Methods for Estimation and Evaluation of Marker-Guided Treatment Rules Based on Multivariate Marker Panels

Veronika Skrivankova

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Patrick J. Heagerty, Chair

Charles L. Kooperberg

Ali Shojaie

Program Authorized to Offer Degree:
Public Health, Biostatistics

University of Washington

Abstract

Methods for Estimation and Evaluation
of Marker-Guided Treatment Rules
Based on Multivariate Marker Panels

Veronika Skrivankova

Chair of the Supervisory Committee:
Ph.D. Patrick J. Heagerty
Biostatistics

Due to vast heterogeneity in patients' responses, a uniformly preferred treatment is often not available. In such cases, clinical practice may be enhanced by use of person-level information that could guide treatment choice and lead to better outcomes for both treated individuals and for the population. The scientific challenge is to identify those factors that can be used to target treatment, and to accurately quantify the expected treatment benefit as a function of candidate markers. The proposed research develops statistical methodology for the generation and evaluation of a single index score, that estimates the expected treatment benefit associated with patient characteristics as a linear combination of the markers. Our methods specifically decouple the model used to generate the treatment benefit score from non-parametric methods that are adopted to evaluate the score. Cross-validation methods ensure honest evaluation of the score performance at the population level, if it was used to guide treatment. We also show that the treatment benefit score can be used for selecting a subset of patients with enriched treatment response. The methods are illustrated on multiple examples, including data

from a randomized trial of steroid injections where baseline clinical and imaging data are candidate measures for guiding the therapy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost I would like to express my deepest gratitude to my adviser, Dr. Patrick Heagerty, for his excellent guidance, caring, patience, and providing me with such an inspirational atmosphere for doing research. He has been exceptionally supportive from the very beginning until the very end, both inside and outside academia, and I have always felt highly privileged to have him as my adviser. Thank you, Patrick, for encouraging me to follow my passions and dreams, and for being the most wonderful adviser one could wish for, my mentor and friend.

I sincerely thank my committee members Drs. Charles Kooperberg, Ali Shojaie and Ken Rice for additional support and guidance throughout the process of developing my research. Their valuable comments and suggestions greatly helped to shape my thesis.

It would not have been possible to write my dissertation without financial, academic and technical support of the Department of Biostatistics at the University of Washington. I would also like to thank all its faculty and staff for creating a friendly, stimulating community and making me feel so welcome in the graduate program.

This work would not be complete without real data examples provided with the support of grants NIH NHGRI U01 HG005157 (GARNET VISP study data), AHRQ R01 HS019222 (LESS data) and 1U1CA180819 (SWOG Multiple Myeloma data). My special thanks goes to Dr. Bryan Comstock for his patient willingness to clear up all my questions and helping me to navigate through the maze of variables, and to Dr. Michael LeBlanc for helping me obtain access to Multiple Myeloma data.

I cannot thank enough to our fantastic Heagerty Working Group – especially Aasthaa Bansal, Leila Zelnick, Jason Liang and Katherine Tan – who enabled me to endlessly explain my research and practice my talks on them. Thank you so much for your support over the last couple of years, your sharp feedbacks, inspiring discussions and for your friendships.

Throughout my years in Seattle, I have met some of the most wonderful people in my life, who helped me maintain my work-life balance, and whom I am so grateful to call friends. Thank you all very much for your cheerful distractions, your compassion, for being my undying source of energy and inspiration, and being here for me all this time.

Finally, I would like to thank my parents for supporting me throughout all my studies at University, encouraging me with their kind words and best wishes, and for providing a home where I could always return.

DEDICATION

To all the people who are trying to make this world a better place.

Chapter 1

BACKGROUND LITERATURE OVERVIEW

1.1 Introduction

Many recent studies of genome-wide markers have sought to identify associations between individual markers such as single nucleotide polymorphisms (SNPs) and key health outcomes such as disease onset. Genome-wide association studies have lead to novel discoveries that have provided insight into the biological pathways associated with both disease onset and progression (Urabe et al., 1997; Peters et al., 2013; Rafiq et al., 2013). In addition, individual markers have been used prognostically to identify subjects who are at increased risk to develop disease, and sets of markers have been combined to construct risk prediction models where the goal is to best predict future risk of disease occurrence or future health impairment characterized using continuous scales (Jacobs Jr et al., 1999; de Mendonça et al., 2000; Inci et al., 2013).

While development of prognostic models is important toward disease prevention or individual counseling, it is becoming increasingly more attractive to investigate whether candidate biomarkers can predict a diseased patient's response to alternative treatment choices. Heterogeneity in patients' responses implies that a uniformly preferred treatment is often not available. In such cases, clinical practice can potentially be enhanced by use of person-level information to guide treatment choice when correlates of response to treatment have been identified. Specifically, if patient characteristics such as demo-

graphics, serum protein markers, or genetic markers can establish those patients who are more likely to respond to certain therapies, then targeted treatment algorithms could be applied and lead to both better outcomes for individuals and for the population of treated patients.

Large randomized clinical trials offer the potential to evaluate biomarkers as predictors of treatment response, and to develop individualized rules to guide treatment. In pharmacogenetic studies it is common to simply focus on treated subjects and then to identify predictors of a favorable outcome (Daly, 2010). Additional analytical approaches include testing the treatment effect in a patient subgroup defined by restriction to a select range of marker values (Freidlin et al., 2010) or testing the marker-treatment interaction (Matsui et al., 2012) using both treated and non-treated subjects. However, ultimate clinical use of a marker leads to consideration of the decision making potential of the marker, and associated statistical methods that characterize the accuracy of medical decision, or the population consequence of using a marker-guided strategy are needed.

Recent statistical methods have focused on the assessment of marker-specific treatment benefit, defined as the difference between expected outcomes associated with a particular treatment for a subgroup defined by the marker (Gunter et al., 2011; Janes et al., 2011). The ultimate goal of individulized or guided treatment is to prescribe each patient with a therapeutic option that leads to the largest expected benefit for the individual. The scientific challenge is to identify those factors that can be used to target treatment, and to accurately quantify the expected treatment benefit as a function of the candidate markers. Available patient information is limited to the variables collected during the trial but often biospecimens are available and can be used to obtain

high-dimensional predictive information. The primary statistical objective is to carefully evaluate candidate information for the potential to reproducibly direct individual treatment toward optimal performance at both the individual and population level.

Statistical methods that can clearly distinguish between groups of patients with differential treatment benefit profiles are particularly of interest in cases where available treatment options are weak (or null), or where enthusiasm for the uniform delivery of treatment would be low due to cost and/or side effects. In these situations if the treatment strongly and qualitatively interacts with one or multiple markers, then it might be plausible to identify a subgroup of patients who would have a large treatment benefit despite a small marginal or overall treatment effect size.

The proposed research will develop statistical methodology for the evaluation and comparison of candidate prescriptive models with focus on the use of high-dimensional biomedical information. Chapter 1 summarizes current literature and key framework for development of marker-guided treatment decision rules. In Chapter 2, we introduce our approach to decision rule development for continuous outcomes and establish an algorithm for marker panel selection that will honestly evaluate and compare derived decision rules with respect to the improvement of the mean outcome in the population. Chapter 3 describes the extension of our methods to settings with survival outcomes. In Chapter 4, we detail and compare design characteristics for a follow-up evaluation study that would assess how a treatment benefit score performs as a predictor of response and as a basis for treatment choice. Finally, in Chapter 5, we outline some of possible directions for future work.

Throughout this document, we will use the following format of notation: upper case

letters denote random variables (e.g., $\boldsymbol{X}, Y$) and lower case letters denote specific values from the value spaces of random variables (e.g., $\boldsymbol{x}$, y), while bold case denotes a vector or matrix (e.g., vector of predictors $\boldsymbol{X}$, $\boldsymbol{x}$) and regular case denotes a single-dimensional variable or value (e.g., the outcome Y, y). The individual predictors are indexed by the letter j (X_j; $j = 1 \ldots, m$), and the letter i is used to index patients and their observed values ($\boldsymbol{x}_i$, y_i; $i = 1, \ldots, n$).

1.2 Prognostic Markers

Biomarkers can be used to diagnose disease, to predict future disease, or to identify individuals who benefit from treatment. In current practice, biomarker-based prognostic models are used to identify high-risk or susceptible patients with a focus on identifying those subjects who are likely to advance to a disease or disabling state if otherwise not treated. In many application such high-risk subjects are then targeted with prevention strategies or with aggressive treatment options (Aaronson et al., 1997; Kamath et al., 2001). The essential statistical property of a prognostic marker is that it can accurately predict a future patient status. Implicit in the use of prognostic markers to choose treatment is the assumption that all patients are better off being treated since no evaluation of treated patients is used for prognostic marker evaluation. Similarly, it is common in pharmacogenomic studies to focus on patient response to treatment and therefore evaluate a marker as a predictor among treated patients. Again such an evaluation makes the implicit assumption that untreated subjects have a poor prognosis.

Statistical methods to evaluate prognostic markers are well established and focus on measures of performance such as minimizing the mean squared error (MSE), or maximizing classification accuracy in out-of-sample performance (Kooperberg et al., 2010).

However, many biomedical scenarios need to consider the performance of a marker as a predictor of both treated and untreated subjects in order to accurately determine if marker-defined subgroups are expected to have a positive treatment benefit.

1.3 Prescriptive Markers

Often, a primary goal of medical research is to determine whether a treatment leads to improvement in a patient outcome or condition. Randomized clinical trials are the standard for establishing whether treatment is effective on average for indicated patients. For those treatments with a weak overall benefit it can be important to evaluate whether certain subgroups of patients with a strong response exist. The utility of a marker in selection of treatment or identifying subgroups has been evaluated by either assessing prognostic performance in treated or untreated subjects, by testing the treatment effect in a subgroup defined by marker values (Freidlin et al., 2010), or by testing the marker-treatment interaction (Matsui et al., 2012).

However, strong interactions alone are not sufficient for a marker to be a good predictor of utility, since marker-based decisions about treatment might be altered for only a small subset of patients (Gunter et al., 2011; Janes et al., 2011). Recent literature has suggested evaluation of prescriptive markers based on criteria that require both large interactions and significant proportions of subjects whose optimal decision about treatment is impacted by the marker (Gunter et al., 2011). In addition to evaluation of individual candidate markers, a summary score has been proposed consisting of a linear combination of k most highly ranked markers and then evaluated with respect to estimated mean response in the population when using the score to guide treatment (Gunter et al., 2011). Discussion by Janes et al. (2011) illustrates why none of the common earlier

approaches provides the necessary information about whether a marker should be used to select treatment. Janes et al. (2011) propose to plot a marker-by-treatment predictiveness curves in order to determine the population effect that results from using a marker to select treatment. Evaluation of performance for alternative marker thresholds can determine an optimal binary treatment choice function. The authors also suggest that the predictiveness curves can be particularly useful for comparing the relative performance of different candidate treatment selection markers.

1.4 Accuracy

Huang et al. (2012) and Sitlani and Heagerty (2014) characterized prescriptive usefulness of a marker by how accurately it can distinguish between patients with positive (cases) and negative (controls) expected benefit from the treatment. In order to estimate classification accuracy at the individual level, one must know subject-specific differences in potential outcomes that would be achieved with and without treatment ("principal strata"). The actual subject-specific treatment effects are however unobservable. Alternatively, (Sitlani and Heagerty, 2014) proposed to use longitudinal data with (multiple) crossover treatment periods, for which an expected or time-averaged subject-specific treatment response can be estimated. A "principal" receiver operating characteristic (p-ROC) curve for the subject-specific treatment response can then be constructed to assess the discriminatory ability of the marker to properly classify, or to separate cases from controls. Evaluation using p-ROC curves requires estimation of conditional probability of case/control status given marker, based on the longitudinal data and achieved through specification of a parametric structural mixed model (LSMM) (Sitlani et al., 2012).

1.5 Population Performance: Continuous Outcome

In settings where only a single outcome per patient is available, prescriptive methods can rely on population-level measures of performance. Descriptive measures to evaluate the potential impact of marker-based treatment selection algorithms are graphical displays that show the population mean under various marker guided treatment rules (Janes et al., 2011; Matsui et al., 2012; Song and Pepe, 2004). A typical goal of prescriptive methods is to develop a decision rule, $d : \mathcal{X} \rightarrow \{0, 1\}$, that maps covariates, $\boldsymbol{X}$, into a treatment or action, A, with $\mathcal{X}$ denoting the space of possible values of covariate vector $\boldsymbol{X}$. The decision rule determines the treatment in an individualized fashion and ideally will yield an overall population mean that is better than static or uniform treatment guidelines (Gunter et al., 2011; Zhang et al., 2012).

The focus is on identification of population subgroups that benefit from a particular treatment A, taking on values 1 (treated) or 0 (non-treated). The classification of patients is based on their baseline characteristics $\boldsymbol{X}$, measured before the treatment is administered, and Y denotes a clinical outcome of interest. Without a loss of generality, larger values of Y are considered "better" and hence, improving the clinical outcome means maximizing Y over available therapies.

In a context of binary treatment, $A \in \{0, 1\}$, and following treatment decision rule d, a patient with covariates $\boldsymbol{X} = \boldsymbol{x}$ would be recommended to receive treatment 1 if $d(\boldsymbol{x}) = 1$ and treatment 0 if $d(\boldsymbol{x}) = 0$. The expected continuous outcome for a patient with covariates $\boldsymbol{X} = \boldsymbol{x}$ under treatment $A = a$ can be written as

$$\mathrm{E}[Y|\, \boldsymbol{X} = \boldsymbol{x}, A = a] = \mathrm{E}[Y|\, \boldsymbol{X} = \boldsymbol{x}, A = 0] + a\, \Delta(\boldsymbol{x})\,,$$

where $\Delta(\boldsymbol{x}) = \mathrm{E}[Y|\,\boldsymbol{X} = \boldsymbol{x}, A = 1] - \mathrm{E}[Y|\,\boldsymbol{X} = \boldsymbol{x}, A = 0]$ is a covariate-specific treatment benefit. A positive estimated treatment benefit $\Delta(\boldsymbol{x}) > 0$ implies a higher expected outcome if the treatment is administered ($A = 1$).

An optimal regimen on $\mathcal{X}$ is defined as one that leads the largest expected outcome from among all regimens on $\mathcal{X}$, for every set of covariates $\boldsymbol{x}$,

$$d^*(\boldsymbol{x}) = \arg\max_d \mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}, A = d(\boldsymbol{x})],\ \forall \boldsymbol{x} \in \mathcal{X}\,,$$

and hence equivalently in the whole population, $d^* = \arg\max_d \mathrm{E}_{\boldsymbol{X}}\{\mathrm{E}_{Y|\boldsymbol{X}}[Y|\,\boldsymbol{X}, A = d(\boldsymbol{X})]\,$. Since the outcome of a patient with covariates $\boldsymbol{X} = \boldsymbol{x}$ is better under the treatment $A = \mathrm{I}\,[\Delta(\boldsymbol{x}) > 0]$, an optimal regimen is then $d^*(\boldsymbol{x}) = \mathrm{I}\,[\Delta(\boldsymbol{x}) > 0]$.

Unfortunately, the individual-level quantities $\Delta(\boldsymbol{x}) = \mathrm{E}[Y|\,\boldsymbol{X} = \boldsymbol{x}, A = 1] - \mathrm{E}[Y|\,\boldsymbol{X} = \boldsymbol{x}, A = 0]$ are unobservable, and so the ultimate optimal rule is not applicable. However, if one can specify a reasonable model for $\Delta(\boldsymbol{X})$, an optimal rule within the class of regimens induced by this model would be still of interest to estimate.

For example, if we assumed a linear additive model for the treatment benefit, $\Delta(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\gamma}$, an optimal decision rule implied by this model would be of the form $d^*(\boldsymbol{X}) = \mathrm{I}\,[\boldsymbol{X}\boldsymbol{\gamma} > 0]$, and its estimation is equivalent to estimation of the parameters $\boldsymbol{\gamma}$. Hence a class of regimens induced by such model is defined as a set of all decision rules that have this form,

$$\mathcal{D} = \{d_{\boldsymbol{\gamma}} : d_{\boldsymbol{\gamma}}(\boldsymbol{X}) = \mathrm{I}\,[\boldsymbol{X}\boldsymbol{\gamma} > 0];\, \boldsymbol{\gamma} \in \boldsymbol{\Gamma} \subseteq \mathbb{R}^m\}\,,$$

where m is the number of covariates. Then, every vector $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ defines one regimen from the class $\mathcal{D}$, and an optimal regimen within the class is such that leads the maximum

mean outcome in the population from amongst all the regimens in $\mathcal{D}$. If we define

$$\mathrm{E}\, Y(d) = \mathrm{E}_{\boldsymbol{X}} \left\{ \mathrm{E}_{Y|\boldsymbol{X},A}[Y|\, \boldsymbol{X}, A = d(\boldsymbol{X})] \right\}$$

to be the mean outcome in the population under a scenario where everybody is assigned to the treatment according to the regimen d, then we can write

$$d^* = \arg\max_{d \in \mathcal{D}} \mathrm{E}\, Y(d)\,,$$

or alternatively, $d^* = d_{\boldsymbol{\gamma}^*}$, where

$$\boldsymbol{\gamma}^* = \arg\max_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} \mathrm{E}\, Y(d_{\boldsymbol{\gamma}})\,.$$

One approach to estimate an optimal decision rule is based on fitting a prediction model for $\mathrm{E}\, Y(d_{\boldsymbol{\gamma}})$. However, minimizing the prediction error might not necessarily result in a regimen that best improves the mean outcome in the population, $\mathrm{E}\, Y(d)$, among all regimens in $\mathcal{D}$, particularly when the model for $\Delta(\boldsymbol{X})$ is misspecified (Qian and Murphy, 2011; Zhang et al., 2012). This mismatch between the loss functions (weighted 0-1 loss and the quadratic loss) was illustrated by Qian and Murphy (2011) on the following simple example, where the chosen prediction model misspecifies the treatment benefit, $\Delta(\boldsymbol{X})$.

Example. Suppose X is uniformly distributed in $[-1, 1]$, A is binary $\{-1, 1\}$ with probability $\frac{1}{2}$ each and is independent of X, and outcome Y is normally distributed with mean $\mathrm{E}(Y|X, A) = A(X - \frac{1}{3})^2$ and variance 1. It is easy to see that the optimal decision

rule satisfies $d_0(X) = 1$ a.s. and $\mathrm{E}Y(d_0) = \frac{4}{9}$. If we consider a linear mean model with interaction between A and X, $\mathrm{E}(Y|X, A) = \theta_1 + \theta_2 X + \theta_3 A + \theta_4 AX$, the class of rules induced by this model is

$$\mathcal{D} = \{d; d(X) = \mathrm{I}\,(\theta_3 + \theta_4 X > 0);\, \theta_3, \theta_4 \in \mathbb{R}\}$$

Note that $d_0 \in \mathcal{D}$ since it can be written as $\mathrm{I}\,(\theta_3 + \theta_4 X > 0)$ for any $\theta_3 > 0$ and $\theta_4 = 0$. However, minimizing the prediction error L_2 yields $\hat{d}^*(X) = \mathrm{I}\,[\frac{2}{3} - X]$, which leads to lower mean population outcome than d_0.

1.6 Estimating Optimal Treatment Regimens: Low-dimensional Markers

An alternative approach to minimizing predictive MSE for model selection is to maximize an estimated mean outcome in the population under all possible decision rules from $\mathcal{D}$, or equivalently for $\forall \boldsymbol{\gamma} \in \boldsymbol{\Gamma}$. In practice it means that $\mathrm{E}\,Y(d_{\boldsymbol{\gamma}})$ needs to be estimated for every $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ and then $\hat{\boldsymbol{\gamma}}^*$ is chosen such that leads the highest $\widehat{\mathrm{E}}\,Y(d_{\boldsymbol{\gamma}})$.

However, the quality of d^* approximation is sensitive to the choice of the estimator for $\mathrm{E}\,Y(d_{\boldsymbol{\gamma}})$. Zhang et al. (2012) illustrated performance of three different estimators of $\mathrm{E}\,Y(d)$ when the search for the optimal treatment is done throughout the whole parameter space $\boldsymbol{\Gamma}$. The class of all regimens is induced by a linear outcome model $Y = \boldsymbol{X}\boldsymbol{\beta} + A\boldsymbol{X}\boldsymbol{\gamma}$, which assumes $\Delta(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\gamma}$ and that $\mathrm{I}\,[\boldsymbol{X}\boldsymbol{\gamma} > 0]$ would optimize $\mathrm{E}Y(d_{\boldsymbol{\gamma}})$. In their example, the model involves interaction between the treatment and covariates X_1 and X_2,

$$\mathrm{E}[Y|\,\boldsymbol{X}, A] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_0 A + \gamma_1 X_1 A + \gamma_2 X_2 A\,,$$

and an optimal treatment regimen within the model-induced class has the form $d(\boldsymbol{X}) = \mathrm{I}\left[\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 > 0\right] = \mathrm{I}\left[\alpha_0 + \alpha_1 X_1 > X_2\right]$, where $\alpha_0 = \gamma_0/\gamma_2$ and $\alpha_1 = \gamma_1/\gamma_2$. Hence, all possible values of $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ define the class of regimens induced by this model. Every value of $\boldsymbol{\alpha}$ corresponds to a unique regimen, d_α, and thus possibly leads to a different mean outcome in the population, $\mathrm{E}\, Y(d_\alpha)$.

It can be easily shown that using the ordinary least square (OLS) estimator for the mean outcome, that assumes the same model as $\mathcal{D}$, leads to the optimal regimen with $\hat{\boldsymbol{\alpha}}^*$ corresponding exactly to the coefficient estimates from the regression model (Zhang et al. (2012)). This property saves the exhaustive search through the whole parameter space. Nevertheless, the authors advocate to use the inverse probability weighted estimator (IPWE) or augmented (A)IPWE to estimate $\mathrm{E}\, Y(d_\alpha)$ for any given d_α, which is then to be maximized with respect to the parameter(s) implied by the model,

$$\hat{d}^*_\alpha(\boldsymbol{X}) = d_{\hat{\alpha}^*}(\boldsymbol{X}), \quad \text{where} \quad \hat{\alpha}^* = \arg\max_{\alpha \in \mathbb{R}^2} \widehat{\mathrm{E}}\, Y(d_\alpha)\,.$$

For a given regimen d_α, the IPWE averages observed outcomes only for the patients who received treatment equivalent to what d_α would prescribe them, weighted by their probability of receiving that treatment.

$$\widehat{\mathrm{E}}_{\mathrm{IPWE}} Y(d_\alpha) = n^{-1} \sum_{i=1}^{n} \frac{C_{\alpha,\,i} Y_i}{\hat{\pi}_c(\boldsymbol{X}_i; \alpha)}\,,$$

where $C_{\alpha,\,i} = A\, d_\alpha(\boldsymbol{X}_i) + (1 - A)(1 - d_\alpha(\boldsymbol{X}_i))$ is an indicator of consistency between observed and prescribed treatment and $\pi_c(\boldsymbol{X}_i; \alpha)$ is the probability of $C_{\alpha,\,i} = 1$. In case of data obtained from a randomized clinical trial, the propensity weights π_c are

(known) constants, but they need to be estimated in case of observational data. For the latter, it is common to posit a logistic model on π_c, which may, however, be subject to misspecification.

The AIPWE approach additionally incorporates the outcomes also for the patients who received treatment different from what d_α would prescribe them, as fitted by a regression model that needs to be speciifed. The AIPWE estimator then equals

$$\widehat{\mathrm{E}}_{\mathrm{AIPWE}} Y(d_\alpha) = n^{-1} \sum_{i=1}^{n} \left(\frac{C_{\alpha,i} Y_i}{\hat{\pi}_c(\boldsymbol{X}_i; \alpha)} - \frac{C_{\alpha,i} - \hat{\pi}_c(\boldsymbol{X}_i; \alpha)}{\hat{\pi}_c(\boldsymbol{X}_i; \alpha)} \widehat{\mathrm{E}}(Y | \boldsymbol{X}_i, A = d_\alpha(\boldsymbol{X}_i); \hat{\beta}) \right) ,$$

where β are coefficients of the assumed regression model for $\mathrm{E}(Y|\boldsymbol{X}, A)$. While the consistency of the regression estimator leans on the correct specification of the model for $\mathrm{E}(Y|\boldsymbol{X}, A)$, and the IPWE estimator requires a correct model for π_c, the AIPWE estimator offers protection against one of these two model misspecifications. It follows from Robins et al. (1994) and Cao et al. (2009) that the augmented estimator has also an increased asymptotic efficiency, and leads to better performance of the estimated optimal regimens, as shown by the simulations in Zhang et al. (2012).

The downside of this approach is that the maximization of $\widehat{\mathrm{E}}\, Y(d_\alpha)$ over the whole parameter space becomes computationally very extensive even for a moderately large number of markers. The dimension of the parameter space in the simulations presented in Zhang et al. (2012) was only 2, and the authors report that the estimation of an optimal regimen based on 8-dimensional parameter took more than 2 minutes. Since the computational time grows exponentially with the number of parameters, hundreds or thousands of markers involved in the decision making function would be unfeasible

to evaluate with this approach.

The inverse probability estimators also have their disadvantages. While AIPWE requires a specification of full regression model for $\mathrm{E}(Y|\boldsymbol{X}, A)$ and the estimation of all its coefficients, the IPWE uses for the estimation of $\mathrm{E}\,Y(d)$ only a subset of the patients who received treatment equivalent to what d would prescribe them. The fluctuation of $\widehat{\mathrm{E}}\,Y(d)$ across regimens is then largely attributable to the difference in subgroups of patients used for estimating $\mathrm{E}\,Y(d)$. Hence, the estimator that does not use the full set of observations suffers from increased variability and lack of efficiency. Therefore, additional statistical methods are needed for analysis of high-dimensional candidate predictors.

1.7 High-dimensional Markers

Analysis of high-dimensional biomedical data often face the problem of many potential spurious variables when only a small portion of the available markers is expected to be useful. Even with OLS estimator, which doesn't require a search through the whole parameter space, identifiability and overestimation become an issue when the number of variables largely exceeds the number of subjects. For treatment decision function estimation, common approaches to reduce the number of candidate models involve ordering markers with respect to their marginal interactions with the treatment (Gunter et al., 2011; Matsui et al., 2012). The result is a sequence of marginally ordered markers and associated nested models when including the top p ranking markers, $p = 1, 2, \ldots$. Such a decrease in the number of candidate models from 2^m to m, or less, is especially desirable when m is very large (e.g., in case of genomic data) and evaluation of all possible models would not be feasible.

Gunter et al. (2011) suggest to select prescriptive markers based on criteria that require both large interactions and large proportions of subjects whose optimal decision about treatment is influenced by the marker. They argue that the standard techniques developed to enhance prediction often downplay the importance of interaction variables, which are key to decision making. Through two different approaches, they rank variables with respect to their marginal "usefulness" to guide the treatment. The quality of a score consisting of a linear combination of k most highly ranked markers is then evaluated with respect to estimated mean response in the population, subject to penalization for the number of markers that compose the score.

A marker X is said to have a *qualitative interaction* with the treatment if the expected outcome across levels of X is not always maximized by the same treatment:

$$\exists\, x_1, x_2 \in \mathcal{X} : \text{argmax}_a \text{E}[Y|X = x_1, A = a] \neq \text{E}[Y|X = x_2, A = a]\,,$$

i.e., decision about which treatment is better for a patient depends on their level of the marker X. The degree to which a variable X_j could be useful in prescribing the treatment is captured by two factors, a marginal magnitude of the interaction,

$$\begin{aligned} D_j &= \max_i(\hat{\text{E}}[Y|X_j = x_{ij}, A = d_0^*] - \hat{\text{E}}[Y|X_j = x_{ij}, A \neq d_0^*]) \\ &- \min_i(\hat{\text{E}}[Y|X_j = x_{ij}, A = d_0^*] - \hat{\text{E}}[Y|X_j = x_{ij}, A \neq d_0^*])\,, \end{aligned}$$

and a proportion of patients affected by knowledge of the variable,

$$P_j = \frac{1}{n}\sum_{i=1}^{n} \text{I}\left[\arg\max_a \hat{\text{E}}[Y|X_j = x_{ij}, A = a] \neq d_0^*\right],$$

where $d_0^* = \arg\max_a \hat{\mathrm{E}}[Y|A=a]$ is the overall optimal decision.

The first ranking method (U) combines these two factors as a product of their values relative to the other variables in $\boldsymbol{X}$. The second ranking procedure (S) looks directly at the expected increase in the estimated mean outcome in the population due to the knowledge of X_j. For both methods, all K variables with positive scores (U or S) are then included in one model, enhanced by their corresponding main effects and important predictive variables. The final array of nested marker subsets is eventually based on the order of entry of the K q-interaction variables in a weighted Lasso, where weights are proportional to the scores (U or S).

For every panel size $k = 1, ...K$, the individual mean outcomes are proposed to be estimated using a regression model with the top k variables from the list. An optimal policy for a panel size k is then the one that maximizes the estimated mean outcome for every individual,

$$d_k^*(\boldsymbol{x}) = \arg\max_a \hat{\mathrm{E}}[Y|\boldsymbol{X}=\boldsymbol{x}, A=a]\,,$$

with the corresponding estimator of the mean outcome in the population

$$\hat{\mathrm{E}}\, Y(d_k^*) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathrm{E}}[Y|\boldsymbol{X}=\boldsymbol{x}_i, A=d_k^*(\boldsymbol{x}_i)] = \hat{V}_k\,.$$

The optimal policy for each individual is however based on the same model that evaluates the policy, which implicitly leads to overly optimistic results. In order to account for the upward bias of the regression estimator for $\mathrm{E}\, Y(d_k^*)$, the optimal panel size k^* is recommended to be selected such that maximizes the relative gain in the estimated

mean response in the population, penalized for the number of markers on the panel,

$$k^* = \arg\max_k \frac{\hat{V}_k - \hat{V}_0}{\hat{V}_{m^*} - \hat{V}_0} \left(\frac{m^*}{k}\right),$$

where $m^* = \arg\max_k \hat{V}_k$ and $\hat{V}_0$ is the estimated mean outcome in the population, having everybody treated with the marginally better treatment, $d_0^* = \arg\max_a \hat{\mathrm{E}}[Y|A=a]$.

1.8 Penalized Likelihood

Another attractive class of techniques for marker selection is based on the computational feasibility of LASSO methods (Tibsirani, 1996). The objective criterion in such penalized likelihood approaches is the probability (or density) placed on the observed data subject to a parameter constraint or penalty.

Qian and Murphy (2011) argue that in order to model the treatment benefit reasonably well, it is of interest to consider rich conditional mean models, but eventually rather fewer than more variables should be used by the decision rule. The hope is that with a decreasing penalty parameter, the subsequent models will contain only those variables that are important toward improving model performance. The proposed estimator was therefore based on the L_1 penalized least squares (L_1-PLS). Specifically, under a considered linear model

$$\mathrm{E}[Y|\,\boldsymbol{X}, A] = \Phi(\boldsymbol{X}, A)\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + A\boldsymbol{X}\boldsymbol{\gamma}\,,$$

where $\Phi(\boldsymbol{X}, A)$ is a 1 by $2m$ vector composed of basis functions on $(\mathcal{X}, \mathcal{A})$ and m is

number of all available markers, the L_1-PLS estimator for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ is

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \mathbb{R}^{2m}} \mathrm{E}_n[Y - \Phi(\boldsymbol{X}, A)\boldsymbol{\theta}]^2 + \lambda \sum_{j=1}^{2m} \hat{\sigma}_j |\, \theta_j \,|\,,$$

where $\hat{\sigma}_j = [\mathrm{E}_n \phi_j(\boldsymbol{X}, A)^2]^{1/2}$ and E_n denotes the empirical expectation.

An estimator based on the L_1 penalized least squares (L_1-PLS) method provides both variable selection and parameter estimation, and which can also result in a sequence of nested models as the value of the penalty parameter is relaxed. A regimen d_λ is then based on the coefficient estimates corresponding to λ.

Since the form of the optimal regimen only depends on the estimated treatment effect $\Delta(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\gamma}$, minimizing the penalized empirical prediction error yields high population mean outcome if the treatment benefit $\Delta(\boldsymbol{X})$ can be well approximated. The authors calculated finite sample upper bounds on the difference in the population mean outcome between the optimal regimen and the one estimated by the L_1-PLS. The bounds imply that the regimen produced by L_1-PLS method leads roughly the same mean outcome in the population as if we knew the sparsity of the oracle model and then estimated its coefficients using OLS (Qian and Murphy, 2011).

The tuning parameter λ was selected such that maximized a cross-validated empirical estimator of the population mean outcome, i.e. the average outcome among those who received the same treatment as would be prescribed by the regimen d_λ. This is equivalent to IPWE in the setting of $1:1$ randomized clinical trial. Even though unbiased, this method only uses part of the patients to estimate $\mathrm{E}\, Y(d)$ and is hence prone to larger variability than methods using the whole data set.

An advantage of the variable ranking methods is the resulting list of nested marker panels. Instead of a search through all possible combinations of m variables, the dimensionality of the problem decreases from 2^m to $K \leq m$. Even though the goal of these methods is not to find the "correct" prescriptive model, both Qian and Murphy (2011) and Gunter et al. (2011) showed that it is possible to systematically navigate through a large space of potential models, leading to an improved decision making in terms of a clinically relevant outcome.

1.9 Survival Outcomes

In a setting of survival outcome, a different approach to evaluation of treatment regimens developed in randomized clinical trials was proposed by Matsui et al. (2012). A binary decision rule d is usually based on a composite score of multiple markers, which dictates whether or not the treatment should be prescribed to a patient. The developed score, S, is typically continuous and represents varying treatment effects among patients. The proposed framework for evaluation of treatment regimens suggests to estimate the underlying variation of the treatment effect as a function of the scores that define the decision rules, $\Delta(S)$.

For any decision rule, this allows us to predict the treatment effect for every individual patient as a function of their score only. Since the information from all the variables $\boldsymbol{X}$ is collapsed into one-dimensional score, it is only needed to capture the main effect of S and the interaction between S and the treatment. In a survival setting, Matsui et al. (2012) considered a measure of treatment effect to be a logarithm of hazard ratio (HR)

between the two treatment arms. For a patient with a score s,

$$\Delta(s) = \text{logHR}(S = s, A = 1) - \text{logHR}(S = s, A = 0)\,.$$

and the assumed Cox proportional-hazards model has the following form,

$$h(t|s, a) = h_0(t)\exp\{\beta_1 a + f_2(s) + a f_3(s)\}\,,$$

where a is an indicator of treatment and f_2, f_3 are continuous, possibly non-linear functions. The treatment effect function can be then expressed as

$$\Delta(s) = \beta_1 + f_3(s)\,.$$

If the score S is truly predictive of the treatment effect Δ, the function f_3 is expected to be non-decreasing in s.

