Harvard University

Harvard University Biostatistics Working Paper Series

Year 2014 *Paper* 173

Predicting the Future Subject's Outcome via an Optimal Stratification Procedure with Baseline Information

Florence H. Yong* Lu Tian† Sheng Yu‡

Tianxi Cai** L. J. Wei††

*Harvard University, florenceyong04@hotmail.com

†Stanford University, lutian@stanford.edu

‡Harvard University

**Harvard University, tcai@hsph.harvard.edu

††Harvard University, wei@hsph.harvard.edu

Predicting the future subject's outcome via an optimal stratification procedure with baseline information

Florence H. Yong,* Lu Tian,† Sheng Yu,* Tianxi Cai,* and LJ Wei*

ABSTRACT

In predictive medicine, one generally utilizes the current study data to construct a stratification procedure, which groups subjects with baseline information and forms stratum-specific prevention or intervention strategies. A desirable stratification scheme would not only have a small intra-stratum variation, but also have a "clinically meaningful" discriminatory capability across strata to avoid unstable and overly sparse categorization. We show how to obtain an optimal stratification strategy with such desirable properties from a collection of candidate models. Specifically, we fit the data with a set of regression models relating the outcome to its baseline covariates. For each fitted model, we create a scoring system for predicting potential outcomes and obtain the corresponding optimal stratification rule. Then, all the resulting stratification strategies are evaluated with an independent dataset to select a final stratification system. Lastly, we obtain the inferential results of this selected stratification scheme with a third independent holdout dataset. If there is only one current study data available and the study size is moderate, we combine the first two steps via a conventional cross-validation process. We illustrate the new proposal using an AIDS clinical trial study for binary outcome and a cardiovascular clinical study for censored event time outcome.

KEY WORDS: Cox regression model; Cross-validation; Dynamic programming; Prediction score; Stratified medicine

1. INTRODUCTION

To construct a prediction procedure for the future subject's outcome via its baseline information, a common practice at the first step is to fit the current data with a parametric or semi-parametric

*Department of Biostatistics, Harvard University, Boston, MA 02115

†Department of Health Research and Policy, Stanford University, Stanford, CA 94305

regression model, which relates the subject's outcome to its covariates. If the model is a reasonable approximation to the true one, the resulting individual predicted value would be close to its outcome value. Such predicted values create a scoring system for all future subjects. In practice, one then groups the scores into several strata and uses the average of observed outcome values in one stratum to predict outcomes from new subjects classified into the same stratum for making targeted prevention or intervention. A desirable stratification scheme would have a small intra-stratum variation and a clinically meaningful discriminatory capability to avoid unstable and overly sparse groupings. To the best of our knowledge, there are no systematic approaches one can take to construct an optimal stratification with such desirable features.

As an example to illustrate the current practice in stratified medicine, we utilize the data from a clinical study for treating HIV diseases (Hammer et al., 1997). This trial (ACTG 320) was a randomized, double-blind, placebo-controlled clinical study conducted by the AIDS Clinical Trials Group. It successfully demonstrated the overall efficacy of a combination of two nucleoside regimen with a protease inhibitor Indinavir for treating HIV-infected patients. The combination treatment concept has since been well adopted for the current HIV patient's management. However, this particular combination therapy may not work well for all patients from risk-cost-benefit perspectives. Here, we show a conventional, ad hoc procedure to construct a stratification scheme using the patients' baseline variables. For this study, there were 537 patients treated by the three drug combination who had complete baseline information. One of the endpoints was a binary outcome Y, indicating whether the patient's HIV-RNA viral level was under an assay detectable level (500 copies/mL) at week 24 or not. A non-responder to the treatment was defined as the RNA level being above 500 copies/mL at week 24 or an informative dropout before week 24 due to treatment failure. The observed overall response rate was 45%. To build a predictive scoring system, for illustration, let us consider a rather simple additive logistic regression model for Y with four baseline covariates: age, sex (female=1, male=0), CD4 count (CD4_0), and the $\log_{10}$ of HIV-RNA ($\log_{10}\text{RNA}_0$). The numerical RNA level is used and all $\log_{10}\text{RNA}_0$ measures below $\log_{10}(500)$ are replaced by $0.5\log_{10}(500) = 1.35$ in our analysis. We then fit the entire dataset with the model and the resulting individual predicted response rate is:

$$\psi(-0.508 + 0.044\text{age} - 0.493\text{sex} + 0.004\text{CD4}_0 - 0.346\log_{10}\text{RNA}_0), \qquad (1.1)$$

where $\psi(s) = \{1 + \exp(-s)\}^{-1}$ is the anti-logit function. The 537 predicted response rates range from 0.09 to 0.93. If the model is reasonably good, a future subject with a high score tends to respond to the treatment. A conventional way to group those patients is to stratify them into, for instance, four consecutive categories with roughly equal sizes by using the quartiles of the predicted scores. The empirical average response rates for these strata are 31%, 42%, 38%, and 67%, respectively. Unfortunately this ad hoc stratification scheme does not have discriminatory capability across all the strata. The average response rates are not monotonically increasing over the ordered strata, potentially due to the inadequate prediction model (1.1) or an improper grouping of the prediction scores.

In this article, we present an optimal and systematic stratification strategy incorporating model selection from a collection of candidates which satisfy certain clinically meaningful criteria. Specifically, in Section 2, we show how to obtain an optimal grouping scheme for each candidate scoring system created from a regression model. For example, with the predicted response rates (1.1), we consider all possible discretization schemes, whose stratum sizes would be at least 10% of the study sample size and any two consecutive stratum-specific average response rates are monotonically increasing with an incremental value of at least 20% to ensure a discrimination capability for future population. We then choose the best stratification, which minimizes a certain overall prediction error (2.2) among all the possible stratification schemes with such desirable features. Dynamic programming techniques are utilized to solve this nontrival optimization problem. With the data from the HIV example and model (1.1), this results in three categories with the stratum-specific average response rates of 11%, 42% and 69% and stratum sizes of 65, 343, and 129, respectively. Note that if model (1.1) is appropriate, future patients classified to the first stratum may not benefit much from this rather costly three-drug combination therapy especially for regions where the resource is limited.

In Section 3, we consider a collection of candidate scoring systems and utilize the method in Section 2 to obtain the optimal stratification for each scoring system. To avoid the overfitting problem, we then use an independent dataset to evaluate all the resulting stratification schemes with respect to a clinically interpretable prediction error measure, which also quantifies the within-stratum heterogeneity, to select the final stratification scheme. The last step is to make inferences for the selected prediction procedure using a third independent dataset. If the data are from a single study with a

moderate size, one may combine the model building and evaluation processes via a cross-validation procedure. In Section 4, we generalize the new proposal to handle censored event time outcomes and illustrate the procedure using the data from a recent cardiovascular study (Braunwald et al., 2004) in Section 5. We conclude with additional observations and potential generalizations in Section 6.

