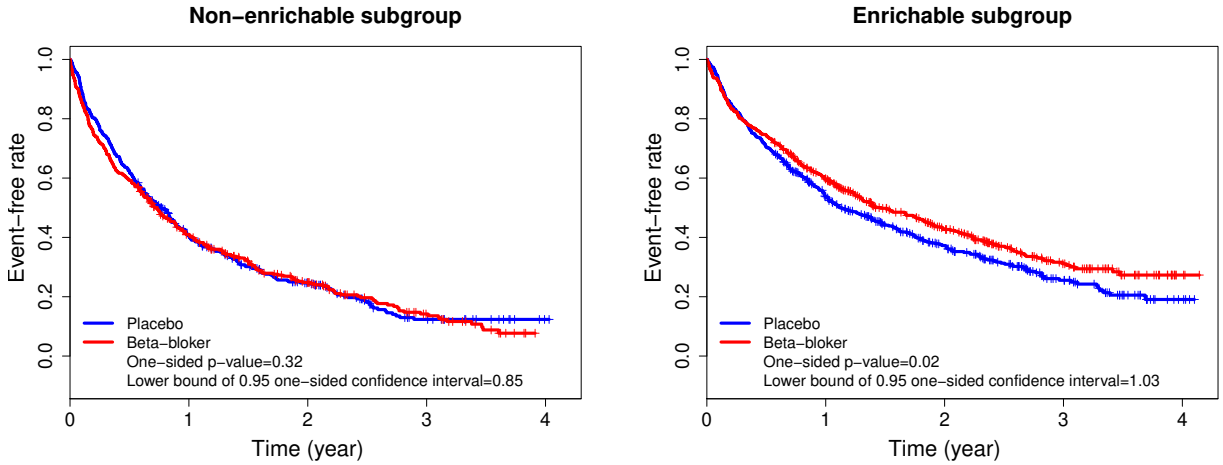

Figure 10: Two-sample inferences for the difference if two RMSTs up to 3.5 year follow-up

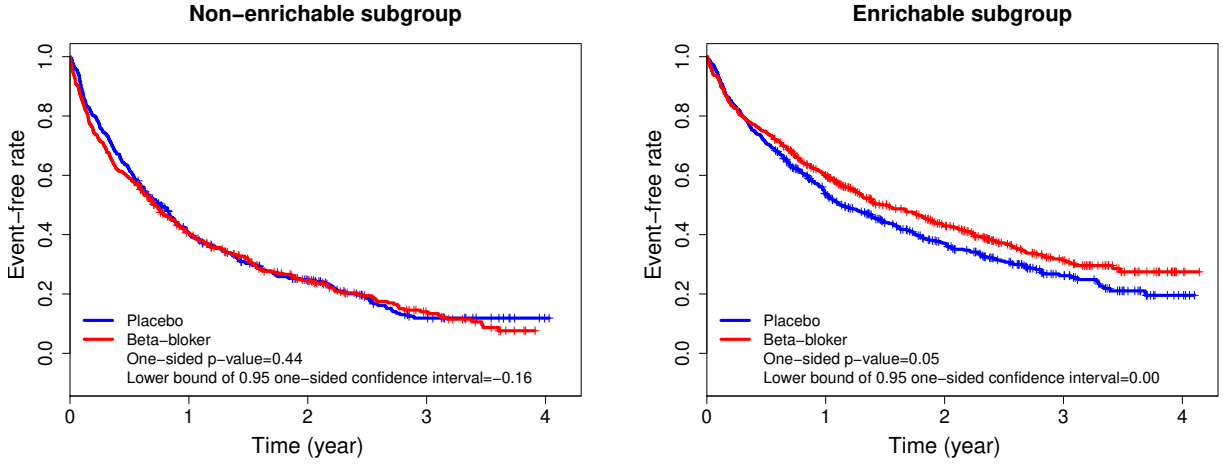

Table 1: Regression coefficient estimates by fitting two Cox models with ridge regression procedure using a randomly chosen subset of 450 patients from the first 900 patients in the BEST study

Covariate	$\widehat{\beta}_0$ (Placebo)	$\widehat{\beta}_1$ (Beta-blocker)
Age	0.003	-0.004
Male	-0.001	-0.102
LVEF	-0.011	-0.016
I($45 < eGFS \leq 60$)	-0.141	-0.353
I($60 < eGFS \leq 75$)	-0.170	-0.599
I($eGFS > 75$)	-0.243	-0.880
SBP	-0.002	-0.005
Class IV Heart Failure	-0.322	1.153
I($BMI > 30$)	-0.0002	0.016
Ever Smoker	-0.088	-0.054
Heart Rate	0.004	-0.007
History of Hypertension	0.051	0.023
History of Diabetes	0.070	0.114
Ischemic Etiology	0.049	0.244
Atrial Fibrillation	0.124	0.035
White Race	0.024	-0.062

Table 2: List of candidate scoring systems

Scoring system	Working model	Regularized estimation
Two.Cox_Lasso	Two Cox models	Lasso
Two.Cox_Ridge	Two Cox models	Ridge
Two.Cox_ALasso	Two Cox models	Adaptive Lasso
Two.Cox_ElaNet	Two Cox models	Elastic Net
One.Cox_Lasso	One Cox interaction model	Lasso
One.Cox_Ridge	One Cox interaction model	Ridge
One.Cox_ALasso	One Cox interaction model	Adaptive Lasso
One.Cox_ElaNet	One Cox interaction model	Elastic Net