Quantitative Methods for Stratified Medicine

Quantitative Methods for Stratified Medicine

A dissertation presented

by

Florence Hiu-Ling Yong

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

April 2015

Dissertation Advisor: Professor Lee-Jen Wei Florence Hiu-Ling Yong

Quantitative Methods for Stratified Medicine

Abstract

Stratified medicine has tremendous potential to deliver more effective therapeutic intervention to improve public health. For practical implementation, reliable prediction models and clinically meaningful categorization of some comprehensible summary measures of individual treatment effect are vital elements to aid the decision-making process and bring stratified medicine to fruitful realization. We tackle the quantitative issues involved from three fronts : 1) prediction model building and selection; 2) reproducibility assessment; and 3) stratification. First, we propose a systematic model development strategy that integrates cross-validation and predictive accuracy measures in the prediction model building and selection process. Valid inference is made possible via internal holdout sample or external data evaluation to enhance generalizability of the selected prediction model. Second, we employ parametric or semi-parametric modeling to derive individual treatment effect scoring systems. We introduce a stratification algorithm with constrained optimization by utilizing dynamic programming and supervised-learning techniques to group patients into different actionable categories. We integrate the stratification and newly proposed prediction performance metric into the model development process. The methodologies are first presented in single treatment case, and then extended to two treatment cases. Finally, adapting the concept of uplift modeling, we provide a framework to identify the subgroup(s) with the most beneficial prospect; wasteful, harmful, and futile subgroups to save resources and reduce unnecessary exposure to treatment adverse effects. The contribution of this dissertation is to provide an operational framework to bridge predictive modeling and decision making for more practical applications in stratified medicine.

Contents

List of Figures

List of Tables

To my husband Victor Lo, our son Joshua
my parents
and
the loving memory of
my sister Virginia Yong

Acknowledgments

I would like to first thank my advisor, Lee-Jen Wei, who has not only served as a great mentor but also supported me with his enormous kindness and wisdom. His words and patience have brought me through some very challenging times. Lu Tian, my committee members Tianxi Cai and Sebastian Schneeweiss, and other collaborators have contributed their knowledge and provided immensely helpful advice on the dissertation. Furthermore, I would like to thank my friends, ex-colleagues, and mentors from the Department of Statistics at the University of Hong Kong, the Departments of Biostatistics at UCLA and Harvard University, the Center for Biostatistics in AIDS Research, and the Genzyme Corporation. The wonderful learning and working experience there have greatly helped my personal growth and shaped my passion in biostatistical research and applications. Finally, I would like to thank my husband Victor Lo, and without his loving support, this journey would have been impossible. Special thanks to my parents, my son, my church, and my friends who have supported me throughout the years especially during my ordeals. I am deeply grateful for all their encouragement and love.

1. Making valid inferences for prediction of survival via Cox's working models with baseline covariates

Florence H. YONG[1], Tianxi CAI[1], Lu TIAN[2], and L.J. WEI[1]

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Health Research and Policy, Stanford University School of Medicine

Abstract

For a longitudinal study with time to event as the endpoint, an important objective is to make prediction of the endpoint distribution, particularly for future population. The Cox regression model is commonly used for such event time analysis. Conventionally, the same dataset is used for model building, model selection and inference. It is unclear how generalizable the results are, or how we can make valid inference about predicting survival with patients' baseline information with such a potentially over-optimistic approach. A more appropriate prediction inference procedure can be constructed using two independent datasets under similar study settings. The first set is used to conduct model building and selection while the second one is used to make inference for prediction with the final selected model. In this article, using the well-known Mayo liver study data for illustration, we show that the conventional practice may produce an overly optimistic model. We further propose a systematic, 3-step-in-1 dataset modeling strategy utilizing cross-validation techniques, predictive performance measures, and holdout samples to efficently obtain valid and more generalizable prediction.

KEY WORDS: Cox's proportional hazard model; C-statistics; Cross-validation; Generalizability; Holdout sample; 3-in-1 dataset modeling strategy; Model selection; Overfitting; Prediction accuracy measures; Survival analysis

1.1 Introduction

There are two major goals for conducting regression analysis: examining the association of the outcome variable and the covariates, and making prediction for future outcomes. The choice of the process for model building, selection and validation depends on the aim of the investigation. Generally it is difficult, if not impossible, that the selected model would be the correct one. On the other hand, a good approximation to the true model can be quite useful for making prediction. Model building and selection should not be a stand-alone procedure, we need valid inference for prediction with the final model.

To establish a prediction model, the same observations in a dataset are often used to build, select, and conduct inference. This traditional practice can lead to quite overly optimistic and unreliable model at the end. To tackle this issue, we present a model development strategy based on the well-known "machine learning" concept with additional inferential component. We recommend splitting the dataset randomly into two pieces. The first part is used for model building and selection via conventional cross-validation techniques; and the second part, often called the "holdout sample," is used for statistical inference based on the final selected model. Although conceptually this strategy is applicable to the general regression problem, our focus is on censored event time outcome variable.

The Cox proportional hazards (PH) model (Cox, 1972) is the most widely used model in analyzing event time data. Its statistical procedures for making inferences about the association between the event time and covariates are well developed and theoretically justified via martingale theory (Andersen and Gill, 1982). For prediction using Cox's model, van Houwelingen and Putter (2008) provide an excellent review of various prediction methods in clinical settings. Other recent development in this area involves high-dimensional data in which the number of observations is much smaller than the number of variables (van Houwelingen et al., 2006; Witten and Tibshirani, 2010). For typical study analyses, a vast majority utilizes all observations in the same dataset for model building and validation, despite a growing concern of false positive findings (Ioannidis, 2005; Simmons et al., 2011).

The goal of our investigation is to establish a Cox prediction model and draw reliable inferences using such a model. We discuss in detail the model-building strategies from a prediction point of view. We will use a well-known dataset from a Mayo Clinic Primary Biliary Cirrhosis (PBC) study (Fleming and Harrington, 1991; Therneau and Grambsch, 2000) to guide us through each step of the process for conducting model building, selection and inference.

The article is organized as follows. In Section 1.2, we describe our study example. Section 1.3 summarizes various model building and selection procedures in the literature for the Cox model. Model evaluation based on predictive accuracy measures such as a censoring-adjusted C-statistic is introduced, which can be used to identify the optimal method among all candidate models to develop a final Cox model. Section 1.4 applies five candidate model selection methods to the PBC dataset to demonstrate the conventional model building, selection, and inference procedure. Section 1.5 presents some challenges using the conventional process. It shows that the conclusion of such inference can be quite misleading due to using an overly optimistic model building procedure. A prediction model development strategy that integrates cross-validation in the model building and validation, utilizes predictive measure to help identify the optimal model selection method, and conducts valid prediction inference on a holdout dataset is proposed. This 3-in-1-dataset modeling procedure is illustrated in detail with the PBC data in Section 1.6. We conclude with discussion of potential issues and interesting research problems on model selection in the Remarks section.

1.2 Mayo Clinic primary biliary cirrhosis (PBC) data

The Mayo clinical trial in PBC of the liver has been a benchmark dataset for illustration and comparison of different methodologies used in the analysis of event time outcome study (Fleming and Harrington, 1991; Therneau and Grambsch, 2000). The trial was conducted between January 1974 and May 1984 to evaluate the drug D-penicillamine versus placebo with respect to survival outcome. There were 424 patients who met the eligibility criteria for the trial, in which 312 cases participated in the double-blinded randomized placebo controlled trial and contained mostly complete information on the baseline covariates. Six patients were lost to followup soon after their initial clinic visit and excluded from our study. The rest of the 106 cases did not participate in the randomized trial but were followed for survival with some basic measurements recorded.

Since there was no treatment difference with respect to the survival distributions at the end of study, the study investigators combined the data from the two treatment groups to establish models for predicting survival. In this article, we utilized data on all 418 patients to establish a prediction model for the patient's survival given their baseline covariates. The average follow-up time of these 418 patients was 5.25 years. Like other studies, there were missing covariate values among the patients ranging from 2 patients missing prothrombin time (protime) to 136 patients missing triglyceride levels. For illustration, we imputed the missing values with their group sample median.

The outcome variable is the time to death ($time_i$). Censoring variable ($death_i$) for each case i has value 1 if the death date is available, or value 0 otherwise. The patient's baseline information consists of

- Demographic attributes: age in years, sex
- Clinical aspects: ascites (presence/absence), hepatomegaly (presence/absence), spiders (blood vessel malformations in the skin, presence/absence), edema (0 no edema and no diuretic therapy for edema, 0.5 edema untreated or successfully treated, 1 edema despite diuretic therapy)
- Biochemical aspects: serum bilirubin (mg/dl), albumin (g/dl), urine copper (μg/day), prothrombin time (standardised blood clotting time in seconds), platelet count (number of platelets $\times 10^{-3}$ per mL^3), alkaline phosphotase (U/liter), ast (aspartate aminotransferase, once called SGOT (U/ml)), serum cholesterol (mg/dl), and triglyceride levels (mg/dl)
- Histologic stage of disease.

We applied logarithmic transformations to albumin, bilirubin, and protime in the process of model building, based on analyses of this dataset in Fleming and Harrington (1991).

To establish a prediction model, ideally one should have three similar but independent datasets, or split the dataset randomly into three subsets. Using the observations from the first subset, we fit the data with all model candidates of interest; using the data from the second piece, we evaluate those fitted models with intuitively interpretable, model-free criteria and choose a final model; and using the data from the third piece, we draw inferences about the selected model. In practice, if the sample size is not large, we may combine the first two steps with a cross-validation procedure.

We will use the PBC dataset to illustrate this model selection strategy in Section 1.6. First we review some classical algorithms for model selection and introduce some model-free, heuristically interpretable criteria for model evaluation.

1.3 Model building procedures and evaluation

Depending on the study question and subject matter knowledge, we may identify a set of potential explanatory variables which could be associated with the survival outcome in a Cox PH model, the hazard function at time t for an individual is:

$$\lambda(t|Z) = \lambda_0(t)\text{exp}(\beta' Z),$$

where $\lambda_0(t)$ is an unknown baseline hazard function, $Z = (z_1, z_2, \ldots, z_p)'$ is the vector of explanatory variables of the individual, and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a $p \times 1$ vector of coefficients of the explanatory variables $Z_1, Z_2, \ldots, Z_p$. We estimate the parameter β by maximizing the partial likelihood:

$$L(\beta) = \prod_{r \in D} \frac{\text{exp}(\beta' Z^{k_r})}{\sum_{k \in R_r} \text{exp}(\beta' Z^{k})},$$

where D is the set of indices of the failures, R_r is the set of indices of subjects at risk at time t_r, and k_r is the index of the failure at time t_r.

1.3.1 Variable selection methods

A classical variable selection method is the stepwise regression using $L(\beta)$ as the objective function and p-value as a criterion for inclusion or deletion of covariates. It combines forward selection and backward elimination methods, allowing variables to be added or dropped at various steps according to different pre-specified p-values for entry to or stay in the model. Variations of stepwise regression method have been proposed. For example, forward stepwise regression starts from a null model with intercept only, while backward stepwise regression starts from a full model. We use forward stepwise procedure, as backward stepwise selection may be more prone to the issues of collinearity.

To reduce overfitting (Harrell (2001); Section 1.5.1), we may introduce a penalty of complexity of the candidate models for the stepwise procedures using Akaike information criterion (AIC; Akaike (1974)) or Bayesian information criterion (BIC; Schwarz (1978)). Both AIC and BIC penalizes degrees of freedom (k) which is the number of nonzero covariates in regression setting, and their objective functions are:

AIC = $-2 * L(\beta) + 2 * k$; and

BIC = $-2 * L(\beta) +$ log(No. of Events)$* k$. The AIC's penalty for model complexity is less than that of BIC's. Hence, it may sometimes over-select covariates in order to describe the data more adequately; whereas BIC penalizes more and may under-select covariates (Acquah and Carlo, 2010). Note that we usually follow the principle of hierarchical models when building a model, in which interactions are included only when all the corresponding main effects are also included; however, this can be relaxed (Collett, 2003).

We can also select a model based on the maximization of a penalized partial likelihood (Verweij and Van Houwelingen, 2006) with different penalty functions including L_2 penalty, smoothing splines, and frailty models, which are studied extensively in the literature. Two commonly used methods are Lasso (Least Absolute Shrinkage and Selection Operator) selection (Tibshirani, 1996) and Ridge regression methods (Van Houwelingen, 2001).

Lasso Selection

Tibshirani (1996) proposed the Lasso variable selection procedures which was extended to the Cox model (Tibshirani, 1997). Instead of estimating β in the Cox model through maximization of the partial likelihood, we can find the β that minimizes the objective function {-log$L(\beta) + \lambda_1||\beta||_1$} (Park and Hastie, 2007), where

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\{\text{-log}L(\beta) + \lambda_1||\beta||_1\}.$$

Lasso imposes an L_1 absolute value penalty, $\lambda_1||\beta||_1 = \lambda_1 \sum_{j=1}^{p} |\beta_j|$ to log $L(\beta)$, with $\lambda_1 \geq 0$. It does both continuous shrinkage and automatic variable selection simultaneously. Notice that Lasso penalizes all $\beta_j (j = 1, \ldots, p)$ the same way, and can be unstable with highly correlated predictors, which is common in high-dimensional data settings (Grave et al., 2011).

Different methods such as path following algorithm (Park and Hastie, 2007), coordinate descending algorithm (Wu and Lange, 2008), and gradient ascent optimization (Goeman, 2009) can be used to select variables and estimate the coefficients in Lasso models. Instead of using cross-validation to select the tuning parameters, we will consistently apply AIC or BIC to select models across various classical model selection methods for illustration.

Ridge regression

The Ridge penalty is a L_2 quadratic function, $\lambda_2||\beta||_2^2 = \lambda_2 \sum_{j=1}^{p} {\beta_j}^2$ in a general penalized regression setting. It achieves better prediction performance through a bias-variance trade-off. Of note, this method always keeps all predictors in the model and hence cannot produce a parsimonious model.

There are other penalized regression methods such as elastic net (Zou and Hastie, 2005) which combines both L_1 and L_2 penalty, the smoothly absolute clipped deviation (SCAD) penalty (Fan and Li, 2001, 2002), and various modification of the Lasso procedures. Other variable selection procedures and different combinations of model selection methods and algorithms to select the tuning parameter(s) have also been developed.

1.3.2 Model evaluation based on prediction capability

Many evaluation criteria can be used to select a model; however, if some covariates are difficult to obtain due to cost or invasiveness, a heuristically interpretable criterion is more informative than a purely mathematical one. Since it is desirable to examine the predictive adequacy of the Cox model for the entire study period, one of such criteria is the C(Concordance)-statistic (Pencina and D'Agostino, 2004).

C-statistics

To select a model with best predictive capability, C(Concordance)-statistics are routinely used to evaluate the discrimination ability and quantify the predictability of working models. Good predictions distinguish subjects with the event outcome from those without the outcome accurately and differentiate long-term survivors from the short-lived in survival context. The traditional C-statistic is a rank-order statistic for predictions against true outcomes (Harrell, 2001), and it has been generalized to quantify the capacity of the estimated risk score in discriminating subjects with different event times. Various forms of C-statistics are proposed in literature to provide a global assessment of a fitted survival model for the continuous event time. However, most of the C-statistics may depend on the study-specific censoring distribution.

Uno et al. (2011) proposed an unbiased estimation procedure to compute a modified C-statistic (C_τ) over a time interval $(0, \tau)$, which also has the same interpretation as Harrell's C-statistic for survival data, except that Uno's method is censoring-independent, and is given by (Uno et al., 2011) equations (5) and (6). This censoring-adjusted C-statistic is based on inverse-probability-of-censoring weights, which does not require a specific working model to be valid. The procedure is valid for both type I censoring without staggered entry, and random censoring independent of survival times and covariates (other conventional C-statistics may not be valid in this situation). A simulation study reported in Uno et al. (2011) did not find the procedure to be sensitive to violation of the covariate independent censoring assumption.

van Houwelingen and Putter (2008) and Steyerberg et al. (2010) provide a very helpful discussion of other assessments of predictive performance such as Brier score (Graf et al., 1999; Gerds and Schumacher, 2006). We show, as an example, the model-free, more recently developed censoring-adjusted C-statistic to evaluate the overall adequacy of the predictive model.

1.4 Application of conventional model development and inferences

The goal of this section is twofold: 1) to show the conventional way of analyzing time-to-event data, using the Mayo clinic PBC dataset described in Section 1.2; and 2) to present some challenges and limitations, which lead us to propose an alternative strategy for selecting a model among several candidate model selection procedures and establishing a more reliable prediction model.

1.4.1 Model Building

We apply five classical model selection algorithms to the PBC dataset for illustration. These candidate methods are: forward selection, backward elimination, stepwise regression, Lasso, and Ridge regression. For each of these methods, we build a model using AIC and BIC as model tuning criteria, respectively. For the Lasso and Ridge regression method, AIC (or BIC) as a function of the regularization parameter λ is plotted and evaluated to find the global minimum. Models are fitted using the λ at which the least AIC (or BIC) is achieved. The results are shown in Table 1.1, which consists of two parts. The first part summarizes all the resulting models via the aforementioned model building processes. The second part of the table summarizes how we obtained these models.

As a reference, all but two covariates (sex and alk.phos) contributed to a "significant" increase or decrease in risk ratio in univariate analysis ($p < 0.005$). Numerous studies in the literature have used Cox models to identify prognostic factors on event outcome. As shown in Table 1.1, the risk ratio estimates for each risk factor of interest can be very sensitive to what other covariates are put in the same model for evaluation.

1.4.2 Selecting procedure using C-statistics

Using the entire PBC dataset, Table 1.2 summarizes the censoring-adjusted C-statistics of the eight models presented in Table 1.1. The higher the measure, the better the model predicts throughout the course of study.

The two penalized regression methods yield slightly higher C-statistics. However, models derived from these two methods use more variables than the classical methods. The best single variable model, M4, has the lowest C-statistic. The predictive measures of M1, M2 and M3 are close to the models derived from the two penalized regression methods while using fewer variables. Inference for the difference in C between the models shows a difference between M1 and M4, using the method proposed by Uno et al. (2011). M1 appears as the most parsimonious model with reasonably good C-statistic of 0.790 among all these models.

Table 1.1: Models derived from various classical model selection methods, using entire PBC data set, hazard ratios $\exp(\hat{\beta})$ are presented.

Covariates	M1	M2	M3	M4	Lasso		Ridge	
					AIC	BIC	AIC	BIC
logbili	2.334	2.372	2.279	2.688	2.213	2.137	2.016	1.651
edema	2.238	2.110	2.107		2.022	1.996	2.182	2.099
age	1.034	1.032	1.034		1.031	1.027	1.029	1.023
stage	1.386	1.394	1.412		1.366	1.326	1.369	1.284
lalb	0.120	0.119	0.128		0.168	0.180	0.166	0.199
lptime	8.164	7.267	8.004		6.535	5.513	7.715	6.898
ast			1.002		1.002	1.001	1.002	1.002
copper		1.002	1.001		1.001	1.001	1.001	1.002
ascites					1.320	1.291	1.407	1.498
trig		0.998	0.998		0.998	0.999	0.998	0.999
hepato					1.049		1.158	1.223
spiders							0.947	1.044
sex							1.057	1.063
chol							1.000	1.000
alk.phos							1.000	1.000
platelet							1.000	1.000

Note: All covariates were treated as continuous effects.