An analogous version of L_1-penalized methods for Cox regression was proposed by Tibsirani (1997) and later implemented by Gui and Li (2005). The proposed methods are suited for identification of important markers that are related to time to event and building models for prediction of the survival of future patients. Simulations showed that in the survival setting, L_1-penalized Cox regression leads to better predictiveness performance than the L_2-penalized regression and a few other dimension-reduction based methods.

1.10 Design of a Validation Study

A proper validation study is necessary to determine whether a pre-specified marker-guided treatment (MGT) truly benefits the patients as suggested by the analysis of the source study. For test of MGT superiority, a design where all patients are randomized to MGT versus standard care would require an unnecessarily large sample size due to average treatment effect being diluted by patients who were not expected to benefit from the treatment, and hence untreated/provided with standard care in either arm. Randomizing only patients for whom the MGT suggests to treat (score $S > 0$) might lead to a substantial reduction in required sample size (Simon, 2008), depending on the proportion of marker positive patients.

Accordingly, the enrichment strategy design suggests to test a null hypothesis of non-superiority of the treatment among marker positive patients:

$$H_0 : \mathrm{E}[Y|A = 1, S > 0] \leq \mathrm{E}[Y|A = 0, S > 0]\,.$$

Rejecting H_0 means there is an evidence that the treatment works better in the subgroup of marker positive patients, which implies superiority of MGT over the standard care.

Lai et al. (2012) discuss a generalized version of enrichment strategy for a binary outcome, ovarian cancer remission within 6 months, with multiple competitive treatment options, $k = 1, ..., K$. The range of the examined marker divided into J+1 categories, where category $j = 0$ corresponds to marker values where no benefit from MGT is expected, and patients are to be randomized only if $j > 0$. Lai et al. (2012) assess the effect of adhering to the marker-guided treatment by comparing the response rate

(within each stratum $j > 0$) under MGT to the response rate under the complementary treatment (not recommended by MGT), averaging over treatments when necessary, and then averaging over strata.

The response rate p_{jk} is the rate of remission within 6 months for patients in the j-th marker group receiving treatment k, while R_j and N_j are the sets of treatment indices recommended and not recommended, respectively, by the MGT for patients in stratum j. If $P_j = \sum_{k \in R_j} p_{kj}/|R_j|$ is the average response rate to treatments recommended by MGT for patients in stratum j, the overall response rate to MGT is $\sum_{j=1}^{J} \pi_j P_j$, where π_j is the prevalence of subgroup j. Similarly, $Q_j = \sum_{k \in N_j} p_{kj}/|N_j|$ is the average response rate to treatment complementary to MGT for patients in stratum j. Lai et al. (2012) then proposed to test the *enriched strategy null hypothesis*

$$H_0 : \sum_{j=1}^{J} \pi_j (P_j - Q_j) \leq 0 \,,$$

where the benefit from the MGT is evaluated only in the strata where it is expected.

Chapter 2

CONTINUOUS OUTCOMES

2.1 Introduction

Most patients diagnosed with a particular health condition are faced with two or more therapeutic options. Treatment recommendations are typically based on high quality clinical trials which summarize average treatment effects. However, often a uniformly preferred therapy is not available due to heterogeneity in patients' responses. Our research is motivated by the common hypothesis that patient-level characteristics might predict which treatment works better for an individual and hence could be used to guide treatment choice. As a result, targeted treatment algorithms could be applied systematically in medical practice and lead to both better outcomes for individuals and for the overall population of treated patients.

The ultimate goal of individualized or guided treatment is to prescribe each patient with a therapeutic option that leads to the largest expected benefit for that individual. The scientific challenge is to identify those factors that can be used to target treatment, and to accurately quantify the expected treatment benefit as a function of the candidate markers. Hence, the associated statistical methods that characterize the accuracy of medical decisions, or the overall population consequence of using a marker-guided strategy are needed. The primary statistical objective is to rigorously and validly evaluate candidate marker information to determine the potential to reproducibly direct individ-

ual treatment toward optimal performance at both the individual and population level. However, existing approaches for development of an optimal decision rule are limited to models with a small number of variables (Zhang et al., 2012), which may be inadequate for composite biomarkers that are reflective of complex disease pathways.

In this chapter, we propose a new approach for the evaluation and comparison of candidate prescriptive models in the setting of continuous outcome, with a focus on the use of potentially high-dimensional biomedical information. The goal of the proposed prescriptive methods is to construct a reliable individual "benefit score" and to evaluate the population potential of using such a benefit score to guide treatment. We outline a simple algorithm that will generate a sequence of nested marker panels, each resulting in treatment benefit scores. The corresponding decision rules based on the benefit scores will be compared with respect to the estimated marker-guided population mean outcome. Although we consider a computationally simple development step, our key contribution is a detailed cross-validation algorithm for evaluation and comparison of derived benefit scores.

In order to evaluate the potential performance of a developed decision rule, one needs to estimate the population characteristic of interest, such as the mean outcome under the hypothetical scenario where all patients follow the prescribed regimen (Qian and Murphy, 2011). Parametric approaches, which require model fitting and estimation of associated parameters, are usually inappropriate for settings with a large number of variables, as they tend to over-fit the data and become computationally intensive with a large set of candidate predictors (Zhang et al., 2012). A commonly used non-parametric approach for evaluation of a treatment-assignment rule performance is based on only

those patients who received treatment equivalent to what would be prescribed to them by the marker-guided regimen (Freidlin et al., 2010; Qian and Murphy, 2011). However, the variability of the estimated population response across different candidate regimens may then be attributable to differences in the corresponding subgroups of patients that contribute to the evaluation of the regimens.

Often, a primary goal of therapeutic medical research is to determine whether a treatment leads to improvement in a patient outcome or condition. Randomized clinical trials are the standard for establishing whether treatment is effective on average for indicated patients. For those treatments with a weak overall benefit it can be important to evaluate whether certain subgroups of patients have a moderate or strong response, since identification of subgroups will permit directed treatment.

The current literature on personalized medicine aims to identify population subgroups that benefit from a particular treatment. The classification of patients is based on their baseline characteristics, measured before the treatment is administered. A popular example is the response of colorectal cancer patients to the epidermal growth factor receptor (EGFR)-inhibiting drugs, such as Vectibix or Ectibux. While patients with a wild-type KRAS gene generally respond well to these drugs, certain mutations in KRAS gene are predictive of a very poor response (Liévre et al., 2006).

The proposed methods focus on developing and evaluating a marker-guided decision rule for treatment assignment in the setting of a continuous outcome and a large number of candidate variables. Our goal is to establish an algorithm for marker panel selection that will honestly evaluate and compare derived decision rules with respect to the improvement of the mean outcome in the population.

2.2 Methods

2.2.1 Individual and Population Treatment Benefit

Let Y denote a continuous clinical outcome of interest, and let A denote the treatment (action). We assume that $\boldsymbol{X}$ denotes a vector of patient characteristics collected prior to treatment, taking values from a space $\mathcal{X} \subseteq \mathbb{R}^m$, where m is the number of covariates. Without loss of generality, larger values of Y are considered "better" and hence, improving the clinical outcome means maximizing it over available therapies. For simplicity, we consider A taking only values 0 or 1, but our framework is directly extendable to multiple treatments, or a treatment with multiple categories, such as dose levels.

As described in Chapter 1, a treatment regimen d is a function, or decision rule, that maps values of $\boldsymbol{X}$ to $\{0, 1\}$, so that a patient with covariates $\boldsymbol{X} = \boldsymbol{x}$ would be recommended to receive treatment 1 if $d(\boldsymbol{x}) = 1$ and treatment 0 if $d(\boldsymbol{x}) = 0$. The expected outcome for a patient with covariates $\boldsymbol{X} = \boldsymbol{x}$ under treatment $A = a$ can be written as

$$\mathrm{E}[Y|\, \boldsymbol{X} = \boldsymbol{x}, A = a] = \mathrm{E}[Y|\, \boldsymbol{X} = \boldsymbol{x}, A = 0] + a\, \Delta(\boldsymbol{x})\,, \tag{2.1}$$

where $\Delta(\boldsymbol{x}) = \mathrm{E}[Y|\, \boldsymbol{X} = \boldsymbol{x}, A = 1] - \mathrm{E}[Y|\, \boldsymbol{X} = \boldsymbol{x}, A = 0]$ is a covariate-specific treatment benefit. In case the interest was in studying a continuous treatment, such as $A =$ dose, the treatment benefit would now also be a function of A, and one would need to specify a model for $\Delta(\boldsymbol{X}, A)$. A positive estimated treatment benefit implies a better expected outcome if the treatment is administered ($A = 1$), suggesting that patients should be treated when their expected benefit $\Delta(\boldsymbol{x}) > 0$.

We consider a class of regimens, $d \in \mathcal{D}$, which contains a set of candidate decision functions. Then the regimen that chooses the largest expected outcome in every individual,

$$d^{opt}(\boldsymbol{x}) = \arg\max_{d\in\mathcal{D}} \mathrm{E}[Y|\boldsymbol{X} = \boldsymbol{x}, A = d(\boldsymbol{x})],\ \forall \boldsymbol{x} \in \mathcal{X}\,,$$

also leads to the largest mean outcome in the whole population,

$$d^{opt} = \arg\max_{d\in\mathcal{D}} \mathrm{E}\, Y(d)\,,$$

where $\mathrm{E}\, Y(d) = \mathrm{E}_{\boldsymbol{X}}\{\mathrm{E}_{Y|\boldsymbol{X}}[Y|\,\boldsymbol{X},\, A = d(\boldsymbol{X})]\}$ is the mean outcome in the population under regimen d. If we use the expression (2.1) for the individual expected outcome, we can write the optimal decision rule for an individual with covatiates $\boldsymbol{x}$ as

$$\begin{aligned} d^{opt}(\boldsymbol{x}) &= \arg\max_d \mathrm{E}[Y|\,\boldsymbol{X} = \boldsymbol{x},\, A = d(\boldsymbol{x})] = \\ &= \arg\max_d \mathrm{E}[Y|\,\boldsymbol{X} = \boldsymbol{x},\, A = 0] + d(\boldsymbol{x})\,\Delta(\boldsymbol{x}) = \\ &= \arg\max_d \{d(\boldsymbol{x})\,\Delta(\boldsymbol{x})\} = \\ &= \mathrm{I}\,[\Delta(\boldsymbol{x}) > 0]\,. \end{aligned}$$

Hence, a theoretically optimal regimen rule on $\mathcal{X}$, that leads the largest expected outcome for every $\boldsymbol{x} \in \mathcal{X}$, also has the form $d^{opt}(\boldsymbol{x}) = \mathrm{I}\,[\Delta(\boldsymbol{x}) > 0]$, which assigns treatment only to those individuals for whom their covariate-defined subgroup yields a positive treatment benefit. We refer to this as the "oracle rule" since it depends on correct specification of the covariate-specific treatment benefit $\Delta(\boldsymbol{X})$, which is a potentially high-dimensional function, mapping $\boldsymbol{x} \in \mathcal{X}$ to $\mathbb{R}$.

In order to simplify the search for the optimal decision rule, we can restrict our estimation space to a smaller class of working models for $\Delta(\boldsymbol{X})$, which is easy to navigate through. The main scientific motivation for prescriptive models comes from the common assumption that treatment "works" differently in some individuals than in others, and that covariates or biomarkers, $\boldsymbol{X}$, can identify subgroups that benefit from the treatment. Statistically, the concept implies that the treatment interacts with some of the individual-level characteristics, i.e. $\Delta(\boldsymbol{X})$ is non-constant.

For high-dimensional markers, $m \gg 0$, we consider a working linear additive mean model

$$\mathrm{E}[Y|\,\boldsymbol{X}, A] = \beta_0 + \gamma_0 A + \sum_{j=1}^{m}(\beta_j X_j + \gamma_j A X_j) = \boldsymbol{X}\boldsymbol{\beta} + A\boldsymbol{X}\boldsymbol{\gamma}\,, \tag{2.2}$$

which is appealing for its simplicity and interpretability. The marker-specific treatment effect $\Delta(\boldsymbol{X})$ can then be expressed as a linear combination of the marginal treatment effect and the individual marker-treatment interactions:

$$\Delta(\boldsymbol{X}) = \mathrm{E}[Y|\,\boldsymbol{X}, A=1] - \mathrm{E}[Y|\,\boldsymbol{X}, A=0] = \gamma_0 + \sum_{j=1}^{m}\gamma_j X_j = \boldsymbol{X}\boldsymbol{\gamma}\,.$$

The restricted class of regiments, $\mathcal{D}$, then consists of decision rules that have the form

$$d(X) = \mathrm{I}\left[\boldsymbol{X}\boldsymbol{\gamma} > 0\right],$$

and within this class, the estimation of the marker-specific treatment effect is equivalent to estimation of the model parameters $\boldsymbol{\gamma}$. Every vector $\boldsymbol{\gamma} \in \boldsymbol{\Gamma} \subseteq \mathbb{R}^m$ defines one regimen

from the class $\mathcal{D}$, and the optimal regimen within this class is one that leads to the maximum mean outcome in the population from amongst all the regimens in $\mathcal{D}$,

$$d^* = \arg\max_{d\in\mathcal{D}} \mathrm{E}\, Y(d)\,,$$

or alternatively, $d^* = d_{\boldsymbol{\gamma}^*}$, where

$$\boldsymbol{\gamma}^* = \arg\max_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \mathrm{E}\, Y(d_{\boldsymbol{\gamma}})\,.$$

2.2.2 Approaches to Benefit Score Development and Variable Selection for Large Marker Panels

Analysis of high-dimensional biomedical data often faces the problem of many potential spurious variables when only a small portion of the available markers is expected to be useful. For treatment decision function estimation a common approach to reduce the number of candidate models involves ordering markers with respect to their marginal interactions with the treatment (Gunter et al., 2011; Matsui et al., 2012). The result is a sequence of marginally ordered markers and associated nested models when including the top p ranking markers. Such a decrease in the number of candidate models from 2^m to m (or less) is especially desirable when m is very large (e.g., in case of genomic data) and evaluation of all possible models would not be feasible.

Another attractive class of techniques for marker selection is based on the computational feasibility of LASSO method (Tibsirani, 1996). The objective criterion in such penalized likelihood approaches is the probability (or density) placed on the observed

data subject to a parameter constraint or penalty. An estimator based on the L_1 penalized least squares (L_1-PLS) method provides both variable selection and parameter estimation. Exploration of the full solution path over a range of penalty parameter values can also result in a sequence of nested models as the value of the penalty parameter is relaxed.

The hope is that with a decreasing penalty parameter, the subsequent models will contain only those variables that are important toward improving model performance. Although common approaches to penalty parameter selection include assessment of cross-validated predictive performance, it was concluded by Qian and Murphy (2011) that if the selection of the penalty parameter λ is based on the improvement of population characteristic of interest, such as the mean outcome, the resulting decision rule performs almost as well as if we knew the sparsity of the oracle model and then estimated coefficients using OLS.

Due to the latter property, we propose to generate a set of candidate decision rules based on L_1-penalized least squares, although other selection methods could also be adopted. The mean model (2.2) induces a class of additive linear decision rules, $d(\boldsymbol{X}) = \mathrm{I}\left[\boldsymbol{X}\boldsymbol{\gamma} > 0\right]$, which correspond to parameters $\boldsymbol{\gamma} \in \boldsymbol{\Gamma} \subseteq \mathbb{R}^m$. Then, for a total number of markers m and $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ of length $J = 2m + 2$, we define

$$\widehat{\boldsymbol{\theta}}(\lambda) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^J} \{ \mathrm{E}_n[Y - (\boldsymbol{X}, A\boldsymbol{X})\boldsymbol{\theta}]^2 + \lambda \sum_{j \in \mathcal{J}} \hat{\sigma}_j |\, \theta_j| \} \,,$$

where E_n denotes the empirical expectation, $\hat{\sigma}_j$ is the estimated standard deviation of the j-th covariate, and $\mathcal{J} = \{2, ..., m{+}1,\ m{+}3, ..., J\}$, so the intercept and main treatment

effect are not penalized. While the interactions (γ's) are the parameters of interest that compose $\widehat{\Delta}$, the main effects (β's) are included in order to improve precision of $\widehat{\boldsymbol{\gamma}}$. The idea is that the LASSO will gradually select parameters that are most relevant for estimation of $\mathrm{E}[Y|\boldsymbol{X}, A]$, including the contrast between the two treatments, $\Delta(\boldsymbol{X})$.

The tuning parameter λ controls the amount of penalization and thus also the number of variables in the model. Since the parameters involved in the decision rules are only $\boldsymbol{\gamma}$, we will measure the size of a panel by the number of non-zero estimates of coefficients $\gamma_1, ..., \gamma_m$, and denote the estimates $\widehat{\boldsymbol{\gamma}}^p$. Each set of estimates, $\widehat{\boldsymbol{\gamma}}^p$, then corresponds to a decision rule

$$d_p(\boldsymbol{X}) = \mathrm{I}\left[\boldsymbol{X}\widehat{\boldsymbol{\gamma}}^p > 0\right]$$

that provides a way of patient classification as to whether they should get treated or not, depending on their individual scores, $s_i^p = \boldsymbol{x}_i\widehat{\boldsymbol{\gamma}}^p$. The main reason why we index models with respect to panel size p is that this indexation is consistent with some other approaches in which the markers are ordered, for example, based on their marginal interactions with treatment, and gradually included in the model, as in (Gunter et al., 2011) or (Matsui et al., 2012).

Using regular regression function and standard marker selection techniques, such as LASSO, allows us to nominate a sequence of candidate rules that can be then compared with respect to their population performance. In case multiple marker panels (rules) have very similar population impact, the usefulness of marker-guided rules can be also considered from perspective of additional aspects such as cost or time to measure the biomarkers, as smaller number of markers that need to be assessed usually means lower cost.

An alternative approach that directly relates the estimated function in the nomination phase and the target parameter from the evaluations phase (e.g., mean population outcome) is targeted learning, in which the targeted MLE is aimed to find an optimal bias-variance tradeoff for the parameter of interest (van der Laan and Rose, 2011). The proposed approach can however be computationally very extensive and also typically results in a single "optimal" rule instead of a set of candidate rules which could be independently evaluated and compared with respect to their population impact.

2.2.3 Estimation of Marker-Guided Population Response

In this section, we first focus on evaluation of decision rules based on a single quantitative marker S and then will later show how these concepts can be applied to multivariate marker panels. As implied previously, we wish to evaluate each candidate marker-guided decision rule d with respect to its population performance. In particular, we ask what would be the population mean response if everybody's treatment was assigned based on d. The marker-guided population response (MGPR), $\mathrm{E}\,Y(d)$, is however not straightforward to assess, since some patients in our data set might have observed (randomized) treatments different from what d would prescribe to them.

If we assume a marker-guided decision rule $d(s) = \mathrm{I}\,[s > 0]$, then the MGPR under the rule d can be rewritten as

$$\begin{aligned}
\mathrm{E}\,Y(d) &= \mathrm{E}_S\{\,\mathrm{E}_{Y|S,A}[Y|\,S,\,A = d(S)]\,\} = \\
&= \mathrm{E}_S\{\,\mathrm{E}_{Y|S,A}[Y|\,S,\,A = 0]\,\mathrm{I}\,[d(S) = 0] + \mathrm{E}_{Y|S,A}[Y|\,S,\,A = 1]\,\mathrm{I}\,[d(S) = 1]\,\} = \\
&= \int_{\mathbb{R}} \mu_0(s)\mathrm{I}\,[s \leq 0] + \mu_1(s)\mathrm{I}\,[s > 0]\,dF(s)\,, \qquad (2.3)
\end{aligned}$$

where $\mu_a(s) = \mathrm{E}\left[Y \middle| S = s, A = a\right]$ is the conditional expectation of outcome Y given the marker value s and treatment $a \in \{0, 1\}$. Common approaches to $\mathrm{E}\, Y(d)$ estimation involve either fitting a parametric regression model (e.g. linear, quadratic) to capture the relationship between the marker S and the expected outcome of interest in the two treatment arms or non-parametric estimation of the population response using only subset of patients whose observed and prescribed treatments match.

Standard Non-parametric Estimator

A commonly used non-parametric estimator of the marker-guided population response under the rule $d(s) = \mathrm{I}\left[s > 0\right]$ is based on the following simplification of the expression (2.3),

$$\mathrm{E}\, Y(d) = \mathrm{E}[Y \middle| S \leq 0, A = 0]\, \mathrm{P}[S \leq 0] + \mathrm{E}[Y \middle| S > 0, A = 1]\, \mathrm{P}[S > 0]\,. \tag{2.4}$$

Under the assumption that our data set is a random sample from the target population, the probability $\mathrm{P}[S > 0] = \pi_1$ can be estimated by the sample proportion of $s-$positive patients,

$$\widehat{\pi}_1 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\left[s_i > 0\right].$$

Further, if the observed treatments a_i are independent of the patients' marker values s_i, $i = 1, ..., n$, then the conditional expectations in (2.4) can be estimated by average outcomes over the patients for whom the observed treatment is the same as what d would prescribe to them, separately for each treatment arm. This leads to

$$\widehat{\mathrm{E}}\, Y(d) = \frac{\sum_{i=1}^{n} y_i \,\mathrm{I}\,[a_i = 0]\,\mathrm{I}\,[s_i \leq 0]}{\sum_{i=1}^{n} \mathrm{I}\,[a_i = 0]\,\mathrm{I}\,[s_i \leq 0]}\,(1 - \widehat{\pi}_1) \;+\; \frac{\sum_{i=1}^{n} y_i \,\mathrm{I}\,[a_i = 1]\,\mathrm{I}\,[s_i > 0]}{\sum_{i=1}^{n} \mathrm{I}\,[a_i = 1]\,\mathrm{I}\,[s_i > 0]}\,\widehat{\pi}_1 \;.$$

If the proportion of observed treatment $a_i = 1$ is the same among $s-$positive and $s-$negative patients, this estimate further simplifies to

$$\widehat{\mathrm{E}}\, Y(d) = \frac{\sum_{i=1}^{n} y_i \,\mathrm{I}\,\{a_i = \mathrm{I}\,[s_i > 0]\}}{\sum_{i=1}^{n} \mathrm{I}\,\{a_i = \mathrm{I}\,[s_i > 0]\}}\,,$$

which was used, for example, by Qian and Murphy (2011).

A potential disadvantage of the standard non-parametric estimator is that it does not use the information from all subjects to estimate the population mean response. The observed outcomes of the patients who received a treatment that was different from what d would prescribe are ignored, which impairs the efficiency of this estimator and suggests potential room for improvement.

Smoothing Methods

Alternatively, one can derive an estimate of the marker-guided population response based on estimates of the expected outcome as a function of the score S under the two treatment options. The conditional expectations μ_0 and μ_1 from expression (2.3) can be modeled as functions of the score S by either parametric (e.g. linear, quadratic, etc.) or non-parametric regression. As one might not have a strong belief in any of the parametric models, it seems natural to approximate the relationship by some non-parametric regression method such as smoothing splines.

We introduce an estimator of the marker-guided population response under regimen d, $\mathrm{E}\,Y(d)$, which is smooth, non-parametric and consistent, yet uses the data set more effectively than the simple non-parametric estimator described above. The idea is related to the predictiveness curves proposed by Janes et al. (2011). In each treatment arm, the conditional expected outcome is considered as a function of the marker score S, $\mu_a(s) = \mathrm{E}\,[Y|A = a, S = s]$, and will be approximated by a separate non-parametric regression such as a smoothing spline $\widehat{\mu}_a(s)$, $a \in \{0, 1\}$, which is a flexible continuous function of s. An estimated expected outcome under a regimen d for a patient with score s_i is then based on the curve $a = d(s_i)$ (see Fig 2.1). Hence, the mean response in the population under the regimen $d(s) = \mathrm{I}\,[s > 0]$ is estimated by

$$\widehat{\mathrm{E}}\,Y(d) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_0(s_i)\,\mathrm{I}\,[s_i \leq 0] + \widehat{\mu}_1(s_i)\,\mathrm{I}\,[s_i > 0]\,.$$

While non-parametric and unbiased, our estimator combines all the observed outcomes to estimate $\widehat{\mathrm{E}}\,Y(d)$ and is hence more efficient than the standard non-parametric estimator discussed above, as we will see later from the simulations in section 3.3.

For our simulations as well as for the examples, we chose smoothing splines with 5 degrees of freedom (df). Even though this selection was somewhat arbitrary, we selected df that seemed to provide a good balance between the flexibility of the fitted curves and their low dimensionality in order to prevent data overfitting. In order to implement more formal selection of the degrees on freedom, one could use the cross-validation approach to determine what degree provides the best fit to the data.

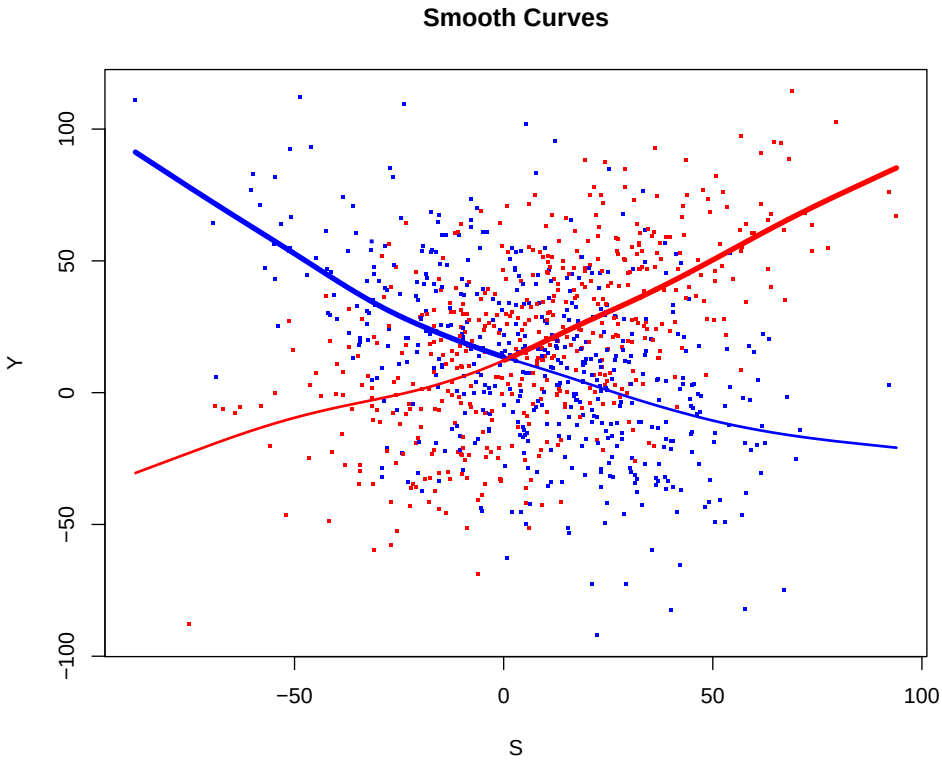

Figure 2.1: Estimation of the expected outcome as a function of the score, S, by smoothing splines, separately for the two treatment arms; $A = 0$ (blue) and $A = 1$ (red). The estimate of the mean population response is based on the thicker parts of the curves; outcomes for score negative patients are approximated by the blue curve and for score positive patients by the red curve.

Penalized B-splines

A spline function is a piece-wise polynomial function and the places where the pieces meet are called knots. The order of the spline function is determined by the order of the polynomial pieces, k, which match at the knots in their derivatives up to order $(k-1)$. Every spline function can be uniquely represented as a linear combination of basis splines, or B-splines, of the same degree and support (knots). Basis splines are local functions with Gaussian-like shape and comprise the columns of a basis matrix $\boldsymbol{B}$.

In order to fit a spline to a set of observations given by (z, y), we first need to select knots $t_1, ..., t_r$, such that $min(z) < t_1 < ... < t_r < max(z)$, at which the polynomial pieces of B-splines are joint. The knots can be chosen, e.g., at all observed values or equidistantly across the range of values. We then construct the spline bases $B_1(z), ..., B_m(z)$, where m is number of knots r plus the order of the splines plus one (e.g. r+4 for cubic splines). The expected response function can be expressed as

$$\mu(z) = \sum_{k=1}^{m} \beta_k B_k(z) = \boldsymbol{B}(z)\boldsymbol{\beta}\,.$$

With a large number of B-splines, or equivalently regression parameters, a penalty can be incorporated into the estimation process to prevent over-fitting and to control the smoothness of $\mu(z)$. Penalized B-splines hence require specification of basis splines and a penalty. The penalty consists of a penalty function $P(.)$ and penalty parameter λ_P, which controls the amount of penalization. We adopt a cubic spline penalty function that is commonly used in practice, defined as

$$P(\boldsymbol{\beta}) = \int \left[\sum_{k=1}^{m} \beta_k B_k^{(2)}(z)\right]^2 dz = \boldsymbol{\beta}'\boldsymbol{D}\boldsymbol{\beta}\,,$$

where $B_k^{(2)}(z)$ is the second derivative of the k-th B-spline at z and $\boldsymbol{D}$ is an m by m matrix with (k, l) entry

$$D_{kl} = \int B_k^{(2)}(z) B_l^{(2)}(z) dz\,.$$

Since the cubic spline penalty function penalizes second derivatives, larger values of λ_P result in smaller curvature, or smoother mean functions. Eventually, as $\lambda_P \to \infty$, the

penalized spline would become a straight line, and $\lambda_P = 0$ corresponds to un-penalized (saturated) spline. The penalty term is added to the standard likelihood, which is being maximized when estimating the coefficients $\boldsymbol{\beta}$. The corresponding estimating equations are

$$(\boldsymbol{B}(\boldsymbol{z})'\boldsymbol{B}(\boldsymbol{z}) + \lambda_P \boldsymbol{D})\boldsymbol{\beta} = \boldsymbol{B}(\boldsymbol{z})'\boldsymbol{y}\,,$$

leading the standard (ridge regression) solution $\widehat{\boldsymbol{\beta}} = (\boldsymbol{B}(\boldsymbol{z})^T\boldsymbol{B}(\boldsymbol{z}) + \lambda_P \boldsymbol{D})^{-1}\boldsymbol{B}(\boldsymbol{z})^T\boldsymbol{y}$. For any set of values $\boldsymbol{s}$, the predicted values at $\boldsymbol{s}$ can be expressed as

$$\widehat{\boldsymbol{y}}(\boldsymbol{s}) = \boldsymbol{B}(\boldsymbol{s})\widehat{\boldsymbol{\beta}} = \boldsymbol{B}(\boldsymbol{s})(\boldsymbol{B}(\boldsymbol{z})^T\boldsymbol{B}(\boldsymbol{z}) + \lambda_P \boldsymbol{D})^{-1}\boldsymbol{B}(\boldsymbol{z})^T\boldsymbol{y} = H(\boldsymbol{s})\boldsymbol{y}\,.$$

One way how to choose the smoothing parameter λ_P is by specifying the amount of smoothing through the effective degrees of freedom (df), which is a trace of the smoother matrix $H(\boldsymbol{z})$. The higher the number of effective degrees of freedom, the more relaxed is the penalty λ_P and hence the more flexible is the spline function.

The penalized B-splines hence offer an effective way to balance the curve fit and smoothness by imposing a penalty on the model coefficient estimates in order to avoid over-fitting. The method of such smoothing that uses a maximal set of knots and fits a spline function subject to a (cubic) spline penalty is also called smoothing spline.

2.2.4 Assessment of Optimal Treatment Regimen

We now consider a sequence of candidate decision rules based on the L_1-PLS estimates of the linear additive model (2.2) with the goal of identifying an optimal panel size. Since the tuning parameter λ in the L_1-PLS method controls the number of variables

in the model, we identified a collection of λ's that coincide with panel sizes $p = 1, 2, ...$ Each set of estimates, $\widehat{\boldsymbol{\gamma}}^p$, corresponds to a decision rule $d_p(\boldsymbol{X}) = \mathrm{I}\left[\boldsymbol{X}\widehat{\boldsymbol{\gamma}}^p > 0\right]$ that provides a way of patient classification through their individual scores $s_i^p = \boldsymbol{x}_i\widehat{\boldsymbol{\gamma}}^p$. The question is, which marker panel should be chosen to ultimately guide the treatment in the population, i.e. which of the candidate decision rules leads the highest mean response in the population.

In order to honestly evaluate the population performance of candidate scoring rules we use a K-fold (e.g. $K = 10$) cross-validation. The out-of-sample performance of a decision rule d_p, based on the (top) p selected markers, is to be measured by a cross-validated estimate of the marker-guided population response under that rule, $\widehat{\mathrm{E}}\, Y(d_p)$. For each panel size $p = 1, 2, ...,$ we estimate the $\mathrm{E}\, Y(d_p)$ by both approaches described in the previous section. The smooth curves from our new approach can be fitted to the observed outcomes as functions of the scores, s_i^p, which collapse the information about $\boldsymbol{X}$ into a one-dimensional quantity. This allows us to use the smoothing spline method in the cross-validation step instead of specifying and estimating a complex model for high-dimensional $\boldsymbol{X}$.

It is vital that the penalized least squares are fitted to each fold separately, so that the scores for every test set are always completely unaffected by the outcomes of the patients in the same test set. Once all the scores are attained for each panel size, the whole data set can be used to fit the smooth curves and gain precision in the estimation of the expected outcomes under alternative treatments. The optimal panel size p^* is then chosen as the smallest panel size that maximizes a cross-validated estimator of the mean response in the population. However, we advise to always inspect the plot

of estimated clinical objective (MGPR) as a function of panel size (such as in Fig.2.3), since sometimes panel sizes much smaller than p^* may provide MGPR very close to the maximum and be hence more appropriate choices than p^*, especially when the additional markers are expensive or difficult to measure.

Another visual aid in assessing quality of a developed decision rule is to graph the fitted smooth curves versus individual scores, as in Fig 2.1. If the scores based on the selected marker panel are truly predictive of the treatment benefit, we expect a clear separation between the two treatment arms. The difference between the fitted smooth curves corresponds to a score-specific treatment benefit and should match the score itself, if generated accurately.

2.2.5 Single Index Cross-validation Algorithm

The out of sample evaluation of nested marker panels is performed by a K-fold (e.g, 10-fold) cross-validation. The L_1-penalized estimator of the marker-specific treatment benefit, $\Delta(\boldsymbol{X})$, provides a variable selection, which depends on a tuning parameter λ. In each training set, we hence select collection of λ's that correspond to panel sizes $p = 1, 2, ...$ The corresponding coefficient estimates $\widehat{\boldsymbol{\gamma}}^p$ are then used to calculate scores $s_i^p = \boldsymbol{x}_i \widehat{\boldsymbol{\gamma}}^p$ for patients in the test set. Once the scores for all patients and each panel size are assessed, we combine the whole data set to fit the smooth curves in order to gain precision in the estimation of $\mu_0(s^p)$ and $\mu_1(s^p)$. Details of the evaluation algorithm are listed below.