2. AN OPTIMAL STRATIFICATION PROCEDURE FOR A SPECIFIC SCORING SYSTEM

In this section, we show how to obtain an optimal grouping system from a single working model such as (1.1). Let Y be the outcome variable and V be a vector of "baseline" covariates. Assume that the conditional mean $\mu(V) = \mathrm{E}(Y|V)$ is the parameter of interest for future prediction. To estimate $\mu(V)$ when the dimension of V is more than one, we generally use a working model which relates Y to Z, a function of V. For example, $\mu(V) = g(\beta' Z)$, where $g(\cdot)$ is a given smooth monotone function. Let the data consist of n independent copies $\{(Y_i, V_i, Z_i), i = 1, \cdots, n\}$ of (Y, V, Z). An estimate $\hat{\beta}$ for β can be obtained via a regularized estimation procedure, for example, the least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996), especially when the dimension of Z is large. If the regression model is a reasonable approximation to the true one, the resulting estimator $\hat{\mu}(V) = g(\hat{\beta}' Z)$ would be close to $\mu(V)$, and a large $\hat{\mu}(\cdot)$ indicates that the subject would have a large outcome value Y.

As an example, for the binary outcome Y in the HIV study discussed in the Introduction, one may use a logistic model with lasso or ridge regularization to obtain $\hat{\mu}(\cdot)$. Here, the regression parameter estimate $\hat{\beta}$ is obtained by minimizing the loss function

$$-\log\{L(\beta)\} + \lambda \|\beta\|_p^p, \tag{2.1}$$

where $L(\beta)$ is the likelihood function, λ is the tuning (penalty) parameter and $\|\beta\|_p^p$ is the p^{th} power of the L_p norm for the vector β. The (2.1) results in the standard lasso and ridge regression with $p = 1$ and 2, respectively. The score (1.1) in the Introduction gives the individual predicted response rate based on the simple additive logistic regression with four baseline covariates and $\lambda = 0$.

Suppose that we group n subjects into K consecutive strata $S_1, S_2, \cdots, S_K$ based on the score $\hat{\mu}(\cdot)$. Let $\bar{Y}_k$ be the empirical mean of Y_i's in the k^{th} stratum, $k = 1, \cdots, K$. For a future sub-

ject being classified to the k^{th} stratum, we predict the individual outcome with the corresponding stratum-specific mean $\bar{Y}_k$. To evaluate the performance of this stratification, one may consider a loss function:

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in S_k}|Y_i-\bar{Y}_k|, \tag{2.2}$$

which also quantifies the average within stratum variation. When all the scores are distinct, an optimal stratification, which minimizes (2.2), would result in n strata with only one member in each stratum and the observed prediction error is zero. However, the prediction error for future observations with such an overly sparse stratification would be unacceptably high. To increase the prediction precision while ensuring stable subgroups, one may group the subjects with a minimal stratum size of at least a certain fraction p_0 of the sample size n. Moreover, to ensure that the stratification scheme has a clinically meaningful discriminatory capability, namely, yielding meaningful differences in group-specific average outcomes between subgroups, we further impose a constraint for all candidate stratification schemes such that

$$\bar{Y}_k-\bar{Y}_{k-1}\geq d;\ \text{ for }\ k=2,\cdots,K, \tag{2.3}$$

where d is a given positive value, representing the minimum clinically meaningful increment.

Consider all possible stratifications which satisfy (2.3) with a minimum stratum size fraction of at least p_0. To obtain an optimal stratification scheme by minimizing (2.2) is a rather challenging problem. In Appendix A, we show how to identify the boundary values $\{\hat{c}_0,\hat{c}_1,\cdots,\hat{c}_{K-1},\hat{c}_K\}$ of the consecutive strata via dynamic programming (Taha, 2003), that is, $S_k=\{i\mid \hat{c}_{k-1}<\hat{\mu}(V_i)\leq\hat{c}_k, i=1,\cdots,n\}, k=1,2,\cdots,K$. Without loss of generality, here we assume that $\hat{c}_0=-\infty$ and $\hat{c}_K=\infty$. Note that we use L_1 norm (2.2) to evaluate the prediction error, which is more heuristically interpretable than, for example, the one with the L_2 norm. Furthermore, if the regularized estimator $\hat{\beta}$ of the working regression model converges to a constant vector β_0 and the resulting score estimate $\hat{\mu}(v)$ converges to a deterministic function $\tilde{\mu}(v)$ for all v, the above empirical optimal stratification scheme, in the limit, minimizes, with respect to

$\mathbf{c} = \{-\infty = c_0 < c_1 < \cdots < c_{K-1} < c_K \ = \infty\}$, the limit of (2.2):

$$L(\mathbf{c}) = \mathrm{E}\big|Y - f(V|\mathbf{c})\big| \quad \text{subject to} \quad \begin{cases} \mathrm{pr}(c_{k-1} < \tilde{\mu}(V) \le c_k) \ge p_0 \\ \bar{\mu}_k - \bar{\mu}_{k-1} \ge d \end{cases}, \tag{2.4}$$

where

$$f(V|\mathbf{c}) = \sum_{k=1}^{K} I(c_{k-1} < \tilde{\mu}(V) \le c_k)\bar{\mu}_k,$$

$I(\cdot)$ is the indicator function and $\bar{\mu}_k = \mathrm{E}(Y \mid c_{k-1} < \tilde{\mu}(V) \le c_k)$. The justification of this asymptotic property is given in Appendix B. Note that using the lasso regularized estimation procedure, the estimators $\hat{\beta}$ and $\hat{\mu}(\cdot)$ are stabilized asymptotically. This large sample property of the optimal stratification scheme is essential to ensure its stability under the cross-validation setting discussed in Section 3.

3. SELECTING AN OPTIMAL STRATIFICATION SCHEME FROM A COLLECTION OF COMPETING SCORE SYSTEMS

For a prediction score system created by a given working regression model, one can obtain its optimal stratified prediction procedure as presented in Section 2. To make inferences about the resulting stratified prediction procedure, for instance, constructing a valid confidence interval estimate for the mean response value of each stratum, one may utilize an independent dataset to avoid overly optimistic inferential conclusions. To this end, with data from a single study, we may split the data into two independent parts, say, I and II. Using the data from Part I, we obtain the optimal stratification scheme. Then we use the data from Part II to make inferences for prediction. Moreover, if there is a collection of competing stratified score systems considered as potential candidates, we further split the data from Part I into two independent parts, say, Ia and Ib. The data from Part Ia are used for obtaining the optimal stratification schemes for each candidate scoring system as we did in Section 2, whereas data from Part Ib are used for evaluating all candidates and selecting the best stratification scheme.

Suppose that there are several optimal stratification schemes available with the data from Part Ia. In this section, we show how to evaluate them and choose the "best" one. Let the data from Part Ib be denoted by n^* independent identically distributed observations $\{(Y_i^*, V_i^*, Z_i^*), i = 1, \cdots, n^*\}$,

where a generic variable "A^*" is defined as "A" in Section 2. For each candidate scoring system, we obtain its optimal stratified counterpart from the data of Part Ia with boundary points $\{\hat{c}_0, \hat{c}_1, \cdots, \hat{c}_K\}$ and the stratum-specific prediction values $\{\bar{Y}_k, k = 1, \cdots, K\}$. To evaluate the predictive performance of such a stratification scheme with the data from Part Ib, we consider the following loss function, reflecting the within-stratum variation for the prediction accuracy:

$$\mathcal{L}^* = \frac{1}{n^*} \sum_{k=1}^{K} \sum_{\hat{\mu}(V_i^*) \in (\hat{c}_{k-1}, \hat{c}_k]} |Y_i^* - \bar{Y}_k|, \tag{3.1}$$

where $n^* = \sum_{k=1}^{K} n_k^*$ and $\hat{\mu}(V_i^*) = g(\hat{\beta}' Z_i^*)$. An optimal stratification scheme is chosen which minimizes (3.1) among all the candidates under consideration. On the other hand, a parsimonious model may be appealing in practice if its (3.1) is greater than but still comparable to the minimum value of (3.1) derived from a complex model.