Model	Selection Method
M1	Several model building procedures using BIC as stopping criterion came up with this same model: a. Forward selection, BIC; b. Backward elimination, BIC; c. Stepwise, BIC
M2	Backward elimination, AIC
M3	a. Forward selection, AIC; b. Stepwise, AIC
M4	Best single variable model, logbili (log(bilirubin)) is the most significant variable ($p < .00001$)

1.4.3 Making statistical inferences for the selected model

Conventionally, once we find a desirable model, the risk score for this model can be estimated and used to differentiate the risk of the subjects in the cohort. These risk scores can be ranked to put subjects into different risk categories such as tertiles (or deciles if there are more data). We choose M1, and the Ridge BIC model which has the highest C-statistic in Table 1.2 for demonstration.

Table 1.2: C-statistic of models using the full PBC dataset.

Model Selection Method	Model Size	C-Statistic
M4	1	0.748
M2	8	0.784
M1	6	0.790
M3	9	0.791
Lasso, AIC	11	0.794
Lasso, BIC	10	0.794
Ridge, AIC	16	0.796
Ridge, BIC	16	0.799

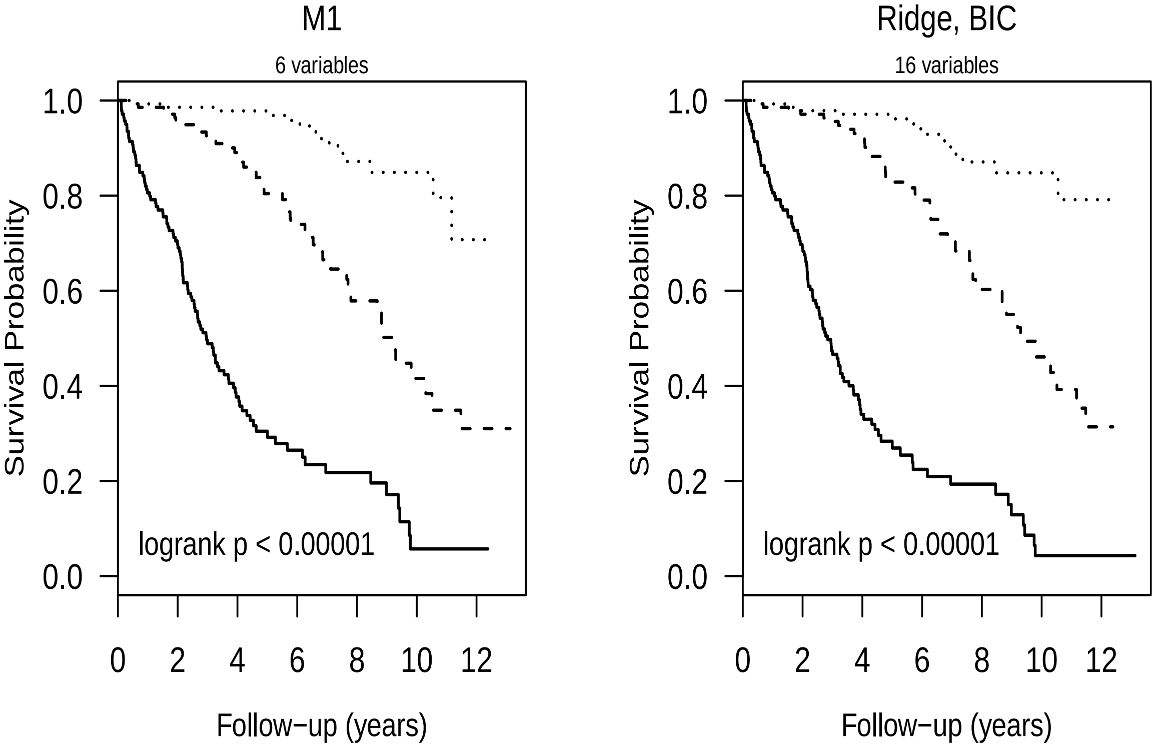


Figure 1.1: Kaplan-Meier curves of the survival time, stratified by tertiles of risk scores from two models: M1 - Six-variable model (left panel), and Ridge, BIC model (right panel).

Table 1.3 presents the summary statistics of the difference in survival distributions depicted in Figure 1.1. The restricted mean survival time is computed as the area under the KM survival curve, over the range from $[0, t_{max}]$, where t_{max} (= 12.5 years) is the maximum time for all KM curves considered and serves as a common upper limit for the restricted mean calculation. The overall logrank test, and the logrank tests for the difference in survival distributions between any two risk categories all yield p-values < 0.00001. Both M1 and Ridge, BIC models produce similar results with little difference in C-statistics, this further illustrates that M1 model is most preferable because it only takes 6 variables to achieve similar predictability.

Table 1.3: Summary statistics of the survival distributions by risk categories, scoring using the entire dataset.

Model Selection Method	Risk Categories	N Events/ Total	Restricted Mean (se) in years	Median (years)	(95% CI)
Stepwise, BIC	Low	14/140	11.31 (0.283)	NA	(NA, NA)
	Medium	45/139	8.66 (0.393)	9.19	(7.70, 11.47)
	High	102/139	4.21 (0.340)	2.97	(2.55, 3.71)
Ridge, BIC	Low	14/140	11.36 (0.278)	NA	(NA, NA)
	Medium	42/139	8.98 (0.381)	9.30	(7.79, NA)
	High	105/139	3.98 (0.320)	2.84	(2.44, 3.55)

1.5 Challenges and a proposal

The aforementioned process of using the same dataset for model building, selection, and inference has been utilized in practice. This conventional process has potential of self-serving problem. In this section, we first summarize the reasons for the issue of overfitting, then we will use the PBC dataset to demonstrate such an overfitting problem with this conventional process.

1.5.1 Overfitting Issue

If we use the same dataset to construct a prediction rule and evaluate how well this rule predicts, the predictive performance can be overstated. Overfitting occurs when a model describes the random variation of the observed data instead of the underlying relationship. Generally an overfit model indicates better fit and smaller prediction error than in reality because the model can be exceedingly complex to accommodate minor random fluctuation in observed data. This leads to the issue of over-optimism. We will demonstrate some covariates can be selected as statistically significant risk predictors of an event outcome even though there is no underlying relationship between them.

1.5.2 Noise variables become significant risk factors

Using the PBC dataset, we first randomly permutate the 418 survival time observations to break the ties between the observed or censored survival time and its covariate vector Z. Then we apply various traditional methods including forward selection, backward elimination, and stepwise regression to fit the data with newly permuted y' and 16 original covariates. These two steps are repeated 5,000 times.

Table 1.4 shows the median, interquartile range, and the range of the number of variables selected in 5,000 simulations. Using AIC as tuning criterion tends to over-select variables to achieve better model fit, while BIC tends to select fewer variables. Stepwise regression with BIC picked up at least one variable 25% of the time.

Table 1.4: Summary statistics of the number of variables selected in 5,000 runs.

Selection Procedure	Tuning Criterion	Median	(1^{st}, 3^{rd}) Quartile	(Min, Max)
Forward Selection	AIC	2	(1, 3)	(0, 9)
	BIC	0	(0, 1)	(0, 3)
Backward Elimination	AIC	3	(2, 4)	(0, 11)
	BIC	0	(0, 1)	(0, 6)
Stepwise	AIC	2	(1, 3)	(0, 10)
	BIC	0	(0, 1)	(0, 4)

Consider one such realization, the risk ratio of the variable log(protime) is 11.5 (se=0.845, p=.0039). Using this single variable model to score the entire dataset, and stratify the risk scores into two strata, we have the left panel of the Kaplan-Meier (KM) plot in Figure 1.2. It appears that this model is a reasonable prediction tool for survival.

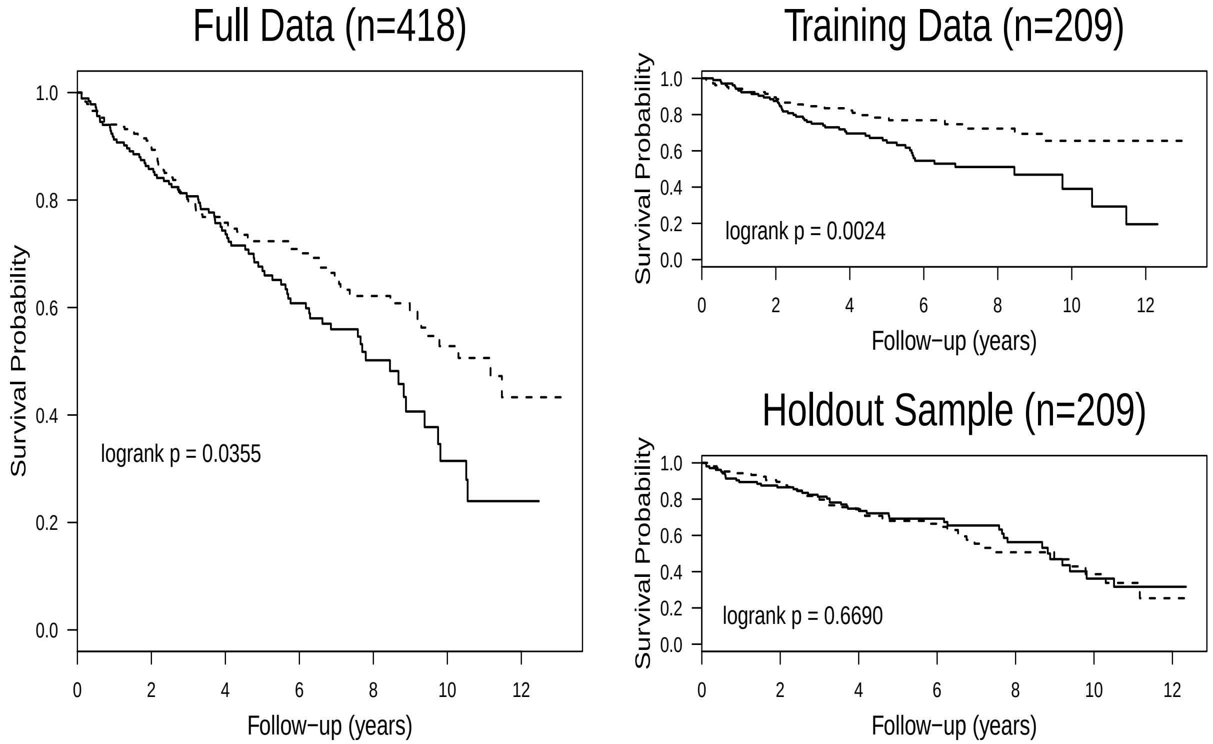


Figure 1.2: An example of using a holdout sample to show the overfitting phenomena, using the entire dataset.

If we randomly split the data into two parts, using the first half (called training data) to fit the model using Stepwise BIC approach, the upper part of the right panel shows some separation again; one may also pause here had we just given the training dataset. However, if we go one step further, using the training model to score the holdout sample (the other half) to evaluate the generalizability of the model, the lower right KM plot shows no separation at all.

1.5.3 Utilizing cross-validation in model selection process

One way to address overfitting is to use cross-validation (CV) techniques. It is preferred to evaluate the prediction error with independent data (validation data) separated from the data used for model building (training data). Two ways of conducting cross-validation are:

1. **K-fold cross-validation**

   - Randomly partition the entire original dataset into K groups
   - For $k = 1, \cdots, K$, do the following:
     - Retain a single k^{th} group as validation data
     - Use the rest $(K - 1)$ groups as training data to estimate β and form prediction rule
     - Evaluate a predictive performance measure (e.g., C-statistic) using the validation data
   - Compute the average of the K predictive performance measures

   All data are used for both training and validation, and each observation is used for validation exactly once.

2. **Monte-Carlo cross-validation**

   - Randomly subsample p percent of the entire dataset without replacement and retain it as validation data
   - Use the rest (1-p percent) data as training data to form prediction rule
   - Evaluate the prediction model using the validation data
   - Repeat the above steps M times
   - Average the M estimates to obtain the final estimate of the predictive accuracy measure

   In this schema, the results may vary if the analysis is repeated with different random splits; some observations may be selected more than once for training, while others may not be selected for validation. However, these can be resolved by increasing M, the number of times the CV is repeated.

CV is especially useful when we do not have enough observations to set aside for test set validation; it may address the overfitting issue.

1.5.4 3-in-1 dataset modeling proposal

We present the following strategy to help us select a model with best predictive accuracy, utilizing CV in model selection process. First we randomly partition a given dataset into two parts, for example, $D_{train.val}$ has 50% of the data and $D_{holdout}$ has the rest.

1. Model Building:

   For each candidate model selection method considered, use dataset $D_{train.val}$ to build a model and apply cross-validation to find the predictive accuracy measures of interest.

2. Model Selection:

   Identify the "optimal" model selection method(s) that gives us the most acceptable or highest predictive accuracy measures with a reasonable model size in Step 1. For the final model, we can either

   (a) Use the "optimal" method to refit the dataset $D_{train.val}$ to obtain the prediction equation for each subject; or

   (b) Apply the average model obtained from the training data portion of $D_{train.val}$ to $D_{holdout}$ to report how good it is, and future data for application (e.g., identifying future study population for intervention).

   While the first approach can provide a simple scoring system (a linear combination of selected covariates in this example), the second approach as the "bagging" version (Breiman, 1996) may have superior performance to the first one for "discrete" procedure such as stepwise regression and Lasso.

3. Statistical Inference:

   Dataset $D_{holdout}$ is not involved in any training process; therefore, it is best suited for testing and reporting the predictive accuracy of the final model derived from $D_{train.val}$ using the optimal model selection method identified in Step 2 without overfitting issue and biases. Additionally, using scores obtained from numerous training models developed during the cross-validation procedure in Step 1, model averaging can be applied to increase predictive accuracy in the holdout sample.

We now use the PBC data to illustrate this proposal.

1.6 Establishing a prediction model

Conventional model building strategies using the entire dataset without external data validation may have limited application. For any given dataset, it will be ideal to be able to

(1) develop a predictive model with validation, and

(2) report how well the model performs externally.

Hereafter, we apply our proposal to the PBC dataset using Monte Carlo cross-validation to illustrate the idea.

1. First, retain a random 50% sample of data from the randomized trial portion of the dataset and another 50% of the follow-up portion of the data. This holdout dataset $D_{holdout}$ consists of 209 observations and will be used to conduct inference, and examine the generalizability of our model developed by the other half of the dataset (called $D_{train.val}$).

2. Apply Monte Carlo CV to $D_{train.val}$ dataset, use 2/3 of the 209 observations as training data, and the rest 1/3 observations (70 in this case) as validation data.

3. In each CV run, evaluate the model selected by each candidate method using a predictive measure of interest: censoring-adjusted C-statistic proposed by Uno et al. (2011).

4. Repeat the model selection and computation of the above performance measure 200 times. The average over all 200 measures is presented in Table 1.5 using various model selection methods. The C-statistic measures how well the model predicts throughout the course of study.

Table 1.5: PBC Cross-validation data $D_{train.val}$: performance measures of different model selection methods, using random cross-validation with 2/3 of observations as training data and 1/3 of data as validation data.

Model Selection Method		C-	Median
Procedure	Tuning Criterion	Statistic	Model Size
Forward	AIC	0.737	6
Selection	BIC	0.733	4
Backward	AIC	0.742	7
Elimination	BIC	0.740	4
Stepwise	AIC	0.736	6
	BIC	0.744	4
Lasso	AIC	0.746	8
	BIC	0.746	6
Ridge	AIC	0.749	16
	BIC	0.757	16

In Table 1.5, the three traditional methods (forward selection, backward elimination, and stepwise regression) have comparable predictive performance in this dataset, with stepwise regression using BIC as stopping criterion yielding the highest censoring-adjusted C-statistic of 0.744. The Lasso and Ridge model selection methods yield slightly higher C-statistic than the other three methods with larger median model size. Ridge regression with BIC as stopping criterion using all 16 covariates yields the best predictive measure.

Stepwise regression using BIC as stopping criterion has comparable predictive performance as Lasso and Ridge regression on this particular dataset. However, the median model size of stepwise, BIC method is 4, compared with 6 in Lasso and 16 in Ridge regression method. The difference in the predictive measures between the traditional methods and the two penalized regression methods are not substantial in this dataset. Since CV-based estimator like all statistics is subject to some variability, the observed differences (if small) may be due to stochastic variation. If one decides that the slight difference in predictive accuracy does not outweigh the ease of implementation and smaller number of variables used by the traditional methods, one may choose the method that gives the highest predictive measures. In this case, stepwise regression using BIC as stopping criterion performs very well.

We should be aware that the CV-based prediction measures, such as the highest C-statistic of 0.757 derived from the Ridge BIC regression model, are optimistically biased because the models with the best estimated prediction accuracy measures are selected. Hence the true C-statistics may actually be lower.

1.6.1 Evaluating model's generalizability

To examine the generalizability of the model fit, we use the holdout sample for the evaluation and reporting. For the model selection method chosen by the CV procedure with best predictive measures, say, stepwise using BIC as stopping criterion, the following scoring algorithm is applied:

1. Use 2/3 of the $D_{train.val}$ dataset to build a model called $\mathscr{M}_1$, the rest 1/3 of this dataset is used to evaluate the predictive performance measures aforementioned.

2. Use β estimates from $\mathscr{M}_1$ to obtain a score for each subject in the holdout sample $D_{holdout}$. This score r_1 = exp($\hat{\beta}Z$), where Z is the covariate of each subject in dataset $D_{holdout}$.

3. Repeat the above two steps 200 times, and we have $r_1, r_2, r_3, \ldots, r_{200}$ for each subject in the holdout sample derived from 200 training models $\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3, \ldots, \mathscr{M}_{200}$.

4. Take the average of $r_1, r_2, r_3, \ldots, r_{200}$, which becomes the final risk score of the subjects in the holdout sample. The distribution of this summary risk score (r) can be obtained and used for risk profiling.

We stratify the summary risk scores (r) using tertiles of r. For the holdout sample, $D_{holdout}$ dataset, Figure 1.3 displays the Kaplan-Meier curves of the survival time stratified by the risk scores using three risk categories. The left panel shows the results using the optimal model selection method stepwise BIC identified in Table 1.5. For reference, the right panel shows the scoring results using Lasso with BIC as stopping criterion, a penalized regression method that used fewer variables than Ridge regression.