Our new estimator, based on smoothing splines, is compared with the standard non-parametric estimator described in section 3.2.4. Although consistent, the standard non-

parametric approach only uses select patients who had the same treatment prescribed by d_p as they actually received in order to compute $\widehat{\mathrm{E}}\, Y(d_p)$. Hence, for any two decision rules that lead to different classifications, their respective estimates of the marker-guided population response are based on different subsets of patients. As we will see later, this results in an increased variability of the standard estimator when compared to the estimator based on smoothing splines.

Algorithm:

1. Divide the sample (n) into K folds (here K=10).

2. For each fold $k = 1, ...K$:

a) Run LASSO for the linear additive model

$$\mathrm{E}[Y \,|\, \boldsymbol{X}, A] = \beta_0 + \gamma_0 A + \sum_{j=1}^{n} (\beta_j X_j + \gamma_j X_j A) = \boldsymbol{X}\boldsymbol{\beta} + A\boldsymbol{X}\boldsymbol{\gamma}$$

on **training set** (9/10 of the data), and obtain ordered sequence of model coefficients.

b) Determine sequence of λ's corresponding to panel sizes $p = 1, 2, ...,$

where a marker panel size equals to the number of non-zero estimates of (interaction) coefficients $\gamma_1, ..., \gamma_N$, and denote the corresponding estimates $\widehat{\boldsymbol{\gamma}}^p$.

c) For each panel size $p = 1, 2, ...$

Calculate scores for individuals in the **test set**, $s_i^p = \boldsymbol{x}_i \widehat{\boldsymbol{\gamma}}^p$, based on the markers with the selected p interactions. Corresponding decision rule $d_p(\boldsymbol{X}) = \mathrm{I}\,[\boldsymbol{X}\widehat{\boldsymbol{\gamma}}^p > 0] = \mathrm{I}\,[S^p > 0]$ assigns the treatment.

3. For each panel size $p = 1, 2, ...$

Estimate the marker-guided population response (MGPR) based on $\{(y_i, a_i, s_i^p),\, i =$

$1, ..., n\}$,

using either a standard non-parametric estimator or a smooth regression estimator.

a) Standard Non-parametric Estimator

This estimator is based on only those patients who received the treatment equivalent to what would be prescribed to them by d_p; $\mathcal{N}_0^p = \{i : a_i = \mathrm{I}[s_i^p > 0] = 0\}$, $\mathcal{N}_1^p = \{i : a_i = \mathrm{I}[s_i^p > 0] = 1\}$.

Target (MGPR):

$$\mathrm{E}\, Y(d) = (1 - \pi_1)\mu_0^- + \pi_1 \mu_1^+$$

where $\mu_0^- = \mathrm{E}[Y | S \leq 0, A = 0]$, $\mu_1^+ = \mathrm{E}[Y | S > 0, A = 1]$ and $\pi_1 = \mathrm{P}(S > 0)$.

- $\widehat{\pi}_1$: proportion of score positive patients
- $\widehat{\mu_0^-}$ average y_i for $i \in \mathcal{N}_0^p$
- $\widehat{\mu}_1^+$ average y_i for $i \in \mathcal{N}_1^p$

$$\begin{aligned} \widehat{\mathrm{E}}\, Y(d_p) &= (1 - \widehat{\pi}_1)\, \widehat{\mu_0^-} + \widehat{\pi}_1\, \widehat{\mu}_1^+ = \\ &= \frac{(1 - \widehat{\pi}_1)}{|\mathcal{N}_0^p|} \sum_{i=1}^{n} y_i\, \mathrm{I}\{i \in \mathcal{N}_0^p\} + \frac{\widehat{\pi}_1}{|\mathcal{N}_1^p|} \sum_{i=1}^{n} y_i\, \mathrm{I}\{i \in \mathcal{N}_1^p\} \end{aligned}$$

b) Smoothing Splines

Estimate smooth curves $\mu_0(s)$, $\mu_1(s)$ as functions of s^p, where $\mu_a(s) = \mathrm{E}[Y | S = s, A = a]$.

Target (MGPR):

$$\mathrm{E}\, Y(d) = \int \mu_1(s)\mathrm{I}[s > 0] + \mu_0(s)\mathrm{I}[s \leq 0]\, dF(s)$$

- $\widehat{F}_S$: empirical distribution function in the data
- $\widehat{\mu}_0(s)$, $\widehat{\mu}_1(s)$ via smoothing splines

$$\widehat{\mathrm{E}}\, Y(d_p) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mu}_1(s_i^p)\, \mathrm{I}\left[s_i^p > 0\right] + \widehat{\mu}_0(s_i^p)\, \mathrm{I}\left[s_i^p \leq 0\right]$$

4. Graph the clinical objective function: $\widehat{\mathrm{E}}\, Y(d_p)$ vs. p and identify a panel size p^* that leads maximum estimated MGPR.

2.2.6 Benefit Among Treated

In this section we focus our attention particularly toward scenarios in which few patients have negative treatment benefits, but some patients simply do not respond to the treatment, i.e. their expected treatment benefit $\Delta(\boldsymbol{x}) = 0$. Then, the target population is a mixture of "responders", who have a range of positive expected benefits, and "non-responders".

In such case, the mean population response $\mathrm{E}\, Y(d)$ does not depend on whether the "non-responders" get treated or not, as their expected outcome is the same under both treatment options. Consequently, the mean population response is maximized under the rule that universally treats everybody and cannot be exceeded by not treating the "non-responders". In such scenarios we might choose to focus on identifying the subgroup of "responders" and evaluate the performance of candidate decision rules via the corresponding benefit among treated patients. We define the mean benefit among

treated under a decision rule d as

$$\begin{aligned} B(d) &= \mathrm{E}_{\boldsymbol{X}}\{\Delta(\boldsymbol{X}) \,|\, d(\boldsymbol{X}) = 1\} \\ &= \mathrm{E}_{\boldsymbol{X}}\{\mathrm{E}[Y|\, \boldsymbol{X}, A = 1] - \mathrm{E}[Y|\, \boldsymbol{X}, A = 0] \,|\, d(\boldsymbol{X}) = 1\}\,, \end{aligned}$$

i.e, a mean difference in the expected responses between the two treatment arms, restricted to the patients who are assigned to treatment under the rule d.

The idea is to identify a subgroup of patients who would benefit from the treatment and avoid treating patients who would not. Such a "concentration of benefit" is often desirable since treating un-responsive patients may be detrimental, especially in cases with limited access to the treatment (for example, organ transplants), in situations with high treatment costs, or potential side effects.

The range of estimated scores indicates the magnitude of variation in the treatment benefit. Also graphing the (cross-validated) smooth curves against the scores might provide information about what kind of treatment benefit can be expected in the target subpopulations. If we indeed do not expect any large negative treatment effects, the focus could be shifted to those patients whose expected benefit is positive, or more generally, greater than some threshold c.

We now introduce a new measure of performance for a general decision rule $d(S; c) = \mathrm{I}\,(S > c)$, which will focus on the expected benefit among only those who are assigned to the treatment under this regimen. The mean benefit among treated under a decision

rule $d(S;c) = \mathrm{I}\,(S > c)$ is defined as

$$B(c) = \int_c^{\infty} [\mu_1(s) - \mu_0(s)]\, dF_{S|S>c}(s)\,,$$

where μ_0 and μ_1 are the expected outcomes in the two treatment arms as functions of the score s; $\mu_a(s) = \mathrm{E}(Y|\, S = s, A = a), a \in \{0, 1\}$.

Considering a more general threshold c could help avoid assigning treatment to a large portion of non-responders whose estimated score is positive but small (for some $c > 0$), or simply when a certain minimal treatment benefit is required in order to achieve an appropriate cost-benefit ratio. Clearly, if the expected treatment benefit is a non-decreasing function of the score, then the mean benefit among treated will increase with the cut-off. Hence, our goal is not to find c that leads to the largest $B(c)$, but rather to investigate whether there exist candidate thresholds c, that corresponds to a subgroup with a scientifically meaningful expected benefit and a subgroups size, $\mathrm{E}(\mathrm{I}\,[S > c])$, that warrants the adoption of a guideline. When considering $B(c)$ as an alternative measure of performance, we can focus on the expected benefit among patients with e.g. top 20% of the scores, in order to aid the panel size selection.

2.2.7 Estimation of Benefit Among Treated

Recall that the mean benefit among treated under a decision rule $d(S;c) = \mathrm{I}\,(S > c)$ is defined as

$$B(c) = \int_c^{\infty} [\mu_1(s) - \mu_0(s)]\, dF_{S|S>c}(s)\,,$$

where μ_0 and μ_1 are the expected outcomes in the two treatment arms as functions of the score s; $\mu_a(s) = \mathrm{E}(Y|\, S = s, A = a), a \in \{0, 1\}$. We will estimate the benefit among treated by replacing the score distribution function by its empirical version and the expected outcome functions $\mu_a(s)$ with the fitted smoothing splines

$$\widehat{B}(c) = \frac{1}{n_c} \sum_{i:\, s_i > c} \left[\widehat{\mu}_1(s_i) - \widehat{\mu}_0(s_i)\right],$$

where n_c is the number of observed scores s_i that are greater than c. The variance of $\widehat{B}(c)$ is then written as

$$\begin{aligned} var\widehat{B}(c) &= \frac{1}{n_c^2}\left[var\left(\sum_{i:\, s_i>c} \widehat{\mu}_0(s_i)\right) + var\left(\sum_{i:\, s_i>c} \widehat{\mu}_1(s_i)\right)\right] = \\ &= \frac{1}{n_c^2} \sum_{i:\, s_i>c} \sum_{j:\, s_j>c} \left[cov(\widehat{\mu}_0(s_i), \widehat{\mu}_0(s_j)) + cov(\widehat{\mu}_1(s_i), \widehat{\mu}_1(s_j))\right], \end{aligned}$$

which can be estimated by replacing the pairwise covariances with their estimates.

We will express the variance-covariance matrix $Var[\widehat{\mu}_k(\boldsymbol{s})]$ at the observed values of scores, $\boldsymbol{s}$, in terms of the matrices for the smoothing splines estimation problem. As already described in section 2.2.3, for any set of values $\boldsymbol{s}$, the predicted values at $\boldsymbol{s}$ can be expressed as

$$\widehat{\mu}(\boldsymbol{s}) = \boldsymbol{B}(\boldsymbol{s})\widehat{\boldsymbol{\beta}} = \boldsymbol{B}(\boldsymbol{s})(\boldsymbol{B}(\boldsymbol{z})^T\boldsymbol{B}(\boldsymbol{z}) + \lambda_P \boldsymbol{D})^{-1}\boldsymbol{B}(\boldsymbol{z})^T\boldsymbol{y} = H(\boldsymbol{s})\boldsymbol{y}\,.$$

If we assume that $var(\boldsymbol{Y}) = \sigma^2 \mathrm{I}$, the covariance matrix of the predicted values $\widehat{\mu}(\boldsymbol{s})$ is then given by

$$Var[\widehat{\mu}(\boldsymbol{s})] = \sigma^2 H(\boldsymbol{s})H(\boldsymbol{s})^T\,.$$

The estimation of $\widehat{\sigma}^2$ is usually based on sample residuals. Since $var(\widehat{\epsilon}_i) = \sigma^2(1 - h_{ii})$, where h_{ii} is the i-the diagonal element of the hat matrix $H(\boldsymbol{x})$, we can estimate σ^2 by the sample variance of standardized residuals. That is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\epsilon}_i^* - \overline{\widehat{\epsilon}^*}_i)^2\,,$$

where $\widehat{\epsilon}_i^* = \widehat{\epsilon}_i/\sqrt{1 - h_{ii}}$ is the i-th standardized residual.

The individual components of matrix $H(\boldsymbol{s})$ can be calculated using, for example, functions `getbasispenalty` (matrix $\boldsymbol{D}$) and `getbasismatrix` (matrices $\boldsymbol{B}(\boldsymbol{z})$ and $\boldsymbol{B}(\boldsymbol{s})$) in R package `fda`, and the estimated parameter $\widehat{\lambda}_P$ is part of the standard output from the function `smooth.spline`.

It is also possible to obtain the matrix $H(\boldsymbol{s})$ directly by use of the function `predict.smooth.spline`. If $\boldsymbol{z}$ is the vector which the smoothing spline is based on and *df* is its desired degrees of freedom, then $H(\boldsymbol{s})$ can be recovered by the following procedure, which uses fake outcome vectors $y_i = (0, 0, ..., 1, 0, ...0)$ with "1" on the i-th position, for $i \in \{1, ..., n\}$, to retrieve columns of the hat matrix $H(\boldsymbol{s})$.

```
> H = matrix(0, length(z), length(s))
> for (i in 1:length(s)) {
      y = rep(0, length(s)); y[i] = 1
      yi = predict(smooth.spline(z, y, df=df), s)$y
      H[,i] = yi;
  }
```

Hence, if we set $\boldsymbol{z} = \boldsymbol{s}[tx = a]$ equal to a subset of scores for patients whose treatment

is a and $\boldsymbol{s}$ the full set of scores, we will obtain the estimates of covariance matrices $Var[\widehat{\mu}_a(\boldsymbol{s})]\,,\ a \in \{0,1\}$, needed for the estimation of $var\widehat{B}(c)$.

2.3 Simulations

In this section we demonstrate properties of our proposed estimator of the marker-guided population response and also compare its performance to the standard non-parametric estimator. For that purpose we will generate data sets in which the outcome follows some specific linear model and will examine both how the LASSO is able to pick up the signal and how well the two estimators are assessing the population response under the developed decision rules.

When adopting a parametric model for data generation, we need to first specify all its parameters. In case of a linear additive model (2.2), the main parameters are the mean outcome among non-treated (β_0), marginal treatment effect (γ_0), marginal marker effects and marker-by-treatment interactions. For the actual data generation, we need to additionally specify a distribution of the markers (e.g., via minor allele frequencies for SNPs) and the variance of random effects (σ^2).

In presence of marker-by-treatment interactions, special care is needed to preserve the marginal treatment effect (γ_0) in the population, as well as the mean outcome among non-treated (β_0). This can be done by centering the values of the markers by their means when generating the expected outcome for a given set of marker values $\boldsymbol{X}$,

$$\mathrm{E}[Y \,|\, \boldsymbol{X}, A] = \beta_0 + \gamma_0 A + \sum_{j=1}^{N} (\beta_j X_j^* + \gamma_j X_j^* A)\,,$$

where $X_j^* = X_j - \mathrm{E}X_j$.

When comparing performance of our methods under various scenarios of marker effects and interactions, we might want to control the total variance of generated outcomes instead of $\sigma^2 = var(Y|\,\boldsymbol{X}, A)$. The total variance in a treatment group a is given by

$$\sigma^2_{Ta} = var(Y|\,A = a) = \sum_{j=1}^{N} \left[(\beta_j + a\gamma_j)^2 \; var(X_j)\right] + \sigma^2$$

Controlling the total variance might be of interest when we wish to assess the behavior of our estimators under multiple hypothesized scenarios for some real-life data, where we only know the total variance based on the observed data set.

2.3.1 Part I: Performance of the Estimators

The first goal of our simulations was to examine how both the standard non-parametric and our new estimator perform in the assessment of the marker-guided population response. Specifically, we wanted to investigate how the bias and variance compare between the two estimators and what is their estimation quality across different marker panel sizes. In order to assess the bias, the estimates $\widehat{\mathrm{E}}Y(d_p)$ were compared to the true population response under d_p's that would be eventually used in the population.

In the 10-fold cross-validation algorithm, a sequence of nested models is estimated using LASSO for each training set and for each marker panel size $p = 1, 2, \ldots$, the scores are calculated for all patients in the corresponding test set. Since the cross-validation mechanism results in a different decision rule d_p for each of the $K = 10$ folds, we calculated a new set of $\widehat{\boldsymbol{\gamma}}^p$ based on the whole sample, as would be done if the rule were to be eventually exported. Then we assessed the true population response under such

d_p based on a large sample of 2×10^5 subjects from the target population, assigned to the treatment according to that d_p.

Data Generation

In the following simulations, we considered data sets of $n = 1000$ subjects, each with $m = 200$ independent markers (SNPs) that take values 0, 1, or 2, representing the number of minor alleles. The minor allele frequencies (MAF) were randomly generated from a uniform distribution with values between 0.1 and 0.5 (rare SNPs were not present here). Then, for $j = 1, \ldots, m$,

$$\begin{aligned} \mathrm{P}\left[X_j = 0\right] &= (1 - \mathrm{MAF}_j)^2 \,, \\ \mathrm{P}\left[X_j = 1\right] &= 2\,\mathrm{MAF}_j(1 - \mathrm{MAF}_j) \,, \\ \mathrm{P}\left[X_j = 2\right] &= \mathrm{MAF}_j^2 \,. \end{aligned}$$

In this first set of simulations, we adopted a slightly different parametrization of the linear model (2.2), where the marker effects were considered separately under the treatment ($\boldsymbol{\beta}^1$) and no treatment ($\boldsymbol{\beta}^0$) arm and generated from a mixture distribution – which we believe might mimic some real-life scenario, as follows.

For 20% of the SNPs the marker effect was moderate to large ($\sim$exp(1/5)) for at least one of the treatment arms, creating moderate to large marker-treatment interactions. For another 15% of markers, the marker effects were small ($\sim$N(0,0.1)) in both arms, which resulted in small interactions. For the remaining 65% there was no marker effect in either of the two arms, and hence consequently no interactions (see example in Fig. 2.2).

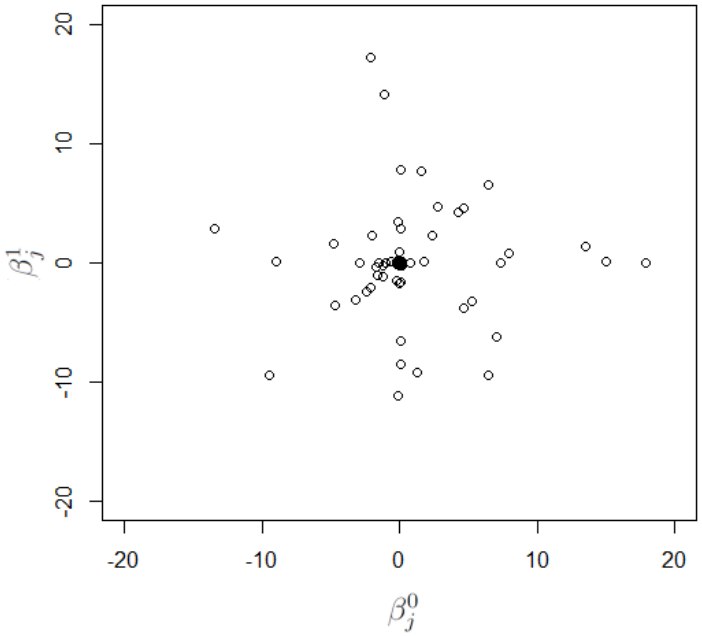

Figure 2.2: Marginal SNP effects in the two treatment arms; β_j^0 is the marginal effect of SNP$_j$ if A=0 and β_j^1 is the marginal effect of SNP$_j$ if A=1. The difference $(\beta_j^1 - \beta_j^0)$ corresponds to the interaction (γ_j) between the SNP j and treatment.

The treatment was assigned to subjects randomly with probability $\frac{1}{2}$ and the mean model for the outcome Y consisted of a linear combination of the marginal treatment effect and the individual SNP effects for the given genotype and treatment:

$$Y_i = \beta_0 + \gamma_0 A_i + (\sum_{j=1}^{N} \beta_j^0 X_{ij}(1 - A_i) + \beta_j^1 X_{ij} A_i) + e_i, \quad \text{where} \quad e_i \sim N(0, \sigma^2), \quad (2.5)$$

$\beta_0 = 10, \gamma_0 = 10$, and $\sigma = 10$. For the data generation, each SNP was centered at its mean in order to preserve the marginal treatment effect, γ_0. The noise added to the outcome had Normal $N(0, \sigma^2)$ distribution, with the standard deviation equal to the

marginal treatment effect.

Results

One set of minor allele frequencies (MAFs) and corresponding $(\boldsymbol{\beta}^0, \boldsymbol{\beta}^1)$ parameters was generated (Fig 2.2) and used for all the following simulations in Part I, as they jointly determine the marker-outcome distribution in the underlying population. From the total of 200 SNPs, the number of SNP-treatment interactions γ_j=$(\beta_j^1 - \beta_j^0)$ larger than 1 in absolute value was 35 and larger than 0 was 70. From the linear model (2.5), we can see that the mean population response if no one gets treated is $\beta_0 = 10$, and it is $(\beta_0 + \gamma_0) = 20$ if we treated everybody. The resulting total standard deviation in each treatment arm was approximately $\sigma_{Ta} = 20$.

Based on a very large sample of 2×10^5 subjects, we assessed the performance of theoretically optimal rule under which everybody was treated optimally based on their marker values (we know the truth!). The mean population outcome under such rule is as high as $\mathrm{E}Y(d^*) = 28.4$ and under the given "nature", defined by the one (from now on fixed) set of parameters $(\boldsymbol{\beta}^0, \boldsymbol{\beta}^1)$ and MAFs, this is the maximum possible marker-guided population response. Now we would like to see how close to "the best" we can get using our cross-validation algorithm, outlined in section 3.2.5.

In order to assess the bias and variance of the two estimators, we generated three hundred data sets of 1000 subjects. In each data set, we first performed the 10-fold cross-validation algorithm with resulting estimates of MGPRs, and then we assessed the true MGPR under the decision rules $d_p, p = 1, 2, ...,$, based on the whole sample, as described above.

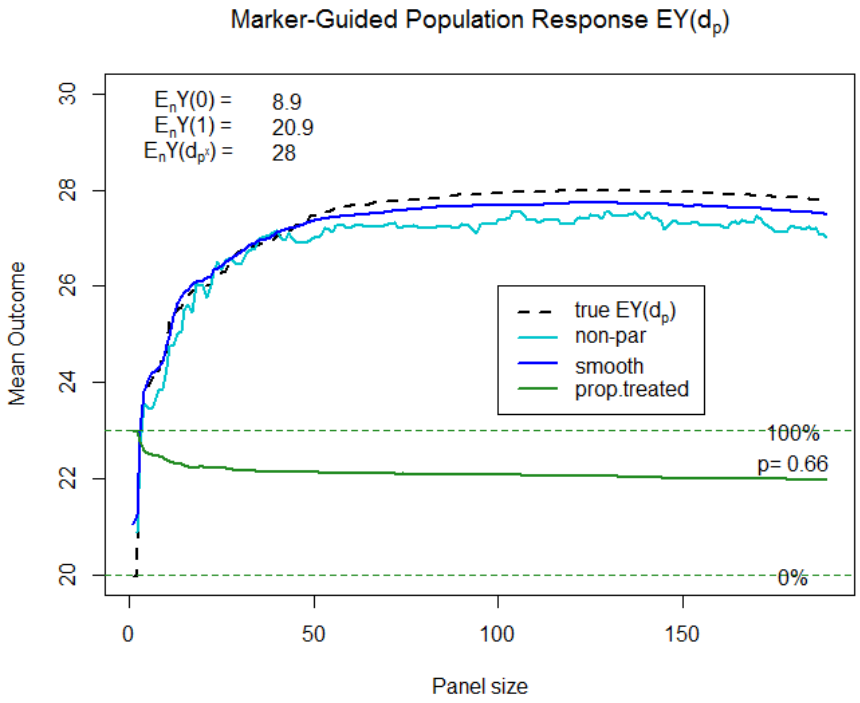

Figure 2.3: One random sample: Marker-Guided Population Response estimated by the standard non-parametric estimator (turquoise) and our new estimator based on smoothing splines (blue), compared to the true population response under the candidate decision rules (black). The green line shows the proportion of treated patients under the candidate decision rules based on different panel sizes p.

Results from one example data set can be seen in Fig. 2.3, which shows the estimated clinical objective, MGPR, as a function of the panel size. The black dashed line shows the population response under decision rules d_p that were based on the whole data set. The turquoise curve is an estimation using the standard non-parametric approach and the blue curve is the estimator using our smoothing spline approach.

For this particular sample, the average response among non-treated patients was $\mathrm{E}_nY(0) = 8.9$ and among treated patients $E_nY(1) = 20.9$. The MGPR estimates by smoothing splines approach suggest that we can improve the population response to,

e.g., as much as $\widehat{\mathrm{E}}_n Y(d_{50}) = 28.0$ if the rule based on just 50 top markers was used to guide the treatment. It is very encouraging to see that marker-guided rules may lead to mean population outcome that is actually fairly close to the theoretical optimum for this population structure.

Additionally, the green line in Fig. 2.3 shows how the proportion of treated patients decreases with larger panel sizes and then stabilizes at around 66%. This implies that by using a marker-guided treatment rule, not only can we largely improve the population response but we can achieve it by prescribing treatment to only a fraction of the target population.

In general, the smoothing spline (SS) estimator seems to approximate the true MGPR more closely that the non-parametric (NP) one, as can be also seen on multiple examples in Fig. 2.4. We moreover observe that the SS estimator is much more stable across panel sizes than the standard NP estimator. As already mentioned earlier, the larger variation is probably occurring due to the fact that the standard non-parametric estimates for different panel sizes not only differ in the quality of the corresponding decision rules but also in the subset of patients classified in agreement with their observed treatment, whom the estimates are based upon.

Figure 2.5 shows the squared bias and variance of the two investigated estimators, based on 300 samples of 1000 subjects. As expected, both estimators proved to be unbiased and the variance of the SS estimator is consistently smaller than the variance of the standard NP estimator across all panel sizes. These results again favor the use of the smoothing spline estimator over the standard non-parametric approach when the target clinical objective MGPR needs to be evaluated under various hypothetical

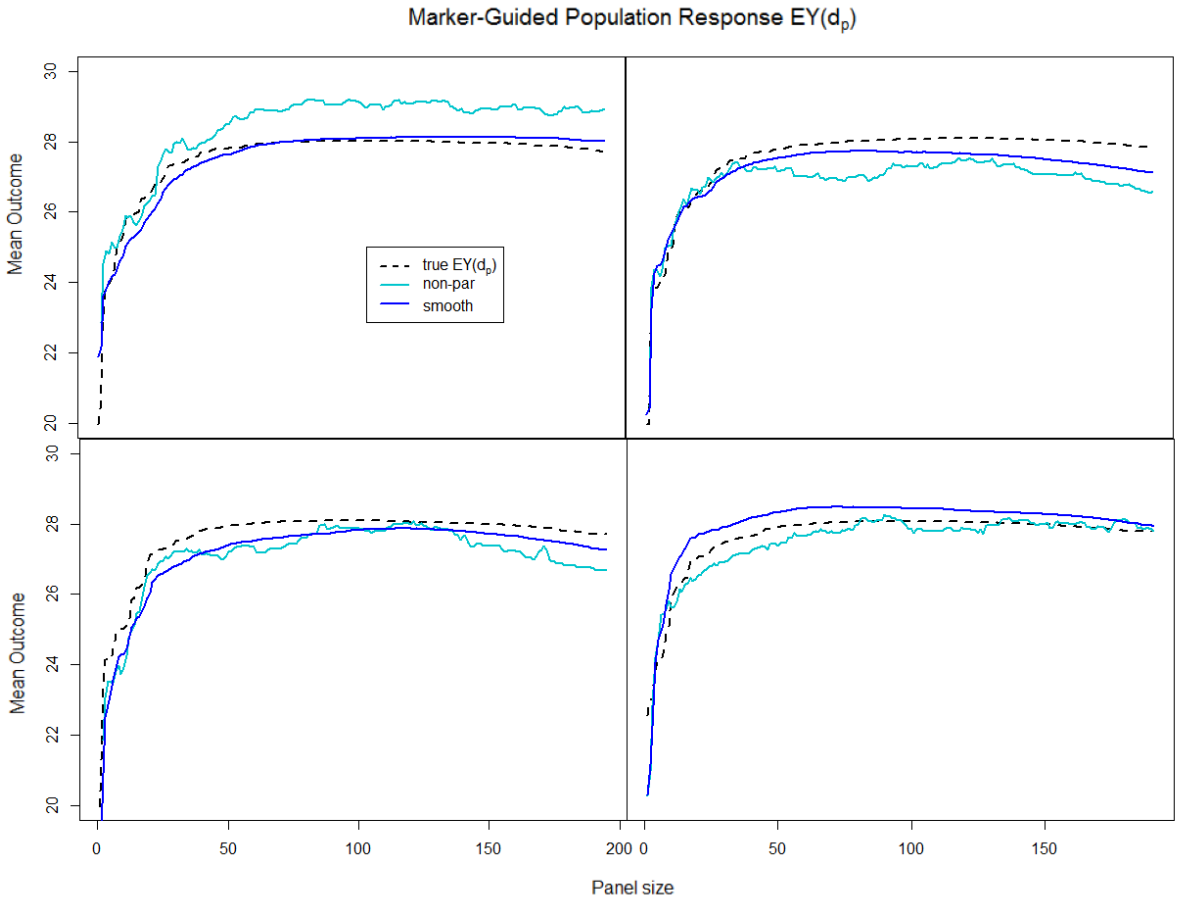

Figure 2.4: Four random samples: Estimated Marker-Guided Population Response (MGPR) as a function of the marker panel size: a) true MGPR (dashed line) b) MGPR estimated by the standard non-parametric approach (turquoise) c) MGPR estimated by smooth curves (blue).

treatment assignment rules.

2.3.2 Part II: Null Markers

In the second part of our simulations we assess how much inclusion of numerous null markers reduces the ability of LASSO to pick up the true signal, and how it consequently affects the performance of the developed decision rules. By null markers or noise we consider markers that have truly no interaction with the treatment and are present in

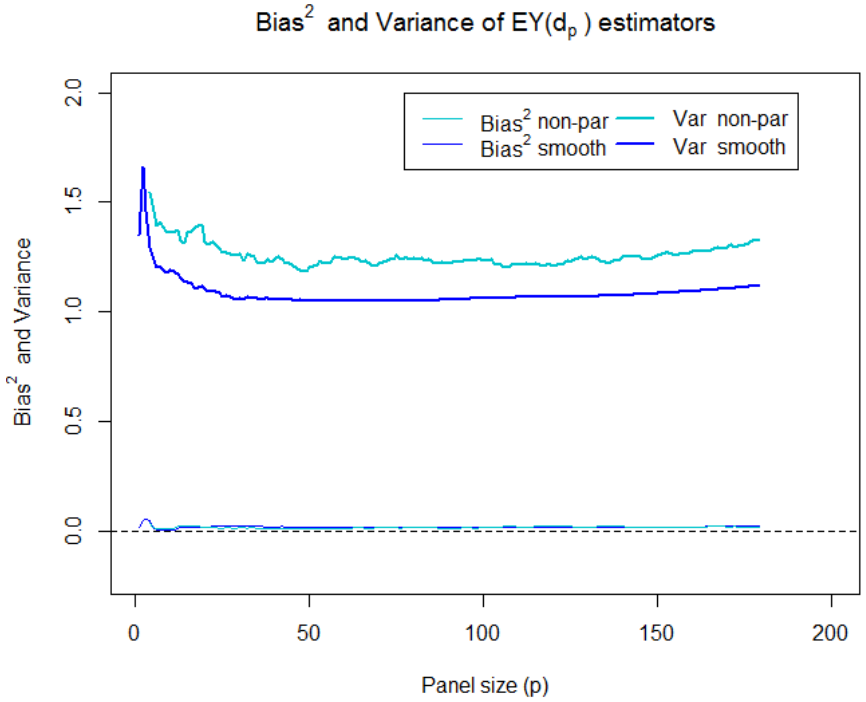

Figure 2.5: Squared bias (thin lines) and variance (thick lines) of the two estimators as functions of the panel size p based on 100 samples from the same population.

the pool of markers considered for the score development.

The performance of the two estimators was now examined across large initial pools containing 1000, 2000 and 3000 markers, from which only 320 had non-zero interactions with the treatment. The estimates $\widehat{\mathrm{E}}Y(d_p)$ were compared to the true population response under a d_p that would be eventually used in the population, as in the previous section.

Data Generation

Again we considered data sets of 1000 subjects, now each with 1000, 2000, or 3000 independent markers (SNPs) that take values 0, 1, or 2. The minor allele frequencies

were randomly generated from a uniform distribution with values between 0.1 and 0.5. As for the parametrization we now switched back to the model (2.2) with marker main effects ($\boldsymbol{\beta}$) and interactions with treatment ($\boldsymbol{\gamma}$).

The main effects (β_j's) for the first 320 markers came from $N(0, 3)$ and the rest was set to 0. The marker-treatment interactions (γ_j's) were divided into four categories; 20 large (effect size in absolute value between 6 and 15), 50 moderate (effect size in absolute value between 1 and 5), 250 small (effect size in absolute value between 0 and 1) and the rest was null. We also kept the effect sizes and the minor allele frequencies the same across the three scenarios. The treatment was assigned to subjects randomly with probability $\frac{1}{2}$ and the model parameters are the same as in Part I, $\beta_0 = 10, \gamma_0 = 10$ and $\sigma^2 = 10$, with $e \sim N(0, \sigma^2)$.

Results

Similarly as in the first part of simulations, we generated data sets of 1000 subjects for each of 1000, 2000 and 3000 marker scenarios. Results from one example data set for each of the three scenarios can be seen in Fig. 2.6, which shows the estimated clinical objective, MGPR, as a function of the panel size.

Red dots in the graphs are values of the true population response under 'oracle' rules (best we can do) based only on top p markers, as ordered by their true size of the interaction with the treatment. For example, if p=20, the classification is based on $d(\boldsymbol{x}) = \mathrm{I}\left(\boldsymbol{x}_{20}\gamma_{20} > 0\right)$, where $\boldsymbol{x}_{20}$ is a set of 20 markers with the largest treatment interactions and γ_{20} is a vector of corresponding true interactions, and so on. If p=0, no markers are used for classification and so the decision is based only on the marginal

treatment effect. Since γ_0=10 is positive, everybody would be prescribed to the treatment which results in the true population response of 20 ($\beta_0 + \gamma_0$). Finally, a decision rule using all 320 relevant markers with the correct values of their interactions would lead to population response of 29.2, which is the best we can do for this population.