Now, since the sizes of Parts Ia and Ib may be small, one may use the Monte-Carlo cross-validation (MCCV) method (Xu and Liang, 2001; Yong et al., 2013) to obtain a more stable (3.1). Specifically, we randomly split Part I dataset into Ia and Ib, say, N times. For the j^{th} split, we repeat the above model building and evaluation procedure for each candidate model and obtain $\mathcal{L}_j^*$ from (3.1). We then compute the average, $\bar{\mathcal{L}}^* = N^{-1} \sum_{j=1}^{N} \mathcal{L}_j^*$. For each candidate model, we refit the entire Part I data and let the final realized stratification rule be denoted by $\mathcal{M}^*$. The pair $(\bar{\mathcal{L}}^*, \mathcal{M}^*)$ reflects the magnitude of the estimated within-stratum variation and the model complexity of each candidate. The selection of an "optimal" stratification rule would be based on such pairs. With the data from Part II, we then construct confidence interval estimates for the stratum-specific mean values of the outcome variables for the final selected stratification scheme. It is important to note that the lasso regularized regression coefficient estimate is stabilized asymptotically for each regression model fitting in the above cross-validation process. It follows that for each candidate regression model, the final refitted stratification scheme would minimize (2.4) in the limit.

We now use the data from the HIV study to illustrate our proposal. For this study, other than the four baseline variables discussed in the Introduction, there are seven baseline covariates and two short-term marker values at week 4 including CD4 count (CD4_4) and $\log_{10}$ RNA ($\log_{10}\text{RNA}_4$), which may be relevant to the outcome and have potential predictive values. The additional baseline covariates are race (non-Hispanic White, African American, other), injection-drug use, hemophilia,

CD8 count, weight, Karnofsky performance score, and months of prior zidovudine therapy. There are very few missing covariate values. Any missing covariates are replaced by their corresponding sample averages from the observed counterparts.

In our analysis, we first randomly split the entire dataset of 537 patients into Part I and Part II evenly with sample sizes of 268 and 269, respectively. The number N of the MCCV is 200 and the sizes of Part Ia and Ib are equal for each cross-validation. For illustration, we consider four different working models in which the first three are various logistic regression models with lasso regularization methods and tuning parameters selected via a 20-fold cross-validation procedure built in the R package $glmnet$ (Friedman et al., 2010). The fourth model is a null model using the overall mean response proportion in Part Ia to predict future outcomes for each cross-validation run. Table 1 summarizes the composition of each model. We also present $\bar{\mathcal{L}}^*$ obtained by averaging the 200 $\mathcal{L}_j^*$ values; and for the corresponding $\mathcal{M}^*$, we report its number of informative baseline covariates needed for computing the score $\hat{\mu}(V)$ and number of nonzero regression coefficients in $\hat{\beta}$ to summarize its complexity. Note that for all candidate stratification schemes, we use the incremental value of $d = 0.2$ and the minimum stratum size fraction of $p_0 = 0.1$ in the Part Ia training data.

From Table 1, Models 1 and 3 have almost the same $\bar{\mathcal{L}}^*$ values, but the $\mathcal{M}^*$ of Model 1 has fewer baseline covariates involved and the resulting predicted response rate is

$$\psi(-0.231 - 0.075 \log_{10} \text{RNA}_0 - 0.459 \log_{10} \text{RNA}_4 + 0.00036\text{CD4}_0 + 0.0028\text{CD4}_4 + 0.0288\text{age}).$$

With this final selected stratification scheme $\mathcal{M}^*$, there are three strata whose $\hat{c}_1 = 0.25$ and $\hat{c}_2 = 0.45$. The stratum-specific means and numbers of observations are 0.06 ($n = 51$), 0.36 ($n = 107$), and 0.62 ($n = 110$), respectively. Note that these stratum-outcome-average estimates may be biased due to the extensive model building, evaluation and selection. To obtain valid inferences for this final prediction procedure, we use the above stratification boundary values $\hat{c}_1$ and $\hat{c}_2$ to group subjects from Part II. The resulting point and 0.95 confidence interval estimates for the three stratum-average response rates are 0.17 (0.06, 0.28), 0.41 (0.31, 0.51) and 0.65 (0.57, 0.73), with stratum size $n = 47, 91$, and 131 respectively as displayed in Figure 1. Note that the above inferential results would be a valid and final assessment on the practical value of this prediction scheme.

4. GENERALIZATION TO CASES WITH EVENT TIME AS THE OUTCOME VARIABLE

If the outcome T is the time to a specific event, potentially this variable T may be censored and the mean or median value of the outcome variable cannot be estimated well. A common summary parameter of interest is the event rate at a specific time point τ. However, this measure does not include information about the event occurrence profile. On the other hand, the restricted mean survival time (RMST) is a clinically meaningful summary for such a distribution (Royston, 2009; Royston and Parmar, 2011; Zhao et al., 2013). Specifically, let $Y = TI(T \leq \tau) + \tau I(T > \tau)$ and $\mu(V) = \mathrm{E}(Y|V) = \int_0^\tau S(t|V)dt$ as defined in Section 2, where $S(t|V) = \mathrm{pr}(T > t \mid V)$. Here, $\mu(V)$ is the average event-free time for all subjects with covariate V, which would be followed up to time point τ. Often, the outcome T (and Y) may be right censored by an independent random variable C. However, one can always observe (X, V, Δ), where $X = \min(T, C)$ and Δ is a binary variable, which is one if $X = T$ and zero otherwise. Therefore, the observed data consist of n independent copies $\{(X_i, V_i, \Delta_i), i = 1, \cdots, n\}$ of (X, V, Δ). Note that $Y_i = min(T_i, \tau)$ is observed when $\Delta_i = 1$ or $T_i \geq \tau$.

Inferences about $\mu(V)$ under the one- and two-sample and regression settings have been extensively studied (Zucker, 1998; Tian et al., 2013). For example, to estimate the RMST for a single group, the area under the Kaplan-Meier curve is a nonparametric consistent estimator. To create a scoring system for $\mu(V)$, one may use the Cox (1972) procedure to model the relationship between the survival function $S(t|V)$ of the event time and its covariates V :

$$\log\{-\log S(t|V)\} = \log\{-\log S_0(t)\} + \beta' Z,$$

where $S_0(\cdot)$ is an unknown baseline survival function, and β is the regression coefficient vector. A regularized estimate $\hat{\beta}$ of β can be obtained by minimizing

$$-\log(PL(\beta)) + \lambda\|\beta\|_p^p,$$

where $PL(\cdot)$ is the partial likelihood function. The $S_0(t)$ can then be estimated by $\exp\{-\hat{\Lambda}_0(t)\}$, where $\hat{\Lambda}_0(t)$ is the Breslow estimate for the underlying cumulative hazard function (Breslow, 1972).