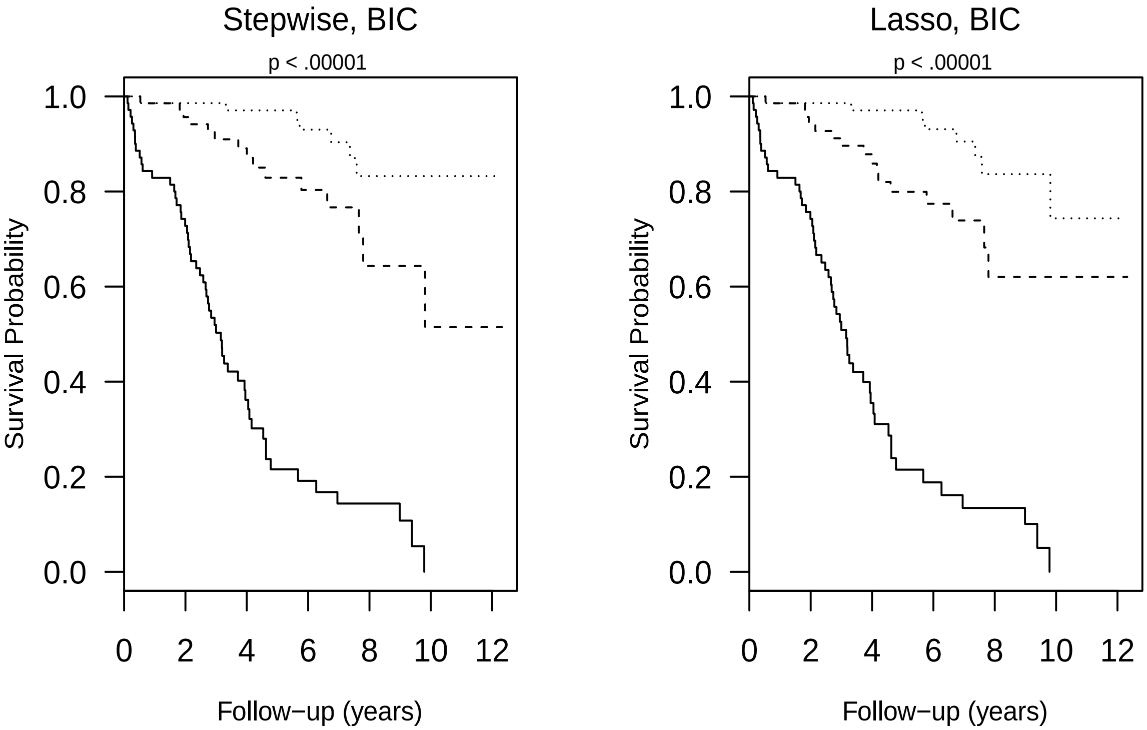


Figure 1.3: Kaplan-Meier curves of the survival time, stratified by tertile of risk scores of the holdout sample.

Table 1.6 and Table 1.7 present the summary statistics of the difference in survival distributions depicted in Figure 1.3. All the reported log-rank test p-values, and confidence intervals based on the holdout sample are valid conditional on the scoring system derived from the training dataset. We also evaluate another schema by stratifying the risk scores using four risk categories (low, medium low, medium high and high). The results (not shown) are similar between the stepwise, BIC method and the Lasso penalized regression method.

Table 1.6: Summary statistics of the survival distributions by risk categories, scoring using 200 training models on $D_{holdout}$ dataset.

Model Selection Method	Risk Categories	N Events/ Total	Restricted Mean (se) in years	Median (years)	(95% CI)
Stepwise, BIC	Low	7/70	11.16 (0.370)	NA	(NA, NA)
	Medium	15/69	9.41 (0.606)	NA	(7.79, NA)
	High	55/70	3.76 (0.365)	3.15	(2.66, 4.05)
Lasso, BIC	Low	8/70	10.96 (0.406)	NA	(NA, NA)
	Medium	16/69	9.48 (0.586)	NA	(7.79, NA)
	High	53/70	3.74 (0.369)	3.15	(2.66, 4.05)

Note: restricted mean with upper limit = 12.1 years

Table 1.7: Logrank test p-values of the difference between the survival distributions by risk categories.

Survival Difference between	Model Selection Method	
Risk Categories	Stepwise, BIC	Lasso, BIC
Overall	p $< 10^{-7}, \chi^2_{(2)}$=124.63	p $< 10^{-7}, \chi^2_{(2)}$=119.47
Pairwise Comparison		
Low vs. Medium	0.0131	0.0151
Low vs. High	$< 10^{-7}$	$< 10^{-7}$
Medium vs. High	$< 10^{-7}$	$< 10^{-7}$

As for the final prediction model, the scoring system presented above used the average model approach, which is an ensemble of 200 training models, no simple formula can be expressed. Figure 1.4 displays the distribution of $\hat{\beta}$ for each covariate obtained from 200 training models. The distribution of $\hat{\beta}$ for log(bilirubin) concentrated around 0.9 with low variability (mean risk ratio is 2.375), the best single prognostic factor. Other covariates have a variety of distributions, with those covariates in the upper right region located closely at zero (5th and 95th percentile equal zero) leading to a risk ratio of 1.000. Alternatively, we can use the stepwise regression method with BIC as stopping criteria to fit the dataset $D_{train.val}$ and obtain a prediction equation based on a linear combination of the selected covariates. These covariates and their risk ratios $exp(\hat{\beta})$ are: log(bilirubin), 2.195; edema, 5.596; age, 1.044; hepato, 1.975 and spiders, 1.943.

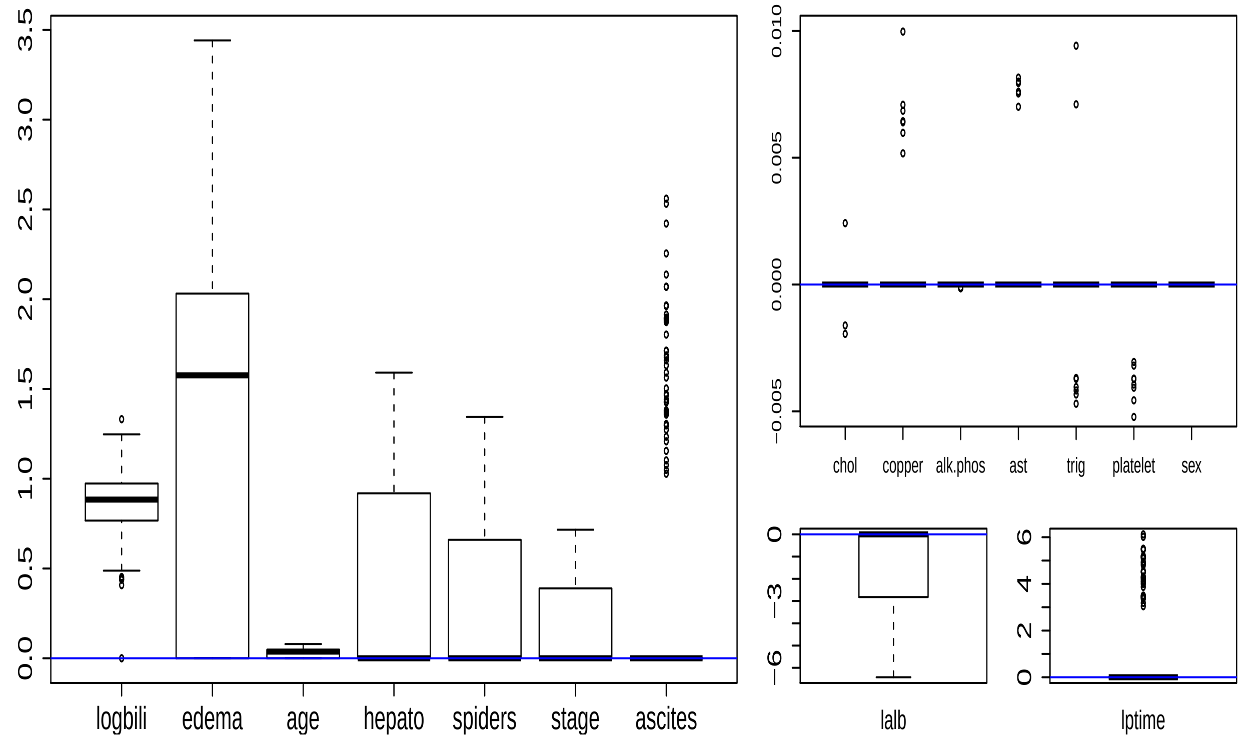


Figure 1.4: Distribution of $\hat{\beta}$ obtained from the 200 training models for each covariate.

1.6.2 Reducing overfitting via 3-in-1 proposal

We examine the performance of different scoring algorithms when there is no underlying relationship between survival time observations and covariates in a high-dimensional setting using simulations on the PBC dataset as follows. We keep the original survival time observations $(time_i, death_i)$, simulate 50 independent binary random variables with event rates ranging from 0.001 to 0.981 with 0.02 increment and 160 independent normal random variables with the same mean and standard deviation as the ten continuous covariates in PBC dataset (16 variables for each covariate distribution).

We randomly partition the dataset into two halves: $D_{train.val}$ and $D_{holdout}$, each has 209 observations ($n < p$). Applying stepwise regression with BIC as stopping criteria to the datasets, we present Figure 1.5 as follows:

- Leftmost panel: using the training data $D_{train.val}$ to build a model, and score on itself (conventional way)
- Middle panel: using the training data $D_{train.val}$ to build a model, but score on the holdout sample $D_{holdout}$
- Rightmost panel: apply our 3-in-1 modeling strategy similar to the procedure described in Section 1.6 with Stepwise BIC as the only candidate method considered, obtain an average model derived from 200 training models obtained during the CV process using $D_{train.val}$, score the holdout sample $D_{holdout}$ using this average model.

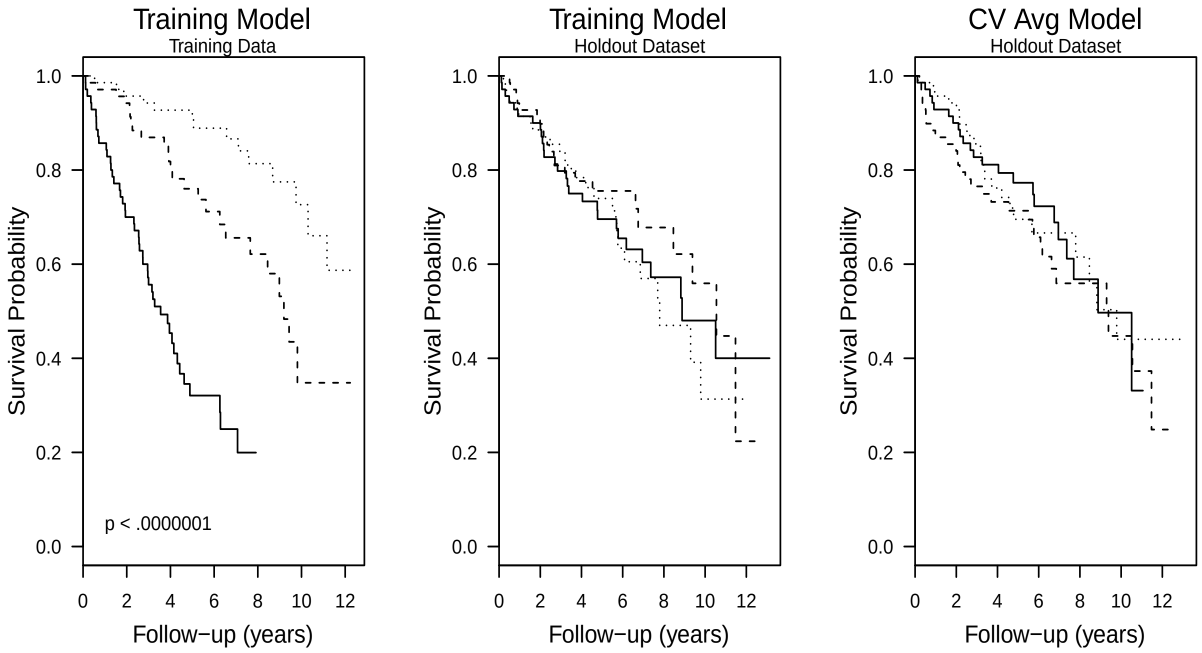


Figure 1.5: Kaplan-Meier curves of three scoring methods, using stepwise regression method with BIC as stopping criterion.

The leftmost panel shows overfitting using conventional same dataset modeling way; the other two panels did not show separation of the KM curves. Applying the average model derived from 200 training models in the spirit of bagging (i.e., bootstrap aggregate), all three KM curves almost overlap each other, leading to the correct conclusion that there is no relationship between the survival time and covariates.

The example shows that a combination of cross-validation and holdout sample is useful in combating overfitting.

1.7 Remarks

It is important to consider model building, selection and inference processes simultaneously as a package. The usual practice of using the same dataset for implementing procedures for these three steps may result in invalid inference as we demonstrated in this chapter. One may question about the efficiency issue for splitting the dataset for the final inference. However, with the conventional method, it is difficult to quantify the reliability of our claim at the end after an extensive, iterative model building process. This is probably why there are numerous false positive findings in all the scientific investigations. In fact, the sizes (or event rates) of most studies in practice may be too small for building reliable models for making valid inference.

We proposed a 3-in-1 dataset modeling strategy, achieving model building, selection, and holdout inference in one dataset. Cross-validation techniques are utilized to provide a sanity check for model fit to assess whether its predictive performance is acceptable. We can then select the optimal model building method to develop the final model and proceed with the inference part using the holdout sample, leading to more reproducible results and better application.

Needless to say, model building does not only depend on statistical grounds, knowledge of subject matter is absolutely essential in selecting the most appropriate model tailored to our needs. We focused on several classical methods and found that careful implementation of these methods could help us find a reasonably good predictive model. While censoring-adjusted C-statistic was used to evaluate predictive performance for illustration, other predictive measures or model evaluation methods can be considered. For example, Tian et al. (2007) proposed model evaluation based on the distribution of estimated absolute prediction error. Uno et al. (2011) looked at the incremental values of predictors. Furthermore, the candidate model selection methods considered were presented mainly under the framework that $n > p$. For the review on high-dimensional regression with survival outcomes, we refer to Sinnot and Cai's Chapter in Klein et al. (2013). They described some of the existing literature on dimension reduction, shrinkage estimation procedures with a range of penalty functions, and some hybrid procedures with univariate screening followed by shrinkage.

As mega datasets (genomic, data warehouse) become increasingly available, together with the ease of data storage and rapid development of data mining methodologies in censored data, these have enabled us to utilize more information for model development. It would be of interest to see how other datasets and predictive measures perform using our scoring algorithm. Moreover, what proportion of samples should we retain for hold-out sample, other cross-validation techniques with different partition schema can also be considered to fine-tune our model building strategies. Additionally, we tend to develop methods separately for each step of the model building, selection and inference process. If one aims at making efficient and valid inference about a parameter, say, the restricted mean survival time, a more consistent and integrated process using criteria to increase the precision of the final inference procedure should be considered. These questions remain an area of active research.

With the advent of the information age and the vast growth in the availability of massive amount of data, the challenges presented a unique window of opportunity for us to re-examine our conventional model selection strategy. Alternative modeling strategies in the analysis of censored outcome data could be considered to utilize the data and increase the overall model predictability in this Big Data era and dawn of personalized medicine age.

Acknowledgments

We thank Dr. Hajime Uno for his advice and Professor Hans C. van Houwelingen's helpful review of the manuscript. This work was supported by NIH grant RC4CA155940.

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

Florence H. YONG[1], Lu TIAN[2], Sheng YU[1], Tianxi CAI[1], and L.J. WEI[1]

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Health Research and Policy, Stanford University School of Medicine

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 dataset 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

2.1 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 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. On the other hand, when the regression model is misspecified, the scoring system can have systematic bias for some score values. To eliminate the bias, one may further calibrate the continuous scoring system nonparametrically (Tian et al., 2014). However, the calibrated prediction for subjects with a given score can be quite unstable due to the sparseness of the observations in the neighborhood. Consequentially the resulting prediction procedure in such a fine level may not perform well for practical usage. To this end, we often then group 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, the combination therapy may not be a good choice for patients who do not have a reasonable chance to respond to the treatment considering the associated economical cost and potential toxicity. Therefore, in the development of personalized medicine, it is an important step to predict the response probability for individual patient receiving the intervention treatment. Here, we show a conventional, ad hoc procedure to construct such 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 ($\log_{10}$ RNA$_0$). The numerical RNA level is used and all $\log_{10}$ 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:

$$\psi(-0.508 + 0.044\text{age} - 0.493\text{sex} + 0.004\text{CD4}_0 - 0.346\log_{10}\text{RNA}_0), \tag{2.1}$$

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 (2.1) 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 2.2, 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 (2.1), 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

(2.3) 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 (2.1), 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 (2.1) 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 2.3, we consider a collection of candidate scoring systems and utilize the method in Section 2.2 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 2.4, we generalize the new proposal to handle censored event time outcomes and illustrate the procedure using the data from a cardiovascular study (Braunwald et al., 2004) in Section 2.5. We conclude with additional observations and potential generalizations in Section 2.6.

2.2 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 (2.1). 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

$$-\log\{L(\beta)\} + \lambda \|\beta\|_p^p, \tag{2.2}$$

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 (2.2) results in the standard lasso and ridge regression with $p = 1$ and 2, respectively. The score (2.1) 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 subject 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:

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in S_k} |Y_i - \bar{Y}_k|, \tag{2.3}$$

which also quantifies the average within stratum variation. When all the scores are distinct, an optimal stratification, which minimizes (2.3), 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. This minimum size requirement may also be clinically meaningful because a general medical guideline for disease prevention or intervention generally is not aiming to a very small subpopulation. 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

$$\bar{Y}_k - \bar{Y}_{k-1} \geq d; \text{ for } k = 2, \cdots, K, \tag{2.4}$$

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

Consider all possible stratifications which satisfy (2.4) with a minimum stratum size fraction of at least p_0. To obtain an optimal stratification scheme by minimizing (2.3) is a rather challenging problem. In Appendix A.1, 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 (2.3) 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 (2.3):

$$L(\mathbf{c}) = \mathrm{E}\big|Y - f(V|\mathbf{c})\big| \quad \text{subject to} \quad \begin{cases} \mathrm{pr}(c_{k-1} < \tilde{\mu}(V) \leq c_k) \geq p_0 \\ \bar{\mu}_k - \bar{\mu}_{k-1} \geq d \end{cases}, \tag{2.5}$$

where

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

$I(\cdot)$ is the indicator function and $\bar{\mu}_k = \mathrm{E}(Y \mid c_{k-1} < \tilde{\mu}(V) \leq c_k)$. The justification of this asymptotic property is given in Appendix A.2. 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 2.3.

2.3 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 2.2. 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 2.2, 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 2.2. 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:

$$\mathcal{L}^* = \frac{1}{n^*} \sum_{k=1}^{K} \sum_{\hat{\mu}(V_i^*) \in (\hat{c}_{k-1}, \hat{c}_k]} |Y_i^* - \bar{Y}_k|, \qquad (2.6).$$

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 (2.6) among all the candidates under consideration. On the other hand, a parsimonious model may be appealing in practice if its (2.6) is greater than but still comparable to the minimum value of (2.6) 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 (2.6). 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 (2.6). 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 (2.5) 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 ($\log_{10}$ 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 2.1 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 2.1, 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

Table 2.1: Regression model candidates, $\mathrm{E}(Y|V) = \beta' Z$, for study ACTG 320, $\bar{\mathcal{L}}^*$ and the complexities of $\mathcal{M}^*$ (#var = the number of informative covariates and $\|\beta\|_0$ = the number of nonzero components of $\hat{\beta}$).

Model	Candidate independent variables	dim(Z)	$\bar{\mathcal{L}}^*$	$\mathcal{M}^*$	
				#var	$\|\hat{\beta}\|_0$
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

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 2.1. Note that the above inferential results would be a valid and final assessment on the practical value of this prediction scheme.

2.4 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 2.2, 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.