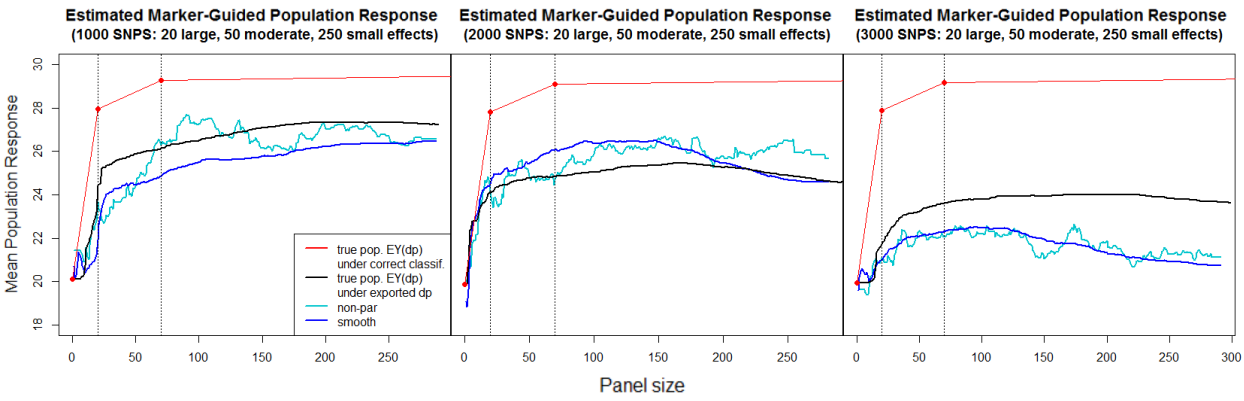

Figure 2.6: Three random samples: Estimated population response as a function of panel size. The initial marker pools contain the same (320) real-effect markers, but have different numbers of nulls: a) total of 1000 markers, b) total of 2000 markers, c) total of 3000 markers.

Naturally, the decision rules developed by the LASSO perform less optimal than the 'oracle' rules and so the population response under the former will be lower. Black lines show the true performance of decision rules developed on our samples, while the turquoise and blue lines are their cross-validated non-parametric and smooth-spline estimates, respectively.

Not surprisingly, the ability of LASSO to pick up the signal from marker pools with more noise is decreasing. We see that in the first scenario, where the marker pool contains only 1000 markers, some of the best developed decision rules can improve the population response to as much as $EY(d^*) = 27.3$. Yet, in the last scenario with initial marker pool

containing extra noise in form of additional 2000 null markers, even the best developed decision rules can only improve the population response to about $\mathrm{E}Y(d^*) = 24$.

Figure 2.7 shows the average proportion of the markers with large, moderate and small interactions selected by LASSO as a function of the panel size. These averages are based on 100 samples of 1000 subjects for each of the three scenarios. As expected, the more null markers are present in the initial pool, the poorer is the ability of LASSO to depict the markers that are truly relevant. When the initial pool contains total of 1000 markers, we see that for panel size $p = 20$, in which all the markers with large effects could have been selected, LASSO actually picked about 50% of those markers on the panel. On average, additional 6% of 50 moderate-effect markers (3) and 2% of 250 small-effect markers (2) were selected, resulting in total of 15/20 markers being "true positives" and 5/20 "false positives". By the time the panel size is $p = 70$, there is about 80% of the markers with large effects (16) in the model and it approaches 100% as the panel size increases even more. For $p = 70$, on average 18% of moderate-effect markers (9) and 6% of small-effect markers (15) enter the model, resulting in total of 40/70 markers being "true positives" and 30/70 "false positives". With initial pool containing additional 1000 and 2000 null markers, the ability of LASSO to pick the signal rapidly decreases.

2.3.3 Part III: Pre-filtering

The last part of the Simulations is devoted to a frequently controversial issue of pre-filtering. In cases when the number of available markers is very large, such as GWAS, one common approach is to pre-filter, or pre-order, markers based on their marginal

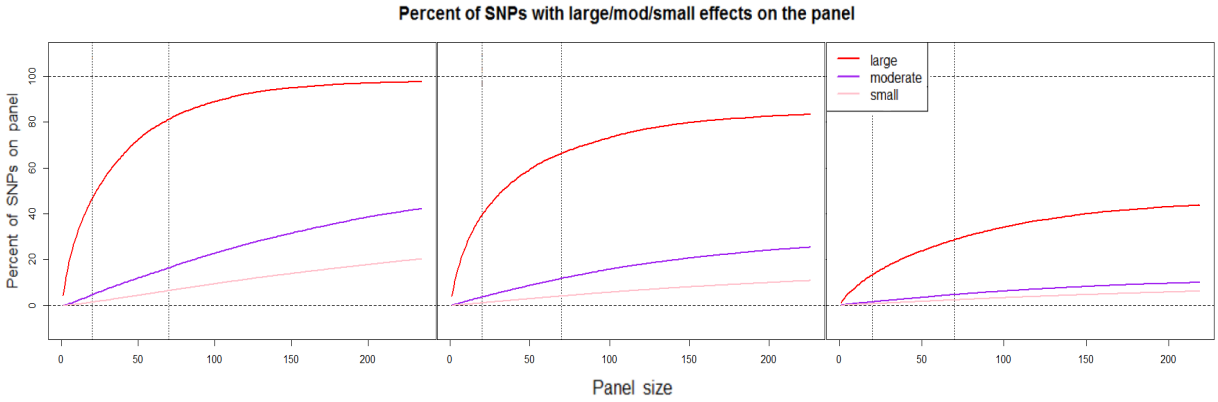

Figure 2.7: Average proportion of large(20)/moderate(50)/small(250) effect markers picked up by the Lasso (based on 100 random samples), as a function of panel size. The initial marker pools contain the same (320) real-effect markers, but different number of nulls: a) total of 1000 markers, b) total of 2000 markers, c) total of 3000 markers.

interactions with the treatment (Gunter et al., 2011; Matsui et al., 2012) and reducing thus the dimensionality of the estimation problem.

It has been pointed out (Kooperberg et al., 2010) that pre-filtering on parameters of interest might lead to overly optimistic results if a careful (e.g., cross-validated) evaluation is not in place. In order to investigate the impact of pre-filtering on the benefit score development, we performed several simulations under various "true" marker effect scenarios.

The settings for the following simulations were inspired by one of our examples, VISP study, which is described in detail later. In the VISP data set, we pre-filtered markers (SNPs) based on their marginal interactions with the treatment and the passing rule for the markers was p-value less than 0.01.

In our simulations we considered the following three scenarios for a set of 4×10^5 markers: a) there are no interactions with the treatment, b) there are 500 moderate

interactions + remainder no associations, c) 50 strong interactions + remainder no associations. The minor allele frequencies and the sizes of interactions and main effects were based on the corresponding distributions in our VISP data set. For scenarios b) and c), we had to generate new outcomes in order to follow the specified models, however, we preserved the marginal treatment effect, total variance of the outcome and the proportion of treated patients in the new data sets. The 50 strong interactions (c) were set to be equal to the top 50 marginal interactions from the VISP data and the 500 moderate interactions (b) were set to be the top 500 marginal interactions divided by two. The threshold for passing the filter was again the marginal interaction p-value of less than 0.01.

As expected with this threshold, about 1% of all 4×10^5 markers passed the filter for each of the three scenarios. In b), however, it was around 70 out of 500 markers (14%) with moderate interaction that passed and in c) 45 out of 50 markers (90%) with strong interaction passed the filter.

As we can see in Fig 2.8a), the population response was not improved by building a score from pre-filtered markers with no effects (noise). On the contrary, prescribing some patients to "no treatment" in a essentially random fashion leads to a lower population response than when we treat everybody. In Fig 2.8b), we see that those few markers with moderate interactions that passed the filter are being picked up by LASSO and the mean population response can be improved slightly. In Fig 2.8c), most of the markers with the strong interactions are still in the marker pool and so our algorithm can build a score that greatly improves the mean population response. Moreover, our estimator(s) assessed the population response under decision rules d_p correctly in all the three scenarios.

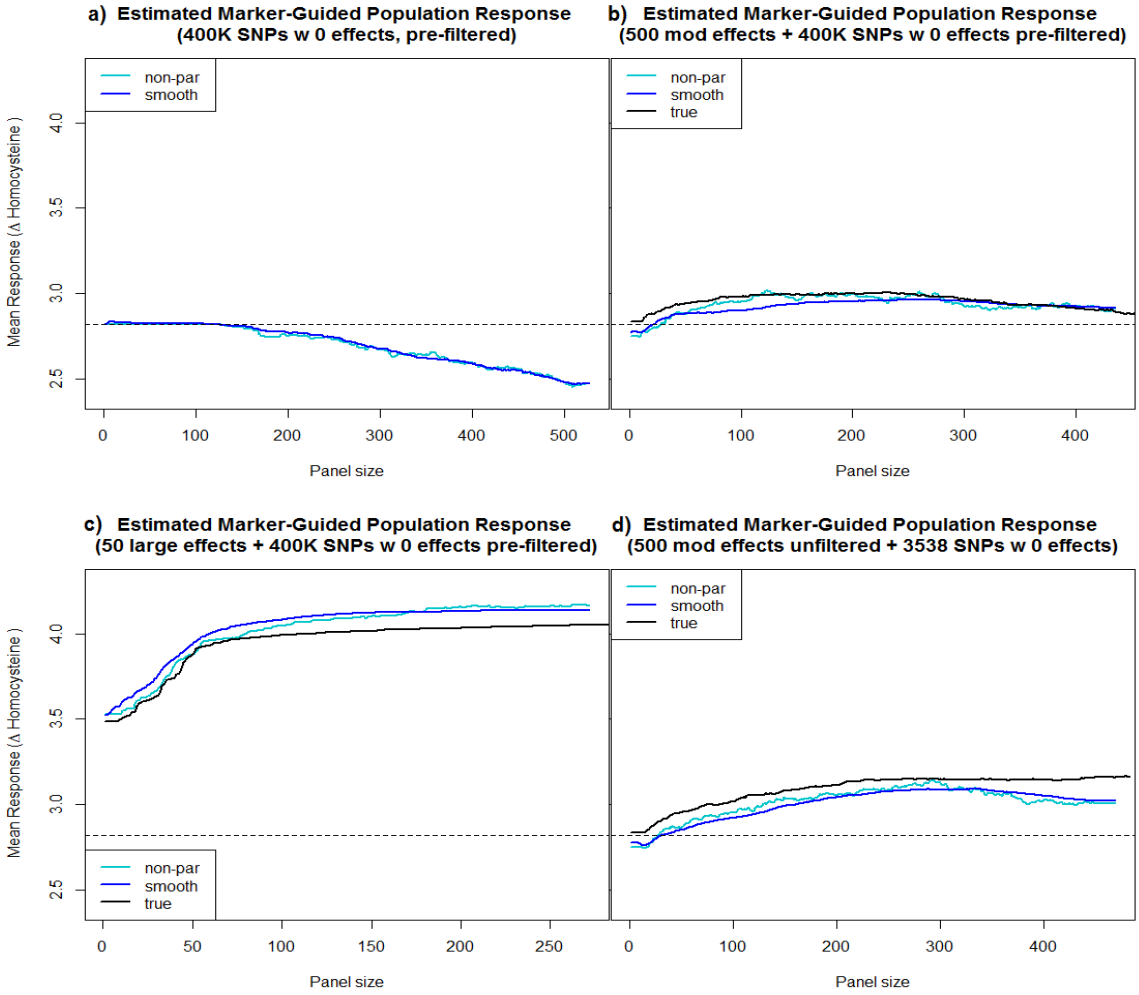

Figure 2.8: Estimated marker-guided population response (MGPR) for the marker panels pre-filtered on interactions under the 3 different scenarios (a-c) and d) panel including all 500 markers of moderate-size interactions (without pre-filtering them).

In addition, we ran a modified analysis of case b), where all 500 markers with moderate interactions were included in the pool of 4000 markers supplied to LASSO, seen in Fig 2.8d). We see that LASSO was able to pick up some of these markers and build decision rules slightly better than those in Fig 2.8b), however, the moderate effects are harder to find and so the performance is not as good as in scenario c) seen in Fig 2.8c).

Even though the potential bias was our primary concern and the reason why we

conducted simulations on pre-filtering at first place, we were very surprised that none of our multiple simulated data sets shown any obvious bias of the estimated population response. An example of four random samples in Figure 2.8 illustrates how the estimated $EY(d_p)$ copy the truth very well in both scenarios b) and c) with either 500 moderate and 50 large effects, respectively.

In conclusion for our methods, it appears that in our setting pre-filtering can harm us a little by removing a large portion of markers with small and moderate interactions from the marker pool and hence leading construction of more poorly performing decision rules. The majority of markers with large interactions should however pass the filters and would likely get integrated into the developed decision rules. Even more importantly, both cross-validated estimators assess the population response under the developed decision rules fairly well.

2.4 Examples

2.4.1 LESS Example

Data for our first illustrative example come from the Lumbar Epidural injections for Spinal Stenosis (LESS) study, an NIH-funded, multi-center, double-blind, randomized, controlled clinical trial. It was designed to evaluate the effectiveness of epidural steroid injections additional to standard treatment by local anesthetics (Lidocaine) in improving pain and function among older adults with lumbar spinal stenosis (Friedly et al., 2012).

The continuous outcome we chose to analyze is a numerical measure of back pain 3 weeks after the treatment was administered. The back pain questionnaire asks patients to rate the intensity of their back pain on the scale between 0-10. Since the higher values

on this scale correspond to worse outcomes, our response variable was taken as 10 minus the back pain scale in order to preserve the goal of maximizing the population response. The new outcome can be thought of as a back "comfort" and is also on the scale 0-10.

Our data set consists of 383 subjects who had the measurement for back pain at 3 weeks available, from which 189 were assigned to the steroid injections plus Lidocaine (A=1) and 194 were assigned to Lidocaine only (A=0). As potential predictors for the analysis we selected all the variables that were considered possibly relevant in a clinical sense and had available values at the baseline (see the list in the table below). All the categorical predictors were modeled using dummy variables and the continuous predictors were modeled linearly, with exception of BMI, for which we additionally included a quadratic term.

Binary variables	Categorical variables	Continuous variables	
Sex	Race	Back pain at baseline	Sssq phfu
Marital status	Education	Roland	Sssq symptoms
Hispanic race	Working status	BPI	Blood glucose
Smoking status	Duration of pain	EQ5D index	Hga 1c
Lawyer	Site	PHQ 8	Cortisol
	MRI screen	FabQ	Age
		Pcs	Hight
		Gad 7	Expectation

The average back "comfort" value at 3 weeks post treatment was 5.75 in the Lidocaine group and 6.27 in the group with additional steroid injections, which makes a difference of about 0.5 point on the 10-point scale. We wanted to examine if using the patient-level characteristics might help us differentiate between individuals who would benefit from the steroid therapy and who would not, in terms of back pain level, and how much improvement in the population we would see if the markers were used to guide the

treatment.

Similarly to the simulations (Section 3.3), in each fold of the cross-validation procedure the L_1-PLS estimator served to develop a sequence of decision rules based on nested set of marker panels, $p = 1, 2, ...$ For each panel size, the marker-guided population response (MGPR) was then estimated based on the whole data set via both standard non-parametric and smoothing spline estimator.

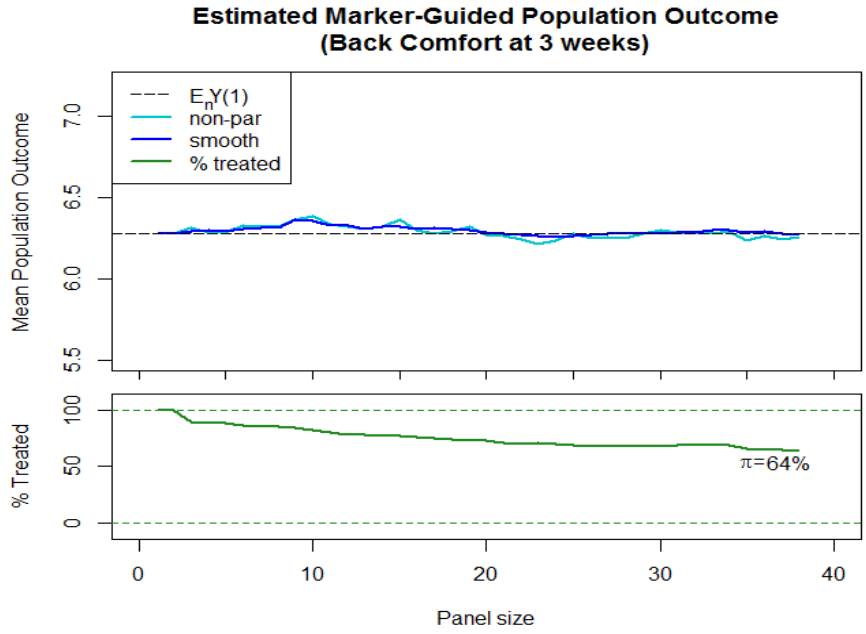

Figure 2.9: LESS example: Estimated marker-guided population response (negative back pain at 3 weeks) vs. panel size using the standard non-parametric estimator (turquoise) and the smoothing splines estimator (blue).

Fig. 2.9 shows how the estimated clinical objective, MGPR, under a rule d_p changes with an increasing panel size p. The SS estimator again provides more stable estimates of the MGPR than the standard NP estimator. The clinical objective function in Fig 2.9 suggests that using patient-level markers to guide the treatment improves the mean back comfort level in the target population only minimally. However, it appears to be able

to select (about two thirds of) patients who would benefit from the addition of steroid injections, while we could withdraw the treatment for a third of the patients who are seemingly unresponsive.

Since the estimated population response does not seem to change across nominated decision rules, we decided to examine our second proposed measure of performance, mean benefit among top 20% treated. Now the compared decision rules are based on the same marker panels and corresponding scores $\boldsymbol{s}^p$ as before, but the cut-offs for the treatment assignment are such that only 20% of the patients with the top scores are assigned to the treatment. A decision rule based on top p markers and corresponding cutoff c_p can be written as

$$d_p(s; c_p) = \mathrm{I}\left[s > c_p\right],$$

where c_p is such that $P(S^p > c_p) = 20\%$. Figure 2.10 shows how the estimated mean benefit among top 20% treated changes with an increasing panel size. The highest value is achieved with the decision rule based on 20 markers and $\widehat{B}_{20}(c_{20}) = 1.3$.

If the mean benefit among treated of 1.3 points on the (0-10) back pain/comfort scale was considered clinically relevant, then our score based on top 20 markers could be useful in the selection of a patient subgroup with such benefit and size 20% of the population. The estimated (non-zero) coefficients of the 20-marker score are listed in the table below and the figure 2.11 shows their relative sizes after standardization of the continuous markers.

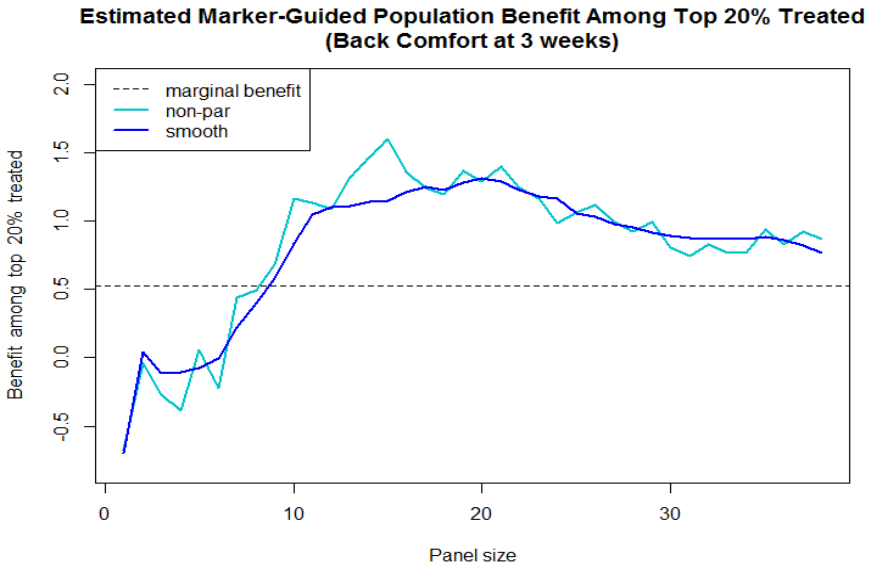

Figure 2.10: LESS example: Benefit among (top 20%) treated patients as a function of the panel size, estimated by the standard NP estimator (turquoise) and our SS estimator (blue).

Coef.	Marker	Coef.	Marker
0.9477 ×	Marital status (married)	0.3448 ×	MRI screen (moderate)
1.1336 ×	Smoking history	**0.9648** ×	MRI screen (severe)
2.0913 ×	Lawyer	**-0.9453** ×	Site (2)
1.7745 ×	Race (other*)	0.0153 ×	Site (3)
0.3297 ×	Working status (retired)	**-1.7316** ×	Site (4)
-0.0277 ×	Working status (disabled)	0.9648 ×	Pcs
0.2038 ×	Education (some college)	-0.0690 ×	Gad7
-0.2012 ×	Education (college)	-0.8177 ×	EQ5D index
0.1179 ×	Duration of pain (3-12mo)	-0.0004 ×	PHQ 8
-0.0114 ×	Duration of pain (1-5yr)	0.0118 ×	Cortisol

* Race other = non-Caucasian and non-African-American

Among the markers with the highest impact on the score S^{20} is whether the patient hires a lawyer, race other than Caucasian and African-American, marital status, smoking history and severity of MRI. A variable that have the largest negative impact on the pre-

dicted treatment benefit is the provider's site, particularly sites 4 and 2. The coefficient estimates are fairly consistent across the 10 folds of our cross-validation procedure.

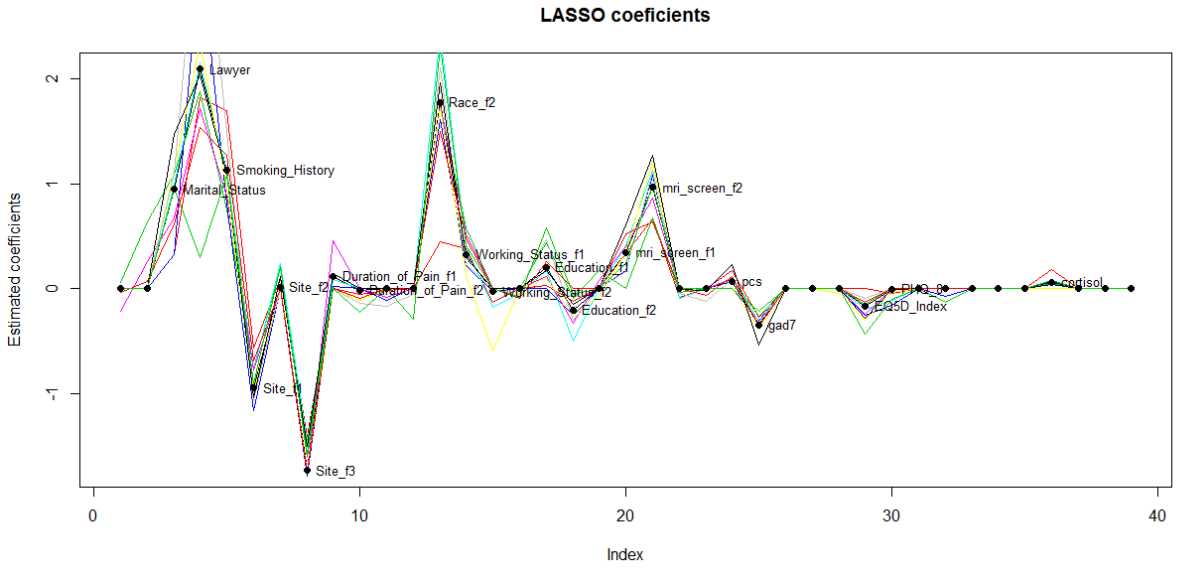

Figure 2.11: LESS example: Relative size of estimated coefficients for the 20-marker score after standardization of the continuous markers.

More generally, we may focus on those patients whose expected benefit is large or greater than some constant c. Our second proposed measure of performance, mean benefit among treated, can be then thought of as an average expected benefit among those patients whose score is greater than c. This is inherently an increasing function of c, however, the question is whether the higher cut-offs result in a mean benefit among treated that is scientifically meaningful in the context of the study, while the percentage of treated patients is still high enough for its practical purposes.

Figure 2.12 shows the relationship between percentage of treated patients as a function of the score cut-off c and the corresponding mean benefit among treated if that cut-off was used. The red dot on the very left corresponds to the mean benefit among

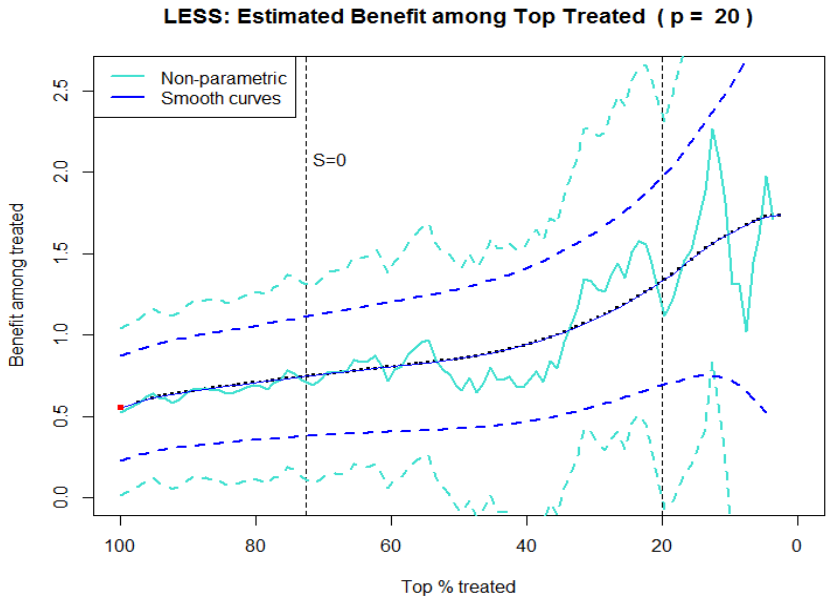

Figure 2.12: LESS example: Benefit among (top) treated patients as a function of the percentage of treated corresponding to various score cut-offs. Red dot on the left represents the marginal benefit from the additional steroid treatment.

treated if everybody was treated, irrespective of the score. This is in other words equivalent to the marginal treatment benefit of 0.5. The vertical line on the left indicates where the score cut-off of zero would be ($c = 0$), which corresponds to about 72% of patients being assigned to the treatment and an estimated mean benefit of 0.75. The vertical line on the right corresponds to the cut-off that leads to treating patients with the top 20% of the scores, and the estimated mean benefit among those top 20% of treated patients is 1.3 points of the (0-10) back pain/comfort scale.

2.4.2 VISP Example

Data for our second example come from the Vitamin Intervention for Stroke Prevention (VISP) study, which was also an NIH-funded, multi-center, double-blind, randomized

clinical trial. This study was designed to determine whether a daily intake of high dose of vitamins B6, B12 and folic acid reduced recurrent cerebral infarction and a combined vascular endpoint (Spence et al., 2001).

The continuous outcome we chose for our example is the blood concentration of homocysteine, high levels of which are associated with cardiovascular disease (Clarke et al., 1991). It was hypothesized that supplementation with folic acid, B6 and B12 may reduce the homocysteine level in the blood. We wanted to see if using the patients' genetic information might help us differentiate between individuals who would benefit from the vitamin therapy and who would not, in terms of homocysteine level, and how much improvement in the population we would see if the genetic markers were used to guide the treatment.

We examine 1670 patients with Caucasian ancestry, from whom 837 were assigned to the high dose and 833 to the low dose of the vitamins. From the 803,122 SNPs that passed the quality control filters, a subset of 3,597 SNPs with the strongest marginal marker-by-treatment interactions and complete data was selected for the analysis. We acknowledge that marker pre-filtering is not an ideal approach, with some of its issues already discussed in section 2.3.3. However, this example serves only as an illustration of the proposed methods and we do not suggest to draw any conclusions from it.

The outcome of interest was taken as the average improvement (negative difference) in the homocysteine level on the visits at 6 months and later after the therapy started compared to the baseline. Notice that for an outcome defined this way, larger values correspond to larger decrease of homocysteine level and are hence considered better. The overall improvement in homocysteine level in the group on the low dose of the vitamins

was $\mathrm{E}_n Y(0) = 0.6$ μmol/L and $\mathrm{E}_n Y(1) = 2.8$ μmol/L in the high dose group. This amounts to an estimated marginal benefit of $\widehat{\gamma}_0 = 2.2$ μmol/L between the high dose and low dose vitamin therapy.

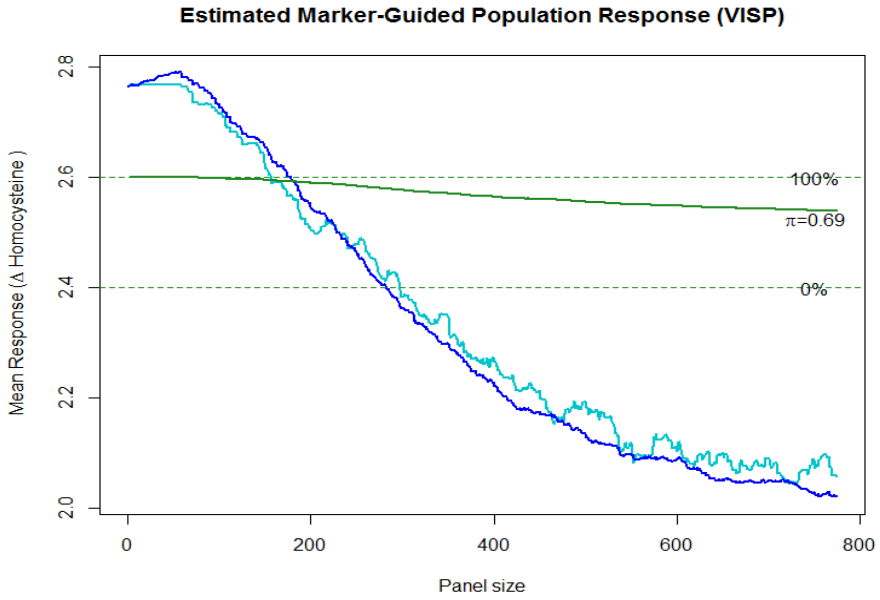

Figure 2.13: VISP example: Estimated Marker-Guided Population Response (homocysteine improvement) vs. panel size using the standard non-parametric estimator (turquoise) and the smoothing splines estimator (blue). The green line shows the proportion of treated patients under the corresponding decision rules.

Similarly to the simulations (Section 3.3), in each fold of the cross-validation procedure the L_1-PLS estimator served to develop a sequence of decision rules based on nested set of marker panels, $p = 1, 2, ...$ For each panel size, the marker-guided population response (MGPR) was then estimated based on the whole data set via both standard non-parametric and smoothing spline estimator.

Fig. 2.13 shows how the estimated clinical objective, MGPR, under a rule d_p changes with an increasing panel size p. The SS estimator again provides more stable estimates

of MGPR that the standard NP estimator. The clinical objective function suggests that using SNPs to guide the treatment does not actually improve the mean homocysteine levels in the target population. Instead, with subscribing higher portion of patients to the low dose vitamin therapy, the estimated population response decreases.

When we compare these results with our simulations in section 2.3.3, we are leaning toward conclusion that the VISP is an example of a study where the collected markers do not provide any information that can help to identify a subgroup of patients benefitting from the marginally inferior therapy. Even if markers predictive of treatment response exist in our data set, the LASSO method in the nomination step was unable to identify them from among a large number of nulls and hence failed in developing decision rules that would lead to improved mean population outcome.

2.5 Discussion

The proposed framework focuses on methods for evaluation of the potential for baseline markers to guide a treatment. With a continuous outcome, the assumed clinical objective is an improvement in the mean population response, which we seek to maximize over a set of candidate marker-based decision rules. On an individual level, the goal is to score patients with respect to their expected treatment benefit. Our algorithm for score development is based on a combination of variable selection methods and careful evaluation of the candidate rules via cross-validated estimation of the marker-guided population response.

For high-dimensional data, we proposed to select candidate marker panels by L_1 penalized least square method, which leads a nested set of marker panels and hence

allows easy navigation through a high-dimensional parameter space. While the LASSO quickly and effectively provides both marker selection and coefficient estimation, in the current literature it is often substituted by more complex algorithms for the nomination of candidate decision rules, which we do not attempt to compete with. In fact, our objective is to provide reliable tools for evaluation and comparison of such decision rules, separately and irrespective of the process of their development.

The working model for the score development was chosen linear and additive for its simplicity and good interpretability. Even though it is not necessarily believed to be the correct model, the resulting decision rules can still be helpful in selecting which patients should get treated and improving the mean population response of interest. Our goal is then to evaluate impact of using the candidate scores to make decision about treatment in the target population. Using the score as a single index predictor and fitting the expected outcome as a smooth function of the score allows us both to evaluate the mean population response more efficiently than through the standard non-parametric estimator and to assess the validity of the estimated benefit.

In situations where the target population is rather a mixture of responders and non-responders, the mean population response cannot be improved and it might be thus more useful to examine a benefit among treated instead. The benefit among treated offers an alternative way of evaluating whether a subgroup with a scientifically meaningful benefit and size exists. For example, researches can focus on the expected benefit among patients with top 20% of the scores, which would be then compared across panel sizes and the corresponding decision rules.

The goal of simulations was to assess the performance of the proposed smoothing-

spline estimator versus the standard non-parametric estimator, examine effects of noise in the initial pool of markers on the quality of developed rules, and examine the impact of pre-filtering. In comparison with the standard non-parametric estimator, the smoothing-spline estimator appeared to be consistently more stable and less variable across various examples and scenarios. Its other major advantage seems to reside in the fact that it additionally offers an assessment of validity of the score, which predicts the treatment benefit. As expected, a large number of noise markers (those with no interaction) in the initial pool decreases the ability of LASSO, and likely all approaches, to pick up the signal, which results in decision rules with poorer performance. Finally, pre-filtering simulations did not suggest any overestimation of the marker-guided population response, as the cross-validated estimators ensure an honest evaluation of the population response under the nominated decision rules.

Our example studies were two NIH-funded clinical trials; Lumbar Epidural injections for Spinal Stenosis (LESS) and Vitamin Intervention for Stroke Prevention (VISP). The estimated population response across nominated decision rules suggested a potential improvement using marker-guided treatment in neither of them. The treatment benefit in the LESS trial appeared to be concentrated in about two thirds of patients who would benefit from the addition of steroid injections, however, both studies we presented in this chapter are likely typical examples of too much enthusiasm in the search of subgroup effects. In such scenarios, we offer statistical tools that allow for honest, cross-validated estimation of population response under the set of nominated treatment prescription rules.

In the next Chapter, we will extend the methods for evaluation of marker-guided

treatment rules to the settings with survival outcomes, where the link between the working regression model and measures of population performance might be even less tight as in the case of a continuous outcome.

Chapter 3

SURVIVAL OUTCOMES

3.1 Introduction

In scenarios when the investigated outcome is an event-free survival time, such as relapse-free or overall survival of patients with cancer, the available therapies might be very aggressive and harmful to those patients who do not respond well. In such cases, the ability to identify a subgroup of patients that would benefit from the treatment is of particular interest and the enthusiasm for development of marker-guided treatment rules that target therapy to responsive patients is growing.