It follows that the RMST for subjects with the covariate V can be estimated as

$$\hat{\mu}(V) = \int_0^\tau \exp\{-\hat{\Lambda}_0(t)e^{\hat{\beta}'Z}\}dt.$$

For any scoring system, we can then use the same technique described in section 3 to obtain an optimal stratification scheme. Specifically, in the limit, we are interested in minimizing (2.4). Assuming that the censoring time is independent of the survival time T and covariates V, the prediction error in (2.4) can be estimated as

$$n^{-1}\sum_{k=1}^{K}\sum_{i\in S_k} w_i|Y_i - \bar{Y}_k|,$$

where $w_i = \{\Delta_i + (1-\Delta_i)I(T_i \geq \tau)\}/\hat{G}(Y_i)$ and $\hat{G}(\cdot)$ is the Kaplan-Meier estimate for the censoring distribution using the entire dataset (Part I and II). Here, $\bar{Y}_k$ is a consistent estimator for the k^{th} stratum-specific RMST, which is the weighted average

$$\frac{\sum_{i\in S_k} w_i Y_i}{\sum_{i\in S_k} w_i}.$$

With the same constraints as described in Section 2, an optimal stratification can be obtained via the dynamic programming technique given in Appendix A. If the estimated RMST converges to a deterministic limit as the sample size increases, it follows from a similar argument in Appendix B, the finite sample stratified scheme would have the same asymptotic property as that for the non-censored case.

To select the "best" scoring model from the competing scoring systems, one can utilize the procedure in Section 3 with the weighted version of (3.1) to evaluate the candidate stratification schemes via the cross-validation using the data from Part Ia and Ib iteratively. The inference of the prediction procedure with the final selected model can then be made accordingly with the data from Part II.

5. AN ILLUSTRATIVE EXAMPLE WITH CENSORED EVENT TIME OUTCOMES

We use the data from a cardiovascular clinical trial "Prevention of Events with Angiotensin Converting Enzyme Inhibition" (PEACE) to illustrate the proposal with an event time outcome variable. The PEACE trial is a double-blind, placebo-controlled study (Braunwald et al., 2004) of 8290 pa-

tients enrolled to investigate if the addition of an Angiotensin-converting-enzyme (ACE) inhibitor therapy trandolapril at a target dose of 4 mg/day to the conventional therapy would provide benefit with respect to, for example, the patient's specific cardiovascular event-free survival. For illustration of our proposal, the outcome variable is assumed to be the time to death, nonfatal myocardial infarction or coronary revascularization, whichever occurred first. There are 2110 patients (25%), who experienced this composite event with the median follow-up time of 54 months. The 0.95 confidence interval estimate for the hazard ratio is (0.86, 1.02) with a p-value of 0.15 based on the logrank test. Since there was no statistically significant treatment effect, we combined the data from the two treatment groups for our illustration. The Kaplan-Meier curve for the entire dataset is given in Figure 2. The overall observed event times in months range between 0.1 and 81.5 with an interquartile range of 12.8 and 42.4. If we let $\tau = 72$ (months), the estimated restricted mean event time for the entire group is 60.4 months. This suggests that for future patients in this study population, one expects to have an average of 60.4 months event-free with a follow-up time of 72 months.

Based on the results by Solomon et al. (2006), we considered the following baseline covariates for prediction: the study treatment indicator, age, gender, left ventricular ejection fraction, history of myocardial infarction, history of hypertension, history of diabetes, and estimated glomerular filtration rate as a 4-category discretized version represented by 3 indicator variables eGFR_1, eGFR_2 and eGFR_3 with cut-points of 45, 60, and 75. We imputed the missing covariate values with their corresponding sample mean counterparts for continuous variables and the most frequently observed category for binary variables. We then randomly split the data evenly into Parts I and II with 4145 patients each. Moreover, for Part I data, we randomly split it evenly for the cross-validation process with 200 iterations. Several candidate models are considered and listed in Table 2. Note that Model 2 is built upon the observation that there is potential treatment and eGFR interaction reported by Solomon et al. (2006).

For each regression candidate model, we use the incremental value of $d = 3$ months and the minimum stratum fraction of $p_0 = 0.1$. Table 2 summarizes the $\bar{\mathcal{L}}^*$ for the optimal stratification based on each regression working model as well as the numbers of informative baseline covariates used in computing the estimated scores and nonzero regression coefficients of $\hat{\beta}$ for $\mathcal{M}^*$. Model 2 has the smallest $\bar{\mathcal{L}}^*$ and yields three strata with cutoff points $\hat{c}_1$ = 56.5 and $\hat{c}_2$ = 60.5 months. The range of

estimated RMST is from 51.0 to 63.8 months in Part I dataset. The corresponding estimated RMSTs for three strata are 54.5, 58.7 and 62.3 months, respectively. To make inferences about the prediction of this selected final stratification scheme, we apply it to the Part II data. The corresponding Kaplan-Meier curves for three strata are given in Figure 3. Based on the restricted area under the Kaplan-Meier curves derived from 1000 bootstrap samples, the point and 0.95 confidence interval estimates for the stratum-specific restricted mean survival times are 54.3 (51.0, 57.2), 58.9 (57.7, 60.0) and 62.0 (61.2, 62.8) months, for the three strata with $n = 245, 1350,$ and 2550 respectively.

6. REMARKS

A common practice in predictive medicine is to create an ordered category system to classify future subjects with their "baseline" information. A desirable quantitative stratification procedure would have both a small overall prediction error and a reasonable discriminatory capability across the strata. In this article, we provide a systematic approach to construct such a stratification rule. To achieve the first goal, we utilize a heuristically interpretable metric (a loss function based on the L_1 norm) for quantifying the prediction error. Moreover, the stratification requires a minimum size of the stratum to avoid having unstable small strata. The choice of the minimum size does not have a rigorous rule. It depends on the amount of "information" of the training set Part Ia, which is usually quantified by the sample size or the observed event rate. To enhance the discriminatory ability of the scheme, we set a minimum incremental value between two consecutive stratum-specific predict values at the model building stage. The choice of this value depends on clinical inputs. For example, for the cardiovascular study, the range of the RMST scores is from 51.0 to 63.8 (months) based on the Part I training data, which is relatively narrow. A choice of an incremental value of 3 months for illustration in Section 5 seems appropriate.

An obvious extension of the new proposal is to construct an optimal stratification procedure for treatment selections based on data either from randomized clinical trials or observational studies. Unfortunately, the L_1 loss function utilized in this article cannot be trivially generalized to deal with this important problem. Further research on the choice of a clinically meaningful metric for quantifying the prediction error for treatment selections is warranted.

APPENDIX A: THE DYNAMIC PROGRAMMING ALGORITHM FOR OPTIMAL STRATIFICATION

We will describe the dynamic programming algorithm for identifying the optimal grouping in this section. Below we first provide a brief introduction to the dynamic programming algorithm. To this end, assume that our objective is to find the minimum total cost of n stages:

$$\min_{\{a_t\}} \sum_{t=1}^{n} c_t(s_t, a_t),$$

where s_t is the state of stage t, $a_t \in A_t(s_t)$ is the action we take at stage t, and $c_t(s_t, a_t)$ is the cost associated with state s_t and action a_t. The state of the next stage s_{t+1} is determined by both s_t and a_t: $s_{t+1} = f_t(s_t, a_t)$, $t = 1, \ldots, n-1$. If we know that the minimum total cost from stage $m+1$ through stage n starting at state s_{m+1} is

$$C_{m+1}(s_{m+1}) = \min_{\{a_t\}} \sum_{t=m+1}^{n} c_t(s_t, a_t),$$

then the optimal cost from stage m starting at state s_m is simply

$$C_m(s_m) = \min_{a_m \in A_m(s_m)} \left\{ c_m(s_m, a_m) + C_{m+1}(f_m(s_m, a_m)) \right\}.$$

Thus we can start from the minimum cost $C_n(s_n) = \min_{a_n} c_n(s_n, a_n)$ at stage n to consecutively find the optimal solutions at stages $n-1, n-2, \cdots, 2$ and 1.