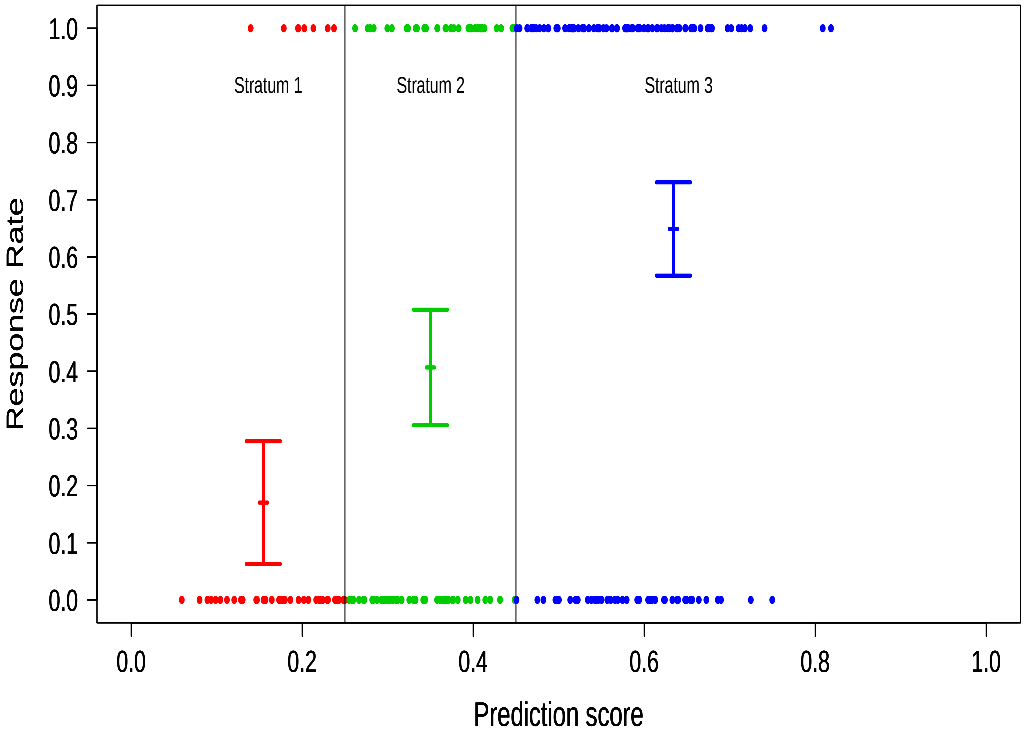


Figure 2.1: Stratum-specific point and 95% confidence intervals for the response rates with the Part II data (denoted by dots) of ACTG 320 with cutoff points $\hat{c}_1 = 0.25$ and $\hat{c}_2 = 0.45$.

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., 2014). 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 2.3 to obtain an optimal stratification scheme. Specifically, in the limit, we are interested in minimizing (2.5). Assuming that the censoring time is independent of the survival time T and covariates V, the prediction error in (2.5) 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 2.2, an optimal stratification can be obtained via the dynamic programming technique given in Appendix A.1. If the estimated RMST converges to a deterministic limit as the sample size increases, it follows from a similar argument in Appendix A.2, 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 2.3 with the weighted version of (2.6) 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.

2.5 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 patients 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 2.2. 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

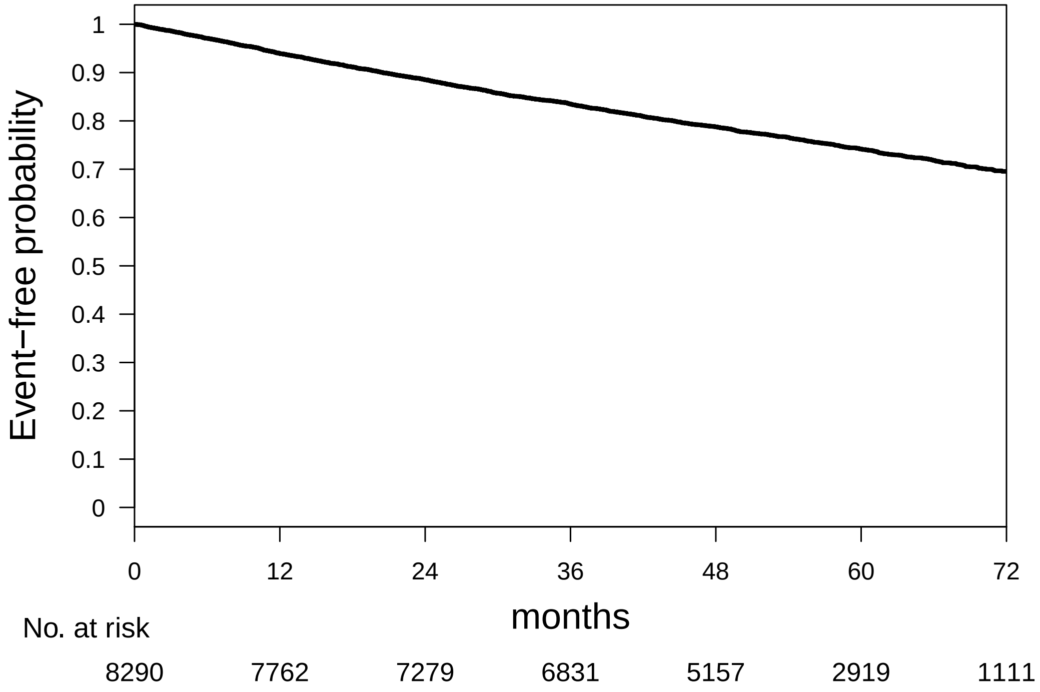


Figure 2.2: The Kaplan-Meier estimate for the time to the composite endpoint with the entire PEACE dataset.

Table 2.2: 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 $\mathcal{M}^*$ (#var = the number of informative covariates and $\|\beta\|_0$ = the number of nonzero components of $\hat{\beta}$).

Model	Candidate independent variables	dim(Z)	$\bar{\mathcal{L}}^*$	$\mathcal{M}^*$	
				#var	$\|\hat{\beta}\|_0$
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

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 2.2. 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 2.2 summarizes the $\bar{\mathcal{L}}^*$ for the optimal stratification based on each regression working model as well as the numbers of informative 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

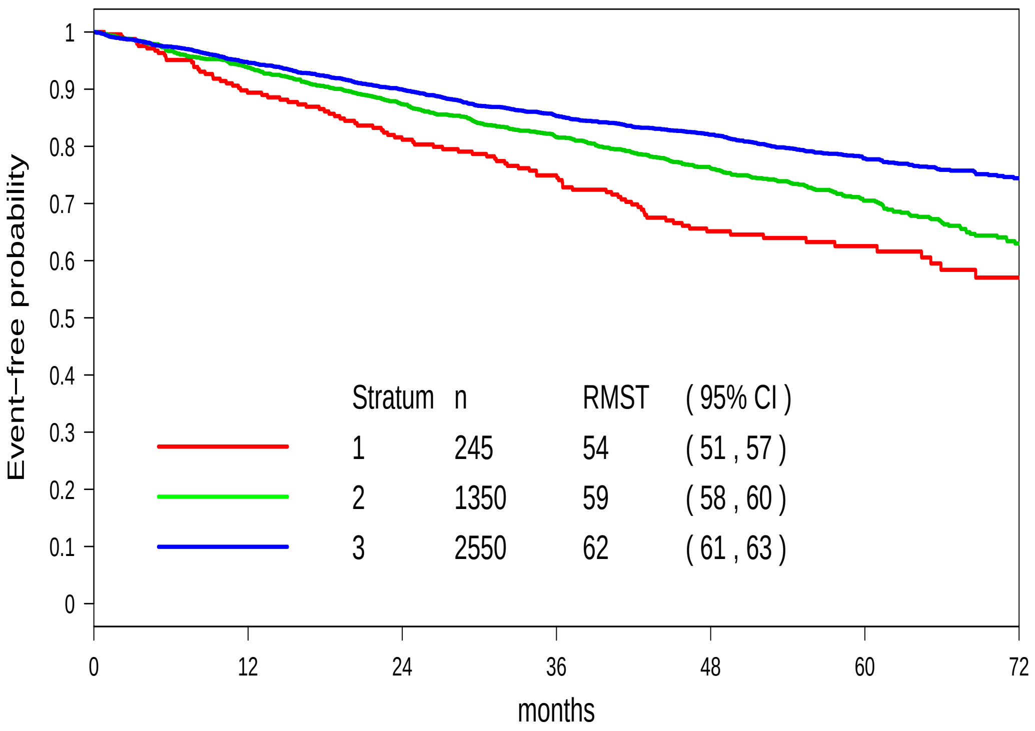


Figure 2.3: Stratum-specific Kaplan-Meier estimates and 95% confidence intervals for RMSTs obtained from Part II data of PEACE study with $\hat{c}_1 = 56.5$ and $\hat{c}_2 = 60.5$.

selected final stratification scheme, we apply it to the Part II data. The corresponding Kaplan-Meier curves for three strata are given in Figure 2.3. 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.

2.6 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 predicted 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 2.5 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. In such a case, one needs to predict the treatment effect measured by the difference of potential clinical outcomes of the patient under different treatments rather than the outcome itself. 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.

Acknowledgments

This research was supported by a grant (RC4CA155940) from the National Institute of Health and a training grant (T32 CA 009337) from the National Cancer Institute.

3. Identifying subgroups with successive average treatment effect via constrained optimization in stratified medicine

Florence H. YONG[1], Lu TIAN[2], and L.J. WEI[1]

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Health Research and Policy, Stanford University School of Medicine

Abstract

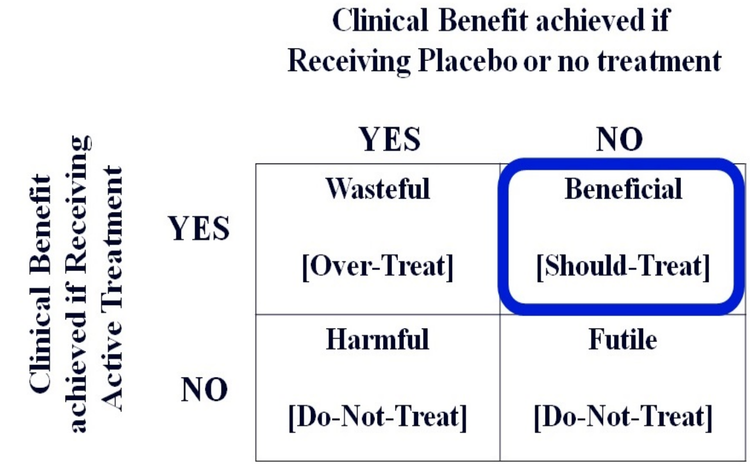
Stratified medicine brings enormous potential and promise to deliver more effective therapy to targeted patient populations. In this article, we propose a systematic procedure to identify subpopulations with a minimal clinically important Successive Average Treatment Effect (SATE) for treatment selection. First, we build a prediction model utilizing baseline information to create a predicted individual treatment effect scoring system. One approach is to fit the data of each treatment group with a separate regression model candidate relating the outcome to its baseline covariates in the training stage. We then apply a constrained optimization algorithm via dynamic programming to categorize the scores into subgroups. The performance of this stratification rule is immediately evaluated using an independent dataset during the validation stage, and the results of the prediction accuracy metric are used for model selection. Cross-validation process is employed for the aforementioned steps when there is only one study dataset available. Lastly, we apply the selected model to obtain the final stratification scheme. The evaluation of the scheme's practicality and corresponding influential results are obtained using a third independent holdout dataset to circumvent the problem of overfitting and enhance reproducibility. We illustrate the proposal using an AIDS clinical trial study for non-censored outcome, and a study on advanced heart failure patients for censored event time outcome to identify subgroups with most beneficial prospect, wasteful, harmful, and futile aspects. Our goal is to help bring stratified medicine one step closer to practical applications.

KEY WORDS: Cox regression model; Cross-validation; Dynamic programming; Predicted Individual Treatment Effect Score (PITES); Reproducibility; Stratified medicine; Successive Average Treatment Effect (SATE); Treatment selection; Uplift modeling

3.1 Introduction

One main goal of stratified medicine is to empower healthcare providers to provide the right therapeutic intervention to the right person based on individual patient's profile instead of vastly one-size-fit-all paradiam approach (WHO, 2013; Trusheim et al., 2007; Hingorani et al., 2013). This quest has become a key global effort to bring more effective and efficient treatment strategy to the patients in need (Medical Research Council, January 29, 2015; The White House, January 30, 2015; Collins and Varmus, 2015). For practical implementation of the personalized treatment strategy, reliable prediction models, clinically meaningful categorization of some comprehensible summary measures of individual treatment effect (ITE) to aid the decision-making process are vital elements to bring stratified medicine to fruitful realization.

First, to build a prediction model on the ITE, one approach (Cai et al., 2010; Zhao et al., 2013; Claggett et al., 2014) is to fit the data of each treatment group 1 or 2 (for control) with a regression model relating the outcome of interest to its baseline covariates to derive an effect score in the training stage. Then the models can be applied to all patients to predict the potential outcome score should that individual be treated by the alternative treatment group j ($j = 1$ or 2). The difference in their predicted scores constitutes a Predicted Individual Treatment Effect Score (PITES). Now, the question is how to utilize this information to identify subpopulations with differential treatment effect. Similar problems emerged in business sectors and have been approached via uplift modeling (Radcliffe and Surry, 1999; Lo, 2002). This subfield of machine learning aims at predicting the causal effect of an action (treatment) such as a marketing action on individuals' response (Sołtys et al., 2014). Its difficulty lies in the fundamental problem of causal inference (Holland, 1986) as a subject's observed outcome is only known after treatment or control is administered, and never both. The approach has been successfully applied to a wide range of problems in business and government sectors including presidential election campaigns (Porter, 2013; Siegel, 2013a) to identify pockets of customers with good prospect for tar-



		Clinical Benefit achieved if Receiving Placebo or no treatment	
		YES	NO
Clinical Benefit achieved if Receiving Active Treatment	YES	Wasteful [Over-Treat]	Beneficial [Should-Treat]
	NO	Harmful [Do-Not-Treat]	Futile [Do-Not-Treat]

Table 3.1: Uplift modeling response concepts in stratified medicine.

geted marketing. Siegel (2013b) provides a thorough overview of uplift modeling, and most recently Jaroszewicz et al. provides some rationale and examples that uplift modeling is a viable approach for selecting treatment in personalized medicine (Jaskowski and Jaroszewicz, 2012; Jaroszewicz and Rzepakowski, 2014). However, the majority of the methodology and tools deal with non-censored data and there has been some effort to turn censored data into non-censored data for the analyses (Štajduhar and Dalbelo-Bašić, 2012; Jaroszewicz and Rzepakowski, 2014).

Adapting the uplift concept (Radcliffe and Surry, 1999; Siegel, 2013b) to our context, Table 3.1 describes the four scenarios of a patient's response to either an active treatment or a placebo. Identifying more targeted subpopulation of patients for future treatment recommendation such that they are more likely to fall in the beneficial quadrant in Table 3.1 can potentially save resources and reduce unneccessary exposure of treatment-related adverse effects. To illustrate the concept, we use the data from a cardiovascular clinical trial "The Beta-blocker Evaluation of Survival Trial" (BEST) with censored outcome variables. The BEST trial is a double-blind, placebo-controlled study (BEST, 2001) of 2708 patients with advanced chronic heart failure enrolled to investigate whether the addition of a beta-adrenergic-receptor antagonist bucindolol (Buc) to the conventional therapy would improve survival (primary endpoint) or provide clinical benefit with respect to hospitalization, for instance. The study did not find any significant difference in mortality between the two groups (unadjusted hazard ratio of 0.90 with p=0.10), but they reported some survival benefit in nonblack patients (BEST, 2001). Data for 2707 patients was available in our analyses.

To identify subpopulations that may have a survival benefit, we first build a predictive ITE scoring system using the restricted mean event time (RMET) up to τ = 36 months, which is an interpretable and clinically meaningful summary of the survival function of a censored event outcome Y (Royston, 2009; Royston and Parmar, 2011; Zhao et al., 2013; Tian et al., 2014). For illustration, let us consider a regression model for all-cause mortality outcome Y truncated at τ = 36 and weighted by the inverse probability weight (w, described in Section 3.4) accounting for the censoring distributions of each treatment group. We consider five baseline covariates: NYHA (New York Heart Association functional class IV vs Class III), ischemic, smoke$_{ever}$ (vs. non-smoker), black (vs. nonblack), and a continuous measure of estimated glomerular filtration rate (eGFR); and fit the entire dataset with the two-model approach using a log link function. An estimate $\hat{\beta}_j$ for β_j (j = 1, 2) can be obtained via a regularized estimation procedure, for example, the least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996). The resulting individual predicted RMET had a patient been treated by Bucindolol is ($\hat{\mu}_1$):

$$exp(3.26-0.28\,\text{NYHA}-0.10\,\text{ischemic}+0.029\,\text{smoke}_{ever}-0.062\,\text{black}+0.0026\,\text{eGFR}). \quad (3.1)$$

The individual predicted RMET had a patient been treated by placebo is ($\hat{\mu}_2$):

$$exp(3.21-0.17\,\text{NYHA}-0.028\,\text{ischemic}-0.0057\,\text{smoke}_{ever}+0.0025\,\text{eGFR}). \qquad (3.2)$$

The 2707 PITES is the difference between $\hat{\mu}_1$ and $\hat{\mu}_2$, which range from -4.4 to 3.9 with a mean of 0.5 month. If the models are reasonably good, a future subject with a high PITES tends to respond to the treatment gaining survival time when being followed up to 36 months. For decision making and patient communication, it would be useful to stratify patients into, for instance, four consecutive categories with roughly equal sizes by using the quartiles of the predicted scores. Estimating the empirical RMET using the area under the Kaplan-Meier curve over $[0, \tau]$, the treatment difference in the average survival time for patients in these strata are -0.8, 0.7, 0.8, and 2.0 months, respectively. The Successive Average Treatment Effect (SATE), defined as the difference in average treatment effect between successive strata, are 1.5, 0.1, and 1.2 months. This conventional ad hoc stratification scheme unfortunately does not have discriminatory capability across all the strata. The small value in SATE between strata may not constitute a clinically meaningful different decision category. This may be due to the inadequate prediction models (3.1) and (3.2) or an improper grouping of the prediction scores.

The authors proposed a systematic procedure to tackle the above undesirable situations in single treatment case (Yong et al., 2014). We extend their approach to treatment selection problems in stratified medicine. Specifically, in Section 3.2, we demonstrate how to obtain an optimal grouping scheme that satisfies certain clinically meaningful criteria for each candidate PITES scoring system created from regression models using the data from an AIDS clinical study for non-censored outcome. Here, with the PITES generated from working models (3.1) and (3.2), we consider all possible discretization schemes, whose minimum stratum sizes would be at least 2.5% of the entire study sample size (n_0=68) and any two consecutive stratum-specific treatment difference in empirical RMET are monotonically increasing with an incremental value of at least 3 months to ensure a discrimination capability for future population. We then choose the best stratification, which minimizes a certain overall prediction error (3.12) 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 BEST mortality example and models (3.1) and (3.2), this results in three categories with the average treatment difference of -6.6, 0.5, and 5.0 months, respectively. Thus the SATE between neighboring strata are 7.1, and 4.5 months. Figure 3.1 displays their Kaplan-Meier estimates of the survival function. Note that if the working models (3.1) and (3.2) are appropriate, and the patients in each stratum are comparable, future patients classified to the stratum 3 will fall into the beneficial quadrant in Table 3.1 and should be our target population for this potentially life-saving treatment. However, patients classified to the stratum 1 and stratum 2 respectively belong to the harmful quadrant, and the wasteful and futile quadrants. No such treatment should be recommended to avoid superfluous exposure to adverse effects and save resources; alternative treatment could be seeked.