Large randomized clinical trials offer the potential to evaluate biomarkers as predictors of treatment response, and to develop individualized rules to guide treatment. Recent statistical literature has focused on estimation of the individual treatment effect as a function of continuous signature scores and predicting patient-level survival curves (Matsui et al., 2012), testing treatment effect in a subset of patients identified by marker values (Freidlin and Simon, 2005; Jiang et al., 2007) and assessing treatment-selection markers using potential outcomes framework (Huang et al., 2012). However, statistical methods for estimation and evaluation of treatment rules that are targeted to improve the net population impact in the context of survival outcomes are not well established.

In this chapter, we extend our proposed methodology for development of marker-guided decision rules to the settings with time-to-event outcomes in a presence of cen-

soring. Our algorithm for marker panel selection will evaluate and compare nominated decision rules with respect to the net event-free survival in the target population. We will derive the individual scores using penalized Cox regression methods and evaluate their net population impact via commonly used univariate summary measures of the survival curve: Area Under the Survival Curve (AUSC); probability of survival at a specific time of interest; and median survival time.

3.2 Methods

In the case of survival data, the outcome of interest is a time to event and it might be unobserved for some patients due to censoring. While our target measure of performance for continuous outcomes was the mean population response, in the presence of censoring, the mean time to event is a tricky/difficult quantity to estimate and hence the methods we developed for the continuous outcomes need to be modified to the survival framework.

A common target in analyses of time to event is a survival curve over time, which can be estimated non-parametrically (Kaplan-Meier method), semi-parametrically (via Cox regression) or assuming a full parametric model (e.g., exponential, Weibull, etc.). In order to evaluate and compare numerous treatment regimens, we would however prefer a univariate measure of performance that will be easy and straightforward to compare across nominated decision rules, as we did in the setting of the continuous outcome.

We let T denote the time to event and Y the measured follow-up time. Then C is an indicator of censoring, which equals 1 if $Y = T$, and 0 if $Y < T$. The survival function $G(t)$ is the complement of the distribution function of the random variable time to event,

$$G(t) = 1 - F(t) = \mathrm{P}[T \geq t]\,.$$

As before, $A \in \{0, 1\}$ denotes the treatment (or, action) and $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^m$ denotes a vector of patient's characteristics collected prior to the treatment administration.

3.2.1 Measures of Performance for Survival Outcomes

Generally, higher survival is consider better, however, it becomes unclear which one of two survival curves is "better" in case they cross. When comparing survival between two or more groups, this problem is often overcome by reducing the survival curve into some one-dimensional quantity. Three commonly used univariate summary measures for evaluation of the survival are: the Area Under the Survival Curve (AUSC), probability of survival at a specific time of interest, $G(t^*)$, and median survival time, $G^{-1}(0.5)$.

Area Under the Survival Curve (AUSC)

The first proposed measure of performance which collapses the whole survival curve into a one-dimensional quantity is the area under the survival curve (AUSC). Mathematically,

$$\text{AUSC} = \int_0^\infty G(t)dt = \text{E}\left[T\right],$$

which brings us back to the expected time to event. However, as we already discussed, mean time to event is difficult to estimate in the presence of censoring without assuming a full parametric model. Instead, a restricted AUSC may be considered for the evaluation of population survival,

$$\text{AUSC}(\tau) = \int_0^\tau G(t)dt\,.$$

The restricted AUSC(τ) has an interpretation of the average time lived within some relevant initial period τ. Besides its good/attractive interpretability, the AUSC is well estimable and has a potential to capture differences along the whole survival curve.

Probability of Survival at Specific Time of Interest

Another way to evaluate the population response is to estimate the survival function G at a particular time of interest t^*. Then, $G(t^*)$ is a probability of event-free survival beyond the time t^*. If there were no observations censored before time t^*, this would become a simpler problem with a binary outcome. We do not discuss the approach to the binary outcome in details, however, the methods for the score development and evaluation would be analogous to those we established for the continuous outcome, with simply replacing the linear models by logistic models.

Median Survival Time

The third one-dimensional quantity which is commonly estimated instead of the population mean survival is the median survival, $G^{-1}(0.5)$. The median survival is the time beyond which the probability of event-free survival is 50%, and it is estimable non-parametrically despite the censoring as long as there are enough events in the dataset.

3.2.2 Individual and Population Treatment Effect

In the survival data framework, the distinction between developmental and evaluation phases becomes even more pronounced than for continuous outcomes. For the development of candidate decision rules, we adopt a proportional hazard model, which will allow us to estimate a single-index score. The score-based decision rules will be then evalu-

ated via cross-validation using population estimates of the above described measures of performance.

Let us first consider a hazard h (risk of event) over time t as a function of covariates $\boldsymbol{X}$ and treatment A; $h(t|\boldsymbol{X}, A)$. Under the assumption of hazards being proportional over time, a log-hazard for a patient with covariates $\boldsymbol{X} = \boldsymbol{x}$ and under the treatment $A = a$ can be written as

$$\log h(t|\boldsymbol{X} = \boldsymbol{x}, A = a) \;=\; \log h(t|\boldsymbol{X} = \boldsymbol{x}, A = 0) \;+\; a\Delta(\boldsymbol{x})\,, \tag{3.1}$$

where $\Delta(\boldsymbol{x})$ is now a covariate-specific treatment effect, and $\exp\{\Delta(\boldsymbol{x})\}$ is a covariate-specific hazard ratio. The proportional-hazards (PH) assumption implies that $\Delta(\boldsymbol{x})$ does not depend on t and we see that a positive treatment effect $\Delta(\boldsymbol{x}) > 0$ corresponds to a higher risk under treatment $A = 1$, while $\Delta(\boldsymbol{x}) < 0$ corresponds to a lower risk under $A = 1$.

The survival at time t for a patient with covariates $\boldsymbol{X} = \boldsymbol{x}$ under the treatment $A = a$ can be then expressed as

$$\begin{aligned} G(t|\boldsymbol{X} = \boldsymbol{x}, A = a) &= \exp\{-\int_0^t h(u|\boldsymbol{X} = \boldsymbol{x}, A = a)du\} \\ &= \exp\left\{-\int_0^t h(u|\boldsymbol{X} = \boldsymbol{x}, A = 0)du \times \exp\{a\Delta(\boldsymbol{x})\}\right\} \\ &= G(t|\boldsymbol{X} = \boldsymbol{x}, A = 0)^{\exp\{a\Delta(\boldsymbol{x})\}} \end{aligned} \tag{3.2}$$

and since the covariate-specific hazard ratio is assumed constant over time, we can say that a lower risk of event under treatment $A = 1$ (i.e., when $\exp\{\Delta(\boldsymbol{x})\} < 1$) is equivalent to better expected survival under $A = 1$. Hence, we would like to treat those patients

who have the negative estimated treatment effect in order to maximize their survival, or equivalently minimize their risk of event.

Similarly as for the continuous outcome, it can be easily seen that a regimen that chooses the smallest risk of event for every individual or, equivalently, the highest event-free survival,

$$d^{opt}(\boldsymbol{x}) = \arg\max_d \, G(t|\boldsymbol{X} = \boldsymbol{x}, A = d(\boldsymbol{x})), \ \forall \boldsymbol{x} \in \mathcal{X}, \ \forall t \geq 0\,,$$

also leads to the highest event-free survival in the population,

$$d^{opt} = \arg\max_d G(t|\, d)\,, \ \forall t \geq 0\,,$$

where $G(t|\, d) = \mathrm{P}(T \geq t|\, d) = \mathrm{E}_{\boldsymbol{X}}\{\mathrm{P}_{T|\boldsymbol{X},A}[T \geq t|\, \boldsymbol{X},\, A = d(\boldsymbol{X})]\}$ is the survival function in the population under a regimen d. Using the expression (3.2) for the individual survival and the fact that the probability of survival is always less or equal to 1, we can show that the optimal decision rule for a patient with covatiates $\boldsymbol{x}$ is

$$\begin{aligned} d^{opt}(\boldsymbol{x}) &= \arg\max_d G(t|\boldsymbol{X} = \boldsymbol{x}, A = d(\boldsymbol{x})) = \\ &= \arg\max_d G(t|\boldsymbol{X} = \boldsymbol{x}, A = 0)^{\exp\{d(\boldsymbol{x})\Delta(\boldsymbol{x})\}} = \\ &= \arg\min_d \{d(\boldsymbol{x})\, \Delta(\boldsymbol{x})\} = \\ &= \mathrm{I}\left[\Delta(\boldsymbol{x}) < 0\right] \equiv \mathrm{I}\left[\exp\{\Delta(\boldsymbol{x})\} < 1\right]. \end{aligned}$$

Thus, the theoretically optimal regimen or oracle rule on $\mathcal{X}$, that leads the best survival in every $\boldsymbol{x} \in \mathcal{X}$, has the form $d^{opt}(\boldsymbol{x}) = \mathrm{I}\left[\Delta(\boldsymbol{x}) < 0\right]$, which assigns treatment to

only those individuals for whom their covariate-specific treatment effect is negative, or equivalently, their covariate-specific hazard ratio is less than 1.

The proposed framework does not necessarily require the that hazards between the two treatments are proportional. We adopted the working Cox PH model, because it allows for easy estimation of the covariate-specific "average" log hazard ratios and nomination of candidate decision rules. For example, a more general case than PH is when the covariate-specific hazard ratio between two treatments is a function of time, $\Delta(\boldsymbol{x}, t)$, but is always either positive or negative, which also assures that the two covariate-specific survival curves do not cross. In both of these cases (and many others), it is hence sufficient to know whether the "average" (over time) risk of event is smaller under the experimental or standard treatment.

The problem with identifying an optimal rule arises when two covariate-specific survival curves do cross, which might result in inconsistent ordering with respect to different measures and the question of optimality becomes more difficult. The concept of optimality relates to the concept of ordering. In case of survival curves crossing, it is necessary to define some unique way of ordering them. For example, one could consider any of the univariate summary measures we listed in section 3.2.1, however, the corresponding optimal rule would be only optimal with respect to the selected quantity, while it might not be optimal with respect to the others. Hence, without assuming completely ordered survival curves, the definition of optimality itself becomes a question that one would need to address based on the particular scientific context, additionally to the problem of treatment benefit estimation.

3.2.3 Approach to Treatment Effect Score Development in Survival Setting

We now consider a class of proportional-hazards models, $\mathcal{D}$, that are linear and additive on the log-hazard scale. The hazard rate at time t as a function of treatment A and markers $\boldsymbol{X}$ is then given by

$$h(t|\boldsymbol{X}, A) = h_0(t) \exp\{\gamma_0 A + \sum_{j=1}^{m} (\beta_j X_j + \gamma_j A X_j)\} = h_0(t) \exp\{\boldsymbol{X}\boldsymbol{\beta} + A\boldsymbol{X}\boldsymbol{\gamma}\}\,, \quad (3.3)$$

and the marker-specific treatment effect $\Delta(\boldsymbol{X})$ can be again expressed as a linear combination of the marginal treatment effect and the individual marker-treatment interactions:

$$\Delta(\boldsymbol{X}) = \log\left\{\frac{h(t|\boldsymbol{X}, A=1)}{h(t|\boldsymbol{X}, A=0)}\right\} = \gamma_0 + \sum_{j=1}^{m} \gamma_j X_j = \boldsymbol{X}\boldsymbol{\gamma}\,.$$

The estimation of the marker-specific treatment effect within $\mathcal{D}$ is hence equivalent to estimation of the model parameters $\boldsymbol{\gamma}$, as it was for the continuous outcomes.

For the development of treatment effect scores, we decided to further exploit the simplicity and computational feasibility of LASSO method. The variable selection and shrinkage in Cox's proportional hazard model is carried out by minimizing the partial log-likelihood, subject to a parameter constraint or penalty (Tibsirani, 1997). The sequence of candidate decision rules is hence again generated based on the L_1 penalty, and the proportional hazard model (3.3) induces a class of additive linear decision rules, $d(\boldsymbol{X}) = \mathrm{I}\,[\boldsymbol{X}\boldsymbol{\gamma} < 0]$, corresponding to parameters $\boldsymbol{\gamma} \in \boldsymbol{\Gamma} \subseteq \mathbb{R}^m$.

Let us assume a random sample of n patients, with their follow-up times y_i, indicators of censoring c_i, observed treatments a_i, and m-dimensional vectors of markers $\boldsymbol{x}_i$, $i =$

$1, \ldots, n$. We denote the distinct failure times by $t_1 < \cdots < t_v$. The parameter estimation in the proportional-hazards model is then based on the partial likelihood

$$L_n(\boldsymbol{\theta}) = \prod_{r \in D} \frac{\exp\{(\boldsymbol{x}_{i_r}^{\mathrm{T}}, a_{i_r}\boldsymbol{x}_{i_r}^{\mathrm{T}})\,\boldsymbol{\theta}\}}{\sum_{j \in R_r} \exp\{(\boldsymbol{x}_i^{\mathrm{T}}, a_i\boldsymbol{x}_i^{\mathrm{T}})\,\boldsymbol{\theta}\}},$$

where D is the set of indices of failures, R_r is the set of indices of the individuals at risk at time t_r, and i_r is the index of the failure at time t_r. Since β_0 is now not in the model, the vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ is of length $J = 2m + 1$ and can be estimated by

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^J} \{ -\frac{1}{n} \log L_n(\boldsymbol{\theta}) + \lambda \sum_{j \in \mathcal{J}} \hat{\sigma}_j |\, \theta_j | \,\}$$

where $\hat{\sigma}_j^2$ is an estimated variance of the j-th covariate, and $\mathcal{J} = \{1, ..., m,\ m{+}2, ..., J\}$, so that the main treatment effect is not penalized.

The panel size p equals to the number of non-zero estimates of coefficients $\gamma_1, ..., \gamma_m$, which will vary for different values of tuning parameter λ. Each set of the parameter estimates $\widehat{\boldsymbol{\gamma}}^p$ then corresponds to a decision rule

$$d_p(\boldsymbol{X}) = \mathrm{I}\,[\boldsymbol{X}\widehat{\boldsymbol{\gamma}}^p < 0]$$

that classifies patients as to whether they should get treated or not, depending on their individual scores, $s_i^p = \boldsymbol{x}_i\widehat{\boldsymbol{\gamma}}^p$.

3.2.4 Estimation of Marker-Guided Population Survival

Analogously as in the case of continuous outcomes, we first present the evaluation of

decision rules based on the single-index score S. For each candidate marker-guided decision rule d, we wish to evaluate what would the population survival be if everybody was treated according to d. Then, we calculate the above proposed measures of performance, which serve for a comparison of the population survival across different panel sizes p.

If we assume a marker-guided decision rule $d(s) = \mathrm{I}\,[s < 0]$, then the marker-guided population survival function (MGPS) under the rule d can be rewritten as

$$\begin{aligned} G(t|\,d) &= \mathrm{P}[T \geq t|\,d] = \mathrm{E}_S\{\,\mathrm{P}_{T|S,A}[T \geq t|\,S,\, A = d(S)]\,\} = \\ &= \mathrm{E}_S\{\,G(t|\,S,\, A = d(S))\,\} \qquad (3.4) \\ &= G(t|\,S \geq 0, A = 0)\mathrm{P}[S \geq 0] + G(t|\,S < 0, A = 1)\mathrm{P}[S < 0] \\ &= G_0(t|\,S \geq 0)\mathrm{P}[S \geq 0] + G_1(t|\,S < 0)\mathrm{P}[S < 0]\,, \qquad (3.5) \end{aligned}$$

where $G_a(t|\,S = s) = \mathrm{P}\,[T \geq t|\,S = s, A = a]$ is the conditional probability of event-free survival beyond time t, given the marker value s and treatment $a \in \{0, 1\}$. We present three different approaches to estimation of the MGPS.

Simple Non-parametric Estimator

The non-parametric estimation of the marker-guided population survival function is similar to the non-parametric estimator of MGPR in the sense that it uses the outcomes only from those patients whose observed and prescribed treatments match. Using Kaplan-Meier method (Kaplan and Meier, 1958), the survival curves are estimated separately for the two treatment arms, and then weighted by the proportion of score-negative patients, according to the equation (3.5).

In particular, the Kaplan-Meier estimator of $G_0(t|\, S \geq 0)$ is given by

$$\widehat{G}_0(t|S \geq 0) = \prod_{t_i < t;\ i \in \mathcal{N}_0} \frac{n_{0i} - d_{0i}}{n_{0i}}\,,$$

where n_{0i} is the number "at risk" and d_{0i} is the number of events corresponding to the event time t_i among those who have both observed and prescribed treatment equal to 0, $\mathcal{N}_0 = \{i : a_i = d(s_i) = 0\}$. Similarly, the Kaplan-Meier estimator of $G_1(t|\, S < 0)$ is

$$\widehat{G}_1(t|S < 0) = \prod_{t_i < t;\ i \in \mathcal{N}_1} \frac{n_{1i} - d_{1i}}{n_{1i}}\,,$$

and the marker-guided population survival is then estimated by

$$\widehat{G}(t|\, d) = \widehat{G}_0(t|\, S \geq 0)(1 - \widehat{\pi}_1) + \widehat{G}_1(t|\, S < 0)\, \widehat{\pi}_1\,,$$

where $\widehat{\pi}_1 = \widehat{\mathrm{P}}[S < 0]$ is a sample proportion of score-negative patients,

$$\widehat{\pi}_1 = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\,[s_i < 0]\,.$$

Semi-parametric Estimator using Cox Model

If we could predict individual survival functions under both treatment arms, the population survival curve under any arbitrary treatment regimen would be then easily estimated by averaging the corresponding predicted values across patients in the sample. Based on expression (3.5), an empirical estimator of the MGPS under a decision rule $d = \mathrm{I}\,[s < 0]$

would be

$$\widehat{G}(t|\, d) \;=\; \frac{1}{n}\sum_{i=1}^{n} \widehat{G}(t|\, S = s_i, A = d(s_i))\,. \tag{3.6}$$

A semi-parametric estimation/prediction of individual survival functions based on a single-index score was proposed by Matsui et al. (2012). The authors suggested using Cox proportional-hazards model in order to fit the risk of event as a flexible function of a composite score. However, while they chose fractional polynomials to model the flexible function of s, we decided to use a cubic spline function with 2 knots (at sample tertiles) instead. Similarly as in Chapter 2, we fit a spline function represented by corresponding B-splines, $\boldsymbol{B}(s)$, to model relationship between the score and the hazard in both treatment arms. For the proportional-hazards model, we considered two following options:

$$\begin{aligned}
&(1) \qquad h_a(t|S=s) = h_0(t)\exp\{\alpha_0^B a + \boldsymbol{B}(s)\boldsymbol{\alpha}_1^B + a\,\boldsymbol{B}(s)\boldsymbol{\alpha}_2^B\}\\
&(2) \qquad h_a(t|S=s) = h_{a0}(t)\exp\{\boldsymbol{B}(s)\boldsymbol{\alpha}_a^B\}\,; \quad a \in \{0,1\}\,,
\end{aligned}$$

where the former assumes common baseline hazard for both treatment arms and models the treatment-score interaction, and the latter assumes separate baseline hazards for the two treatments.

There are multiple established methods for estimation of a baseline survival curve. We adopted a commonly used estimator proposed by Kalbfleisch and Prentice (1972), which is based on a non-parametric full likelihood. Suppose that there are no ties among the failure times and let $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ denote k distinct event times in

the observed data set. For model (1), the estimated common baseline survival curve is then

$$\widehat{G}_0(t) = \prod_{t_{(i)}<t} \left[1 - \frac{\exp\{\widehat{\boldsymbol{\alpha}}\boldsymbol{z}_{(i)}\}}{\sum_{j\in R_i}\exp\{\widehat{\boldsymbol{\alpha}}\boldsymbol{z}_j\}}\right]^{\exp\{-\widehat{\boldsymbol{\alpha}}\boldsymbol{z}_{(i)}\}}$$

where $\widehat{\boldsymbol{\alpha}}$ is the vector of parameter estimates, $\boldsymbol{Z} = [\boldsymbol{a}, \boldsymbol{B}(\boldsymbol{s}), \boldsymbol{a} * \boldsymbol{B}(\boldsymbol{s})]$ is a matrix of covariates and R_i is a risk set at time $t_{(i)}$. Considering either of the two models, the predicted individual survival curves as functions of the single-index score and prescribed treatment can be then obtained by multiplying the estimated baseline survival curve by corresponding covariate-specific hazard ratios. For example, under the model (1),

$$\widehat{G}(t|\, S = s_i, A = d(s_i)) = \widehat{G}_0(t)^{\exp\{\widehat{\alpha}_0^B\, d(s_i) + \boldsymbol{B}(s_i)\widehat{\boldsymbol{\alpha}}_1^B + d(s_i)\,\boldsymbol{B}(s_i)\widehat{\boldsymbol{\alpha}}_2^B\}}$$

is an estimated survival curve for a patient with score s_i under the treatment regimen d.

Weighted Kaplan-Meier Curves Estimator

Our third proposed estimator is again non-parametric, but unlike the simple non-parametric estimator, it additionally exploits continuity of the single-index score. Instead of estimating Kaplan-Meier curves for only two groups as in the case of the first estimator, we use nearest neighbors in order to estimate the individual survival curves locally. For a patient with score s_i and assigned treatment $d(s_i)$, the predicted survival curve in now based on a local neighborhood of the patients with their observed treatment $a = d(s_i)$ and scores closest to s_i.

General locally weighted Kaplan-Meier curves have been proposed previously by

Lumley and Heagerty (2000). For our estimation, we particularly chose symmetric neighborhoods of size $n * n^{-1/3}$, i.e., a neighborhood of s_i contains those patients whose observed treatment $a = d(s_i)$ and distance from s_i is less or equal to $\frac{n}{2}\, n^{-1/3}$ on the rank scale. The locally weighted Kaplan-Meier curve for a patient with score s_i is then given by

$$\widehat{G}(t|\, S = s_i, A = d(s_i)) = \prod_{t_k < t;\ k \in \mathcal{N}_i} \frac{n_{ik} - d_{ik}}{n_{ik}}\,,$$

where $\mathcal{N}_i$ is the neighborhood of s_i, n_{ik} is the number "at risk" and d_{ik} is the number of events corresponding to the event time t_k among those in the neighborhood $\mathcal{N}_i$. Following the expression (3.6), we define the weighted Kaplan-Meier estimator of MGPS under a decision rule d as

$$\widehat{G}(t|\, d) \;\; = \;\; \frac{1}{n} \sum_{i=1}^{n} \widehat{G}(t|\, S = s_i, A = d(s_i))\,,$$

where $\widehat{G}(t|\, S = s_i, A = d(s_i))$ are the locally weighted estimates of individual survival curves.

3.2.5 Assessment of Optimal Treatment Regimen via Cross-validation

We again consider a sequence of candidate decision rules based this time on the LASSO coefficient estimates of the proportional-hazards model (3.3) and identify a collection of λ's that coincide with panel sizes $p = 1, 2, \ldots$. Each set of estimates, $\widehat{\boldsymbol{\gamma}}^p$, corresponds to a decision rule $d_p(\boldsymbol{X}) = \mathrm{I}\,[\boldsymbol{X}\widehat{\boldsymbol{\gamma}}^p < 0]$ that provides a way of patient classification through their individual scores $s_i^p = \boldsymbol{x}_i \widehat{\boldsymbol{\gamma}}^p$. In order to identify an optimal panel size p^*, the candidate decision rules are now compared with respect to one of the three proposed

measures of performance (MoP); AUSC, survival at time t^*, $G(t^*)$, and median survival, $G^{-1}(0.5)$.

The optimal panel size is selected based on its estimated out-of-sample performance, assessed by a K-fold (e.g, 10-fold) cross-validation. The performance of a decision rule d_p is measured by the selected MoP, derived from a cross-validated estimate of the marker-guided population survival function under that rule, $\widehat{G}(t|d_p)$. For each panel size $p = 1, 2, ...$, we estimate $G(t|d_p)$ by one (or multiple) of the three approaches described in the previous section; simple non-parametric, semi-parametric using Cox model, and weighted Kaplan-Meier estimator.

The LASSO estimates are calculated for each fold separately. In each training set, we select collection of tuning parameters λ's that correspond to panel sizes $p = 1, 2, ...$ and the corresponding coefficient estimates $\widehat{\boldsymbol{\gamma}}^p$ are used to calculate scores $s_i^p = \boldsymbol{x}_i\widehat{\boldsymbol{\gamma}}^p$ for patients in the test set. After all the scores are attained for each panel size, the whole data set is used to estimate the marker-guided population survival curves.

Generally, the optimal panel size p^* is chosen as the smallest panel size that maximizes a cross-validated estimator of the selected measure of performance. Similarly as before, the plot of estimated clinical objective as a function of panel size serves as a good visual aid in assessing relative quality of the developed decision rules, and possibly selecting a smaller panel size than p^* with estimated MoP close to the maximum. The single-index cross-validation algorithm is outlined below.

Algorithm:

1. Divide the sample (n) into K folds (here K=10).

2. For each fold $k = 1, \ldots, K$:

a) Run LASSO for the proportional-hazards Cox model

$$h(t|\boldsymbol{X}, A) = h_0(t) \exp\{\gamma_0 A + \sum_{j=1}^{n} (\beta_j X_j + \gamma_j A X_j)\} = h_0(t) \exp\{\boldsymbol{X}\boldsymbol{\beta} + A\boldsymbol{X}\boldsymbol{\gamma}\}$$

on **training set** (9/10 of the data), and obtain ordered sequence of model coefficients.

b) Determine sequence of λ's corresponding to panel sizes $p = 1, 2, ...,$

where a marker panel size equals to the number of non-zero estimates of (interaction) coefficients $\gamma_1, ..., \gamma_N$, and denote the estimates $\widehat{\boldsymbol{\gamma}}^p$.

c) For each panel size $p = 1, 2, \ldots$

Calculate scores for individuals in the **test set**, $s_i^p = \boldsymbol{x}_i \widehat{\boldsymbol{\gamma}}^p$, based on the markers with the selected p interactions. Corresponding decision rule $d_p(\boldsymbol{X}) = \mathrm{I}\left[\boldsymbol{X}\widehat{\boldsymbol{\gamma}}^p < 0\right] = \mathrm{I}\left[S^p < 0\right]$ assigns the treatment.

3. For each panel size $p = 1, 2, \ldots$

Estimate the marker-guided population survival curve (MGPS) based on $\{(y_i, c_i, a_i, s_i^p),\ i = 1, \ldots, n\}$, using either a simple non-parametric estimator, a semi-parametric estimator based on a Cox model, or a weighted Kaplan-Meier estimator.

a) Simple Non-parametric Estimator

Based on only those patients who received the treatment equivalent to what would be prescribed to them by d_p; $\mathcal{N}_0^p = \{i :\ a_i = \mathrm{I}\left[s_i^p < 0\right] = 0\}$, $\mathcal{N}_1^p = \{i :\ a_i = \mathrm{I}\left[s_i^p < 0\right] = 1\}$.

Target (MGPS):

$$G(t|\, d) = (1 - \pi_1)\, G(t|\, S \geq 0, A = 0) + \pi_1\, G(t|\, S < 0, A = 1)\,,$$

where $\pi_1 = \mathrm{P}(S < 0)$.

- $\widehat{\pi}_1$: proportion of score negative patients
- $\widehat{G}(t|\, S \geq 0, A = 0)$ survival curve among $i \in \mathcal{N}_0$
- $\widehat{G}(t|\, S < 0, A = 1)$ survival curve among $i \in \mathcal{N}_1$

$$\begin{aligned} \widehat{G}(t|\, d_p) &= (1 - \widehat{\pi}_1^p)\, \widehat{G}(t|\, S^p \geq 0, A = 0) \,+\, \widehat{\pi}_1^p\, \widehat{G}(t|\, S^p < 0, A = 1) = \\ &= (1 - \widehat{\pi}_1^p) \prod_{t_i < t;\ i \in \mathcal{N}_0^p} \frac{n_{0i}^p - d_{0i}^p}{n_{0i}^p} \,+\, \widehat{\pi}_1^p \prod_{t_i < t;\ i \in \mathcal{N}_1^p} \frac{n_{1i}^p - d_{1i}^p}{n_{1i}^p} \end{aligned}$$

b) Semi-parametric Estimator Based on a Cox Model (1)

Models the hazard rates $h(t|s,a)$, $a \in \{0, 1\}$, as smooth functions of s in a Cox proportional-hazards model (1) and estimates the baseline survival curve $G_0(t)$ by one of the established methods listed above.

Target (MGPR):

$$G(t|\, d) = \int_{\mathbb{R}} \{\, G(t|\, S = s, A = 0)\mathrm{I}\,[s \geq 0] + G(t|\, S = s, A = 1)\mathrm{I}\,[s < 0]\,\}\, dF_S(s)$$

- $\widehat{F}_S$: empirical distribution function of s in the data
- $\widehat{G}(t|\, s, a) = \widehat{G}_0(t)^{\widehat{\mathrm{HR}}(s,a)}$
- $\widehat{G}_0(t)$: baseline survival estimator by e.g., ref: Kalbfleisch and Prentice (1973)
- $\mathrm{HR}(s, a)$: hazard ratio for treatment a and score s
- $h(t|\, s, a)$ modeled via Cox model (1) or (2) with splines in s

$$\widehat{G}(t|\, d) = \frac{1}{n} \sum_{i=1}^{n} \widehat{G}(t|\, S = s_i, A = 0)\mathrm{I}\,[s_i \geq 0] + \widehat{G}(t|\, S = s_i, A = 1)\mathrm{I}\,[s_i < 0]\,.$$

c) Weighted Kaplan-Meier Estimator

Uses nearest neighbors in order to estimate the individual survival curves locally.

Target (MGPR):

$$G(t|\, d) = \int_{\mathbb{R}} G(t|\, S = s, A = d(s))\; dF_S(s)$$

- $\widehat{F}_S$: empirical distribution function of s in the data
- $\widehat{G}(t|\, S = s, A = d(s))$: locally weighted Kaplan-Meier curve for the score s

$$\begin{aligned} \widehat{G}(t|\, d) &= \frac{1}{n}\sum_{i=1}^{n} \widehat{G}(t|\, S = s_i, A = d(s_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} \prod_{t_k<t;\; k\in\mathcal{N}_i} \frac{n_{ik} - d_{ik}}{n_{ik}}\,, \end{aligned}$$

where $\mathcal{N}_i$ is a subset of indexes $\{k : a_k = d(s_i)\}$ of patients from a symmetric neighborhood of s_i that has size $n * n^{-1/3}$.

4. Graph the clinical objective function for the selected measure of performance ($AUSC(\tau)$, $G(t^*)$, or $G^{-1}(0.5)$) vs. p and identify a panel size p^* that leads maximum estimated MoP.

3.2.6 Benefit Among Treated

In order to evaluate population impact of candidate decision rules, we can again focus alternatively on the treatment benefit among treated patients. Since the improvement in the survival is measured by one of the three summary quantities ($AUSC(\tau)$, $G(t^*)$, or $G^{-1}(0.5)$), we suggest to assess the benefit among treated also with respect to one

these measures.

Let us consider a general decision rule $d^c(s) = \mathrm{I}[s < c]$, which assigns treatment $A = 1$ to only those patients whose risk of event is estimated to increase by less than a constant c on the log-scale or, equivalently, whose estimated (covariate-specific) hazard ratio $\exp\{s\} < \exp\{c\}$, for some $c \leq 0$. Then, we define the benefit among treated under the rule d^c with respect to a summary measure M as follows:

$$B^M(c) = M[\,G(t|S < c, A = 1)] - M[\,G(t|S < c, A = 0)] = M_1^c - M_0^c\,,$$

where $M_a^c = M[\,G(t|S < c, A = a)]$, $a \in \{0, 1\}$, can be for example any of the three summary measures of the survival curve: area under the survival curve $G(t|S < c, A = a)$ restricted to time τ, $AUSC(\tau)$; probability of survival up to time t^*, $G(t^*|S < c, A = a)$; or median survival time $G^{-1}(0.5|S < c, A = a)$.

We estimate the benefit among treated $B^M(c)$ with respect to the selected summary measure M by replacing the survival functions $G(t|S < c, A = a)$, $a \in \{0, 1\}$, with their respective estimates. Hence, the estimated benefit among treated is

$$\widehat{B}^M(c) = M[\,\widehat{G}(t|S < c, A = 1)] - M[\,\widehat{G}(t|S < c, A = 0)]\,,$$

where $\widehat{G}(t|S < c, A = a)\,,\ a \in \{\,0, 1\}\,,$ are based one of the three proposed estimators and a subset of patients who have scores $s_i < c$. In particular, using the simple non-parametric estimator, the two survival curves for $a \in \{0, 1\}$ can be estimated by

$$\widehat{G}(t|S < c, A = a) = \prod_{t_i < t;\ s_i < c;\ a_i = a} \frac{n_{ai}^p - d_{ai}^p}{n_{ai}^p}\,.$$

Using the weighted Kaplan-Meier or Cox-model estimator, each of the two survival curves is based on the average of individual estimated survival curves over the patients whose score is $s_i < c$. That is,

$$\widehat{G}(t|S < c, A = a) = \frac{1}{n_c} \sum_{i:\, s_i < c} \widehat{G}(t|S = s_i, A = a)\,, \quad a \in \{0, 1\}\,,$$

where n_c is the number of scores below c and the individual $G(t|S = s_i, A = a)$ are estimated by either weighted Kaplan-Meier or Cox-model based estimator as described in the section 3.2.4.

3.3 Simulations

3.3.1 Part I: Under Proportional-hazards Model

The purpose of our simulations is to examine and compare properties of the three proposed estimators. In order to assess their biases and variances, we generated 300 data sets from the same population and calculated how well the estimators estimate the population measures of performance under the developed decision rules.

For a generation of the variable time to event T, we adopted the exponential distribution with covariate-specific hazard rates following the linear additive and proportion-hazards model (3.3). The parameters we control are baseline hazard for T (h_0), main treatment effect (γ_0), main marker effects ($\boldsymbol{\beta}$) and marker-by-treatment interactions ($\boldsymbol{\gamma}$). Additionally, we need to specify the hazard rate for time of censoring C (h_C) and a distribution of the markers.

In order to assure a desired marginal hazard ratio between treated and non-treated

patients in the presence of marker effects and marker-by-treatment interactions, we needed to adjust two components. First, we centered the values of the markers by their means, similarly as in the simulations for the continuous outcome, and hence the covariate-specific hazard rates were given by

$$h(t|\,\boldsymbol{X}, A) = h_0(t)\,\exp\{\gamma_0 A + \sum_{j=1}^{N}(\beta_j X_j^* + \gamma_j X_j^* A)\}\,,$$

where N is a number of markers and $X_j^* = X_j - \mathrm{E}X_j$. Second, the parameter γ_0 does not directly correspond to the marginal treatment log HR, since the hazard rate is non-linear in γ_0. We therefore used an iterative method to evaluate what value of γ_0 should be used in order to obtain the desired marginal HR.