Our problem is more complicated than the formulation above due to the presence of constraints, but the basic principle remains the same. Without loss of generality, we assume that the data consists of $\{(Y_i, w_i, \hat{\mu}(V_i)), i = 1, 2, \cdots, n\}$, with $\hat{\mu}(V_1) < \hat{\mu}(V_2) < \cdots < \hat{\mu}(V_n)$. Here Y_i and w_i are response and associated nonnegative weight for the i^{th} observation. The objective is to group n observations into K strata: $S_k, k = 1, \cdots, K$, such that

$$\sum_{k=1}^{K} \sum_{i \in S_k} |Y_i - \bar{Y}(S_k)| w_i,$$

is minimized under the constraints that

$$n_k \geq np_0 \quad \text{and} \quad \bar{Y}(S_k) - \bar{Y}(S_{k-1}) \geq d,$$

where p_0 is the minimum stratum fraction, S_i denotes the set of observations in the i^{th} stratum: $S_1 = \{1, 2, \cdots, n_1\}$, $S_k = \left\{\sum_{j=1}^{(k-1)} n_j + 1, \sum_{j=1}^{(k-1)} n_j + 2, \cdots, \sum_{j=1}^{(k-1)} n_j + n_k\right\}, k = 2, \cdots, K$ and

$$\bar{Y}(S) = \frac{\sum_{i \in S} w_i Y_i}{\sum_{i \in S} w_i}, \text{ for } S \subset \{1, \cdots, n\}.$$

Here d and p_0 are given a priori but K is unknown. To this end, we consider the optimal grouping for the last m observations $\{n - m + 1, n - m + 2, \cdots, n\}$ with the first stratum S_{mj1} comprised of j observations, where $m \geq np_0$. That is, $S_{mj1} = \{n - m + 1, \cdots, n - m + n_1\}$, $S_{mjk} = \left\{n - m + \sum_{j=1}^{(k-1)} n_j + 1, n - m + \sum_{j=1}^{(k-1)} n_j + 2, \cdots, n - m + \sum_{j=1}^{(k-1)} n_j + n_k\right\}, k = 2, \cdots, K_m$ minimizes

$$\sum_{k=1}^{K_m} \sum_{i \in S_{mjk}} |Y_i - \bar{Y}(S_{mjk})| w_i,$$

under the constraints that $n_1 = j$,

$$n_{mjk} \geq np_0 \quad \text{and} \quad \bar{Y}(S_{mjk}) - \bar{Y}(S_{mj(k-1)}) \geq d.$$

Here $j = 1, 2, \cdots, m$. Let L_{mj} be the minimum L_1 loss for grouping the last m observations with j observations in the first stratum under the constraint above. Let the corresponding optimal grouping $S_{mj1}, \cdots, S_{mjK_m}$ be denoted by $\mathcal{G}_{mj}$. If there is no stratification satisfying the constraints, e.g., when $j < np_0$, then $L_{mj} = +\infty$. In such a case, we let $\mathcal{G}_{mj} = \phi$ for convenience in notations. Also, denote $\bar{Y}(S_{mj1})$ by $\bar{Y}_{mj}$.

Like the standard dynamic programming algorithm, we start from the last observation and $(\mathcal{G}_{11}, L_{11})$ can be obtained easily since $(\mathcal{G}_{11}, L_{11}) = (\{n\}, 0)$ if $1 \geq np_0$ and $(\phi, +\infty)$ otherwise. Assume that

for $1 \leq m < n$ we have obtained

$$\begin{array}{lll}
(\mathcal{G}_{11}, L_{11}, \bar{Y}_{11}) & & \\
(\mathcal{G}_{21}, L_{21}, \bar{Y}_{21}) & (\mathcal{G}_{22}, L_{22}, \bar{Y}_{22}) & \\
(\mathcal{G}_{31}, L_{31}, \bar{Y}_{31}) & (\mathcal{G}_{32}, L_{32}, \bar{Y}_{32}) & (\mathcal{G}_{33}, L_{33}, \bar{Y}_{33}) \\
\cdots & & \\
(\mathcal{G}_{m1}, L_{m1}, \bar{Y}_{m1}) & (\mathcal{G}_{m2}, L_{m2}, \bar{Y}_{m2}) & \cdots \qquad (\mathcal{G}_{mm}, L_{mm}, \bar{Y}_{mm}).
\end{array}$$

We can construct $\{\mathcal{G}_{(m+1)j}, L_{(m+1)j}\}$ based on the previous set of optimal solutions as follows. If $j < np_0$, then

$$(\mathcal{G}_{(m+1)j}, L_{(m+1)j}) = \{\phi, +\infty\}.$$

For $np_0 \leq j \leq m+1$, since the first stratum of size j is fixed, we should choose the optimal grouping strategy that minimizes the loss of the remaining $m+1-j$ observations. To examine the minimum incremental constraint between consecutive groups, we need and only need to consider the first two strata. Let i be the number of members of the second group, i.e., the group after the first j observations, we may define

$$i^* = \arg\min_i c_j(i), i = 1, \cdots, m+1-j$$

where

$$c_j(i) = \begin{cases} \sum_{i=n-m}^{n-m-1+j} |Y_i - \bar{Y}_{(m+1)j}| w_i + L_{(m+1-j)i} & \text{if } \bar{Y}_{(m+1)j} - \bar{Y}_{(m+1-j)i} \geq d \\ \infty & \text{if } \bar{Y}_{(m+1)j} - \bar{Y}_{(m+1-j)i} < d \end{cases}.$$

This step of finding i^* is not difficult since it involves only $O(m+1-j)$ summations. However, we can further simplify the computation by keeping the ranks of $\{L_{l1}, L_{l2}, \cdots, L_{ll}\}$ for all $l \leq m$. To identify i^*, we only need to examine the constraint of the grouping with the smallest $L_{(m+1-j)i}$: $\bar{Y}_{(m+1)j} - \bar{Y}_{(m+1-j)i} \geq d$. If the constraint is satisfied, then i^* is identified, otherwise we examine the constraint of the grouping with the second smallest $L_{(m+1-j)i}$ and et al. Normally, we can find i^* well before exhausting all $L_{(m+1-j)i}$. Once i^* is identified, $L_{(m+1)j} = c_j(i^*)$ and if $L_{(m+1)j} < \infty$, $\mathcal{G}_{(m+1)j} = \{S_{(m+1)j1}\} \cup \mathcal{G}_{(m+1-j)i^*}$. Therefore, one may construct $(\mathcal{G}_{(m+1)j}, L_{(m+1)j}), j =$

$1, 2, \cdots, m+1$ by tracking $(\mathcal{G}_{\tilde{m}\tilde{j}}, L_{\tilde{m}\tilde{j}}), 1 \le \tilde{m} \le m$ and $1 \le \tilde{j} \le \tilde{m}$, for $m = 1, 2, \cdots, n-1$. In the end, once $(\mathcal{G}_{nj}, L_{nj}), j = 1, \cdots, n$ are obtained, the optimal stratification is simply $\mathcal{G}_{nj^*}$, where $j^* = \arg\min_j L_{nj}, j = 1, 2, \cdots, n$.