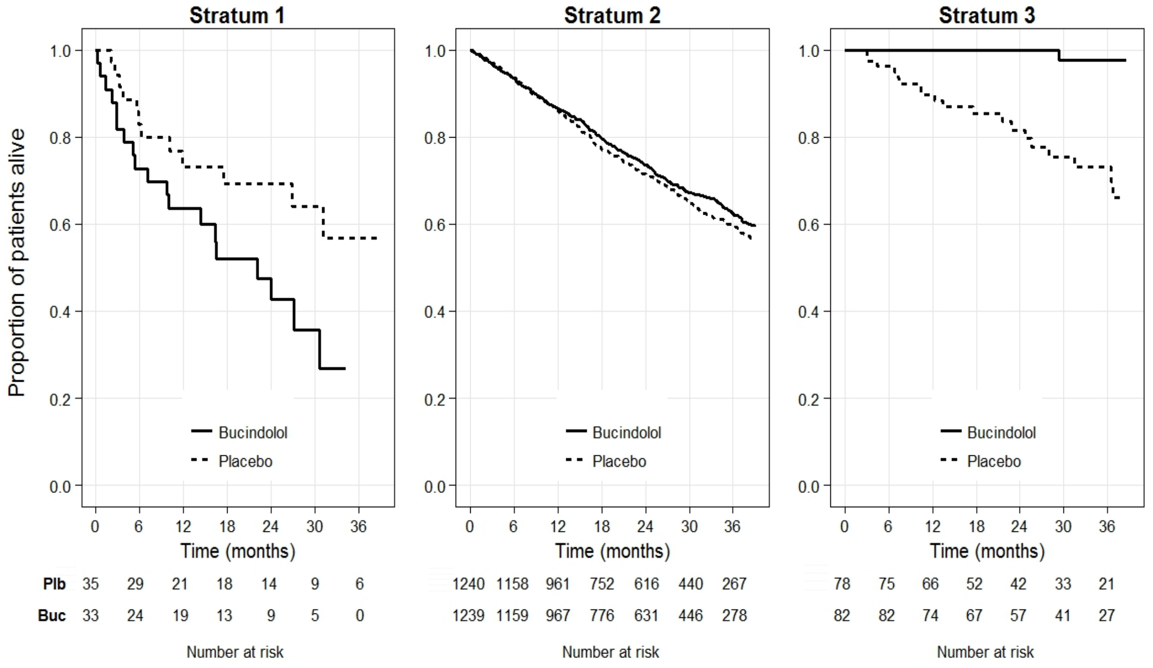


Figure 3.1: The stratum-specific Kaplan-Meier estimates for the time to death. Stratum obtained by a dynamic programming stratification algorithm on a candidate model scoring system derived from the entire BEST study data.

In Section 3.3, we aim at selecting the best model among a collection of candidate scoring systems by utilizing the method in Section 3.2 to obtain the optimal stratification for each scoring system. To avert the problem of overfitting, an independent validation dataset is used to evaluate all the resulting stratification schemes with respect to an empirical estimate of SATE and a prediction error measure quantifying the within-stratum heterogeneity. Once a prediction procedure is selected, a third independent dataset can be used to make inferences and assess the reproducibility of this scheme. 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 3.4, we generalize the proposal to handle censored event time outcomes and continue to illustrate the proposal using the BEST study data in Section 3.5. We conclude with remarks and potential generalizations in Section 3.6.

3.2 An optimal stratification procedure for an individual treatment effect scoring system

The authors showed how to obtain an optimal grouping scheme from a working model in the single treatment case (Yong et al., 2014). We now extend the procedure to the two treatment case for non-censored outcomes in randomized trial settings utilizing dynamic programing, supervised-learning approaches, and the uplift modeling concept that subjects in a well-constructed stratum are comparable. Let Y_1 and Y_2 be the outcome variables of treatment group 1 and control group 2 respectively, with V_1 and V_2 be the corresponding vector of "baseline" covariates. We can use a working model relating Y_j to Z_j, a function of V_j, to estimate the conditional mean response $\mu_j(V_j) = \mathrm{E}(Y_j|V_j)$, where the group index j = 1 or 2. For example, $\mu_j(V_j) = g_j(\beta_j' Z_j)$, where $g_j(\cdot)$ is a given smooth monotone function. Let the data consist of n_j independent copies $\{(Y_{ij}, V_{ij}, Z_{ij}), i = 1, \cdots, n_j\}$ of (Y_j, V_j, Z_j) for each group j=1, 2. An estimate $\hat{\beta}_j$ for β_j can be obtained via different variable selection procedures. If the regression model is a reasonable approximation to the true one, the resulting estimator $\hat{\mu}_j(V_j) = g_j(\hat{\beta}_j' Z_j)$ would be close to $\mu_j(V_j)$. Thus for an individual with baseline covariates V, we can predict the individual treatment effect by the difference in the two conditional means, $\nu(\mu_1, \mu_2, V) = g_1(\hat{\beta}_1' Z) - g_2(\hat{\beta}_2' Z)$. This $\nu(\cdot)$ is the Predicted Individual Treatment Effect Score (PITES), and a large PITES indicates that a large treatment effect on the individual's outcome is expected.

To illustrate the procedure for non-censored outcome, we utilize the data from a randomized, double-blind, placebo-controlled clinical trial (ACTG 320) for treating HIV-infected patients (Hammer et al., 1997). This study successfully demonstrated the overall efficacy of a combination of two nucleoside regimen with a protease inhibitor Indinavir, and 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, alternative treatment option may be recommended if the clinical benefit is not high enough. For this study, one of the endpoints was a binary outcome Y_j, indicating whether the patient's HIV-RNA viral level in group j was under an assay detectable level (500 copies/mL) at week 24 or not. A nonresponder 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. Complete baseline RNA information in terms of the $\log_{10}$ of HIV-RNA copies/ml ($\log_{10}$ RNA) are available for 1080 patients (537 treated vs 543 placebo). Other baseline variables available are age, male (vs female), race (non-Hispanic White, African American, other), injection-drug use, hemophilia, CD4 count (CD4), CD8 count, weight, Karnofsky performance s-core, and months of prior zidovudine therapy. Any missing covariates (very few) are replaced by their corresponding sample averages from the observed counterparts. The numerical RNA level is used and all $\log_{10}$ RNA measures below $\log_{10}(500)$ are replaced by $0.5\log_{10}(500) = 1.35$ in our analysis. A total of 242 subjects (of those 48% in treatment group) do not have RNA values at week 24 and their outcomes are treated as a failure. The observed overall difference in response rates between the two treatment groups is 0.39 (95% CI 0.34-0.44).

To build a predictive individual treatment effect scoring system, one may use a logistic model to estimate $\hat{\mu}_j(\cdot)$ for each treatment group and then obtain $\hat{\nu}(\cdot)$. Here, the regression parameter estimate $\hat{\beta}_j$ is obtained by minimizing the loss function

$$-\log\{L(\beta_j)\} + \lambda_1\|\beta_j\|_p^p + \lambda_2\|\beta_j\|_q^q, \tag{3.3}$$

where $L(\beta_j)$ is the likelihood function, λ_1 and λ_2 are the tuning (penalty) parameters, $\|\beta_j\|_r^r$ ($r = p$ or q) is the r^{th} power of the L_r norm for the vector β_j. When λ_2 = 0, the (3.3) results in the standard lasso and ridge regression with $p = 1$ and 2, respectively. When $p = 1$ and $q = 2$, the (3.3) results in the elastic-net regularizaton paths (Zou and Hastie, 2005). In this article, we consider an "elastic-net" candidate model with both λs' equal to 0.5. For each candidate model formulation, a PITES can be obtained for each subject, leading to a candidate PITES scoring system.

To identify subgroups with different treatment responses, we aim at grouping the subjects into K consecutive strata $S_1, S_2, \cdots, S_K$ based on the PITES $\nu(\cdot)$ and their actual responses such that a minimum clinically meaningful treatment effect increment d is achieved between successive strata with respect to the stratum-specific mean difference in treatment responses. Let $\bar{Y}_{1k} = \sum_{i \in S_k} Y_{i1}/n_{1k}$ and $\bar{Y}_{2k} = \sum_{i' \in S_k} Y_{i'2}/n_{2k}$ be the k^{th} s-tratum's empirical mean of outcome Y_{i1}'s and $Y_{i'2}$'s in group 1 and 2 with stratum size n_{1k} and n_{2k} respectively, $k = 1, \cdots, K$. For a future subject in group j (j = 1 or 2) being classified into the k^{th} stratum, we first predict the individual's treatment effect with the aforementioned model based score $\nu(\cdot)$. If the predictive model is adequate and the stratification is grouping "response-alike" subjects successfully, the individual treatment effect (ITE) response can also be non-parametrically estimated as the corresponding stratum-specific treatment difference $\delta_k = \bar{Y}_{1k} - \bar{Y}_{2k}$. One may consider the following loss function to evaluate the performance of this stratification:

$$\mathcal{L} = \frac{1}{n} \sum_{k=1}^{K} \{ \frac{n_{1k} + n_{2k}}{n_{1k} * n_{2k}} \} \sum_{i,i' \in S_k} |(Y_{i1} - Y_{i'2}) - \delta_k|, \tag{3.4}$$

where $n = \sum_{k=1}^{K} (n_{1k} + n_{2k})$ is the total number of observations. This loss function considers all possible pairwise treatment differences and quantifies their average difference from the corresponding stratum-specific treatment effect reflecting within stratum variation.

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 total sample size n leading to n_0. Furthermore, to make sure each stratum has a reasonable representation of each treatment group, one may impose a constraint that the minimum of the two treatment group sizes in each stratum to be at least m. Finally, to ensure that the stratification scheme has a clinically meaningful discriminatory capability, namely, yielding meaningful differences in group-specific average treatment differences between subgroups, we further impose a constraint for all candidate stratification schemes such that

$$\delta_k - \delta_{k-1} \geq d; \text{ for } k = 2, \cdots, K, \tag{3.5}$$

where d is a pre-specified positive value, representing the minimum clinically meaningful Successive Average Treatment Effect (SATE) that may warrant a different decision.

An optimal stratification scheme would minimize (3.4) while satisfying the aforementioned constraints. There could be an enormous number of possible stratification schemes and the problem of finding the optimal solution numerically is rather challenging. In Appendix A.1, the authors show how to identify the boundary values of the optimal stratification scheme $\hat{c} = \{\hat{c}_0, \hat{c}_1, \cdots, \hat{c}_{K-1}, \hat{c}_K\}$ of the K consecutive strata, where $S_k = \{i \mid \hat{c}_{k-1} < \hat{\nu}(V_i) \leq \hat{c}_k, i = 1, \cdots, n\}$ for $k = 1, 2, \cdots, K$; $\hat{c}_0 = -\infty$ and $\hat{c}_K = \infty$. Extended from the single treatment methodology developed in Yong et al. (2014), we propose a solution via dynamic programming algorithm (see examples in Taha (2003)). The concept is illustrated through a simple hypothetical example in the Appendix A.3. Specifically, the additional complexity of handling two treatment cases lies upon the determination of all possible groupings that satisfy the constraints. The Dynamic Programming Stratification (DPS) algorithm can be run more efficiently if the dimension of stratification operation can be reduced, for instance, by pregrouping subjects with very similar scores numerically and/or clinically to control for oversplitting. For the HIV study example, patients with an PITES of 0.1731 can be pregrouped with those with 0.1734 for a d of 0.1. Note that we use L_1 norm (3.4) to evaluate the prediction error because of its more interpretable quality. Other measures such as L_2 norm can be used instead without added difficulty.

3.3 Selecting an optimal stratification scheme from a collection of competing PITES systems

To circumvent the problem of overfitting and enhance reproducibility in prediction models, Figure 3.2 summarizes the data-splitting approach and objectives as presented in Yong et al. (2013, 2014) when there are data from a single study only. Consider any collection of candidate PITES systems generated from the Part Ia training data, we show how to evaluate them and choose the "best" one in this section. Let the validation 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 3.2. For each candidate scoring system, we obtain its optimal stratification scheme from the Part Ia training data with boundary points $\hat{c} = \{\hat{c}_0, \hat{c}_1, \cdots, \hat{c}_K\}$ and the stratum-specific treatment difference $\{\delta_k = \bar{Y}_{1k} - \bar{Y}_{2k}; k = 1, \cdots, K\}$. To evaluate the predictive performance of this stratification scheme with the data from Part Ib, we consider the following loss function that captures the incosistency between training and validation samples, and reflects the deviation of pairwise difference from the stratum-specific treatment differences:

$$\mathcal{L}^* = \frac{1}{n^*} \sum_{k=1}^{K} \{\frac{n_{1k}^* + n_{2k}^*}{n_{1k}^* * n_{2k}^*}\} \sum_{i,i' \in S_k^*} |(Y_{i1}^* - Y_{i'2}^*) - \delta_k|, \tag{3.6}$$

where the PITES is $\nu(V_i^*) = g_1(\hat{\beta}_1' Z_i^*) - g_2(\hat{\beta}_2' Z_i^*)$ for the two-model approach; $\hat{\beta}_j (j = 1, 2)$ and δ_k are derived from the Part Ia training data. The k^{th} stratum S_k^* contains Part Ib observations such that their scores $\nu(V_i^*) \in (\hat{c}_{k-1}, \hat{c}_k]$, for $k = 1, \ldots, K$. To obtain a more stable (3.6), Monte-Carlo cross-validation (MCCV) method is employed. Specifically, stratified by treatment assignment, we randomly split Part I dataset into Ia and Ib, say, N times, as shown in Figure 3.2. For the r^{th} split, we repeat the above model building and evaluation procedure for each candidate model and obtain $\mathcal{L}_r^*$ from (3.6). We then compute the average, $\bar{\mathcal{L}}^* = N^{-1} \sum_{r=1}^{N} \mathcal{L}_r^*$.

Given a stratification scheme obtained from a training sample that satisfies our pre-

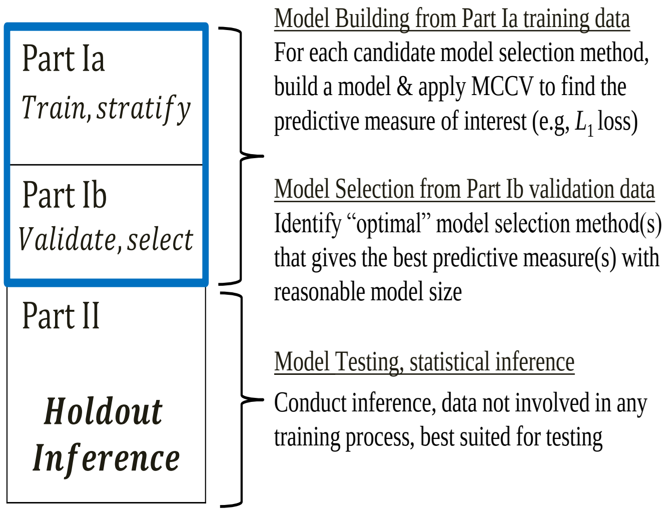

Part Ia
$Train, stratify$
Part Ib
$Validate, select$
Part II
$Holdout$
$Inference$
Model Building from Part Ia training data
For each candidate model selection method, build a model & apply MCCV to find the predictive measure of interest (e.g, L_1 loss)
Model Selection from Part Ib validation data
Identify "optimal" model selection method(s) that gives the best predictive measure(s) with reasonable model size
Model Testing, statistical inference
Conduct inference, data not involved in any training process, best suited for testing

Figure 3.2: 3-in-1 dataset modeling scheme. A dataset is randomly partitioned into Part I and Part II for modeling building and selection, and statistical inference.

specified constraints, we can also evaluate how well the scheme performs empirically with respect to the Successive Average Treatment Effect (SATE) between neighboring strata in a new sample. To this end, we define a weighted SATE measure $\mathcal{W}$ to be the weighted average of the empirical SATE estimates d_k:

$$\mathcal{W} = \begin{cases} 0, & \text{if } K = 1\text{, since there is zero successive difference in one stratum} \\ \sum_{k=2}^{K} w_k d_k & \text{if } K \geq 2 \end{cases} \tag{3.7}$$

where

$$\begin{aligned} \mathbf{w} = (w_2, \cdots, w_K) \quad &= \begin{cases} 1 & \text{if } K = 2 \\ \frac{1}{\sum_{k=1}^{3} n_k}(n_1 + \frac{n_2}{2}, \frac{n_2}{2} + n_3) & \text{if } K = 3 \\ \frac{1}{\sum_{k=1}^{K} n_k}(n_1 + \frac{n_2}{2}, \frac{n_2}{2} + \frac{n_3}{2}, \cdots, \frac{n_{K-1}}{2} + n_K) & \text{if } K \geq 4 \end{cases} \\ d_k = \delta_k - \delta_{k-1} \end{aligned}$$

Note that $\sum_{k=2}^{K} w_k = 1$ and the k^{th} stratum size is $n_k = n_{1k} + n_{2k}$, for $k = 1, \cdots, K$. Let us denote the $\mathcal{W}$ obtained from the training and validation data by $\mathcal{W}_{train}$ and $\mathcal{W}_{val}$ respectively.

Lastly, the "best" model should not only have a small $\mathcal{L}^*$ in (3.6), big $\mathcal{W}$ in (3.7), but also a parsimonious attribute for pragmatic application. Hence we refit the entire Part I data for each treatment group and let the final realized models be denoted by $\mathcal{M}_1^*$ and $\mathcal{M}_2^*$. These measures reflect the model complexity of each candidate scoring system. The selection of an "optimal" stratification rule would be based on the mean values of $\mathcal{L}^*$ ($\bar{\mathcal{L}}^*$), $\mathcal{W}$ derived from the validation data during the N MCCVs ($\bar{\mathcal{W}}_{val}$), and $\mathcal{M}_j^*$s'. Confidence interval estimates for the stratum-specific treatment differences can be constructed using the data from Part II holdout data based on the selected "optimal" stratification rule.

We now use the HIV study data to elucidate our proposal. First, we randomly split the entire dataset of 1080 patients evenly by treatment group to compile Part I (269 patients on treatment) and Part II holdout data (268 patients on treatment) with sizes of 540 each. A total of $N = 200$ MCCVs with 270 observations each in Part Ia and Part

Table 3.2: Study ACTG 320 regression model candidates, $\mathrm{E}(Y_j|V) = \beta_j' Z$, where $j = 1$ for treatment and $j = 2$ for control group; the complexities of $\mathcal{M}_j^*$ ($\Vert\hat{\beta}_j\Vert_0 =$ the number of nonzero components of $\hat{\beta}_j$), and the mean values of $\mathcal{L}^*$, $\mathcal{W}_{train}$, and $\mathcal{W}_{val}$ derived from 200 CV runs are shown.

Model	Candidate independent	dim	$\mathcal{M}_1^*$		$\mathcal{M}_2^*$		Results from 200 CV		
ID	variables at baseline	(Z)	#var	$\Vert\hat{\beta}_1\Vert_0$	#var	$\Vert\hat{\beta}_2\Vert_0$	$\bar{\mathcal{L}}^*$	$\bar{\mathcal{W}}_{train}$	$\bar{\mathcal{W}}_{val}$
1	age,sex,CD4,$\log_{10}$RNA	4	4	4	4	4	0.495	0.208	0.072
2	all baseline covariates	77	10	10	11	11	0.501	0.239	0.022
3	plus their first-order	77	12	75	12	75	0.496	0.424	0.030
4	interaction terms	77	12	28	12	25	0.494	0.317	0.037
5	none	0	0	0	0	0	0.516	0	0

Ib for every cross-validation is used to find the "optimal" model among several candidate PITES scoring systems. For illustration, we consider five working models in which the first one is the simple logistic regression model with four baseline variables. The second, third, and fourth models with all baseline covariates and their first-order interaction terms are logistic regression models using lasso, ridge, and elastic-net regularization methods respectively. Their tuning parameters fitted with are selected via a 10-fold cross-validation procedure built in the R package *glmnet* (Friedman et al., 2010). The fifth model is a null model using the overall mean treatment and control response proportions in Part Ia to predict the future outcomes for each cross-validation run. Table 3.2 summarizes the model candidate form, the complexity of the corresponding $\mathcal{M}_j^*$ (j = 1, 2) with respect to the number of informative baseline covariates (#var) needed to compute the score $\nu(\cdot)$, and the number of nonzero regression coefficients in $\hat{\beta}_j$. The $\bar{\mathcal{W}}_{train}$, $\bar{\mathcal{L}}^*$ and $\bar{\mathcal{W}}_{val}$ are the averages of the 200 corresponding measures derived from the Part Ia training and Part Ib validation data respectively during the CV process. For all stratification schemes, we use the SATE value d of 0.1 (or 10%) and the minimum stratum proportion of 0.1 (or $n_0 = 54$) to enhance the stability of the estimates.