For evaluation of bias and variance of the three estimators, the estimates of all three measures of performance (AUSC, $G(t^*)$ and $G^{-1}(0.5)$) were compared with their true counterparts under decision rules d_p that would be eventually used in the population. Since the cross-validation mechanism causes a d_p to be different for each of the K folds, we calculated a new set of $\widehat{\boldsymbol{\gamma}}^p$ based on the whole sample, as would be done if the rule were to be eventually exported. Then we assessed the true population survival (and its summary measures) under such d_p's based on a large sample of 10^5 subjects from the target population, assigned to the treatment according to those rules.

Data Generation

In the following simulation, we considered data sets of 1000 subjects, each with 200 independent markers (SNPs) that take values 0, 1, or 2. The minor allele frequencies (MAF) were randomly generated from a uniform distribution with values between 0.1

and 0.5 (rare SNPs were not allowed). A SNP j then has a distribution as a sum of two independent binary variables with probability of success equal to MAF_j.

The marker effects were again considered separately under the treatment ($\boldsymbol{\beta}^1$) and no treatment ($\boldsymbol{\beta}^0$) arm and generated from a mixture distribution as follows. From total of $N = 200$ SNPs, 10% had moderate to large ($\sim$ U(-1,1)) main marker effect under both treatment arms, 10% had them both small ($\sim$ U(-$\frac{1}{2}$,$\frac{1}{2}$)), and 20% of the SNPs had moderate main effect under one arm and small under the other. The reminder of SNPs had both β_j^0 and β_j^1 equal to 0 (see example in Fig. 3.1).

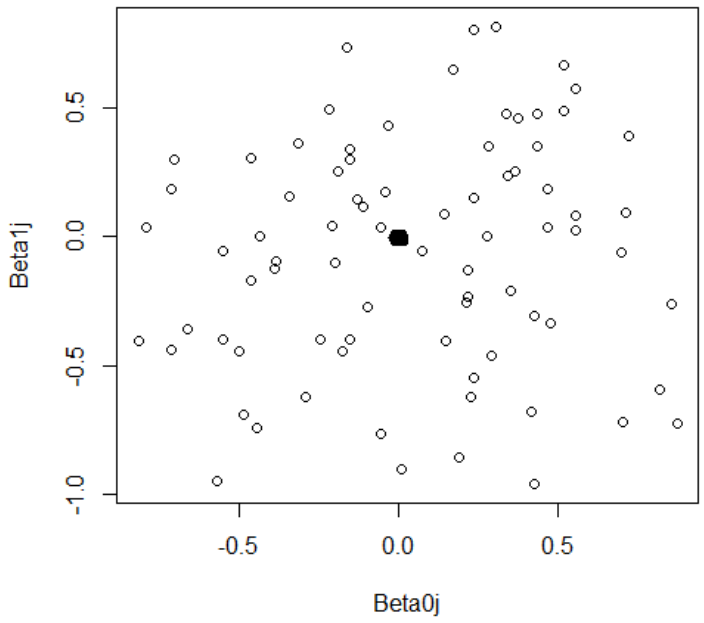

Figure 3.1: SNP effects (on the log-scale) in the two treatment arms; β_j^0 is the main marker effect of SNP_j if A=0 and β_j^1 is the main effect of SNP_j if A=1. The difference $(\beta_j^1 - \beta_j^0)$ corresponds to the interaction (γ_j) between the SNP j and treatment.

We chose a marginal hazard ratio of 0.8, which corresponds to a 20% decrease in risk

of event among treated patients. While $\log(0.8) = -0.22$, the corresponding γ_0 (obtained iteratively) that leads to the desired marginal HR was equal to -0.6. The treatment was assigned to patients randomly with probability $\frac{1}{2}$ and the individual times to event were generated from an exponential distribution with hazard rates

$$h(t|\,\boldsymbol{X}, A) = h_0(t)\,\exp\{\gamma_0 A + \sum_{j=1}^{N}(\beta_j^0 X_{ij}^*(1-A_i) + \beta_j^1 X_{ij}^* A_i\}\,, \tag{3.7}$$

where $h_0(t) = 1$. The individual censoring times were generated from an exponential distribution with a hazard rate $h_C(c) = 0.2$, which resulted in approximately 20% of events being un-observed due to censoring. The estimated marginal Kaplan-Meier curves from one random sample together with the estimated survival curves under one of the developed marker-guided treatment rules can be seen in Fig. 3.2.

Results

The underlying population of patients was characterized by one random set of minor allele frequencies (MAFs) and corresponding $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ parameters and used to generate all the samples in the simulation. From the total of 80 non-zero SNP-treatment interactions $\gamma_j = (\beta_j^1 - \beta_j^0)$, 46 were larger than 0.3 in absolute value and 8 were larger than 1 (shown in Fig. 3.1).

Based on a large sample of 10^5 subjects, we assessed a survival curve of theoretically optimal rule under which everybody was treated optimally based on their marker values, and calculated the corresponding measures of performance. Now we would like to see how close to the optimal MoPs we can get using our cross-validation algorithm outlined in section 3.2.5.

In order to evaluate the bias and variance of the three estimators, we generated a 300 samples of 1000 patients. In each data set, we first performed a 10-fold cross-validation algorithm with resulting estimates of MGPS curves and corresponding measures of performance (MoP), and then assessed the true MGPS curve (and MoP) under the exported decision rules $d_p, p = 1, 2, \ldots,$ as described previously.

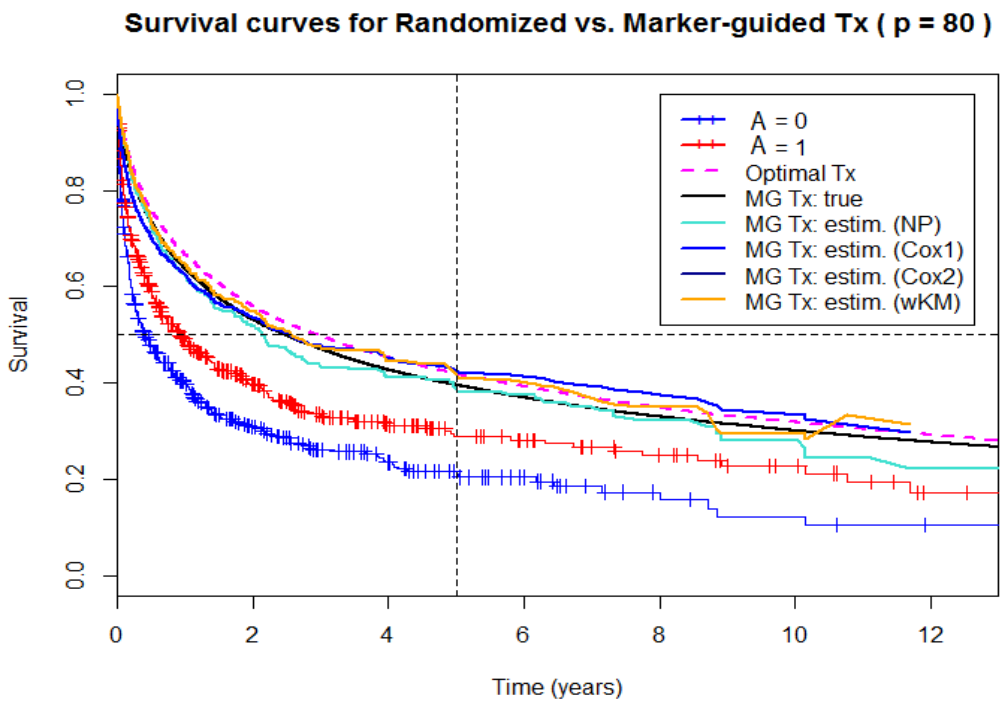

Figure 3.2: Estimated survival curves under d_{80}. The red and blue curves on the bottom are marginal Kaplan-Meier curves for treated and non-treated patients, respectively. The pink dashed line represents the population survival curve under an optimal decision rule (when everybody is treated correctly with respect to their covariates), while the black line shows the population survival curve under the exported decision rule d_{80}.

Results from one random data set can be seen in Fig. 3.3, which shows the estimated measures of performance as functions of the panel size. The black lines shows the (true) population MoP under the exported decision rules d_p. The turquoise line is the estimator

using the simple non-parametric approach, the light-blue and dark-blue lines correspond to the estimator based on a Cox model with common baseline hazard and separate baseline hazards, respectively, and the orange line represents the weighted Kaplan-Meier estimator. All three shown estimators seem to perform very similarly and copy the true MoP fairly well. Notice that the two versions of the Cox model-based estimator are almost identical, which implies that the choice of model is not crucial, and they result in smoother curves for estimated AUSC and 5-year survival than the other two estimators. Finally, all three estimators suggest to select a panel size between 50-100.

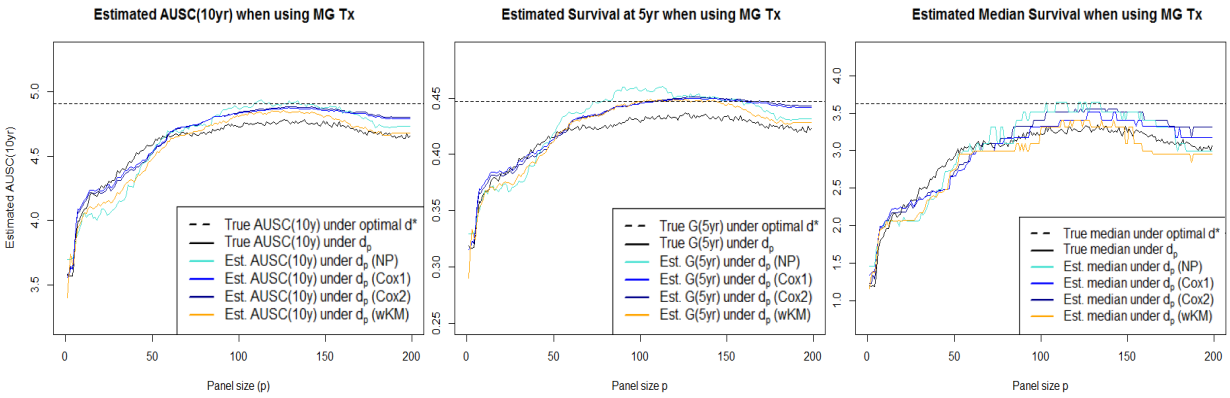

Figure 3.3: Estimated Measures of Performance (MoP) from one simulated data set: AUSC(10yr), probability of 5-year survival and median survival time, as functions of the marker panel size. The MoP are based on the underlying Marker-Guided Population Survival curves (MGPS), corresponding to 1) the true MGPS (black dashed line) and estimators by 2) simple non-parametric approach (turquoise), 3) PH Cox model with common baseline hazard (light blue) and separate baseline hazards (dark blue), and 4) weighted Kaplan-Meier curves (orange). The horizontal dashed line shows the population MoP under an optimal decision rule.

The estimated survival curves under decision rule based on panel size $p = 80$ is shown in Fig. 3.2. The red and blue curves on the bottom of the graph are marginal Kaplan-Meier curves for treated and non-treated patients, respectively. The pink dashed

line represents the population survival curve under an optimal decision rule (when everybody is treated correctly with respect to their covariates), while the black line shows the population survival curve under the exported decision rule d_{80} (for which the $\widehat{\gamma}^{80}$ coefficients are calculated based on the whole data set). It is encouraging to see that, for this population structure, the population survival under a developed marker-guided rule is actually very close to the optimal survival and provides a large improvement over the regimen that uniformly prescribes everybody with the marginally superior treatment $(A = 1)$.

The sample median survival time among non-treated patients was $G_n^{-1}(0.5|\, A = 0) = 0.6$ years and among treated patients $G_n^{-1}(0.5|\, A = 1) = 1.0$ years, while using the marker-guided treatment rule based on panel size $p = 80$ would lead to median survival in the population of $G^{-1}(0.5|\, A = d_{80}) = 2.6$ years. The sample probability of 5-year survival among non-treated patients was $G_n(5|\, A = 0) = 0.19$ and among treated patients $G_n(5|\, A = 1) = 0.28$, while using d_{80} would lead to probability of 5-year survival in the population $G(5|\, A = d_{80}) = 0.43$. Similarly, the sample AUSC(10yr) among non-treated and treated patients was AUSC$(10|\, A = 0)$= 2.4 years and AUSC$(10|\, A = 1)$= 3.3 years, respectively. Using d_{80} would lead to the population AUSC(10yr) of 4.7 years. Additionally, the sample proportion of treated patients for d_{80} is only about 60%, which implies that by applying a marker-guided treatment, not only can we largely improve the population survival but we can achieve it by prescribing treatment to only a fraction of the target population.

Figures 3.6-3.8 (included at the end of the chapter) show the bias and standard deviation of the investigated estimators (including both Cox models 1 and 2), based on 300

samples of 1000 subjects. Both non-parametric estimators tend to be more conservative and underestimate the survival for smaller panel sizes. However, the estimator based on the Cox model tends to overestimate the survival for larger panel sizes. For the latter, we speculate that the consistently observed bias for larger panels might be a result of the non-linear transformation of the score "measurement error", i.e., the error with which the score estimates the marker-specific treatment benefit.

The key findings are summarized in the following list:

- For this population structure, the developed decision rules based on the LASSO result in highly improved population survival as measured by all three summary quantities; restricted AUSC(10yrs), survival at 5 years and median survival.

- All three investigated estimators appear to be similarly effective in assessing the net population impact of the developed decision rule, however, the Cox-model based estimator results in both smoother estimated survival curves and more stable population measures of performance across sample sizes p.

- In terms of empirical bias based on 300 samples, the two non-parametric estimators tend to be more conservative for smaller panel sizes, while the estimator using Cox model overestimates all three measures of population survival for larger marker panels.

3.3.2 Part II: Under Non-proportional-hazards Model

We performed an additional simulation for survival outcome under the data-generating model that has hazards between two treatment arms non-proportional over time. In

order to assess the biases and variances of the three proposed estimators, we generated 100 data sets from the same population and calculated how well the estimators estimate the population measures of performance under the developed decision rules.

Data Generation

For a generation of the variable time to event T, we adopted the Weibull distribution with shape parameters different for the two treatment arms. For each set of covariates (patient), we sat the covariate-specific argument equal to linear combinations of markers (using the same marker effects as before), while the shape parameter was equal 1.2 in the control arm ($A = 0$) and 1 in the experimental arm ($A = 1$). Different shape parameters might result in covariate-specific survival curves to deviate or cross – depending on the covariate attributed hazard ratio. The experimental treatment arm ($A = 1$), with smaller shape parameter, corresponds to more "aggressive" treatment that may result in more subjects dying early on, but then those that survive live longer than in the control arm. The parameters for baseline hazard for T (h_0), main treatment effect (γ_0), main marker effects ($\boldsymbol{\beta}$), marker-by-treatment interactions ($\boldsymbol{\gamma}$), hazard rate for time of censoring C (h_C) and a distribution of the markers were the same as in the simulations with proportional-hazards generating model.

Results

Results from one random data set can be seen in Figure 3.9, which shows the estimated measures of performance as functions of the panel size. The black lines represent the (true) population MoP under the exported decision rules d_p. The turquoise line is

the estimator using the simple non-parametric approach, the light-blue and dark-blue lines correspond to the estimator based on a Cox model with common baseline hazard and separate baseline hazards, respectively, and the orange line represents the weighted Kaplan-Meier estimator.

All three shown estimators seem to copy the curve shape well, but underestimate the true population impact of the developed decision rules in terms of all three measures of performance. Notice that the two versions of the Cox model-based estimator are still almost identical, indicating that the choice of model does not seem crucial even under non-proportional hazards. The Cox models again result in smoother curves for estimated AUSC and 5-year survival than the other two estimators. Finally, all three estimators suggest to select a panel size around 50, which is the true number of markers that have moderate to large interaction with the treatment.

Figures 3.10-3.12 show the bias and standard deviation of the investigated estimators (including both Cox models 1 and 2), based on 100 samples of 1000 subjects. In all our simulations, both empirical standard deviation and bias seem to be fairly small and the standard deviation appears to be about the same across all three estimators. For the AUSC (10yr), the bias is less than 0.2 years, for 5-year survival probability, the bias is less than 2%, and for the median survival, which is around 3 years under the marker-guided treatment, the empirical bias is less than 0.3 years across all panel sizes.

Similarly as in the setting with proportional hazards, both non-parametric estimators tend to be more conservative than the Cox-model based estimators. Surprisingly, however, they seem to remain underestimating the survival even for larger panel sizes under this scenario, while the bias of the estimator based on the Cox model decreases for larger

panels. In order to better understand performance of the three estimators under various scenarios, more data-generating models with both proportional and non-proportional hazards can be investigated in the future.

3.4 Example

We illustrate our methods on data from a prospective randomized phase III trial (S9321) of multiple myeloma, launched by three North American cooperative groups. The aim of the study was to assess an effectiveness of high-dose chemoradiotherapy (HDT) by melphalan (MEL) 140 mg/m2 plus total-body irradiation 12 Gy compared with standard-dose therapy (SDT) using the vincristine, carmustine, MEL, cyclophosphamide, and prednisone regimen in patients diagnosed with multiple myeloma (Barlogie et al., 2006).

The outcome we analyzed was overall survival since the study enrollment. Our data set consists of 392 patients, after excluding those with missing values on some of the examined markers. The HDT (A=1) was randomly assigned to 191 patients and the SDT (A=0) was assigned to 201 patients. The available baseline markers are listed in the table below and include both host and tumor features.

Available baseline	variables	
Age (yrs)	Creatinine limit	Calcium (mg/dL)
Weight (kg)	Serum B2M (mg/L)	Serum M-component
Height (cm)	Serum LDH (U/L)	Hemoglobin
Albumin (μL)	White blood cell count	Bone lesions
Creatinine (μL)	Platelets	Performance status

The marginal results suggest only a small difference between the two treatments,

and actually slightly favor SDT over HDT. The estimated (unadjusted) hazard ratio between HDT and SDT using Cox regression was $\widehat{\text{HR}} = 1.03$ (95% CI: (0.81, 1.30)). For the comparison of survival curves under various decision rules, we focus on the three summary measures: probability of 5-year survival, median survival time, and the area under the survival curve restricted at 10 years, AUSC (10yrs).

The overall 5-year survival was 53% in the SDT group and 45% in the HDT group, while the median survival time was 5.2 years (95% CI: (4.3, 6.6)) and 4.5 years (95% CI: (3.6, 6.0)), and the AUSC(10) was 5.5 and 5.4 years, respectively. The Kaplan-Meier curves by treatment can be seen in Figure 3.4 and show minimal difference in the overall survival between the two treatment arms. However, Barlogie et al. (2006) discussed that there was a differential response to the HDT treatment associated with some of the baseline markers and we were therefore interested to explore if we can develop a marker-guided therapy that would improve survival of the patient population, i.e. whether we can identify a subgroup of patients benefitting from the high-dose chemoradiotherapy despite its null marginal effect.

For our analysis, we used a leave-one-out cross-validation algorithm, and in each fold the LASSO for Cox model served to develop a sequence of decision rules based on nested set of marker panels, $p = 1, 2, ...$ The variables that tend to be very skewed, such as albumin, creatinine, white blood cell count and LDH, were log-transformed, and the "working" Cox model was linear and additive with interactions between individual variables and the treatment, as outlined in (3.3). For each panel size, the marker-guided population survival (MGPS) was then estimated based on the whole data set via all three proposed estimators, as described in section 3.2.4.

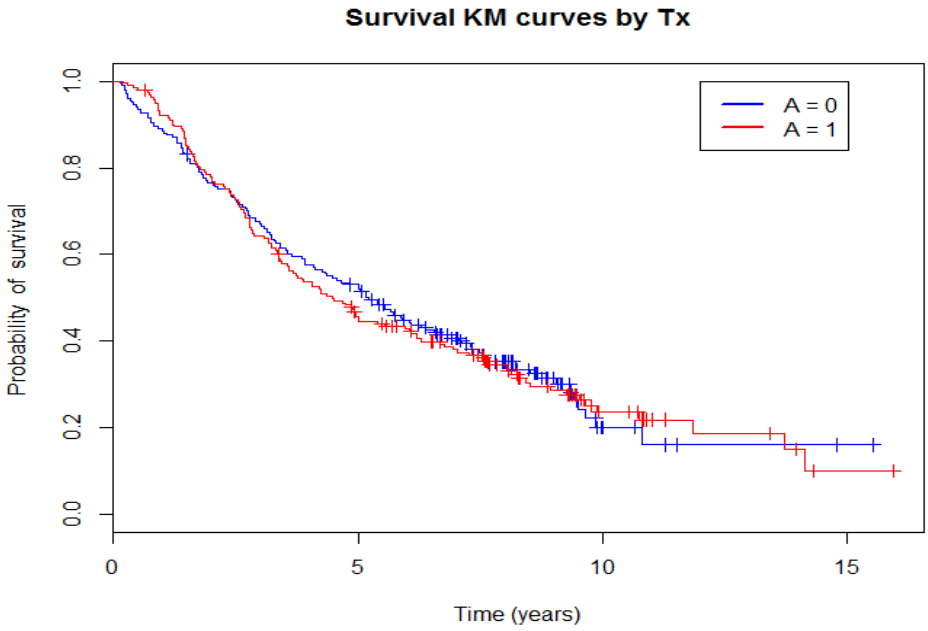

Figure 3.4: Multiple Myeloma Example: Estimated survival by Kaplan-Meier curves for the two treatment arms; high-dose chemoradiotherapy ($A = 1$) and standard-dose therapy ($A = 0$).

The comparison of all three summary measures across panel sizes is shown in Figure 3.5 (included at the end of the chapter). The developed decision rules based on LASSO method appear to perform even more poorly than the uniform standard-dose treatment ($A = 0$) in terms of all three measures. The proportion of patients treated with HDT ($A = 1$) in the bottom plot shows that the decision rules based on more than 3 markers suggest to treat about 50% of the population, however the resulting AUSC(10), 5-year survival and median are all lower than when uniformly treating everybody with the less invasive SDT ($A = 0$). The negative results are supported consistently by all the three estimators.

This example again implies that a search for patient subgroups with strong benefit

from the experimental treatment should be carried out carefully. Marginal marker-by-treatment interactions can give and illusive idea of existence of such subgroup benefits, which might be often an artifact of numerous comparisons across multiple markers and corresponding subgroups. It is therefore crucial that the population impact of the developed treatment rules is properly evaluated in order to avoid hasty enthusiasm from spurious discoveries.

3.5 Discussion

In this chapter, the proposed methods focus on the evaluation of marker-guided treatment rules with a time-to-event outcome and in a presence of censoring. The assumed clinical objective is an improvement in the population survival, which is being summarized via three common one-dimensional quantities: Area Under the Survival Curve (AUSC); survival at a specific time of interest, $G(t^*)$; and median survival, $G^{-1}(0.5)$. The goal is to score patients with respect to their expected treatment benefit so that the population measures of performance are maximized over a set of candidate marker-guided treatment rules.

Similarly as in the chapter for continuous outcomes, our algorithm for the development of treatment decision rules combines variable selection methods and careful evaluation of the candidate rules via cross-validated estimation of their net population impact. For high-dimensional $\boldsymbol{X}$, we proposed to select candidate marker panels by L_1 penalized least square method for Cox regression, which again leads to a nested set of marker panels. The working model for the score development was chosen Cox proportional hazards model which is linear and additive on the log scale.

We proposed three different estimators of the net population survival under the marker-guided treatment rules. The simple non-parametric estimator assesses the net population survival by a Kaplan-Meier curve based only on the patients who have their randomized treatment consistent with the rule-based assignment. The weighted Kaplan-Meier estimator combines individual survival curves estimated locally based on neighborhoods of patients with similar scores. The third approach is semi-parametric and adopts the Cox PH model with splines in the score, assuming either a common or separate baseline hazards for the two treatment arms.

While the first two estimators do not require a model specification, the Cox-model based estimator seems to be insensitive to the model selection and provides a smoother fit of the population survival curve. Using simulations, we evaluated bias and variance of all three estimators based on 300 samples drawn from the same population. The two non-parametric estimators appear to perform very similarly in terms of both bias and variance across all three examined survival summary measures (AUSC, $G(t^*)$, and $G^{-1}(0.5)$). The Cox-model based estimator shows slightly smaller variance than the non-parametric estimators when the data-generating model is correctly specified, however, it tends to overestimate the marker-guided population survival for larger panel sizes, where the non-parametric estimators showed only minimal bias.

Our methods are illustrated on an example from a prospective randomized phase III trial of multiple myeloma, which showed no marginal benefit from the high-dose chemoradiotherapy as compared to the standard-dose therapy. Despite hypothesized differential response to the HDT associated with some of the baseline markers, the developed decision rules appeared to perform even more poorly than the uniform standard-dose

treatment in terms of all three measures, and these results were suggested consistently by all three proposed estimating approaches.

In summary, we offer statistical tools that allow for honest, cross-validated estimation of the net population survival under the set of nominated treatment prescription rules. In the next chapter, we would like to detail and compare design characteristics for a follow-up evaluation study that would assess how a treatment benefit score performs as a predictor of response and as a basis for treatment choice.

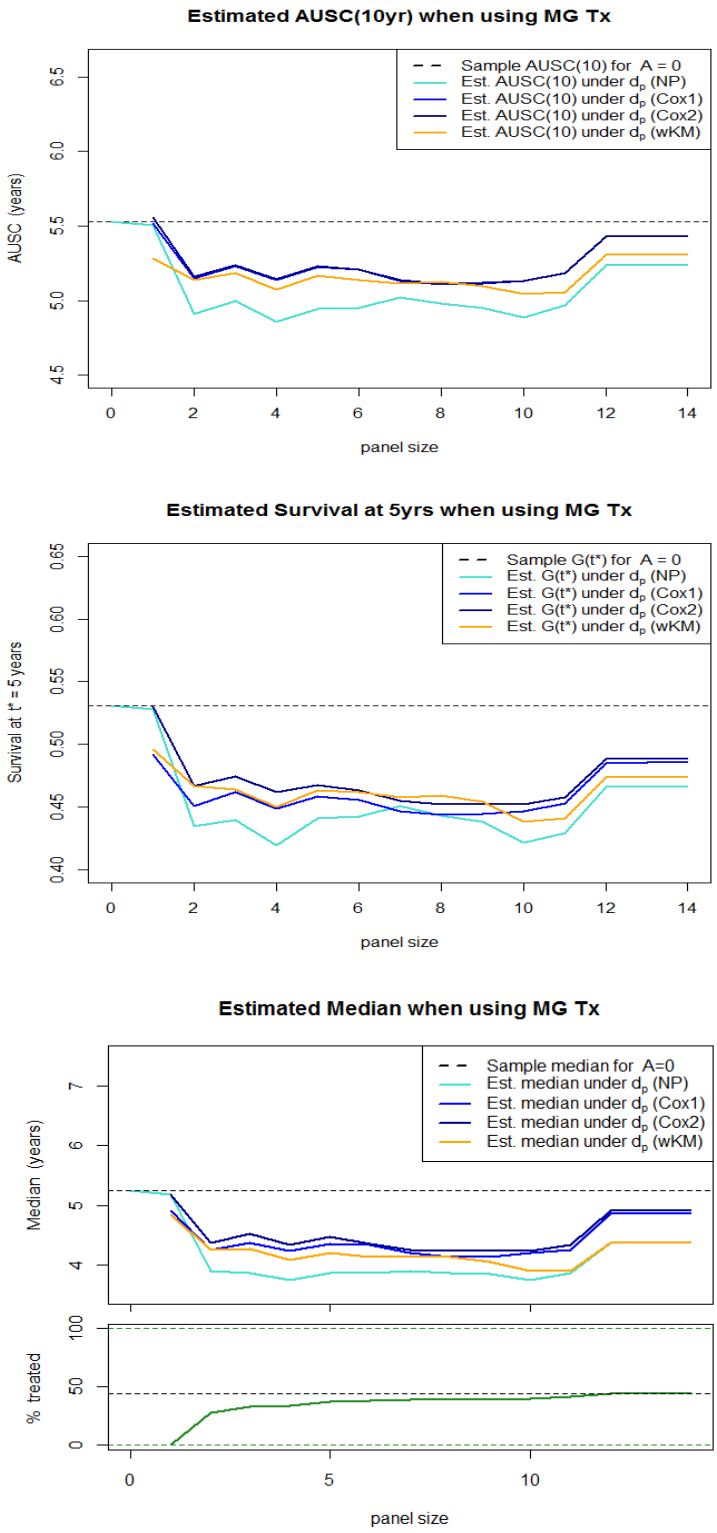

Figure 3.5: Estimated Measures of Performance (MoP) from Multiple Myeloma Example: AUSC(10yr), probability of 5-year survival and median survival time, as functions of the marker panel size. The MoP are based on the underlying Marker-Guided Population Survival curves (MGPS) estimated under the corresponding decision rules by 1) simple non-parametric approach (turquoise), 2) PH Cox model with common baseline hazard (light blue) and separate baseline hazards (dark blue), and 3) weighted Kaplan-Meier curves (orange). The green line in the bottom plot shows proportion of patients suggested to be treated by HDT across decision rules based on different marker panels.

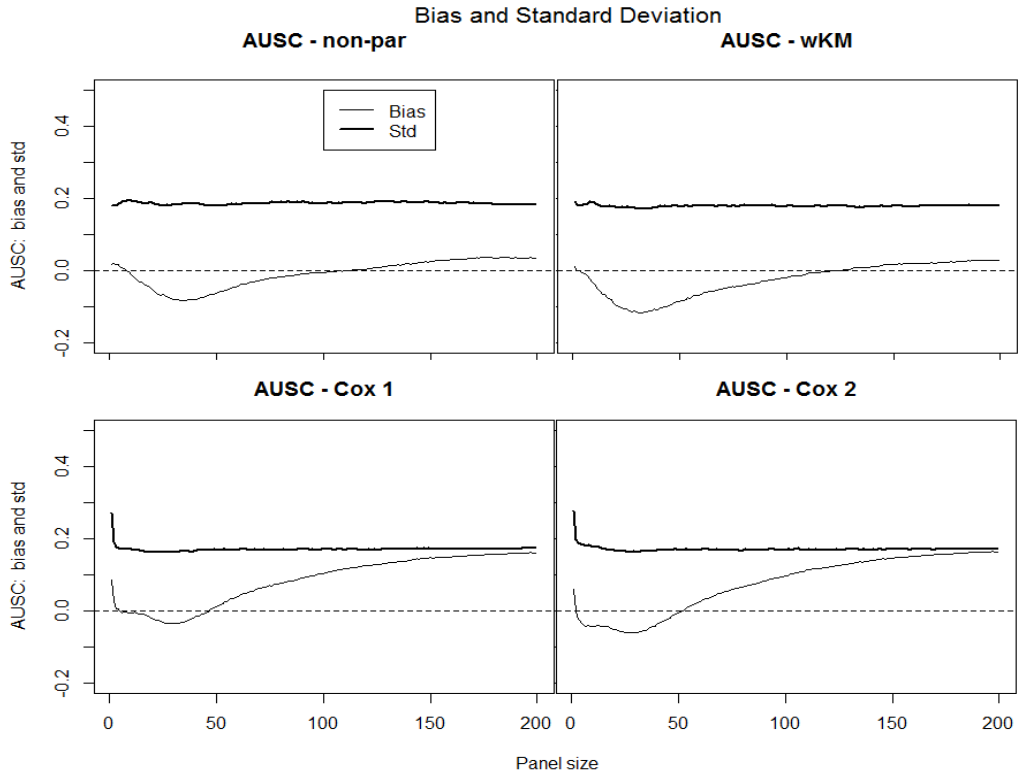

Figure 3.6: Bias (thin lines) and standard deviation (thick lines) of the proposed estimators as functions of the panel size p : population AUSC(10yr) under marker-guided treatment rules.

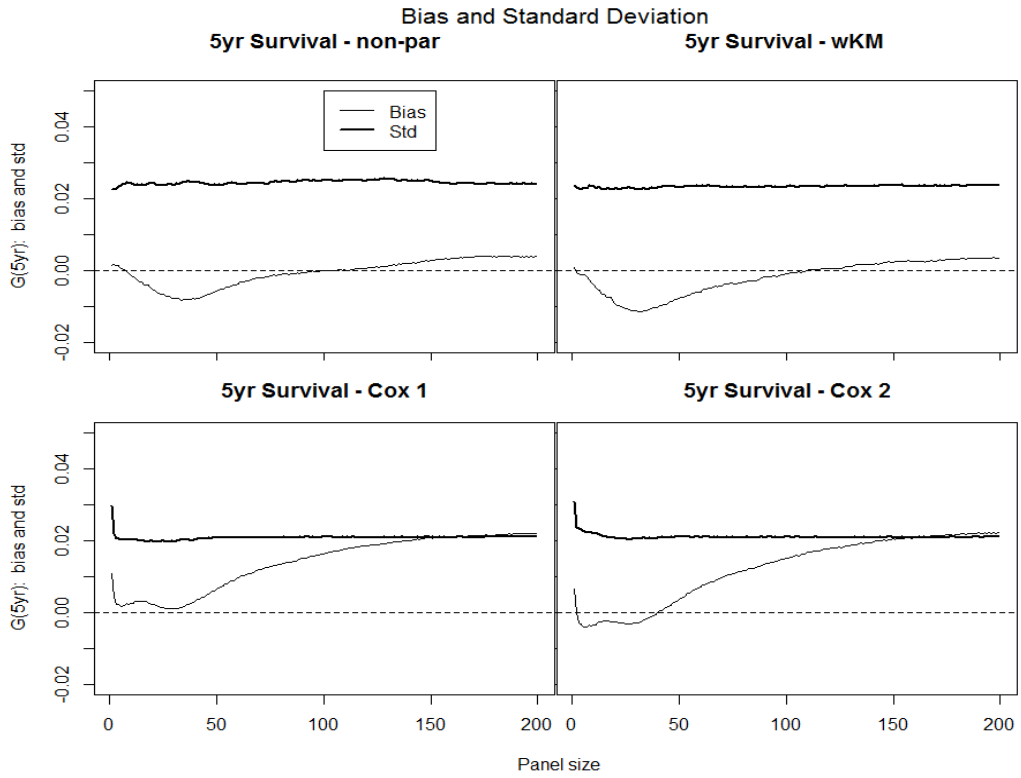

Figure 3.7: Bias (thin lines) and standard deviation (thick lines) of the proposed estimators as functions of the panel size p : population survival at 5 years under marker-guided treatment rules.