The complexity of the algorithm is $O(n^3)$ and therefore the computation can be slow when n is big. In such a case, one may pre-group observations with similar $\hat{\mu}(V_i)$s together before applying the dynamic programming. One way to achieve this is to divide the interval containing all the estimated scores into subintervals and represent all the $\hat{\mu}(V_i)$s in the same subinterval by its center. In this way, we effectively reduce the choices of potential grouping while using the original Y_is and w_is to calculate the $\bar{Y}_k$ and prediction error. The computation speed can be substantially improved without sacrificing much precision in locating the optimal stratification scheme.

APPENDIX B: ASYMPTOTIC PROPERTIES FOR THE OPTIMAL STRATIFICATION SCHEME

We first assume that $\hat{\beta} - \beta_0 = o_p(1)$ for properly chosen λ, where β_0 belongs to a compact set, the parameter space of interest. Without loss of generality, we also assume that the score $\mu(V_i)$ is a continuous random variable with a bounded support and the joint density function of the continuous components of (Y_i, V_i) is continuously differentiable. Furthermore, we assume that the outcome Y_i is bounded. Let

$$L_n(\mathbf{c}) = n^{-1} \sum_{i=1}^{n} \left| Y_i - f(V_i|\mathbf{c}) \right|$$

and

$$L(\mathbf{c}) = \mathrm{E}\left| Y_i - f_0(V_i|\mathbf{c}) \right|,$$

where $\mathbf{c} = (-\infty = c_1 < c_2 < \cdots < c_K = \infty)'$, $\hat{\mu}(V_i) = g(\hat{\beta}' Z_i)$, $\mu(V_i) = g(\beta_0' Z_i)$,

$$f(V_i|\mathbf{c}) = \sum_{k=1}^{K} \hat{\mu}_Y(c_{k-1}, c_k) I(\hat{\mu}(V_i) \in (c_{k-1}, c_k]),$$

$$f_0(V_i|\mathbf{c}) = \sum_{k=1}^{K} \mu_Y(c_{k-1}, c_k) I(\mu(V_i) \in (c_{k-1}, c_k])$$

$$\hat{\mu}_Y(a, b) = \frac{n^{-1} \sum_{i=1}^{n} Y_i I(\hat{\mu}(V_i) \in (a, b])}{n^{-1} \sum_{i=1}^{n} I(\hat{\mu}(V_i) \in (a, b])} \quad \text{and} \quad \mu_Y(a, b) = \mathrm{E}(Y|\mu(V_i) \in (a, b]).$$

Firstly, we will show that

$$\sup_{\mathbf{c}} |L_n(\mathbf{c}) - L(\mathbf{c})| = o_p(1),$$

where the sup is over all $\mathbf{c}$ such that $\text{pr}(\mu(V_i) \in (c_{k-1}, c_k]) \geq \delta_0 > 0$. Since K is bounded and takes only finite number of possible values, it is sufficient to show the above uniform convergence for fixed any fixed K. To this end, we note that the coverage number $N_{[]}(\epsilon, \mathcal{F}, L_1) < \infty$ for the class of functions $\mathcal{F} = \{yI(g(\beta' z) \in (a, b]) \mid \max(|a|, |b|, \|\beta\|_1) < C_0\}$ or $\{I(g(\beta' z) \in (a, b]) \mid \max(|a|, |b|, \|\beta\|_1) < C_0\}$, where $C_0 < \infty$ is a constant. Thus it follows from the Glivenko-Cantelli theorem that

$$\sup_{\max\{|a|,|b|,\|\beta\|_1\}<C_0} \left| n^{-1} \sum_{i=1}^{n} Y_i I(g(\beta' Z_i) \in (a, b]) - \text{E}\left\{Y_i I(g(\beta' Z_i) \in (a, b])\right\} \right| = o_p(1)$$

and

$$\sup_{\max\{|a|,|b|,\|\beta\|_1\}<C_0} \left| n^{-1} \sum_{i=1}^{n} I(g(\beta' Z_i) \in (a, b]) - \text{pr}(g(\beta' Z_i) \in (a, b]) \right| = o_p(1),$$

which implies that

$$\sup_{(a,b,\beta)\in\Omega_0} \left| \frac{n^{-1} \sum_{i=1}^{n} Y_i I(\hat{g}(\beta' Z_i) \in (a, b])}{n^{-1} \sum_{i=1}^{n} I(\hat{g}(\beta' Z_i) \in (a, b])} - \text{E}(Y | g(\beta' Z_i) \in (a, b]) \right| = o_p(1), \tag{B.1}$$

where $\Omega_0 = \{a, b, \beta \mid \text{pr}(g(\beta' Z_i) \in (a, b]) \geq \delta_0, \max\{|a|, |b|, \|\beta\|_1\} < C_0\}$. Next, consider

$$U_n(a, b, \beta) = n^{-1} \sum_{i=1}^{n} I(g(\beta' Z_i) \in (a, b]) \left| Y_i - \frac{n^{-1} \sum_{i=1}^{n} Y_i I(g(\beta' Z_i) \in (a, b])}{n^{-1} \sum_{i=1}^{n} I(g(\beta' Z_i) \in (a, b])} \right|.$$

It follows from (B.1) that

$$\sup_{(a,b,\beta)\in\Omega_0} \left| U_n(a, b, \beta) - n^{-1} \sum_{i=1}^{n} I(g(\beta' Z_i) \in (a, b]) \big| Y_i - \text{E}(Y | g(\beta' Z_i) \in (a, b]) \big| \right| = o_p(1).$$

Now, consider the class of functions $\mathcal{F} = \{I(g(\beta' v) \in (a, b]) | y - \tilde{\mu}(a, b, \beta)| \mid (a, b, \beta) \in \Omega_0\}$, where $\tilde{\mu}(a, b, \beta)$ has continuous partial derivatives with respect to a, b and β. The covering number

of the class is finite as well, and it follows from the Glivenko-Cantelli theorem that

$$n^{-1}\sum_{i=1}^{n} I(g(\beta' Z_i) \in (a,b])\big|Y_i - \mathrm{E}(Y|g(\beta' Z_i) \in (a,b])\big|$$

uniformly converges to $u(a,b,\beta) = \mathrm{E}\left\{ I(g(\beta' Z_i) \in (a,b])\big|Y_i - \mathrm{E}(Y|g(\beta' Z_i) \in (a,b])\big|\right\}$ over the set Ω_0 and thus

$$\sup_{(a,b,\beta)\in\Omega_0} \Big| U_n(a,b,\beta) - u(a,b,\beta) \Big| = o_p(1).$$

Coupled with the fact that $u(a,b,\hat{\beta}) - u(a,b,\beta_0) = o_p(1)$, it suggests that

$$\sup_{(a,b,\beta)\in\Omega_0} \Big| U_n(a,b,\hat{\beta}) - u(a,b,\beta_0) \Big| = o_p(1).$$