From Table 3.2, Models 1, 3, and 4 have similar $\bar{\mathcal{L}}^*$ values. However, Model 1 is a simpler model involving fewer baseline covariates as reflected by $\mathcal{M}_1^*$ and $\mathcal{M}_2^*$. It also has the highest $\bar{\mathcal{W}}_{val}$ of 0.072, and thus is our selected model to proceed. Based on 200

CVs, with at least 50% of the chance, Model 1 selects 3 strata and Models 2-4 select 3 or 4 strata. More complex model candidates appear to fit the training data better attaining higher $\bar{\mathcal{W}}_{train}$ values which are expected to be at least 0.1 according to our pre-specified constraint. However, when the training stratification schemes are applied to the Part Ib validation data, the 0.1 differences between strata may not hold. Overfitting in the training data and the relatively small sample sizes may contribute to this phenomenon.

We fit Model 1 to the entire Part I data for each treatment group to obtain the following models: had the subject been treated with the three drug combination, the response rate is:

$$\psi(-0.0233 + 0.0373\,\text{age} + 0.370\,\text{male} - 0.443\ \log_{10}\text{RNA}_0 + 0.00212\,\text{CD4}_0); \qquad (3.8)$$

had the subject been treated by placebo, a two-drug combination, the response rate is:

$$\psi(-16.424 + 0.010\,\text{age} + 16.696\,\text{male} - 0.672\ \log_{10}\text{RNA}_0 - 0.00192\,\text{CD4}_0), \qquad (3.9)$$

where $\psi(s) = \{1 + \exp(-s)\}^{-1}$ is the anti-logit function.

The PITES $\nu(\cdot)$ is the difference in the individual predicted response rates generated by (3.8) and (3.9). Applying our DPS algorithm to these $\nu(\cdot)$ scores in Part I data, a stratification scheme with $\hat{c} = (\hat{c}_1, \hat{c}_2, \hat{c}_3)$ = $(0.262, 0.327, 0.524)$ gives rise to four strata with sizes 63, 119, 302, and 56. The stratum-specific treatment difference δ_ks′ are 16.5%, 27.4%, 40.4%, and 64.6% for $k = 1, 2, 3,$ and 4. These point estimates may be biased due to the extensive model training, validation, and selection. To obtain valid inferences for this final prediction procedure and stratification scheme, we fit models (3.8) and (3.9) using Part II holdout data to generate PITES, and then group the subjects by $\hat{c}$. The resulting point and 0.95 confidence interval estimates based on 1000 bootstrap samples for the four stratum treatment differences $\delta_k, k = 1, \ldots, 4,$ are 4.0% (0.0%, 13.6%), 34.7% (19.9%, 48.9%), 42.1% (34.1%, 50.9%) and 61.7% (39.7%, 81.0%), with stratum size $n_k = 49, 114, 333,$ and 44 respectively as displayed in the top panel of Figure 3.3. The three SATE point estimates $d_k = \delta_k - \delta_{k-1}$ for $k = 2, 3, 4,$ and the corresponding 95% bootstrap confidence intervals are 30.7% (13.1%, 47.9%), 7.4% (-9.6%, 24.1%), 19.6% (-3.5%, 40.8%) as shown in the bottom panel. The $\mathcal{L}$ and $\mathcal{W}$ derived from Part II data are 0.48 and 16.7% respectively. Table 3.3 summarizes the baseline characteristics of the four strata obtained from this final stratification scheme. To examine whether there is any imbalance between the treatment groups within each strata, Fisher's exact test and chi-square test are performed for treatment comparison association for categorical variables depending on the sample size. For continuous variables, Wilcoxon rank-sum test is performed. There is no statistically significant difference between the treatment groups within each stratum at α = .05 level of significance. The results suggest that future patients classified into Stratum 1 may consider alternative treatment options, while the Stratum 4 patients are more likely to receive higher than average treatment benefit.

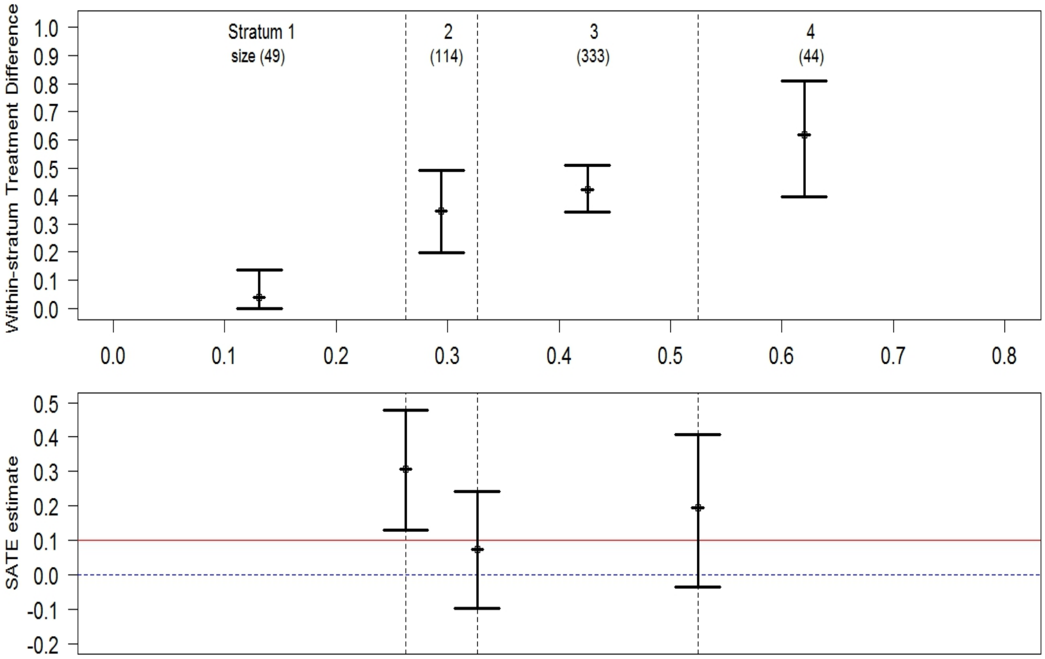


Figure 3.3: ACTG 320 Part II data: Stratum-specific point estimates and 95% confidence intervals for the treatment difference in response rates $\delta_k, k = 1, \ldots, 4$ (top panel) and for the empirical Successive Average Treatment Effect $d_k, k = 2, 3, 4$ (bottom panel).

Table 3.3: Study ACTG320: Baseline Characteristics of the patients in the Part II holdout data by strata and treatment.

Baseline	Stratum 1 ($n_1 = 49$)		Stratum 2 ($n_2 = 114$)		Stratum 3 ($n_3 = 333$)		Stratum 4 ($n_4 = 44$)	
Characteristic	TRT	Control	TRT	Control	Treatment	Control	Treatment	Control
	$n_{11} = 25$	$n_{12} = 24$	$n_{21} = 49$	$n_{22} = 65$	$n_{31} = 170$	$n_{32} = 163$	$n_{41} = 24$	$n_{42} = 20$
Male sex n (%)	20 (80)	17 (71)	38 (78)	57 (88)	146 (86)	146 (90)	21 (88)	16 (80)
Age (year)	28 ± 4	29 ± 3	34 ± 4	35 ± 4	41 ± 8	41 ± 7	49 ± 10	50 ± 9
Race or ethnic group n (%)								
White, non-Hispanic	10 (40)	8 (33)	25 (51)	32 (49)	91 (54)	82 (50)	11 (46)	6 (30)
African American	9 (36)	9 (38)	13 (27)	21 (32)	39 (23)	47 (29)	7 (29)	9 (45)
Other	6 (24)	7 (29)	11 (22)	12 (18)	40 (24)	34 (21)	6 (25)	5 (25)
Injection-drug use n (%)	1 (4)	2 (8)	4 (8)	7 (11)	28 (16)	30 (18)	6 (25)	7 (35)
Hemophilia n (%)	1 (4)	1 (4)	0 (0)	2 (3)	3 (2)	4 (2)	0 (0)	1 (5)
Karnofsky score≥90 n (%)	20 (80)	17 (71)	41 (84)	47 (72)	136 (80)	132 (81)	19 (79)	15 (75)
CD4 count	28 ± 33	36 ± 36	48 ± 44	41 ± 43	100 ± 64	99 ± 67	169 ± 67	164 ± 92
log_{10} HIV-1 RNA copies/ml	5.6 ± 0.3	5.4 ± 0.3	5.3 ± 0.4	5.4 ± 0.4	4.9 ± 0.6	4.8 ± 0.8	3.8 ± 1.2	4.4 ± 0.7
Months of prior therapy*	21 ± 15	18 ± 12	32 ± 31	29 ± 23	32 ± 32	28 ± 32	34 ± 31	41 ± 30

Plus-minus values are means ± SD. *prior zidovudine therapy, alone or in combination.

3.4 Generalization to cases with event time as the outcome variable

We extend the proposal to handle censored outcome in this section. Let T_{ij} be the time to an event of interest for the i^{th} subject in the j^{th} treatment group G_1, where $i = 1, \ldots, n_j; j = 1$ for treatment group 1 and $j = 2$ for control group G_2. The restricted mean event time (RMET) is a clinically meaningful summary measure to describe the event profile, and will be used as the scores to evaluate the treatment differences for subjects being followed up to a pre-specified time point τ. Let $Y_{ij} = T_{ij} I(T_{ij} \le \tau) + \tau I(T_{ij} > \tau)$, then the average event-free time for subjects in group j with covariate V_{ij} is $\mu_j(V_{ij}) = \mathrm{E}(Y_{ij}|V_{ij}) = \int_0^\tau S_j(t|V_{ij})dt$, where $S_j(t|V_{ij}) = \mathrm{pr}(T_{ij} > t \mid V_{ij})$. Suppose the outcome T_{ij} (and Y_{ij}) is right censored by an independent random variable C_{ij}, one can still observe $(X_{ij}, V_{ij}, \Delta_{ij})$, where $X_{ij} = \min(T_{ij}, C_{ij})$ and Δ_{ij} is one if $X_{ij} = T_{ij}$ indicating the event is met, and zero otherwise for censored outcome $T_{ij} > C_{ij}$. The observed data now consist of n_j independent copies $\{(X_{ij}, V_{ij}, \Delta_{ij}), i = 1, \cdots, n_j\}$ of (X_j, V_j, Δ_j) for each group $j = 1, 2$.

To obtain the $\nu(\cdot)$ based on the difference between $\mu_1(\cdot)$ and $\mu_2(\cdot)$, we need to formulate how to estimate $\mu_j(V_j)$. Extensive studies (Zucker, 1998; Andersen et al., 2004; Tian et al., 2014) have been conducted and we will present three major approaches to create our candidate PITES scoring systems. First, as summarized in Yong et al. (2014), one may use the Cox (1972) procedure to model the relationship between the survival function $S_j(t|V_j)$ of the event time and its covariates V_j for each treatment group j:

$$\log\{-\log S_j(t|V_j)\} = \log\{-\log S_{0j}(t)\} + \beta_j' Z_j,$$

where $S_{0j}(\cdot)$ is an unknown baseline survival function, and β_j is the regression coefficient vector. A regularized estimate $\hat{\beta}_j$ of β_j can be obtained by minimizing a loss function similar to (3.3), with $L(\beta_j)$ being the partial likelihood function instead. The $S_{0j}(t)$ can then be estimated by $\exp\{-\hat{\Lambda}_{0j}(t)\}$, where $\hat{\Lambda}_{0j}(t)$ is the Breslow estimate for the underlying cumulative hazard function (Breslow, 1972). It follows that the RMET for subjects with the covariate V can be estimated as

$$\hat{\mu}_j(V) = \int_0^\tau \exp\{-\hat{\Lambda}_{0j}(t)e^{\hat{\beta}_j{}'Z}\}dt. \tag{3.10}$$

For the second and third approaches, we use an approach similar to the accelerated failure time model and adopt Tian et al. (2014)'s proposal to estimate the RMET via direct modeling with the baseline covariates instead of channeling through the hazard function for each treatment group j:

$$\hat{\mu}_j(V) = \eta^{-1}\{\hat{\beta}_j'Z\}, \tag{3.11}$$

where $\eta(\cdot)$ is a given smooth and strictly increasing link function and $Z' = (1, V')$. We employ the identity link $\eta(v) = v$ and the log link $\eta(v) = log(v)$ functions to create some of our candidate PITES scoring systems. Notice that (3.11) can exceed τ for an individual with covariates V because of model misspecification; here our focus is on the ranking of $\nu(\cdot)$ derived from the difference between $\hat{\mu}_1$ and $\hat{\mu}_2$.

To obtain an optimal stratification scheme using the DPS algorithm in Section 3.3 given a scoring system, we desire to minimize the prediction error estimated by the L_1 loss in a measure of differential event time difference $(Y_{i1} - Y_{i'2})$ from its average empirical estimate, accounting for the group censoring distribution. Let $\delta_k = \bar{Y}_{1k} - \bar{Y}_{2k}$ be the difference in the two consistent estimators $\bar{Y}_{1k}$ and $\bar{Y}_{2k}$ for the k^{th} stratum-specific RMET in treatment group and control group, respectively. The two estimators are the weighted average of the event times defined as:

$$\bar{Y}_{1k} = \frac{\sum_{i \in S_k} w_{i1} Y_{i1}}{\sum_{i \in S_k} w_{i1}}, \quad \text{and} \quad \bar{Y}_{2k} = \frac{\sum_{i' \in S_k} w_{i'2} Y_{i'2}}{\sum_{i' \in S_k} w_{i'2}}$$

where $w_{i1} = \{\Delta_{i1} + (1 - \Delta_{i1})I(T_{i1} \geq \tau)\}/\hat{G}_1(Y_{i1})$ and $w_{i'2} = \{\Delta_{i'2} + (1 - \Delta_{i'2})I(T_{i'2} \geq \tau)\}/\hat{G}_2(Y_{i'2})$; $\hat{G}_j(\cdot)$ is the Kaplan-Meier estimate for the j^{th} group's censoring distribution derived from the Part I data, $j = 1, 2$. For a stratification scheme with K strata with the k^{th} stratum S_k comprising of n_{1k} and n_{2k} observations from treatment group 1 and control group 2 respectively, where $k = 1, \cdots, K$, the effective k^{th} stratum size for each group can be estimated by $m_{1k} = \sum_{i \in S_k} w_{i1}$ and $m_{2k} = \sum_{i' \in S_k} w_{i'2}$ accounting for the censoring distribution. Similar to (3.4), we want to minimize the following loss function:

$$\mathcal{L} = \frac{1}{n} \sum_{k=1}^{K} \frac{(m_{1k} + m_{2k})}{(m_{1k} * m_{2k})} \sum_{i,i' \in S_k} w_{i1} * w_{i'2} * |(Y_{i1} - Y_{i'2}) - \delta_k| \tag{3.12}$$

If we have a collection of competing scoring systems, one can evaluate each candidate stratification scheme via the cross-validation using the data from Part Ia and Ib iteratively. In particular, once a stratification scheme with boundary values $(\hat{c}_0, \hat{c}_1, \cdots, \hat{c}_K)$ leading to the stratum-specific treatment difference estimates δ_k $(k = 1, \cdots, K)$ is derived from Part Ia data, we can assess the predictive performance of this stratification scheme using Part Ib validation data $\{(X^*_{ij}, V^*_{ij}, \Delta^*_{ij})$ and the derived $\nu(V^*_{ij}); i = 1, \cdots, n^*; j = 1, 2\}$. First, we identify the stratum membership of each Part Ib observation by finding its corresponding stratum S^*_k such that $\nu(V^*_{ij}) \in (\hat{c}_{k-1}, \hat{c}_k], \exists k \in \{1, \cdots, K\}$. Then the following loss function can be estimated:

$$\mathcal{L}^* = \frac{1}{n^*} \sum_{k=1}^{K} \frac{(m^*_{1k} + m^*_{2k})}{(m^*_{1k} * m^*_{2k})} \sum_{i,i' \in S^*_k} w^*_{i1} * w^*_{i'2} * |(Y^*_{i1} - Y^*_{i'2}) - \delta_k| \tag{3.13}$$

where $n^* = \sum_{k=1}^{K} (n^*_{1k} + n^*_{2k})$ is the total number of observations in Part Ib data.

An alternative way to estimate RMET consistently is by the area under the Kaplan-Meier (KM) curve over $[0, \tau]$. Let $\mathcal{R}_{jk}$ be these empirical measures derived from group j in stratum k. We impose an additional constraint (3.14) other than those described in Section 3.2 to account for the difference in estimated RMET based on the inverse probability weight measures w due to small sample sizes and the within-stratum variation in w estimates.

$$[\mathcal{R}_{1k} - \mathcal{R}_{2k}] - [\mathcal{R}_{1(k-1)} - \mathcal{R}_{2(k-1)}] \geq d \tag{3.14}$$

An appealing feature of this constraint is to make sure that the graphical display using KM curves is coherent with our stratification results, especially when the sample size is small. Lastly, one can impose another constraint to control the minimum observed event rate in each stratum to be at least 0.05 (say), such that there is minimal information to generate strata with higher quality and stability. An optimal stratification can then be obtained via the dynamic programming technique as described in Section 3.2 and the Appendix A.1 of Yong et al. (2014). For evaluation, the empirical estimate of weighted SATE $\mathcal{W}$ will be based on the treatment difference derived from the KM estimated RMET $\mathcal{R}_{jk}$ in the next section.