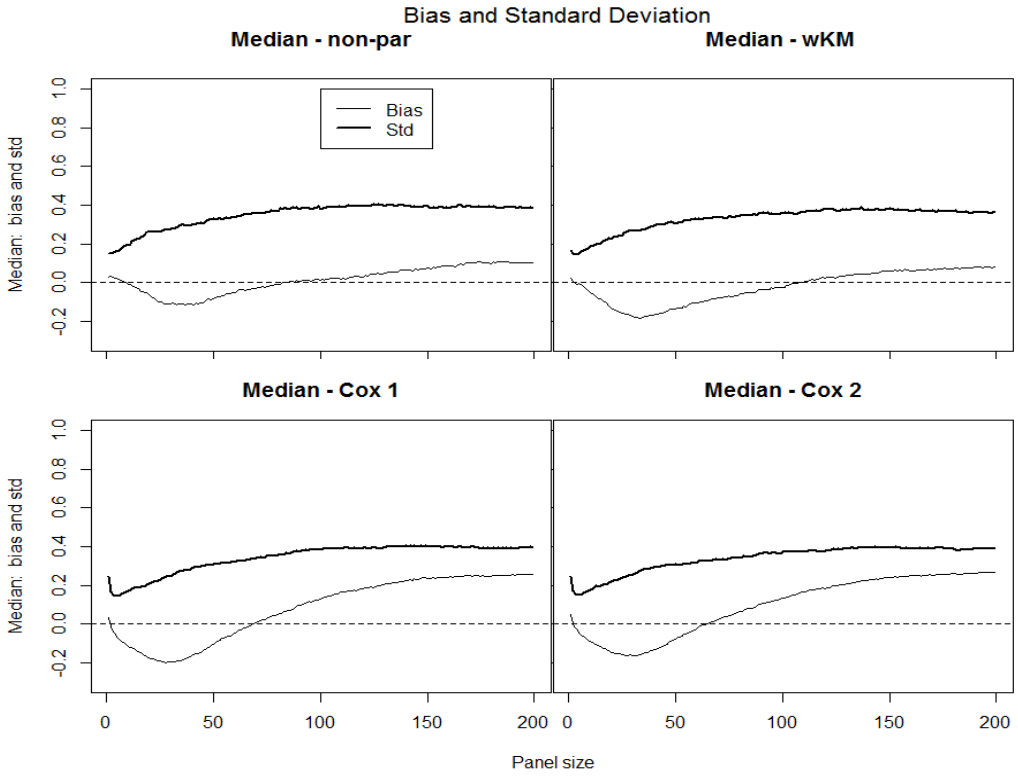

Figure 3.8: Bias (thin lines) and standard deviation (thick lines) of the proposed estimators as functions of the panel size p : median population survival under marker-guided treatment rules.

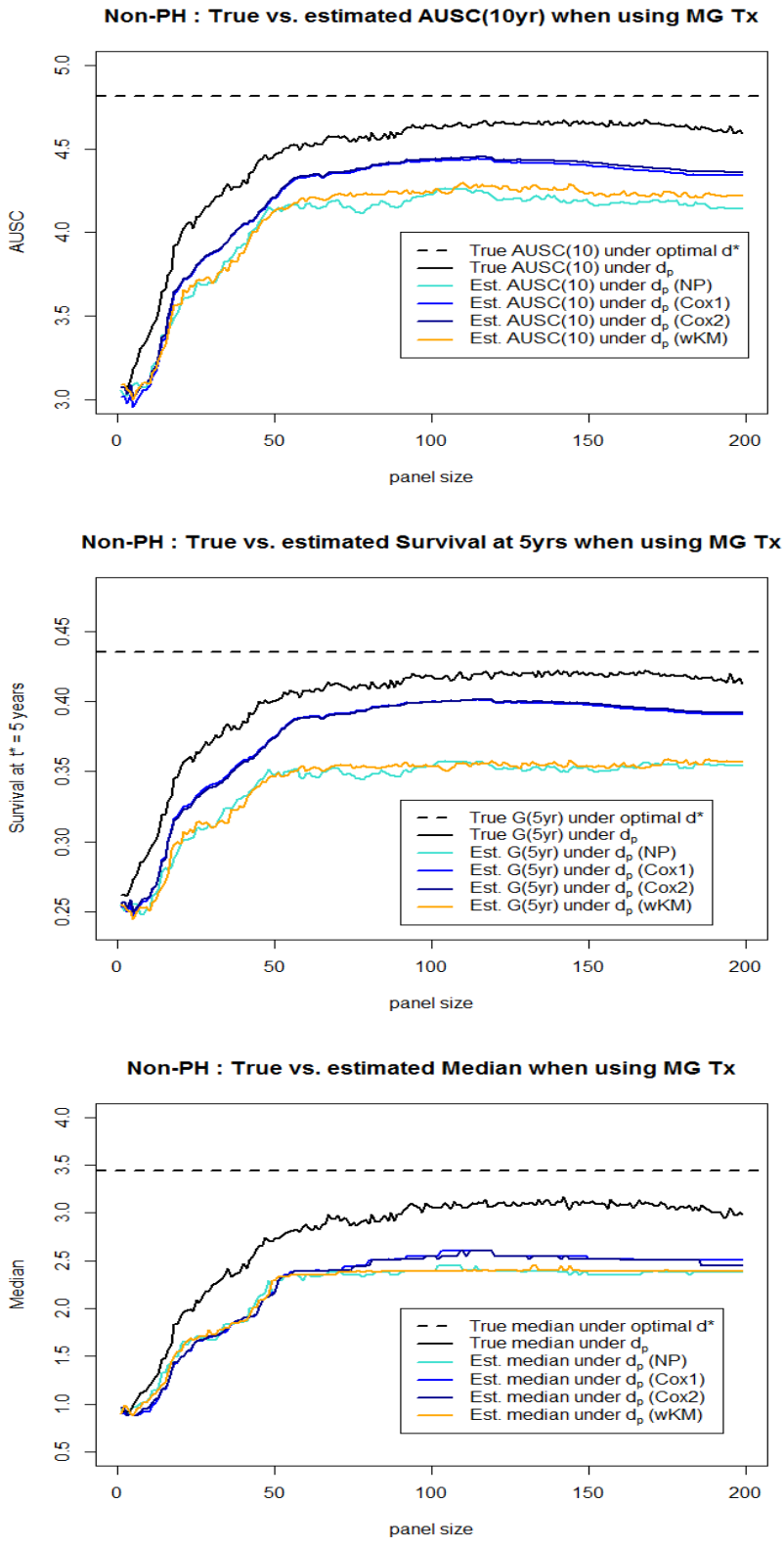

Figure 3.9: Estimated Measures of Performance (MoP) from one simulated data set under non-PH: AUSC(10yr), probability of 5-year survival and median survival time, as functions of the marker panel size. The MoP are based on the underlying Marker-Guided Population Survival curves (MGPS), corresponding to 1) the true MGPS (black dashed line) and estimators by 2) simple non-parametric approach (turquoise), 3) PH Cox model with common baseline hazard (light blue) and separate baseline hazards (dark blue), and 4) weighted Kaplan-Meier curves (orange). The horizontal dashed line shows the population MoP under an optimal decision rule.

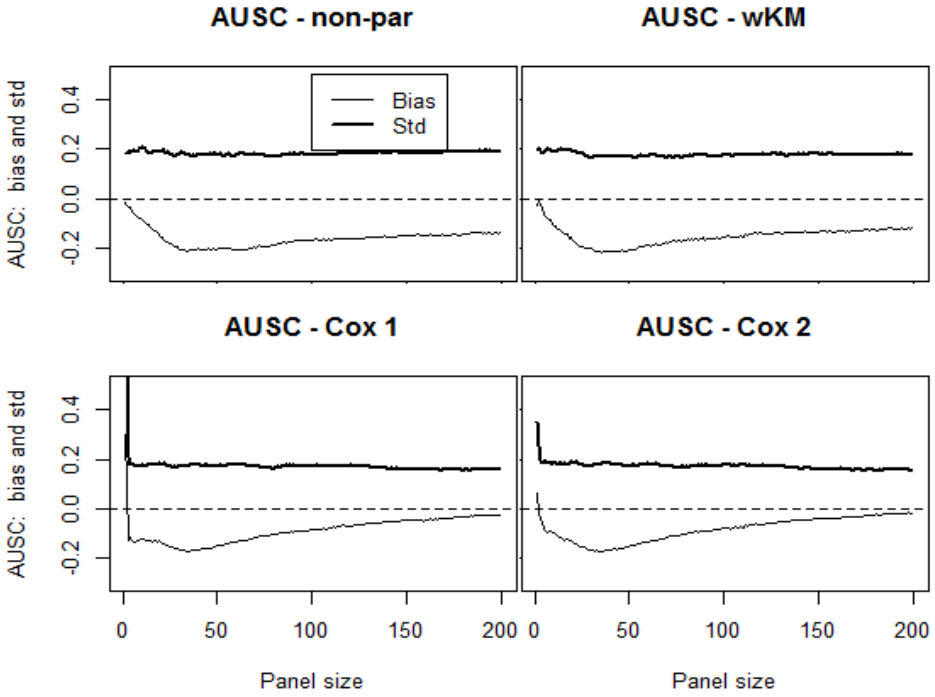

Figure 3.10: Non-PH data-generating model: Bias (thin lines) and standard deviation (thick lines) of the proposed estimators as functions of the panel size p : population AUSC(10yr) under marker-guided treatment rules.

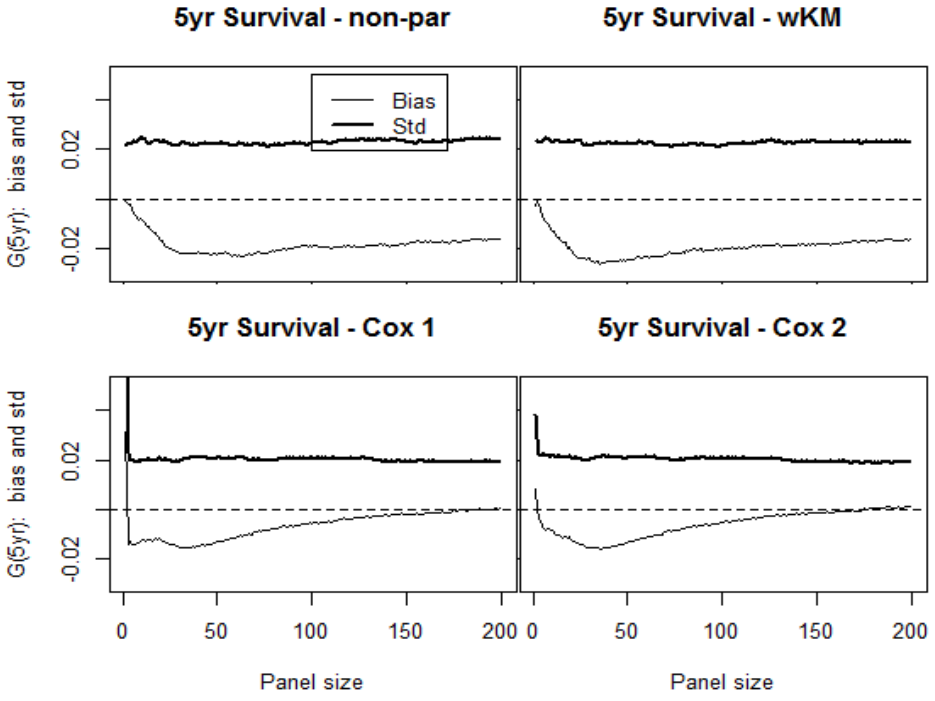

Figure 3.11: Non-PH data-generating model: Bias (thin lines) and standard deviation (thick lines) of the proposed estimators as functions of the panel size p : population survival at 5 years under marker-guided treatment rules.

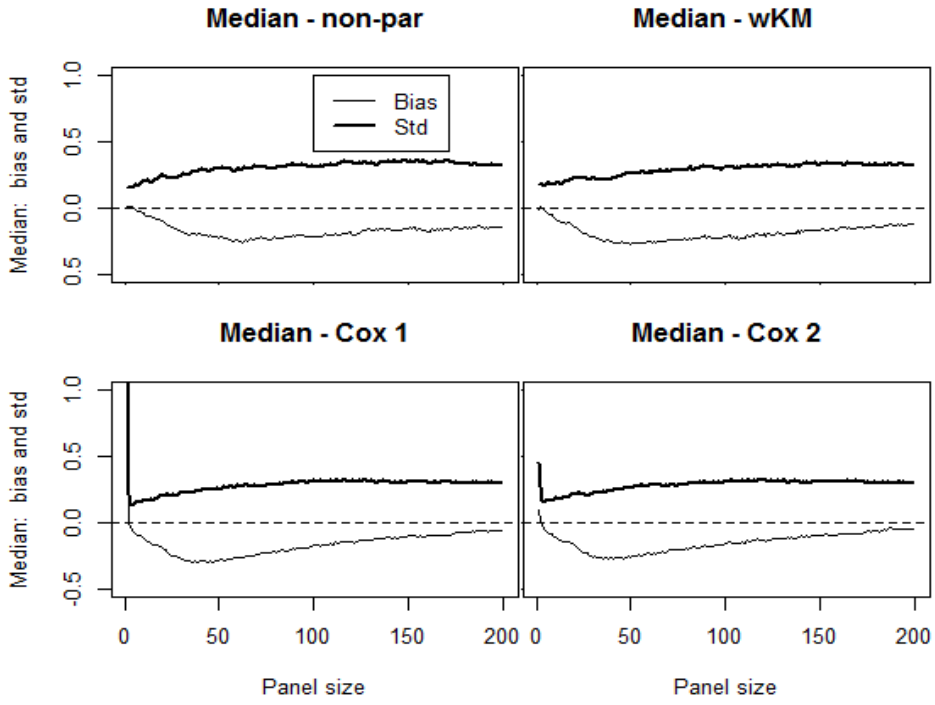

Figure 3.12: Non-PH data-generating model: Bias (thin lines) and standard deviation (thick lines) of the proposed estimators as functions of the panel size p : median population survival under marker-guided treatment rules.

Chapter 4

INDEPENDENT VALIDATION STUDY FOR MARKER-GUIDED TREATMENT

4.1 Introduction

The primary focus of our work is on the development of a single-index marker score that could be used to guide clinical decision making. Subsequently to the "discovery" of a score a careful validation study would be important in order to determine whether the specified marker-guided treatment (MGT) rule truly benefits the patients as suggested by the analysis of the source clinical trial data. Despite the protective cross-validation approach for the score development, data-driven variable selection is prone to imply overly optimistic results and out of sample generalization may not hold due to different population characteristics. A properly conducted confirmatory study is necessary to demonstrate the therapeutic relevance and robustness of the developed treatment rule. The goal of such external validation is hence to assess both quality and clinical impact of using a completely specified decision rule to guide the treatment.

While standard randomized clinical trials focus on evaluation of marginal treatment effect in a broad population, the validation of the clinical utility of a marker-guided treatment rule is more complex, as it requires evaluation of the relationship between the corresponding score and the score-specific treatment effect. Howeverm such comprehensive assessment is not always feasible, as we will discuss later, and the ultimate study design needs to compromise between validating the clinical utility of the score,

evaluating the score-indicated subgroup treatment benefit and doing so effectively.

In this chapter, we will detail and compare design characteristics for a follow-up evaluation study that aims to assess how the proposed treatment benefit score performs as a predictor of response and as a basis for treatment choice. On the individual level, we wish to evaluate whether the developed score is a strong and accurate predictor of the individual treatment benefit while at the population level we seek to establish superiority over the standard of care therapy.

4.2 Validation of Marker-Guided Treatment Rule

In order to determine whether a pre-specified marker-guided treatment (MGT) truly benefits the patients as suggested by the analysis of the source study, we consider validating three important aspects of the developed treatment decision rule: 1) superiority of the MGT over the standard therapy in the target population, 2) positive correlation between the treatment-guiding score and the actual treatment benefit corresponding to the score, and 3) accuracy of the score in assessing the associated treatment benefit and classifying the patients.

4.2.1 Superiority

In a successive prospective (phase III) clinical trial, the primary goal is to properly and efficiently test for superiority of the MGT when compared to standard care at the population level. It implies to test a null hypothesis that the mean outcome in the population under MGT is no better than under the scenario where everybody would be

treated according to the standard of care (SoC),

$$H_0 : \mathrm{E}\, Y(d^{\mathrm{MGT}}) \leq \mathrm{E}\, Y(d^{\mathrm{SoC}})\,,$$

where d^{MGT} is the mean outcome in the population under MGT and d^{SoC} is the mean outcome in the population under SoC. If the null is rejected, we can conclude that the proposed treatment decision rule (significantly) improves the mean population outcome and is therefore validated for guiding the treatment choice.

In the situation where the standard care corresponds to one of the treatment arms (e.g., $A = 0$) and the score-based decision rule is of the previously assumed form, $d(S) = \mathrm{I}\,[S > 0]$, the MGT and standard care treatment are equal for part of the population (here, for the score-negative patients). Any potential difference in the mean outcome between the two treatment regimens is therefore only due to the treatment effect in the subgroup of patients that have MGT different from the standard of care treatment. As a result, a more efficient way to assess the population benefit of MGT over SoC is to test the difference in the mean outcome between the two treatment arms among the score-positive patients only, as proposed by Simon and Maitournam (2004). The null hypothesis of non-superiority of MGT over SoC is then equivalent to

$$H_0 : \mathrm{E}[Y|A = 1, S > 0] \leq \mathrm{E}[Y|A = 0, S > 0]\,,$$

or

$$H_0 : \mathrm{E}[\Delta(S)|S > 0] \leq 0\,,$$

where $\Delta(s) = \mathrm{E}[Y|A = 1, S = s] - \mathrm{E}[Y|A = 0, S = s]$ is an expected treatment benefit among patients with score $S = s$. Rejecting H_0 means there is evidence that the treatment works better in the subgroup of score positive patients, which implies superiority of MGT over the standard care (assuming $s < 0$ corresponds to choosing the standard therapy).

4.2.2 Correlation

At the individual level, we wish to evaluate whether the score S is a strong and accurate predictor of the individual treatment benefit

$$\Delta(S) = \mathrm{E}[Y|S, A = 1] - \mathrm{E}[Y|S, A = 0].$$

However, the actual treatment benefit is not observable at an individual level and hence the correlation between S and $\Delta(S)$ cannot be quantified via standard sample measures such as Pearson or Spearman correlation coefficient.

Instead, we assess the linear correlation between S and $\Delta(S)$ by adopting a parametric model and estimating the associated parameters. Since the score is supposed to predict the actual benefit, we assume a linear relationship $\Delta(S) = \alpha_0 + \alpha_1 S$. Testing whether $\Delta(S)$ correlates well with S is then equivalent to the test of interaction between treatment and score in the semi-parametric model:

$$\mathrm{E}[Y|S, A] = \mathrm{E}[Y|S, A = 0] + \alpha_0 A + \alpha_1 AS\,. \tag{4.1}$$

Rejecting the null hypothesis $H_0 : \alpha_1 \leq 0$ would imply that the examined treatment

benefit score positively correlates with the actual treatment benefit.

Similarly to the developmental stage, estimation of the mean outcome as a function of the score will be based on smooth curves. In the assumed model (4.1), the expected outcome $\mathrm{E}[Y|S, A = a] = \mu_0(S)$ under the two treatment arms $a \in \{0, 1\}$ will be approximated by a smoothing method, such as penalized splines. Such an approach allows the estimated mean outcome in each treatment arm to be a flexible function of the score, while their difference is a linear function of the score, parametrized by (α_0, α_1).

The test of association between S and $\Delta(S)$ will be based on regression coefficient of the score-treatment interaction α_1 from model (4.1). If the association proves to be significant, it then also becomes of interest to evaluate how accurate the score S is as a predictor of the corresponding treatment benefit and as a patient classifier.

4.2.3 Accuracy

The accuracy of the score can be assessed through estimation of the coefficients α_0 and α_1 in (4.1). Under the assumed linear relationship, the parameter value $\boldsymbol{\alpha} = (\alpha_0, \alpha_1) = (0, 1)$ corresponds to an optimally calibrated predictor. If S is fairly accurate, $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ should be close to 0 and 1, respectively. While $\alpha_0 > 0$ suggests that some of the patients with negative scores are actually expected to benefit from the treatment, $\alpha_0 < 0$ would indicate that not all the patients with positive scores are necessarily expected to benefit from the treatment. Similarly, an estimated slope α_1 greater or less than 1 suggests that the expected benefit grows with the increasing score faster or slower, respectively, than is predicted by the score S.

Often, secondary interest might be in testing whether the classification of patients

to receive treatment $A = 0$ vs. $A = 1$ is accurate. Assuming a monotone relationship between the treatment benefit and the score, this reduces to testing whether the score-based classification threshold of 0 really corresponds to treatment benefit $\Delta(0) = 0$, or, in the terms of our model, whether $\alpha_0 = 0$. However, even under these assumptions, such test of score classification accuracy does not have a simple null hypothesis. We wish to reject the null in favor of an alternative, $H_A : \alpha_0 = 0$, hence an "inaccurate" score thereshold will be considered such that α_0 is far from H_A. It leads to testing a composite null hypothesis

$$H_0 : |\alpha_0| \geq c_0,$$

where the constant $c_0 > 0$ should be chosen based on the particular clinical context.

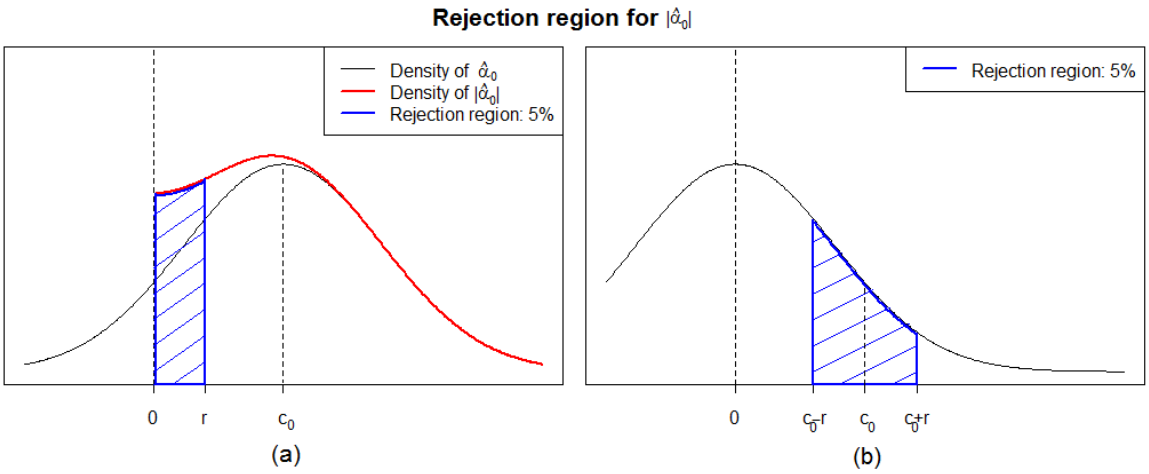

Figure 4.1: Rejection region for the null hypothesis $H_0 : |\alpha_0| \geq c_0$ graphed in (a) the distribution of $|\widehat{\alpha}_0|$ and (b) when transformed into the Normal distribution centered at 0 with the same variance as $\widehat{\alpha}_0$, $N(0, \sigma_0^2)$.

In order to calculate the rejection region for $\widehat{\alpha}_0$, we assume that its distribution under H_0 is $\widehat{\alpha}_0 \sim N(c_0, \sigma_0^2)$. Then the distribution of $|\widehat{\alpha}_0|$ is as shown in red in the

Figure 4.1(a) and we wish to reject H_0 for small values of $|\widehat{\alpha}_0|$. If we consider rejection at 5% significance level, the area of rejection region which is highlighted under the density function curve in blue needs to be 0.05. This area is equivalent to the highlighted area in the Figure 4.1(b), where the center of the density was just shifted from c to 0. Hence the rejection region for $|\widehat{\alpha}_0|$ is $[\,0, r)$, and the border value r is given by the equation

$$F(c_0 + r) - F(c_0 - r) = 0.05\,,$$

where F is a probability function of $N(0, \sigma_0^2)$. As the parameter σ_0 is usually unknown, we approximate r by the solution of

$$F_{N(0,1)}((c_0 + r)/\widehat{\sigma}_0) - F_{N(0,1)}((c_0 - r)/\widehat{\sigma}_0) = 0.05\,,$$

where $F_{N(0,1)}$ is the probability function of standard normal distribution $N(0, 1)$, and we reject the null hypothesis H_0 of the score threshold inaccuracy if $|\widehat{\alpha}_0| < r$.

Alternatively, one might want to test whether the whole estimated linear relationship between S and $\Delta(S)$ represented by the vector $(\widehat{\alpha}_0, \widehat{\alpha}_1)$ is consistent with an accurate score. We translate this idea into a composite null hypothesis

$$H_0 : (\alpha_0, \alpha_1) \in \Theta_0\,,$$

where Θ_0 is a pre-specified set of vectors $\boldsymbol{\alpha}$ that correspond to "inaccurate" scores. Since the parameters α_0 and α_1 determine the linear relationship jointly, we would like to define Θ_0 so that small deviations in one allow for larger deviations in the other and

vice versa. For example, if $\alpha_0 = 0$ then α_1 would be permitted to vary between $1 \pm c_1$, but if $|\alpha_0|$ is close to c_0, the slope would need to be more accurate. For an illustration, we set the null region to be an ellipse (see Figure 4.2) defined by

$$\Theta_0 = \left\{ (\alpha_0, \alpha_1) \in \mathbb{R}^2 \, : \, \left(\frac{\alpha_0}{c_0}\right)^2 + \left(\frac{\alpha_1 - 1}{c_1}\right)^2 \geq 1 \right\},$$

where the selected constants $c_0 > 0$ and $1 > c_1 > 0$ depend on the clinical context.

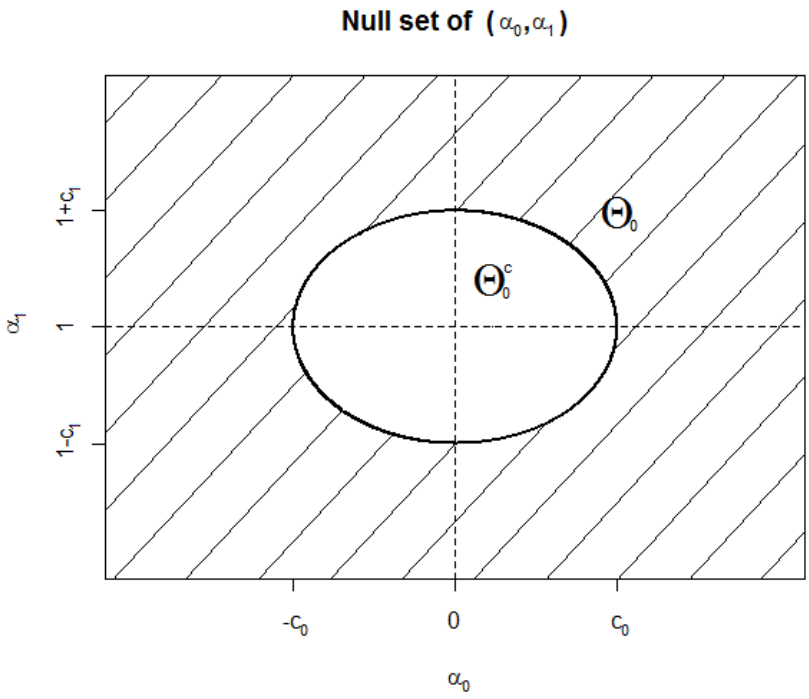

Figure 4.2: Null set of the composite hypothesis of an inaccurate score S under the model (4.1), $H_0 : \boldsymbol{\alpha} \in \Theta_0$, where Θ_0 is given by $\{(\alpha_0, \alpha_1) \in \mathbb{R}^2 \, : \, \left(\frac{\alpha_0}{c_0}\right)^2 + \left(\frac{\alpha_1 - 1}{c_1}\right)^2 \geq 1\}$. The central point $\boldsymbol{\alpha}^A = (0, 1)$ corresponds to an optimally calibrated score S as a predictor of the treatment benefit $\Delta(S)$ under the model (4.1).

The composite null hypothesis can be then tested against the alternative $H_A : \boldsymbol{\alpha} \in \Theta_0^c$ using a likelihood ratio test, which compares the largest likelihood of parameter values

from the null space with the largest likelihood from the whole space $\Theta = \Theta_0 + \Theta_0^c$. If the maximum likelihood from the null set is as high, or almost as high, as the overall maximum likelihood, and their ratio is hence close to 1, we do not reject the null. On the other hand, if the maximum likelihood from the alternative set is much higher than the one from the null set, we would reject H_0 in favor of the alternative. The likelihood ratio test (LRT) of our composite null is therefore based on the following statistics

$$\Lambda(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\alpha}\in\Theta_0,\boldsymbol{\beta}} L(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{x})}{\sup_{\boldsymbol{\alpha}\in\Theta,\boldsymbol{\beta}} L(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{x})},$$

where $\boldsymbol{x}$ is the observed data, $\boldsymbol{\beta}$ are the parameters corresponding to parametric splines for μ_0, and the likelihood $L(\boldsymbol{\alpha},\boldsymbol{\beta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\beta})$. It was shown by Wilks (1938) that $-2\log(\Lambda)$ has asymptotically $\chi^2_{(p)}$ distribution, with p being the dimension reduction of tested submodel, or the number of parameters defining Θ_0, so in our case $p = 2$. Asymptotic properties of the likelihood-ratio test in the context of semi-parametric models are discussed by Murphy and Van Der Vaart (1997).

The supremum in the denominator of Λ is the likelihood at the maximum-likelihood estimate, $(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}})$, and hence it equals to $L(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}}|\boldsymbol{x})$. If moreover the MLE $\widehat{\boldsymbol{\alpha}} \in \Theta_0$, then we know that the supremum over space Θ_0 is also at $\widehat{\boldsymbol{\alpha}}$ and the LR statistics $\Lambda(\boldsymbol{x}) = 1$ (which means we do not reject H_0). If $\widehat{\boldsymbol{\alpha}} \in \Theta_0^c$, then the maximum likelihood over the null set is attained on the boundary of Θ_0 and we thus only need to calculate $L(\boldsymbol{\alpha},\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})|\boldsymbol{x})$ over the boundary Θ_0^B, where $\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ is the MLE of $\boldsymbol{\beta}$ for a specific fixed value of $\boldsymbol{\alpha}$.

Now we will show how the likelihood at any given vector of parameters $\boldsymbol{\alpha}'$ can be

calculated using standard software, for example, via function `logLik` in R package `stats`. Based on the assumed model (4.1), we can write the conditional expected outcome as

$$\begin{aligned} \mathrm{E}[Y|A,S] &= \mu_0(S) + \alpha_0 A + \alpha_1 AS \pm (\alpha_0' A + \alpha_1' AS) \\ &= \mu_0(S) + (\alpha_0 - \alpha_0')A + (\alpha_1 - \alpha_1')AS + (\alpha_0' A + \alpha_1' AS) \\ \Longrightarrow & \\ \mathrm{E}[Y^*|A,S] &= \mathrm{E}[Y|A,S] - (\alpha_0' A + \alpha_1' AS) = \mu_0(S) + \alpha_0^* A + \alpha_1^* AS\,, \end{aligned}$$

where $\mu_0(S) = \mathrm{E}[Y|A=0,S]$, $Y^* = Y - (\alpha_0' A + \alpha_1' AS)$ and $\boldsymbol{\alpha}^* = \boldsymbol{\alpha} - \boldsymbol{\alpha}'$. Hence the likelihood at $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$ in model (4.1) is equivalent to the likelihood at $\boldsymbol{\alpha}^* = \mathbf{0}$ in the model

$$\mathrm{E}[Y^*|A,S] = \mu_0(S) + \alpha_0^* A + \alpha_1^* AS\,,$$

which can be obtained from `logLik(lm(`$y^* \sim \mu_0(s)$`))`. Since the boundary of the null set Θ_0 is closed, the supreme likelihood over Θ_0 when $\widehat{\boldsymbol{\alpha}} \in \Theta_0^c$ is simply the maximum likelihood over the boundary set

$$\Theta_0^B = \left\{ (\alpha_0, \alpha_1) \,:\, \left(\frac{\alpha_0}{c_0}\right)^2 + \left(\frac{\alpha_1 - 1}{c_1}\right)^2 = 1 \right\}.$$

Hence, our LRT statistics simplifies to

$$\Lambda(\boldsymbol{x}) = \frac{\max_{\boldsymbol{\alpha}\in\Theta_0^B} L(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})|\boldsymbol{x})}{L(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}|\boldsymbol{x})}\,,$$

and $L(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})|\boldsymbol{x})$ for $\boldsymbol{\alpha} \in \Theta_0^B$ can be easily calculated as described above.

4.3 Designs of Validation Study

In order to examine estimation and testing properties of the outlined score-validation procedures, we focus on some of the most commonly adopted study designs and the corresponding data analyses. For testing the MGT superiority a typical design is where all patients are randomized to MGT versus standard care, and the treatment is assigned based on the score only in the MGT arm. It is however well recognized that such a design would require an unnecessarily large sample size due to use of subjects for whom marker-guided and standard treatments do not differ. Randomizing only patients for whom the MGT and standard are discordant, or an "enrichment strategy design", can lead to a substantial reduction in required sample size (Simon, 2008), depending on the proportion of score positive patients.

4.3.1 Marker-guided Treatment Strategy Design

Randomized controlled phase III trials are the gold standard for evaluation of a new (experimental) therapy. In order to evaluate whether a (fully specified) marker-guided treatment is more beneficial than the standard care, patients diagnosed with the studied disease or health condition are randomized to either marker-guided treatment (MGT) arm or standard of care (SoC) treatment arm. Subsequently, the biomarkers are measured for the patients who were randomized to MGT arm and their treatment assignment is based on the resulting treatment benefit score. In this simplest version, the two treatment arms are then compared with respect to the average outcome either by a standard t-test or some non-parametric alternative, such as Wilcoxon rank-sum test.

Marker-guided treatment strategy design allows for the most direct comparison be-

tween the MGT and SoC therapy in the overall population and provides the most realistic estimate of the MGT effectiveness, as was already discussed by Freidlin et al. (2012). It naturally accounts for the measurement error of the score as well as issues with non-compliance. Particularly in the SoC arm, patients do not have their scores evaluated at the time when treatment is administered and so the compliance with the treatment is not influenced by their score knowledge.

Validation study designs that enroll all patients irrespective of their marker-based score have, however, been criticized by, for example, Simon (2008) for being possibly very inefficient and requiring huge sample sizes in order to assess the difference between the two contrasted treatment strategies. The primary analysis in such trials is based on all randomized patients, which often fails to recognize an effective therapy due to "dilution" of its overall effect by patients who would receive the same treatment under both arms. This inefficiency might be particularly high if the proportion of patients receiving differential treatments under the two arms is small.