Now, note the fact that

$$L_n(\mathbf{c}) = \sum_{k=1}^{K} U_n(c_{k-1}, c_k, \hat{\beta}) \quad \text{and} \quad L(\mathbf{c}) = \sum_{k=1}^{K} u(c_{k-1}, c_k, \beta_0),$$

we have

$$\sup_{\mathbf{c}} \Big| L_n(\mathbf{c}) - L(\mathbf{c}) \Big| = o_p(1).$$

Secondly, we will derive the upper bound of $L(\hat{\mathbf{c}})$ as $n \to \infty$. To this end, let the constraint be written as $\mathcal{S}_n(\mathbf{c}) \geq 0$, where

$$\mathcal{S}_n(\mathbf{c}) = \begin{pmatrix} \bar{Y}_2 - \bar{Y}_1 - d \\ \cdots \\ \bar{Y}_K - \bar{Y}_{K-1} - d \\ n^{-1}\sum_{i=1}^{n} I(c_1 \leq \hat{\mu}(V_i) \leq c_2) - p_0 \\ \cdots \\ n^{-1}\sum_{i=1}^{n} I(c_{K-1} \leq \hat{\mu}(V_i) \leq c_K) - p_0 \end{pmatrix}.$$

We also define the limiting constraint by $\mathcal{S}_0(\mathbf{c}) \geq 0$, where

$$\mathcal{S}_0(\mathbf{c}) = \begin{pmatrix} \bar{\mu}_2 - \bar{\mu}_1 - d \\ \cdots \\ \bar{\mu}_K - \bar{\mu}_{K-1} - d \\ \text{pr}(c_1 \le \mu(V_i) \le c_2) - p_0 \\ \cdots \\ \text{pr}(c_{K-1} \le \mu(V_i) \le c_K) - p_0 \end{pmatrix}.$$

Let $\mathbf{c}_0$ be the minimizer of $L(\mathbf{c})$ subject to the constraint $S_0(\mathbf{c}) \ge 0$ and $\hat{\mathbf{c}}$ be the minimizer of $L_n(\mathbf{c})$ subject to the constraint $\mathcal{S}_n(\mathbf{c}) \ge 0$. Furthermore, we let

$$\hat{\mathbf{c}}_\epsilon = \underset{\mathbf{c}:\mathcal{S}_0(\mathbf{c})\ge\epsilon}{\arg\min}\, L_n(\mathbf{c}) \quad \text{and} \quad \mathbf{c}_\epsilon = \underset{\mathbf{c}:\mathcal{S}_0(\mathbf{c})\ge\epsilon}{\arg\min}\, L(\mathbf{c}).$$

Under a rather mild condition that the numbers of strata of both stratification rules $\mathbf{c}_{\tilde{\epsilon}}$ and $\mathbf{c}_0$ are the same for some $\tilde{\epsilon} > 0$,

$$L(\mathbf{c}_\epsilon) \to L(\mathbf{c}_0) = L_0, \quad \text{as} \quad \epsilon \to 0.$$

A sufficient condition for the existence of such a $\tilde{\epsilon}$ is that the optimal grouping $\mathbf{c}_0$ is unique and the set $\{(c_1, \cdots, c_{K_0}) \mid \mathcal{S}_0(\mathbf{c}) \ge 0\}$ is not contained by a $K_0 - 1$ dimensional hyperplane in R^{K_0}, where $K_0 + 1$ is the dimension of the vector $\mathbf{c}_0$. Now, since $\mathcal{S}_n(\mathbf{c}) - S(\mathbf{c}) = o_p(1)$,

$$\text{pr}\Big[\{\mathbf{c} \mid \mathcal{S}_0(\mathbf{c}) \ge \varepsilon\} \subseteq \{\mathbf{c} \mid \mathcal{S}_n(\mathbf{c}) \ge 0\}\Big] \to 1, \quad \text{as } n \to \infty,$$

which implies that

$$\text{pr}\left\{L_n(\hat{\mathbf{c}}) \le L_n(\hat{\mathbf{c}}_\varepsilon)\right\} \to 1 \quad \text{as } n \to \infty.$$

Furthermore, by the definition of $\hat{\mathbf{c}}_\varepsilon$ which minimizes $L_n(\mathbf{c})$ under the constraint $\mathcal{S}_0(\mathbf{c}) \ge \epsilon$,

$$L_n(\hat{\mathbf{c}}_\varepsilon) \le L_n(\mathbf{c}_\varepsilon).$$

From the uniform convergence, for any $\delta > 0$,

$$\text{pr}\left\{L_n(\mathbf{c}_\varepsilon) > L(\mathbf{c}_\varepsilon) + \delta/2\right\} \to 0 \quad \text{and} \quad \text{pr}\left\{L(\hat{\mathbf{c}}) - \delta/2 > L_n(\hat{\mathbf{c}})\right\} \to 0 \quad \text{as } n \to \infty.$$

Therefore, for any $\delta > 0$, there exists an ε_0 such that $L(\mathbf{c}_{\varepsilon_0}) \leq L_0 + \delta$ and

$$
\begin{aligned}
&\text{pr}(L(\hat{\mathbf{c}}) \leq L_0 + 2\delta) \\
&\geq \text{pr}\left\{L(\hat{\mathbf{c}}) \leq L_n(\hat{\mathbf{c}}) + \delta/2 \leq L_n(\hat{\mathbf{c}}_{\varepsilon_0}) + \delta/2 \leq L_n(\mathbf{c}_{\varepsilon_0}) + \delta/2 \leq L(\mathbf{c}_{\varepsilon_0}) + \delta\right\} \\
&\geq 1 - \text{pr}\left\{L(\hat{\mathbf{c}}) > L_n(\hat{\mathbf{c}}) + \delta/2\right\} - \text{pr}\left\{L_n(\hat{\mathbf{c}}) > L_n(\hat{\mathbf{c}}_{\varepsilon_0})\right\} - \text{pr}\left\{L_n(\mathbf{c}_{\varepsilon_0}) > L(\mathbf{c}_{\varepsilon_0}) + \delta/2\right\} \to 1,
\end{aligned}
$$

as $n \to \infty$. It follows that the finite sample optimal stratification scheme minimizes the limit of the total of intra-stratum predicted error. The estimated stratification scheme approaches that of the optimal stratification scheme as the sample size goes to infinity.

REFERENCES

Braunwald, E., Domanski, M., Fowler, S., Geller, N., Gersh, B., Hsia, J., Pfeffer, M., Rice, M., Rosenberg, Y., and Rouleau, J. (2004), "Angiotensin-converting-enzyme inhibition in stable coronary artery disease.," *The New England journal of medicine*, 351(20), 2058–2068.

Breslow, N. E. (1972), "Discussion of Professor Cox's paper," *Journal of the Royal Statistical Society - Series B*, 34, 216–217.

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, 33(1), 1. http://www.stanford.edu/~hastie/Papers/glmnet.pdf.

Hammer, S. M., Squires, K. E., Hughes, M. D., Grimes, J. M., Demeter, L. M., Currier, J. S., Eron Jr, J. J., Feinberg, J. E., Balfour Jr, H. H., Deyton, L. R. et al. (1997), "A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less," *New England Journal of Medicine*, 337(11), 725–733.

Royston, P. (2009), "Explained variation for survival models," *Stata Journal*, 6(1), 83–96.

Royston, P., and Parmar, M. (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.

Solomon, S. D., Rice, M. M., Jablonski, K. A., Jose, P., Domanski, M., Sabatine, M., Gersh, B. J., Rouleau, J., Pfeffer, M. A., Braunwald, E. et al. (2006), "Renal function and effectiveness of angiotensin-converting enzyme inhibitor therapy in patients with chronic stable coronary disease in the Prevention of Events with ACE inhibition (PEACE) trial," *Circulation*, 114(1), 26–31.

Taha, A. H. (2003), *Operations research.* Pearson Education.