3.5 An illustrative example with censored event time outcomes

We use the BEST study on advanced chronic heart failure patients described in Section 3.1 to illustrate the proposal with another event time outcome variable. Clinical benefit may be defined by a combination of mortality and morbidity. Here we consider the event time as the time to a composite of all-cause mortality or hospitalization because of heart failure, whichever occurred first. There were 1420 patients (660 treated vs 760 placebo) experienced this composite event with an average follow-up time of 2 years. If we let $\tau = 36$ (months), the KM estimated RMETs for the entire treatment group and control group are 23.2 and 21.2 months, respectively. This suggests that patients treated with Bucindolol expect to have an average of 23.2 months heart failure hospitalization event-free survival when they are being followed for up to 36 months, compared with 21.2 months in placebo group. Fitting the inverse-probability weighted regression models with a log link using the five baseline covariates as in Section 3.1 and (3.11) to the entire data and then stratify the PITES using our algorithm with d = 3 months and n_0 = 68 (or p_0 =.05 of Part I data) gives rise to four strata depicted in Figure 3.4. The estimated treatment differences in RMETs for the four strata are -0.9, 2.3, 6.8, and 10.1 months with a SATE estimate of 3.2, 4.5, and 3.3 months respectively. The $\mathcal{L}$ and weighted SATE $\mathcal{W}$ are 14.9 and 3.6 respectively. However, we do not know whether this stratification scheme

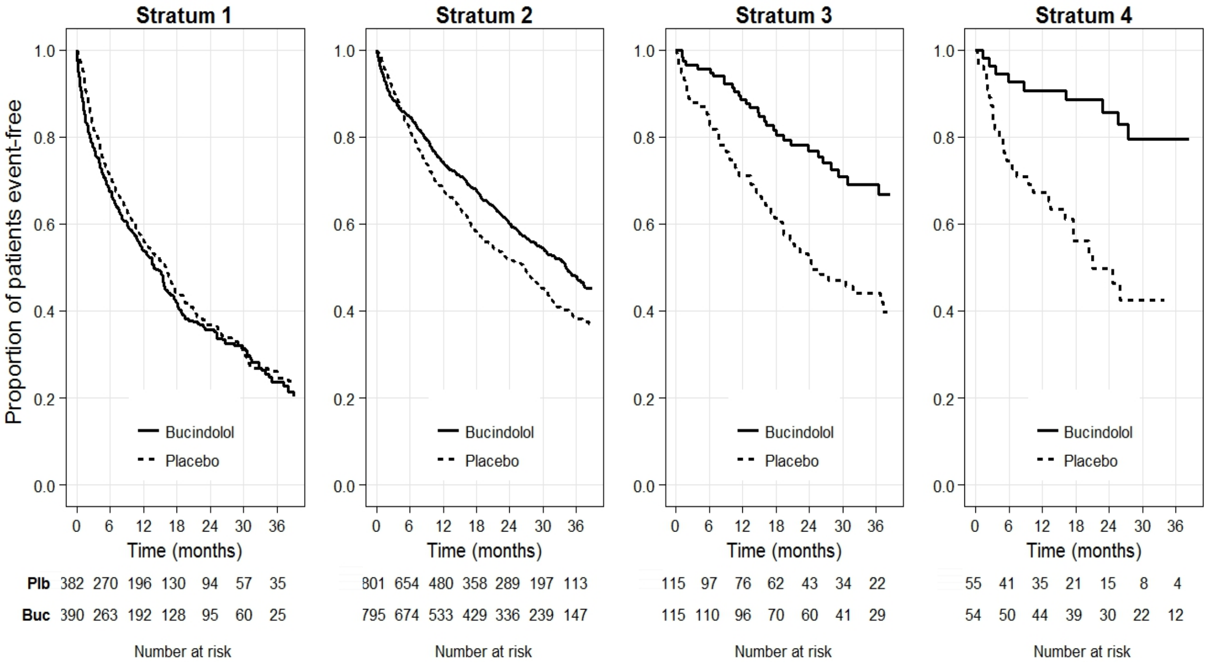


Figure 3.4: The stratum-specific Kaplan-Meier estimates for the time to death or hospitalization due to heart failure. Stratum obtained by DPS algorithm on a candidate model scoring system derived from the entire BEST study data.

can be improved.

To select the best performing model from a collection of candidate working models, we consider the candidate variables as described in Table 3.4. The null model does not contain any variables and serves as a background reference model. The other 14 candidate PITES $\nu(\cdot)$ scoring systems are generated by the three aforementioned approaches, with the $\hat{\mu}_j$ estimated via (3.10) for Cox PH models, or via (3.11) for the identity link and log link models. For candidate independent variables, other than the five baseline variables described in Section 3.1, we consider the following baseline covariates for prediction: age, male (vs female), body mass index >30, current smoker (smoke$_{current}$), history of hypertension (hx_hyp), history of diabetes (diab), ischemic as the cause of heart failure, atrial fibrillation (afib), left ventricular ejection fraction (lvef), systolic blood pressure (sbp), heart rate, 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.

Table 3.4: Study BEST regression model candidates description

Model ID	 Candidate independent variables at baseline	Model Description Link function; variable selection procedure	dim(Z)
1	eGFR, NYHA, smoke$_{ever}$, ischemic, black	identity link; Lasso	5
2	eGFR, NYHA, smoke$_{ever}$, ischemic, black	log link; Lasso	5
3	eGFR, NYHA, smoke$_{ever}$, ischemic, black	Cox PH; Lasso	5
4	main effects:	identity link; Lasso	18
5	age, male, black, lvef, sbp,	log link; Lasso	18
6	heart_rate, NYHA, smoke$_{current}$,	Cox PH ; Lasso	18
7	smoke$_{ever}$, hx_hyp, diab,	Bagged ensemble* of 100 model 4	18
8	ischemic, afib, bmi>30,	Bagged ensemble* of 100 model 5	18
9	eGFR, eGFR$_1$, eGFR$_2$, eGFR$_3$	Bagged ensemble* of 100 model 6	18
10	all baseline covariates plus	identity link; Lasso	150
11	their first-order	log link; Lasso	150
12	interaction terms	Cox PH; Lasso	150
13		Cox PH; Ridge penality	150
14		Cox PH; Elastic-net penality	150
15	None	NA	0

*mean of a bagged ensemble of 100 models
Model 4 to 9 use the same set of variables listed; same for Model 10 to 14

We randomly split the data evenly into Parts I and II with 1353 and 1354 patients each. Moreover, the Part I data are evenly split for the cross-validation process with 200 iterations. For each regression model candidate, we use a SATE value of $d = 3$ months and the minimum stratum fraction of $p_0 = 0.05$ (or n_0 = 68) during the CV training and validation process. Figure 3.5 depicts the distribution of $\mathcal{W}$ measures derived from Part Ia training ($\mathcal{W}_{train}$) and Part 1b validation ($\mathcal{W}_{val}$) data during the CV process. Notice that $\mathcal{W}_{val}$ values are generally smaller than the $\mathcal{W}_{train}$ values, indicating that over-training may lead to more irreproducible results. Table 3.5 summarizes the $\bar{\mathcal{L}}^*$ for the optimal stratification of each candidate PITES scoring system, and the $\bar{\mathcal{W}}_{train}$ and $\bar{\mathcal{W}}_{val}$ obtained during the 200 CVs. The numbers of informative baseline covariates used in computing the estimated treatment difference scores and nonzero regression coefficients of $\hat{\beta}_j$ for $\mathcal{M}_j^*(j = 1, 2)$ are also shown. Model 1 is an inverse-probability-weighted regression model

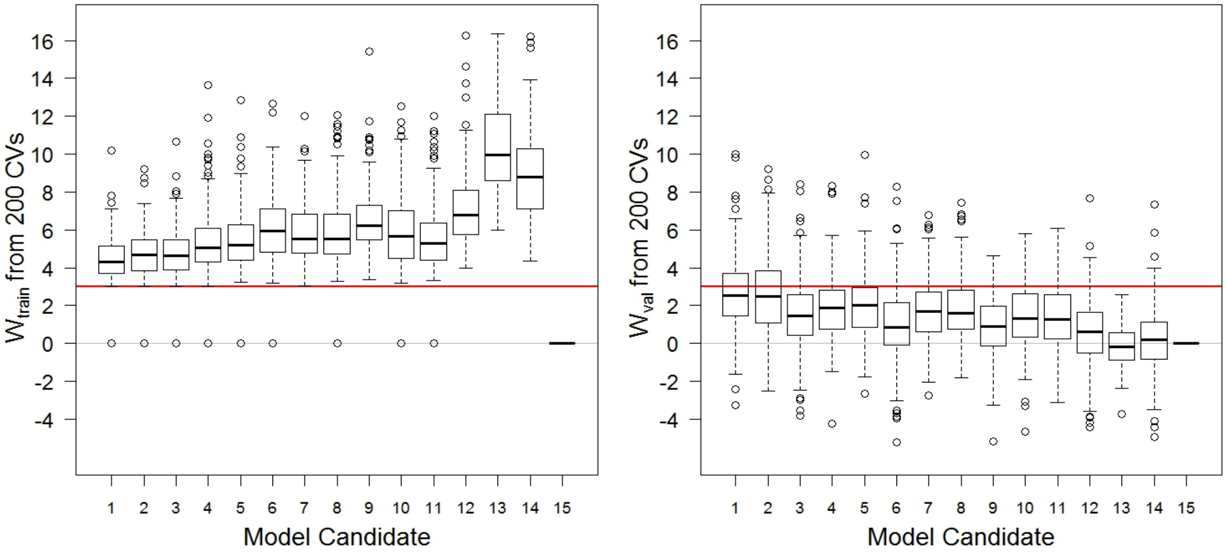


Figure 3.5: The weighted Successive Average Treatment Effects (weighted SATE) obtained from 200 cross-validations. The left and right panel respectively depict the results derived from part Ia training data and part Ib validation data.

on truncated Y at τ = 36 with the identity link. It has the smallest $\bar{\mathcal{L}}^*$ and similar $\bar{\mathcal{W}}_{val}$ of 2.6 as Model 2, and is our selected model to proceed. The final working models used to generate individual predicted RMET score had a patient been treated by Bucindolol is ($\hat{\mu}_1$):

$$11.36 - 7.41\,\text{NYHA} - 2.79\,\text{ischemic} + 3.55\,\text{smoke}_{ever} - 2.01\,\text{black} + 0.16\,\text{eGFR}. \tag{3.15}$$

The individual predicted RMET score had a patient been treated by placebo is ($\hat{\mu}_2$):

$$17.44 - 3.27\,\text{NYHA} - 1.77\,\text{ischemic} - 0.73\,\text{smoke}_{ever} - 2.90\,\text{black} + 0.08\,\text{eGFR}. \tag{3.16}$$

The PITES derived from $\hat{\mu}_2$ - $\hat{\mu}_1$ has a mean of 1.6 month and a range from -9.6 to 15.1 months in Part I dataset. Stratified the PITES by our DPS algorithm, three strata with cutoff points $\hat{c}_1$ = -3.7 and $\hat{c}_2$ = 5.1 months are obtained. To make inferences about the prediction of this selected final stratification scheme, we apply models (3.15) and (3.16) to the Part II holdout data. The mean estimated ITE scores is 1.7 month with a range from -9.5 to 17.5 months. Table 3.6 summarizes the stratum-specific treatment effects and the estimated successive average treatment effects d_k in Part I training data and that of Part II holdout data. The corresponding KM curves for the three strata with n_k^* = 61, 1096, and 197 for k = 1, 2, and 3 respectively are given in Figure 3.6. Applying the stratification scheme of $(\hat{c}_1, \hat{c}_2)$ = (-3.7, 5.1), models (5.1) and (5.2) to 1000 bootstrap holdout samples, the point estimates and 0.95 confidence intervals for the SATE are 2.7 (-4.6, 10.7), and 4.2 (0.3, 8.4) months. The $\mathcal{L}$ and $\mathcal{W}$ derived from the Part II data are 15.1 and 3.5 respectively. The baseline characteristics for the patients are shown in Table 3.7 by stratum and treatment. There is no significant difference in the characteristics shown between treatment and placebo groups within each stratum. However, across strata, Stratum 3 appears to have more males, more blacks, more obese as indicated by BMI>30, and much younger patients (a mean age of 49 vs around 70 years old in Stratum 1) with a very high percentage of ever smokers. There are almost no NYHA class IV classification in Stratum 3, and their kidney functions are fairly normal with a mean eGFR of around 108 vs around 37 for the older patients in Stratum 1 or 61 in Stratum 2. These results suggest a reasonable differentiation between Stratum 2 and Stratum 3. In particular, future patients with similar scores and characteristics as Stratum 3 patients fall into the "Should-Treat" quadrant in Table 3.1, such treatment are more likely to yield beneficial clinical benefit defined above in this target subpopulation.

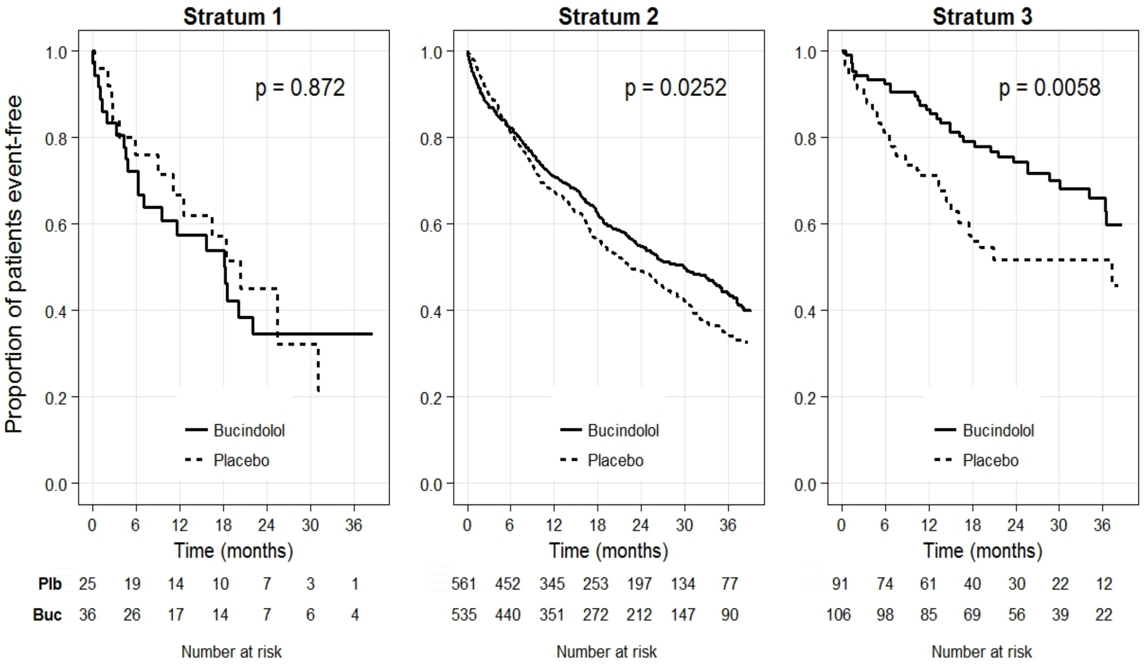


Figure 3.6: Stratum-specific Kaplan-Meier estimates obtained from Part II data of BEST study with $\hat{c}_1 = -3.7$ and $\hat{c}_2 = 5.1$.

Table 3.5: Study BEST regression model candidates, the complexities of $\mathcal{M}_j^*$ ($\|\hat{\beta}_j\|_0 =$ the number of nonzero components of $\hat{\beta}_j$ for j = 1, 2); and the results from 200 CV runs including $\bar{\mathcal{L}}^*$, $\bar{\mathcal{L}}^{**}$ and mean values of weighted SATE derived from Part Ia data ($\bar{\mathcal{W}}_{train}$) and Part Ib data ($\bar{\mathcal{W}}_{val}$).

Model	Treatment model $\mathcal{M}_1^*$		Control model $\mathcal{M}_2^*$		Results from 200 CV		
ID	# covariates	$\|\hat{\beta}_1\|_0$	# covariates	$\|\hat{\beta}_2\|_0$	$\bar{\mathcal{L}}^*$	$\bar{\mathcal{W}}_{train}$	$\bar{\mathcal{W}}_{val}$
1	5	5	5	5	15.59	4.42	2.59
2	5	5	1	1	15.62	4.71	2.60
3	5	5	5	5	15.60	4.75	1.53
4	16	16	5	5	16.05	5.44	1.92
5	15	15	5	5	16.02	5.43	2.02
6	15	15	12	12	15.99	5.97	1.03
7	18	18	18	18	16.27	5.89	1.73
8	18	18	18	18	16.25	5.95	1.90
9	18	18	18	18	16.14	6.51	0.90
10	16	16	17	22	16.24	5.95	1.47
11	17	17	14	15	16.20	5.61	1.43
12	15	16	16	17	16.17	7.19	0.57
13	18	150	18	150	18.10	10.48	-0.15
14	18	40	18	41	16.75	8.94	0.14
15	0	0	0	0	15.90	0	0

Table 3.6: Stratification results of Model 1, the empirical Kaplan-Meier estimated RMET for each strata, the corresponding treatment difference, and the estimated SATE values.

Data	Stratum k		1		2		3	
Part I	Treatment $\hat{\mu}_{1k}$	(#event/n_{1k})	12.4	(31/40)	21.8	(287/547)	30.8	(22/90)
(Training)	Control $\hat{\mu}_{2k}$	(#event/n_{2k})	14.2	(35/48)	20.4	(306/535)	24.0	(43/93)
	$\delta_k = \hat{\mu}_{1k} - \hat{\mu}_{2k}$		-1.8		1.4		6.8	
	$d_k = \delta_k - \delta_{k-1}$		NA		3.2		5.4	
Part II	Treatment $\hat{\mu}_{1k}$	(#event/n^*_{1k})	18.6	(21/36)	23.4	(267/535)	29.0	(32/106)
(holdout)	Control $\hat{\mu}_{2k}$	(#event/n^*_{2k})	19.7	(15/25)	21.8	(319/561)	23.2	(42/91)
	$\delta_k = \hat{\mu}_{1k} - \hat{\mu}_{2k}$		-1.1		1.6		5.8	
	95% Confidence Interval for δ_k		(-8.2, 6.0)		(-0.0, 3.3)		(2.1, 9.6)	
	$d_k = \delta_k - \delta_{k-1}$		NA		2.7		4.2	

Table 3.7: BEST Study: Baseline Characteristics of the patients in the Part II holdout data by strata and treatment.

Baseline	Stratum 1 ($n_1 = 61$)		Stratum 2 ($n_2 = 1096$)		Stratum 3 ($n_3 = 197$)	
Characteristic	TRT	Control	TRT	Control	TRT	Control
	$n_{11} = 36$	$n_{12} = 25$	$n_{21} = 535$	$n_{22} = 561$	$n_{31} = 106$	$n_{32} = 91$
Male sex n(%)	17 (47)	15 (60)	418 (78)	435 (78)	94 (89)	75 (82)
Age (year)	70 ± 11	69 ± 12	62 ± 11	62 ± 11	49 ± 11	49 ± 9
Black, not Hispanic n(%)	3 (8)	3 (12)	122 (23)	119 (21)	32 (30)	27 (30)
BMI>30 n(%)	7 (19)	5 (20)	154 (29)	147 (26)	61 (58)	53 (58)
Current smoker n(%)	1 (3)	1 (4)	97 (18)	81 (14)	28 (26)	23 (25)
Ever smoked n(%)	2 (6)	4 (16)	396 (74)	399 (71)	98 (92)	88 (97)
NYHA class IV n(%)	14 (39)	12 (48)	46 (9)	34 (6)	1 (1)	1 (1)
Ischemic n(%)	30 (83)	19 (76)	327 (61)	358 (64)	32 (30)	25 (27)
Hypertension n(%)	20 (56)	16 (64)	327 (61)	324 (58)	54 (51)	52 (57)
Diabetes n(%)	15 (42)	10 (40)	193 (36)	194 (35)	31 (29)	33 (36)
Heart rate	84 ± 17	81 ± 12	81 ± 13	81 ± 13	83 ± 13	83 ± 12
Systolic BP mmHg	115 ± 19	121 ± 22	118 ± 19	117 ± 18	118 ± 17	117 ± 15
LVEF (%)	23 ± 7	22 ± 7	23 ± 7	23 ± 7	24 ± 8	23 ± 7
Atrial fibrillation n(%)	4 (11)	6 (24)	54 (10)	72 (13)	16 (15)	7 (8)
eGFR	36 ± 14	37 ± 14	61 ± 20	61 ± 20	109 ± 30	108 ± 18

Plus-minus values are means ± SD. BMI denotes body-mass index; NYHA denotes New York Heart Association; LVEF denotes left ventricular ejection fraction; and eGFR denotes estimated Glomerular Filtration Rate.

3.6 Remarks

Stratified medicine, the grouping of patients based on disease risk or response of therapy (WHO, 2013), has tremendous potential to deliver more effective and efficient therapeutic intervention to improve public health. We worked on several aspects to promote its likelihood of implementation in clinical setting. Incorporating the concept of uplift modeling (a.k.a. true-lift modeling), an advanced data-mining subfield with successful business applications and established guidelines (Kane et al., 2014), we provided a framework to utilize baseline information to identify the subgroup(s) with most beneficial prospect; wasteful, harmful, and futile subgroups to save resources and reduce unnecessary exposure to treatment adverse effects. We tackled the issue from three fronts: 1) stratification; 2) model building and selection; and 3) reproducibility assessment. First, we proposed a stratification algorithm with constrained optimization by utilizing dynamic programming and supervised-learning techniques. Second, we proposed several metrics to evaluate the prediction performance during training and cross-validation stages to select the best model from a collection of competing scoring systems for the predicted individual treatment effect scores (PITES). Lastly, the final stratification system was evaluated using an independent holdout dataset to draw inferential conclusions.

Our objective is to predict and stratify patients into actionable categories by optimizing a pre-defined clinically meaningful benefit as the Successive Average Treatment Effect (SATE) for future treatment recommendations. Importantly, the existence of optimal classifier depends on the magnitude of SATE, the adequacy of the prediction model, the treatment effect, the number of observations, and the information available as reflected by the observed event rate in censored outcomes. It is possible that the study population cannot be further subgrouped with respect to the desirable clinically meaningful SATE if there is no heterogeneous treatment effect among the study population, the prediction model is inadequate, or if the desirable clinical benefit is higher than the reality. Furthermore, while our HIV study and cardiology examples did not contain genetic marker information, any prediction models with any baseline covariates can form a candidate PITES scoring system to compete. As pointed out in Yong et al. (2013), association can vary depending on what variables are put in a prediction model, it would be important to evaluate competing models with respect to our common final goal.

We incorporated reproducibility evaluation via an independent holdout dataset because we only have one such dataset for illustration while attempting to accomplish model building, selection, and evaluation for reproducibility. The ideal approach to evaluate the final model is to use an independent dataset that is not involved in the model building and selection stage. Otherwise, the true prediction error can be underestimated, sometimes substantially (Hastie et al., 2009; Siontis et al., 2015). For discussions of the holdout dataset rationale and cross-validation, the practical implementation and issues involved can be found in the cross-validation literature such as Schorfheide and Wolpin (2013); Rao and Fung (2008); and Esbensen and Geladi (2010). Our focus is to propose an analysis framework to enable some sort of reproducibility assessment in light of a growing concern of irreproducibility in scientific research (Ioannidis, 2005; Loscalzo, 2012; Collins and Tabak, 2014) and the emergence of preemptive medicine based on prediction. To this end, we also imposed a minimum stratum sample size constraint which can be relaxed to n_0=1. We found that the reproducibility performance is reduced when there are too many strata based on a handful of observations. In fact, the more personalized and unique the subgroups become via overfitting, the less likely the results can be reproduced without finding a bigger pool of subjects who behave like the training samples. Hence there is a trade-off between very personalized intervention strategy versus reproducibility. Fortunately, with the advent of technology and Big data era, finding patients with similar characteristics have become more achievable every day. Thus this type of intensive machine-learning techniques can be useful.

The data analytics team for the 2012 presidential election used uplift modeling to identify likely voters early on for fundraising and voter-mobilization efforts (Siegel, 2013b; Porter, 2013). In a similar spirit, early identification of sub-populations who are (or are not) likely to experience a treatment benefit can potentially save lives and resources, while alleviating adverse treatment effects. Our proposed concepts and procedures could be adapted in vast areas of application, ranging from identifying people with life-threatening diseases who can be offered more targeted treatment strategy, to finding those with risky behavior who can be trained to improve wellness. Intervention or therapeutic treatment programs can then be developed via multidisciplinary collaborations. With an ever-growing bank of available data, it would be important to extend similar method to observational data. More targeted treatment and cost-effective strategies can be developed to improve public health. The contribution of this paper is to provide an operational framework to bridge predictive modeling and decision making for more practical applications in stratified medicine.

Acknowledgments

We thank Dr. Brian Claggett for providing the datasets. This work was supported by NIH grant RC4CA155940 and a training grant (T32 CA 009337) from the National Cancer Institute.

A. Appendices

A.1 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

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

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

$$i^* = \underset{i}{\text{argmin}}\, c_j(i), i = 1, \cdots, m+1-j$$

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 $(\mathcal{G}_{\tilde{m}\tilde{j}}, L_{\tilde{m}\tilde{j}}), 1 \leq \tilde{m} \leq m$ and $1 \leq \tilde{j} \leq \tilde{m}$, 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 $j^* = \text{argmin}_j\, L_{nj}, j = 1, 2, \cdots, n$.

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.

A.2 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

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

and

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

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 $\mathrm{pr}(\mu(V_i) \in (c_{k-1}, c_k]) \geq \delta_0 > 0$. 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

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

and

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

which implies that

$$\sup_{(a,b,\beta)\in\Omega_0} \left| \frac{n^{-1}\sum_{i=1}^{n} Y_i I(\hat{g}(\beta' Z_i)\in(a,b])}{n^{-1}\sum_{i=1}^{n} I(\hat{g}(\beta' Z_i)\in(a,b])} - \mathrm{E}(Y|g(\beta' Z_i)\in(a,b]) \right| = o_p(1), \qquad \text{(A.2.1)}$$

where $\Omega_0 = \{a,b,\beta \mid \mathrm{pr}(g(\beta' Z_i)\in(a,b])\geq\delta_0, \max\{|a|,|b|,\|\beta\|_1\}<C_0\}$. 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 (A.2.1) that

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

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

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

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

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

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

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

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

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

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

$$\hat{\mathbf{c}}_\epsilon = \operatorname*{argmin}_{\mathbf{c}:\mathcal{S}_0(\mathbf{c}) \geq \epsilon} L_n(\mathbf{c}) \quad \text{and} \quad \mathbf{c}_\epsilon = \operatorname*{argmin}_{\mathbf{c}:\mathcal{S}_0(\mathbf{c}) \geq \epsilon} L(\mathbf{c}).$$

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 $\{(c_1, \cdots, c_{K_0}) \mid \mathcal{S}_0(\mathbf{c}) \geq 0\}$ 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 $\mathcal{S}_n(\mathbf{c}) - S(\mathbf{c}) = o_p(1)$,

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

which implies that

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

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

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

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.

A.3 An example to illustrate the dynamic programming stratification algorithm

Dynamic programming is employed to identify the optimal stratification scheme by finding the solution of similar subproblems through recursive computations and Bellman's principle of optimality (Bellman, 1952). To illustrate the concept, we first rank all observations in the ascending order of the Predicted Individual Treatment Effect Score (PITES) $\nu(\cdot)$. Using (2.2) and an optional pregrouping scheme that may group observations with similar values of $\nu(\cdot)$ (say, ±.0005), one can construct a hypothetical prediction error matrix containing the stratum-specific $\mathcal{L}_{ij}$ derived from each potential data combination by grouping observations from candidate stratum i to stratum j, for $i \leq j \leq n$:

Candidate Stratum Membership						Optimal solution
i \ j	1	2	3	4	5	$\mathcal{L}(i)$
1	0.5	0.1	0.4	0.2	0.6	0.5
2		0.4	0.2	0.7	0.5	0.4
3			∞	0.2	0.4	0.4
4				0.1	0.2	0.2
5					0.3	0.3

An ∞ denotes the cost of a stratum that violates the constraints. Let $\mathcal{L}(i)$ be the $\mathcal{L}$ of the optimal stratification scheme containing observations from stratum i to n(=5, in this example), the optimal solution can be obtained from the following stages:

1. $\mathcal{L}(5) = \mathcal{L}_{5,5} = 0.3$

2. $\mathcal{L}(4) = min\{\mathcal{L}_{4,4} + \mathcal{L}(5), \mathcal{L}_{4,5}\} = min\{0.1 + 0.3, 0.2\} = 0.2$

3. $\mathcal{L}(3) = min\{\mathcal{L}_{3,3} + \mathcal{L}(4), \mathcal{L}_{3,4} + \mathcal{L}(5), \mathcal{L}_{3,5}\} = min\{\infty, 0.2 + 0.3, 0.4\} = 0.4$

4. $\mathcal{L}(2) = min\{\mathcal{L}_{2,2} + \mathcal{L}(3), \mathcal{L}_{2,3} + \mathcal{L}(4), \mathcal{L}_{2,4} + \mathcal{L}(5), \mathcal{L}_{2,5}\}$
   $= min\{0.4 + 0.4, 0.2 + 0.2, 0.7 + 0.3, 0.5\} = 0.4$

5. $\mathcal{L}(1) = min\{\mathcal{L}_{1,1} + \mathcal{L}(2), \mathcal{L}_{1,2} + \mathcal{L}(3), \mathcal{L}_{1,3} + \mathcal{L}(4), \mathcal{L}_{1,4} + \mathcal{L}(5), \mathcal{L}_{1,5}\}$
$= min\{0.5 + 0.4, 0.1 + 0.4, 0.4 + 0.2, 0.2 + 0.3, 0.6\} = 0.5$

The optimal solution has an $\mathcal{L}$ of 0.5. There are two possible stratum combinations: 1) {1,2} and {3,4,5}; or 2) {1,2,3,4} and {5}. In this article, the first index that gives the smallest $\mathcal{L}(i)$ among all possible combinations is chosen as the optimal cutoff point. Hence the optimal stratification scheme is $\hat{c} = \hat{c}_1$ = the largest $\nu(\cdot)$ of all observations in {1,2}.

References

Acquah, H., and Carlo, M. (2010), "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship," *Journal of Development and Agricultural Economics*, 2(1), 001–006.

Akaike, H. (1974), "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, 19(6), 716–723.

Andersen, P., and Gill, R. (1982), "Cox's regression model for counting processes: a large sample study," *The annals of statistics*, 10(4), 1100–1120.

Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004), "Regression analysis of restricted mean survival time based on pseudo-observations," *Lifetime data analysis*, 10(4), 335–350.

Bellman, R. (1952), "On the theory of dynamic programming," *Proceedings of the National Academy of Sciences of the United States of America*, 38(8), 716.

BEST (2001), "A trial of the beta-blocker bucindolol in patients with advanced chronic heart failure," *The New England Journal of medicine*, 344(22), 1659. Beta-Blocker Evaluation of Survival Trial Investigators and others.

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.

Breiman, L. (1996), "Bagging predictors," *Machine learning*, 24(2), 123–140.

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

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

Claggett, B., Tian, L., Castagno, D., and Wei, L.-J. (2014), "Treatment selections using risk–benefit profiles based on data from comparative randomized clinical trials with multiple endpoints," *Biostatistics*, p. kxu037.

Collett, D. (2003), *Modelling survival data in medical research*, Vol. 57 Chapman & Hall/CRC.

Collins, F. S., and Tabak, L. A. (2014), "NIH plans to enhance reproducibility," *Nature*, 505(7485), 612.

Collins, F. S., and Varmus, H. (2015), "A new initiative on precision medicine," *New England Journal of Medicine*, 372(9), 793–795.

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

Esbensen, K. H., and Geladi, P. (2010), "Principles of Proper Validation: use and abuse of re-sampling for validation," *Journal of Chemometrics*, 24(3-4), 168–187.

Fan, J., and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96(456), 1348–1360.

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.

Fleming, T. R., and Harrington, D. P. (1991), *Counting Processes & Survival Analysis*, Applied probability and statistics Wiley, New York.

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

Gerds, T., and Schumacher, M. (2006), "Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times," *Biometrical Journal*, 48(6), 1029–1040.

Goeman, J. (2009), “L_1 penalized estimation in the Cox proportional hazards model,” *Biometrical Journal*, 52(1), 70–84.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999), “Assessment and comparison of prognostic classification schemes for survival data,” *Statistics in medicine*, 18(17-18), 2529–2545.

Grave, E., Obozinski, G. R., and Bach, F. R. (2011), Trace lasso: a trace norm regularization for correlated designs, in *Advances in Neural Information Processing Systems*, pp. 2187–2195.

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.

Harrell, F. E. (2001), *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer Science & Business Media.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009), *The elements of statistical learning*, Vol. 2 Springer.

Hingorani, A. D., van der Windt, D. A., Riley, R. D., Abrams, K., Moons, K. G., Steyerberg, E. W., Schroter, S., Sauerbrei, W., Altman, D. G., Hemingway, H. et al. (2013), “Prognosis research strategy (PROGRESS) 4: stratified medicine research,” *BMJ*, 346, e5793.

Holland, P. W. (1986), “Statistics and causal inference,” *Journal of the American statistical Association*, 81(396), 945–960.

Ioannidis, J. (2005), “Why most published research findings are false,” *PLoS medicine*, 2(8), e124.

Jaroszewicz, S., and Rzepakowski, P. (2014), "Uplift modeling with survival data". In: ACM SIGKDD Workshop on Health Informatics.

Jaskowski, M., and Jaroszewicz, S. (2012), "Uplift modeling for clinical trial data". In: ICML Workshop on Clinical Data Analysis, Edinburgh, Scotland.

Kane, K., Lo, V. S., and Zheng, J. (2014), "Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods," *Journal of Marketing Analytics*, 2(4), 218–238.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., and Scheike, T. H. (2013), *Handbook of Survival Analysis.* CRC Press.

Lo, V. S. (2002), "The true lift model: A novel data mining approach to response modeling in database marketing," *ACM SIGKDD Explorations Newsletter*, 4(2), 78–86.

Loscalzo, J. (2012), "Irreproducible Experimental Results Causes,(Mis) interpretations, and Consequences," *Circulation*, 125(10), 1211–1214.

Medical Research Council, U. K. (January 29, 2015), *Minister announces £14m investment in stratified medicine.* News, events and publications. Retrieved from `http://www.mrc.ac.uk/news-events/news/minister-announces-14m-investment-in-stratified-medicine/`, last accessed March 7, 2015.

Park, M., and Hastie, T. (2007), "L_1-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4), 659–677.

Pencina, M., and D'Agostino, R. (2004), "Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation," *Statistics in medicine*, 23(13), 2109–2123.

Porter, D. (2013), *Pinpointing the Persuadables: Convincing the right voters to support Barack Obama*. Presented at Predictive Analytics World; Oct, Boston, MA. Retrieved from http://www.predictiveanalyticsworld.com/patimes/pinpointing-the-persuadables-convincing-the-right-voters-to-support-barack-obama/, accessed 1 March 2013.

Radcliffe, N., and Surry, P. (1999), Differential response analysis: Modeling true response by isolating the effect of a single action, in *Proceedings of Credit Scoring and Credit Control VI, Credit Research Centre, University of Edinburgh Management School, Scotland*.

Rao, R. B., and Fung, G. (2008), On the Dangers of Cross-Validation. An Experimental Evaluation, in *Proceedings of the SIAM International Conference on Data Mining, SDM 2008, April 24-26, 2008, Atlanta, Georgia, USA*, pp. 588–596.

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.

Schorfheide, F., and Wolpin, K. (2013), To Hold Out or Not to Hold Out, National bureau of economic research working paper series, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania. Retrieved from http://www.nber.org/papers/w19565.

Schwarz, G. (1978), "Estimating the dimension of a model," *The annals of statistics*, 6(2), 461–464.

Siegel, E. (2013a), *The real story behind Obama's election victory.* The Fiscal Times January 21 2013. Retrieved from http://www.thefiscaltimes.com/Articles/2013/01/21/The-Real-Story-Behind-Obamas-Election-Victory.aspx#page1, last accessed March 7, 2015.

Siegel, E. (2013b), *Predictive analytics: The power to predict who will click, buy, lie, or die.* John Wiley & Sons. Chapter 7. Persuasion by the Numbers. pp.187-217.

Simmons, J., Nelson, L., and Simonsohn, U. (2011), "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22(11), 1359–1366.

Siontis, G. C., Tzoulaki, I., Castaldi, P. J., and Ioannidis, J. P. (2015), "External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination," *Journal of clinical epidemiology*, 68(1), 25–34.

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.

Sołtys, M., Jaroszewicz, S., and Rzepakowski, P. (2014), "Ensemble methods for uplift modeling," *Data Mining and Knowledge Discovery*, pp. 1–29.

Štajduhar, I., and Dalbelo-Bašić, B. (2012), "Uncensoring censored data for machine learning: A likelihood-based approach," *Expert Systems with Applications*, 39(8), 7226–7234.

Steyerberg, E., Vickers, A., Cook, N., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M., and Kattan, M. (2010), "Assessing the performance of prediction models: a framework for traditional and novel measures," *Epidemiology*, 21(1), 128–138.

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

The White House, U. S. (January 30, 2015), *FACT SHEET: President Obama's Precision Medicine Initiative.* Office of the Press Secretary. Retrieved from `http://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative`, last accessed March 7, 2015.

Therneau, T. M., and Grambsch, P. M. (2000), *Modeling Survival Data: Extending the Cox Model* Springer, New York.

Tian, L., Cai, T., Goetghebeur, E., and Wei, L. (2007), "Model evaluation based on the sampling distribution of estimated absolute prediction error," *Biometrika*, 94(2), 297–311.

Tian, L., Zhao, L., and Wei, L. (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)*, pp. 267–288.

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

Trusheim, M. R., Berndt, E. R., and Douglas, F. L. (2007), "Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers," *Nature Reviews Drug Discovery*, 6(4), 287–293.

Uno, H., Cai, T., Pencina, M., D'Agostino, R., and Wei, L. (2011), "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in medicine*, 30(10), 1105–1117.

van Houwelingen, H. C., Bruinsma, T., Hart, A. A. M., van't Veer, L. J., and Wessels, L. F. A. (2006), "Cross-validated Cox regression on microarray gene expression data," *Statistics in Medicine*, 25, 3201–3216.

van Houwelingen, H. C., and Putter, H. (2008), *Dynamic Prediction in Clinical Survival Analysis*, Monographs on statistics and applied probability CRC Press, Boca Raton.

Van Houwelingen, J. (2001), "Shrinkage and penalized likelihood as methods to improve predictive accuracy," *Statistica Neerlandica*, 55(1), 17–34.

Verweij, P., and Van Houwelingen, H. (2006), "Penalized likelihood in Cox regression," *Statistics in Medicine*, 13(23-24), 2427–2436.

WHO (2013), *Priority Medicines for Europe and the World Update Report, Chapter 7.4 Stratified medicine and pharmacogenomics.* World Health Organization. Retrieved from `http://www.who.int/medicines/areas/priority_medicines/Ch7_4Stratified.pdf`, last accessed March 7, 2015.

Witten, D., and Tibshirani, R. (2010), "Survival analysis with high-dimensional covariates," *Statistical methods in medical research*, 19(1), 29–51.

Wu, T., and Lange, K. (2008), "Coordinate descent algorithms for lasso penalized regression," *The Annals of Applied Statistics*, pp. 224–244.

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.

Yong, F. H., Tian, L., Yu, S., Cai, T., and Wei, L. (2014), "Predicting the Future Subject's Outcome via an Optimal Stratification Procedure with Baseline Information," *Harvard University Biostatistics Working Paper Series. Working Paper*, (173). `http://biostats.bepress.com/harvardbiostat/paper173`.

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.

Zou, H., and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.

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.