Tian, L., Zhao, L., and Wei, L. (2013), "On the Restricted Mean Event Time in Survival Analysis," *Harvard University Biostatistics Working Paper Series*, . http://biostats.bepress.com/harvardbiostat/paper156.

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

Xu, Q.-S., and Liang, Y.-Z. (2001), "Monte Carlo cross validation," *Chemometrics and Intelligent Laboratory Systems*, 56(1), 1–11.

Yong, F., Cai, T., Tian, L., and Wei, L. (2013), "Making valid inferences for prediction of survival via Cox's working models with baseline covariates," *Klein, J., van Houwelingen, H., Ibrahim, J. and T.H., S. (2013), Handbook of Survial Analysis*, pp. 265–283; as Chapter 13. Classical Model Selection.

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.

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.

TABLES and FIGURES

Table 1: Regression model candidates, $\mathrm{E}(Y|V) = \beta' Z$, for study ACTG 320, $\bar{\mathcal{L}}^*$ and the complexities of $\mathcal{M}^*$ ($\|\beta\|_0 =$ the number of nonzero components of $\hat{\beta}$).

Model	Candidate independent variables	dim(Z)	$\bar{\mathcal{L}}^*$	$\mathcal{M}^*$	
				# covariates	$\|\hat{\beta}\|_0$
1	age, sex, CD4 count and $\log_{10}$RNA at baseline and week 4	6	0.415	5	5
2	all baseline covariates plus their first-order interaction terms	78	0.465	12	14
3	all baseline covariates and CD4 and $\log_{10}$RNA at week 4 plus their first-order interaction terms	105	0.414	13	20
4	none	0	0.484	0	0

Table 2: Regression model candidates, $\log\{-\log S(t|V)\} = \log\{-\log S_0(t)\} + \beta' Z$, for study PEACE, $\bar{\mathcal{L}}^*$ and the complexities of $\mathcal{M}^*$ ($\|\beta\|_0 =$ the number of nonzero components of $\hat{\beta}$).

Model	Candidate independent variables	dim(Z)	$\bar{\mathcal{L}}^*$	$\mathcal{M}^*$	
				# covariates	$\|\hat{\beta}\|_0$
1	age, gender, left ventricular ejection fraction, history of myocardial infarction, history of hypertension, history of diabetes, eGFR, ACE inhibitor treatment	10	16.919	6	7
2	variables in Model 1 plus three treatment and eGFR interaction terms	13	16.903	6	9
3	variables in Model 1 plus their first-order interaction terms	55	16.966	6	9
4	none	0	18.649	0	0

Figure 1: Stratum-specific point and 95% confidence intervals for the response rates with the Part II data (denoted by dots) of ACTG 320 with cutoff points $\hat{c}_1 = 0.25$ and $\hat{c}_2 = 0.45$.

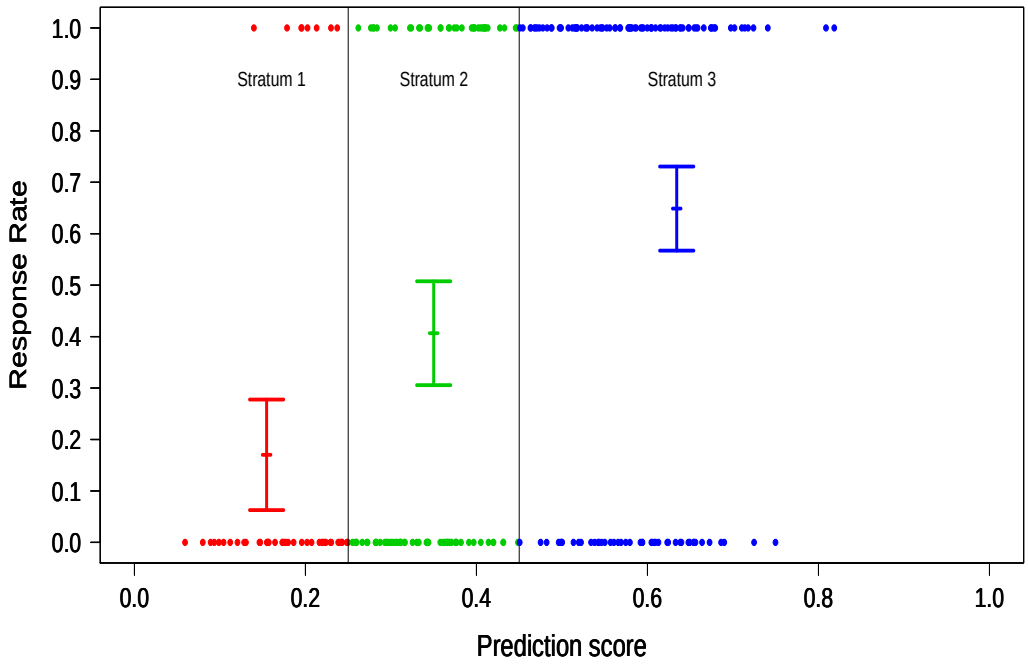

Figure 2: The Kaplan-Meier estimate for the time to the composite endpoint with the entire PEACE dataset.

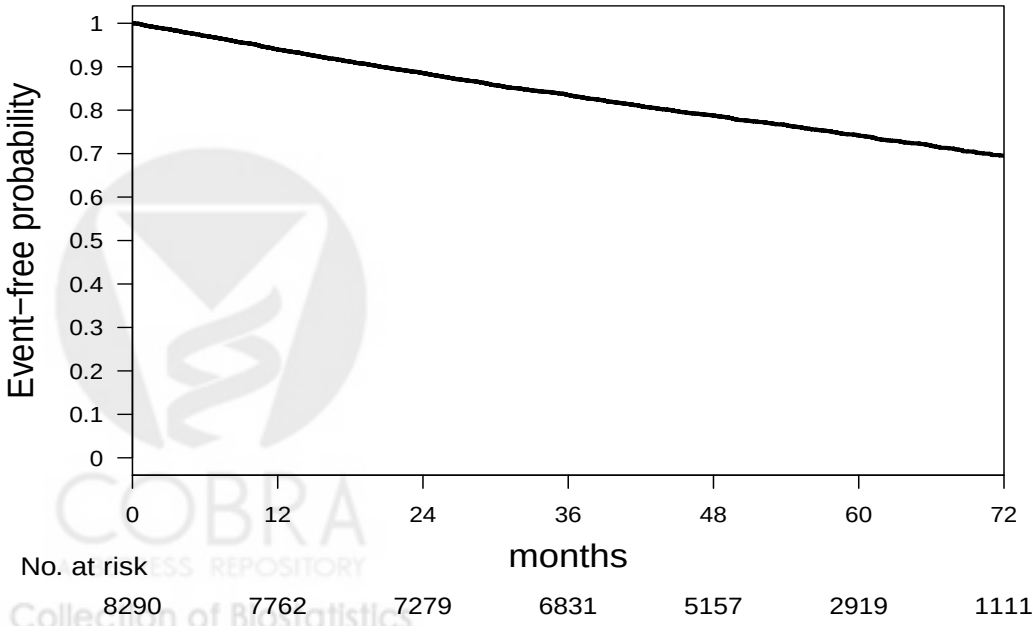

Figure 3: Stratum-specific Kaplan-Meier estimates and 95% confidence intervals for RMSTs obtained from Part II data of PEACE study with $\hat{c}_1 = 56.5$ and $\hat{c}_2 = 60.5$.