4.3.2 Enrichment Strategy Design

A more efficient study design proposed by Simon and Maitournam (2004) is based on the idea that the score-negative patients are assigned to the same treatment (in our case it is $A = 0$) under both MGT and SoC arms and they therefore do not contribute any information about the contrast between the two strategies. The enrichment strategy design hence suggests to first assess the marker-based score on all incoming patients and subsequently randomize only those who have a positive score $s > 0$. The authors showed that such a targeted clinical trial design can largely reduce the number of patients required

for the validation study as compared to the previous design. If the score reliably identifies the subgroup of patients who are likely to benefit from the experimental treatment ($A = 1$), then the analysis based on the score-positive patients only provides an efficient test of its efficacy (Simon and Maitournam, 2004). Moreover, with this approach one needs to randomize smaller portion of the patients than under the first (MGT strategy) design, which might be convenient if the treatments are more expensive than the score assessment.

The corresponding analysis in the enrichment design hence focuses on testing the standard null hypothesis among the score positive patients:

$$H_0 : \mathrm{E}[Y|A = 1, S > 0] \leq \mathrm{E}[Y|A = 0, S > 0]$$

or, equivalently

$$H_0 : \mathrm{E}[\Delta(S)|S > 0] \leq 0\,,$$

where $\Delta(s) = \mathrm{E}[Y|A = 1, S = s] - \mathrm{E}[Y|A = 0, S = s]$ is an expected treatment benefit among patients with score $S = s$. Rejecting H_0 implies there is an evidence that the experimental treatment $A = 1$ works better in the subgroup of score-positive patients and suggests superiority of MGT over the standard care therapy.

The main disadvantage of both study designs described above is that we do not learn about the treatment effect in the score-negative patients. Even if the results are positive in the subgroup of score-positive patients, it does not prove the clinical utility of the score, and we cannot rule out that the experimental treatment is not efficacious also among at least part of the score-negative patients. Since the evaluated score is a continuous

measure, its quality as a predictor of the treatment benefit can be partially (even though very inefficiently) assessed among the score-positive patients. The extrapolation of the estimated relationship to the score-negative patients would however be dangerous.

4.3.3 Score–stratified Design

Apart from assessing superiority of the marker-guided treatment strategy over the standard care, there is often an interest in evaluating the clinical utility of the proposed score. The first two outlined study designs do not allow for assessment of the treatment benefit among the score-negative patients, as those patients are either always assigned to the standard care (MGT strategy design) or left out of the study (Enrichment design). It has been therefore advocated to use a Score-stratified design, which assigns all patients randomly to either treatment $A = 0$ or $A = 1$ regardless of their score (Freidlin et al., 2012). The analysis of the data is then focused on testing the treatment-score interaction, i.e., assessing whether the treatment benefit is different across score-defined subgroups.

In order to evaluate whether the score S is a strong and accurate predictor of the individual treatment benefit, randomizing all the patients to one of the two treatments might be therefore a more appropriate strategy than the previous two designs. As the restricted range of scores is known to decrease the efficiency of the test of association between S and $\Delta(S)$, one might consider adopting a score-stratified design and achieve precision in estimation of the parameters of interest.

The ultimate consideration choice of the study design needs to account for the potential to efficiently address the scientific questions of interest, as well as for the related

ethical aspects. In particular, it is our concern whether it would be ethical to assign patients with large negative scores to the experimental treatment, since the negative score suggests it might be harmful to them. One option to balance these opposing concerns is to assign the score-negative patients to the experimental treatment with probability lower than the usual 50%, or equivalently, to the control treatment with probability π_0 higher than 50%. In the following section, we will illustrate a relative efficiency of the three outlined designs on a simulation study and additionally examine the modified score-stratified approach with probabilities π_0 varying over a range of values.

4.4 Simulations: Power in Select Designs

Our last goal is to examine the considered study designs with respect to their efficiency to assess the quality of a marker-based score and the population impact of the corresponding treatment decision rule. In this section, we empirically evaluate power to detect significant superiority, correlation and accuracy of a pre-specified treatment benefit score under various scenarios and compare it across all three discussed validation study designs. The empirical power of individual tests for all considered designs is based on 1000 samples of $n = 500$ patients. The number of patients in each design is the total number of patients involved (enrolled). In the MGT strategy design, n is hence the number of patients randomized to MGT versus SoC arm, while in the enrichment design it is the number of patients screened, from whom only the score-positive patients were randomized to one of the treatment options. In the score-stratified design, n is the total number of patients randomized to one of the treatments, all of whom were also screened.

In case of the MGT strategy design, patients were randomized to either MGT strategy or SoC therapy ($A = 0$), each with probability 50%. For the patients randomized to the MGT arm, the treatment was determined based on the score, according to the score-guided treatment rule $d(S) = \mathrm{I}[S > 0]$. We assume that the biomarkers were eventually assessed on all randomized patients and so the score value is available for everybody. With this design, the experimental treatment $A = 1$ was only assigned to patients randomized to the MGT arm who were score-positive, and so the score-negative patients all received the control treatment $A = 0$. Data distributions under all three study designs are illustrated on an example shown in Figure 4.3. For the enrichment design, all patients had the score evaluated first and only the score-positive patients were randomized to one of the treatments $A = 0$ or $A = 1$. Consequently, the treatment and outcome data are not available for the score-negative patients. Lastly, the score-stratified design randomizes all patients to one of the treatment options with probability 50%, irrespective of the score, and the scores are then evaluated retrospectively.

We discussed earlier that assigning score-negative patients to the experimental treatment in the score-stratified design might raise some ethical concerns, since the negative score predicts worse outcome under $A = 1$ than under the standard care, or a potential harm. In order to find balance between the score-stratified and MGT design, we propose to consider assigning the score-negative patients to the experimental treatment with probability lower than 50%, and thus reducing the number of patients potentially experiencing negative benefit from the treatment $A = 1$. In our simulations, we evaluated the score-stratified designs with probabilities of "proper" treatment assignment in score-negative patients, $\pi_0 = \mathrm{P}(A = 0|\, S < 0)$, at additional levels 60%, 70%, 80% and

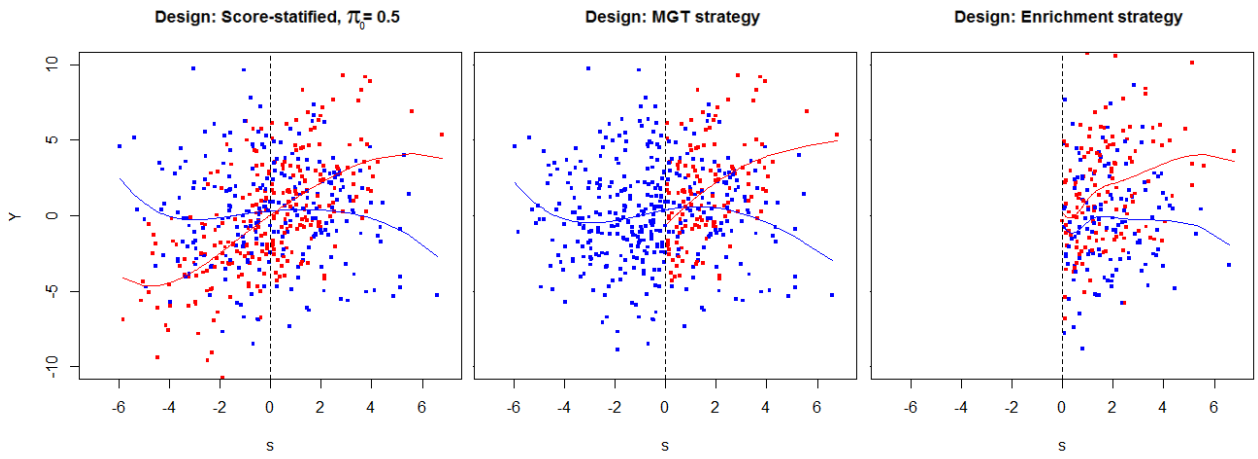

Figure 4.3: Data distribution under the three considered study designs. The blue dots represent patients on the control treatment $A = 0$, red dots represent patients on the experimental treatment $A = 1$, and the corresponding smooth curves are fitted to the data using model (4.1).

90%. Examples of the data distributions for $\pi_0 = 50\%$, 70% and 90% are illustrated in Figure 4.4.

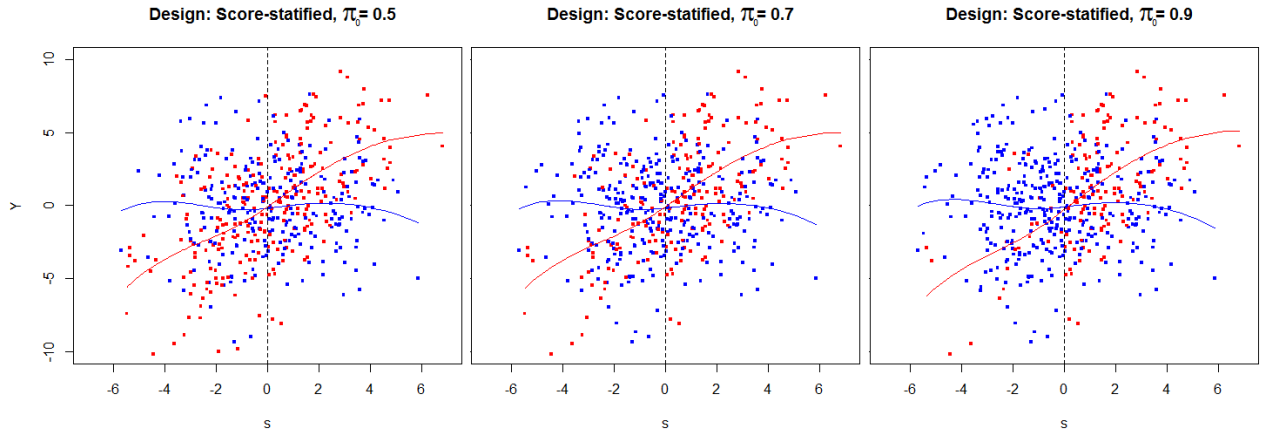

Figure 4.4: Examples of the data distributions under score-stratified design for probabilities of "proper" treatment assignment in score-negative patients $\pi = \mathrm{P}(A = 0|\, S < 0) = 50\%, 70\%$ and 90%. The blue dots represent patients on the control treatment $A = 0$, red dots represent patients on the experimental treatment $A = 1$, and the corresponding smooth curves are fitted to the data using model (4.1).

4.4.1 Data Generation

Throughout all the simulations we maintained the same marginal treatment effect between $A = 0$ and $A = 1$ equal to 0, but we considered 4 different scenarios for the true relationship between the score S and the expected treatment benefit $\Delta(S)$. The first scenario, consistent with the null, was such that $\Delta(S) = S$. The second scenario allowed the slope of the relationship to be attenuated by 50%, hence the corresponding data-generating model (4.1) had parameters $(\alpha_0, \alpha_1) = (0, 0.5)$. In the third scenario, the relationship between the score S and $\Delta(S)$ was shifted by $\alpha_0 = 1$ and in the last scenario, we imposed both shift and attenuation so that (α_0, α_1) was equal to $(-0.5, 0.75)$. The four scenarios are illustrated in Figure 4.5.

The scores were generated from a Normal distribution with $var(S) = 5$ and the mean determined by the data-generating model. Since all the four considered relationships between S and $\Delta(S)$ are symmetric around $\Delta(S) = 0$ (i.e., the treatment benefit is negative in the score-negative patients as much as it positive in the score-positive patients), the corresponding distributions of the scores were centered at $\Delta^{-1}(0)$, in order to preserve the 0 marginal effect. The additional variation of the outcome Y conditional on the score and treatment was set to $var[Y \,|\, S, A] = 10$ across all the scenarios, resulting in a data-generating model

$$Y = \mu_0(S) + A\Delta(S) + \epsilon\,,$$

where $\epsilon \sim N(0, 10)$, $\Delta(S)$ is determined by scenario (1)-(4), $S \sim N(\Delta^{-1}(0), 5)$ A depends on the study design, and we set $\mu_0(s) = \mathrm{E}[Y \,|\, S = s, A = 0] \equiv 0$. Three examples of sample-to-sample variation for scenarios (1): $(\alpha_0, \alpha_1) = (0, 1)$ and (4): $(\alpha_0, \alpha_1) =$

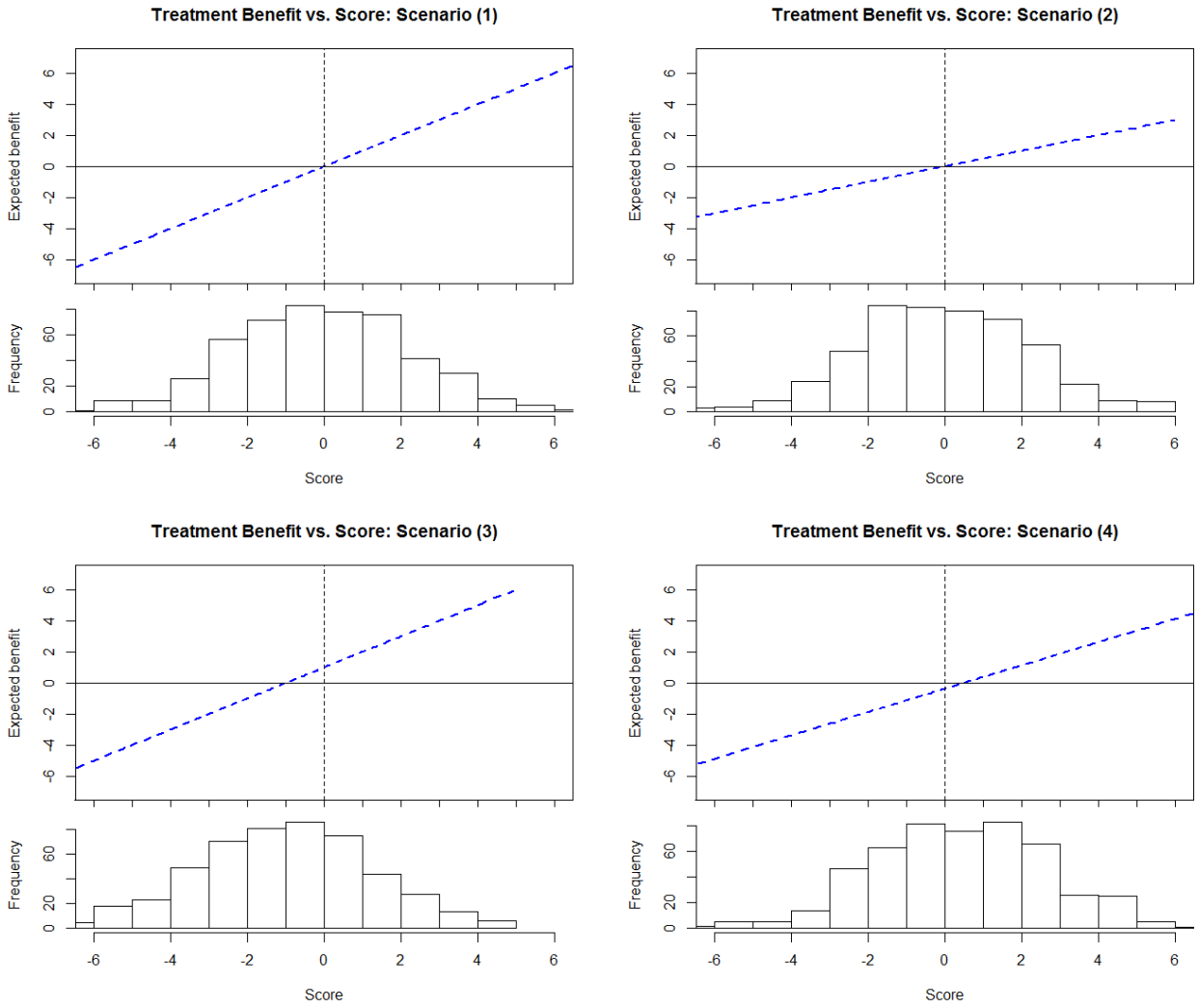

Figure 4.5: Four different scenarios for the true relationship between S and $\Delta(S)$; corresponding to parameter values (1): $(\alpha_0, \alpha_1) = (0, 1)$, (2): $(\alpha_0, \alpha_1) = (0, 0.5)$, (3): $(\alpha_0, \alpha_1) = (1, 1)$ and (4): $(\alpha_0, \alpha_1) = (-0.5, 0.75)$ in the model (4.1). The score distribution is Normal with variance $var(S) = 5$ and the mean $\mathrm{E}(S) = 0$ (scenarios 1,2), $\mathrm{E}(S) = -1$ (scen. 3), or $\mathrm{E}(S) = 0.5$ (scen. 4).

$(-0.5, 0.75)$ are shown in Figures 4.6 and 4.7, respectively.

4.4.2 Tests Specifications

For the MGT strategy design, the test of superiority of MGT over the standard care was based on the whole data set, comparing the average outcome among the patients

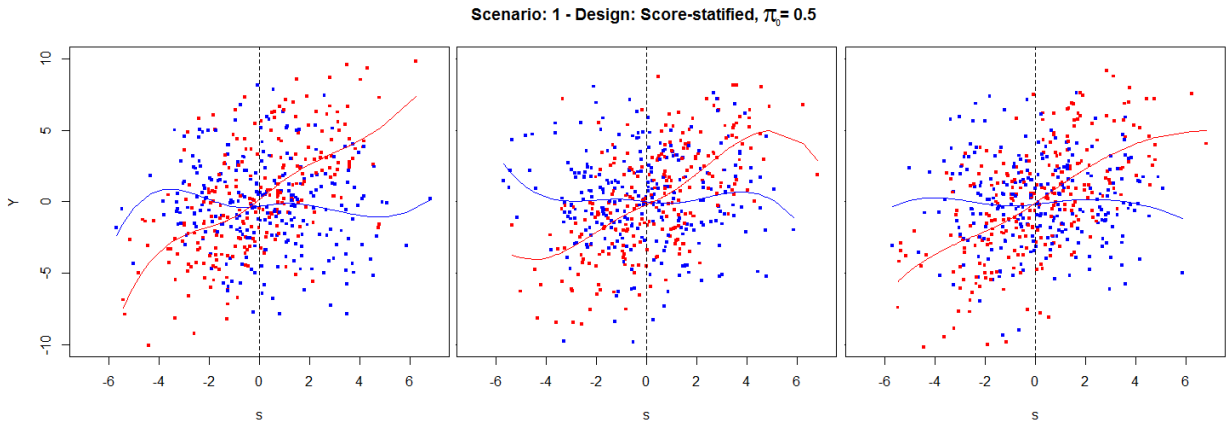

Figure 4.6: Three examples of data generated under the scenario (1): $(\alpha_0, \alpha_1) = (0, 1)$ and the score-stratified design, with the corresponding fitted smooth curves using model (4.1).

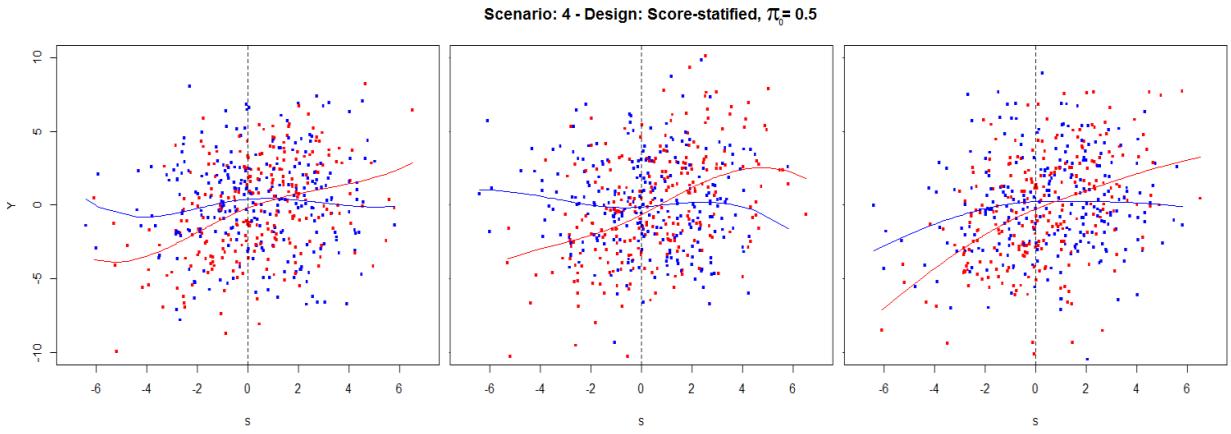

Figure 4.7: Three examples of data generated under the scenario (4): $(\alpha_0, \alpha_1) = (-0.5, 0.75)$ and the score-stratified design, with the corresponding fitted smooth curves using model (4.1).

randomized to the MGT arm and the patients randomized to the SoC arm by a standard one-sided *t-test* at 2.5% significance level. For both enrichment and score-stratified designs, the superiority of MGT was assessed indirectly by evaluating the efficacy of the experimental treatment $A = 1$ among the score-positive patients only, using again a

standard one-sided *t-test* at 2.5%. As already discussed earlier, this indirect assessment of MGT superiority is much more effective than the first approach as we will soon also see in the simulation results. For the score-stratified design, we additionally carried out a test of difference in the mean outcome between treatment 0 and 1 among the score-negative patients (by two-sided *t-test* at 5% level), which is not possible to do with the MGT strategy or enrichment design.

The test of correlation ($H_0 : \alpha_1 \leq 0$) and the tests of accuracy were all evaluated based on the coefficient estimates $(\widehat{\alpha}_0, \widehat{\alpha}_1)$ in the linear model (4.1), where $\mathrm{E}[Y \,|\, S = s, A = 0] = \mu_0(s)$ was approximated by a cubic spline with 2 knots. The test of threshold accuracy had a null hypothesis $H_0 : |\alpha_0| \geq c_0$ with the constant $c_0 = 1$ and was carried out using the approach described in section 4.2.3 at 5% significance level. The test of overall score accuracy was based on the composite null hypothesis $H_0 : (\alpha_0, \alpha_1) \in \Theta_0$, where the null set $\Theta_0 = \left\{(\alpha_0, \alpha_1) \in \mathbb{R}^2 : \left(\frac{\alpha_0}{c_0}\right)^2 + \left(\frac{\alpha_1 - 1}{c_1}\right)^2 \geq 1\right\}$ and the constants were set to $c_0 = 1$, $c_1 = 1/2$. If $\widehat{\boldsymbol{\alpha}} \in \Theta_0^c$, the corresponding likelihood ratio test statistic compared the likelihood at the MLE $\widehat{\boldsymbol{\alpha}}$ and the maximum likelihood over the border of Θ_0, which was calculated as also described in section 4.2.3.

4.4.3 Results

The results from the simulations are summarized in Table 4.1. The presented empirical power based on 1000 samples is shown in percentages and organized by study designs (in columns) and performed tests (in row blocks), each with 4 rows corresponding to the four different scenarios.

Looking at the first row block, we see that the power to detect significant superior-

Test	Sc.	MGT str.	Enr.D	SS: $\pi = 50\%$	$\pi = 60\%$	$\pi = 70\%$	$\pi = 80\%$	$\pi = 90\%$
Super-	1	81.0	97.9	98.1	98.1	98.1	98.1	98.1
iority	2	33.2	53.4	53.3	53.3	53.3	53.3	53.3
at	3	75.5	98.8	98.9	98.9	98.9	98.9	98.9
2.5%	4	64.3	87.1	87.1	87.1	87.1	87.1	87.1
Diff.	1	-	-	99.5	99.7	98.4	92.0	70.0
in S-	2	-	-	58.7	56.9	52.3	41.3	27.3
pts.	3	-	-	93.4	90.5	84.2	76.5	48.9
at 5%	4	-	-	91.5	90.4	85.2	76.9	48.9
Correl.	1	90.5	5.80	99.1	99.1	99.0	99.0	98.9
(H_0 :	2	36.7	5.10	96.2	95.5	94.6	91.2	76.5
$\alpha_1 \leq 0$)	3	69.7	4.20	98.9	99.0	99.0	98.9	99.1
at 2.5%	4	86.7	2.80	98.9	99.0	99.2	99.1	98.9
Accur.	1	19.1	3.20	93.3	93.1	90.0	86.0	70.6
thr. (H_0 :	2	14.2	8.30	94.6	92.0	91.6	84.9	71.5
$\lvert\alpha_0\rvert \geq 1$)	3	5.90	0.10	6.7	6.80	5.70	6.00	5.20
at 5%	4	12.1	3.00	69.3	67.5	67.1	63.8	51.1
Accur.	1	0.30	0.20	60.1	47.1	38.1	26.0	2.90
(H_0 :	2	0.10	0.10	3.90	2.20	0.40	0.30	0.10
$\boldsymbol{\alpha} \in \Theta_0$)	3	0.10	0.00	1.20	0.60	0.20	0.10	0.00
at 5%	4	0.20	0.10	10.1	10.2	7.20	3.10	0.10

Table 4.1: Empirical power of all performed tests based on 1000 samples. Each test was performed (if possible) under four different data-generating scenarios (1)-(4) and for all the considered study designs: MGT strategy design, Enrichment design, and Score-stratified design with probabilities π_0 of "proper" treatment assignment among the score-negative patients varying from 50-90%.

ity of MGT strategy over the standard care is impaired by inclusion of score-negative patients in the analysis, as we expected. For example, under scenario (1), the test of non-superiority was rejected 81% of the time with the MGT strategy design, while it was rejected about 98% of the time under the enrichment and score-stratified designs, both of which focus on the difference in the mean outcome among the score-positive patients only. This pattern is consistent across all 4 scenarios, even if the power level is lower under the scenarios with weaker population impact of the MGT strategy.

The second row block shows the power to detect difference in the mean outcome

between treatment 0 and 1 among the score-negative patients. A test of such contrast can not be evaluated with the MGT strategy or enrichment designs, as the score-negative patients are either always assigned to the $A = 0$ (MGT strategy design) or left out of the study (Enrichment design). For the score-strategy design, the power to detect a significant difference is, not surprisingly, gradually decreasing as the proportion of patients randomized to each treatment is getting further from 50%.

While the enrichment design is very efficient at detecting the superiority of score-based MGT over the standard therapy, it is very inefficient in assessing the clinical utility of the score. As the range of available scores is restricted to positive values, the ability to detect a trend in score is greatly diminished. For both correlation (third row block) and accuracy tests (row block 4 and 5), the enrichment design results in power that is unacceptably small, and makes these tests practically useless. The power is slightly higher, but still very low in case of accuracy tests, under the MGT strategy design, as the inclusion of score-negative patients improves precision in estimating one of the smooth curves, but not the other.

The most effective from all three considered designs is the score-stratified design, which allows for estimation of treatment effect across the whole range of scores and consequently leads to the most precise estimates of the parameters of interest. In our simulations, its power to detect correlation between $\Delta(S)$ and S is generally above 95% and stays relatively high even as the proportion π_0 of score-negative patients treated with $A = 0$ increases toward 0.90. Similar pattern holds for the test of threshold accuracy, even though the deviation from the optimally calibrated score ($\boldsymbol{\alpha} = (0, 1)$) in scenarios (3) and (4) has larger impact on the ability to reject the null, particularly in scenario (3).

Not surprisingly, we notice that across all performed tests, the power is always lowest under the scenarios which put the tested (or most influential) parameter closest to the border, such as scenario (2): $\boldsymbol{\alpha} = (0, 0.5)$ for the test of correlation ($H_0 : \alpha_1 \leq 0$) or scenario (3): $\boldsymbol{\alpha} = (1, 1)$ for the test of threshold accuracy ($H_0 : |\alpha_0| \geq c_0$), where the rejection frequency approaches type I error.

Since the test of overall accuracy (the last row block) has stricter null hypothesis than the previous test of threshold accuracy, the power levels are all lower in the former. Under the optimal scenario (1): $\boldsymbol{\alpha} = (0, 1)$, the power to detect score accuracy with the score-stratified design and $\pi_0 = 0.5$ seem to be around 60%, but rapidly decreases as the proportion of "properly" treated patients π_0 approaches 1. Under the scenarios (2) and (3), where $\boldsymbol{\alpha}$ is directly on the null set border, rejection frequency is again reduced to type I error. In the last scenario (4), with $\boldsymbol{\alpha}$ in between the optimal and null value, power to reject is slightly higher than under the scenarios (2) and (3), but still very low. Finally, the MGT strategy and enrichment design seem to have virtually no power to reject the null hypothesis of the overall score "inaccuracy" as defined by Θ_0.

4.5 Discussion

In this chapter, we detailed and compared characteristics of selected validation study designs. Our main criteria for a favorable study design were its ability to efficiently evaluate how a pre-specified score performs as a basis for treatment selection on the population level and as a predictor of response on the individual level. We hence considered validation of three aspects of the score and corresponding treatment decision rule: superiority; correlation; and accuracy. Each of these properties can be assessed using

one the outlined statistical tests.

Besides the standard test of superiority of marker-guided treatment strategy over the standard care treatment, we considered a model-based test of correlation between the score and the corresponding treatment benefit, and proposed two tests of accuracy of the score as a predictor of the treatment benefit and as a patient classifier. The commonly adopted study designs which we discussed in this chapter were marker-guided treatment (MGT) strategy, enrichment strategy and a score-stratified design. Additionally, we proposed an altered version of the score-stratified design which randomizes score-negative patients to the experimental treatment with probability lower than the standard 50%. The goal of such design is to address possible concerns related to ethical conduct of the clinical trial in the presence of evidence indicating potential harm of the score-negative patients by the experimental treatment.

The discussion of advantages and disadvantages of different approaches is informed by a table of simulation-based power levels evaluated empirically across alternative study designs and under four different data-generating scenarios in order to aid the understanding of relative efficiency of the proposed statistical tests in various settings. As expected, both enrichment and score-stratified designs are more efficient at detecting the MGT superiority through an indirect assessment of treatment effect in the subgroup of score-positive patients. However, in order to evaluate the clinical utility of a validated score, the treatment effect needs to be assessed and compared across the whole range of score values. Study designs that only randomize score-positive patients do not allow for assessment of such contrasts and are confined to the evaluation of efficacy in the subgroup of patients indicated by the score. In the absence of randomization among

the score-negative patients, the ability to assess the clinical utility of the corresponding marker-guided treatment rule is hence largely limited. From the three considered designs, only the score-stratified design allows for evaluation of the treatment effect among the score-negative patients, while the other two designs tend to be very inefficient in assessment of the clinical utility of the score, which holds in particular for the enrichment design.

As the score-stratified design is equally efficient at detecting the MGT superiority as the enrichment design and far more effective in assessing the clinical utility of a prespecified score, it might seem as a plausible choice for the follow-up study. However, if the previous evidence suggests that the experimental treatment might be harmful to the score-negative patients, clinical trials randomizing such patients would likely raise some ethical concerns. Selection of an optimal design for the validation study therefore requires both thoughtful consideration of the scientific questions and careful balance between efficiency and additional aspects such as timeline of the trial, total cost and ethics.

Chapter 5

FUTURE WORK

In the last chapter, we outline remaining open questions in the domain of this work and discuss some of the possible directions for the future work. We have introduced methods for development of marker-guided treatment rules based on a combination of variable selection methods and careful evaluation of the candidate rules via cross-validated estimation of their population impact. For a large number of available markers, we proposed to select candidate marker panels by L_1 penalized least square method, which leads to a nested set of marker panels and hence allows easy navigation through a high-dimensional parameter space. However, using the regular regression function and standard marker selection techniques, such as LASSO, does not reflect ultimate goal of maximizing the population performance. An alternative approach that directly relates the estimated function in the nomination phase and the target parameter from the evaluations phase (e.g., mean population outcome) is targeted learning, in which the targeted maximum-likelihood estimator is aimed to find an optimal bias-variance tradeoff for the parameter of interest (van der Laan and Rose, 2011). Incorporating this approach into the nomination phase might improve the quality of candidate decision rules and result thus in a developed treatment regimen with higher population impact. For example, with aggressive experimental treatments – that result in a fair number of subjects dying early on, but then those that survive live considerably longer – a working

proportional-hazards model might not be very appropriate and lead to poor candidate decision rules. In such cases, it would be important to determine clinically relevant measure with respect to which we would wish to optimize. If the aggressive treatment is considered "better" than standard therapy, we might consider a long-term (e.g. 10-year) survival as the target parameter and based the nomination phase on different approach than Cox regression.

Based on our simulations for continuous outcome, the smoothing spline estimator appears to be more efficient and also more stable across panels than the standard non-parametric estimator. We however have not performed simulations that would show whether this relative efficiency diminishes with increasing sample size. In case of survival outcomes, the simulations suggest that the efficiency of all three proposed estimators is very similar, yet the bias varies across estimators. Therefore, our methods for both continuous and survival outcomes could be enhanced by establishing the asymptotic properties of the proposed estimators for the target parameters.

Another possible extension of our methods for survival outcomes could concern settings in which two covariate-specific survival curves for the two treatments potentially cross. It would be then needed to establish optimality criteria and nomination process that are reflective of the target summary measure of the survival curve. In the nomination phase, we adopted the "working" Cox proportional-hazards model, which allows for easy estimation of the "average" covariate-specific treatment effect and nomination of candidate decision rules. The problem with identifying an optimal rule arises when survival curves cross and might thus result in inconsistent ordering with respect to different summary measures.

Our simulations for data-generating model that violates proportional-hazards assumption surprisingly showed less bias in case of the Cox-model based estimators. Similarly as in the setting with proportional hazards, both non-parametric estimators tend to be more conservative than the semi-parametric one. However, they seem to remain underestimating the survival even for larger panel sizes under this scenario, while the bias of the estimator based on the Cox model decreases for larger panels. In order to better understand performance of the three estimators under various scenarios, more data-generating models with both proportional and non-proportional hazards can be investigated in the future.

In Chapter 4, we detailed and compared design characteristics of selected validation study designs for continuous outcomes. Similar comparisons can be made between study designs for survival outcomes, with focus on their ability to efficiently evaluate how a pre-specified score performs as a basis for treatment selection on the population level and as a predictor of individual survival.

As randomized clinical trials are often very expensive and the large databases (such as Electronic Medical Records) are becoming more available, there is also a large potential for broadening of the proposed estimation and evaluation methods to studies with observational data.

BIBLIOGRAPHY

Aaronson, K. D., Schwartz, J. S., Chen, T., Wong, K., Goin, J. E., and Mancini, D. M. (1997), "Development and prospective validation of a clinical index to predict survival in ambulatory patients referred for cardiac transplant evaluation." *Circulation*, 95.

Barlogie, B., Kyle, R. A., Anderson, K. C., Greipp, P. R., Lazarus, H. M., Hurd, D. D., McCoy, J., Moore Jr, D. F., Dakhil, S. R., Lanier, K. S., Chapman, R. A., Cromer, J. N., Salmon, S. E., Durie, B., and Crowley, J. C. (2006), "Standard Chemotherapy Compared With High-Dose Chemoradiotherapy for Multiple Myeloma: Final Results of Phase III US Intergroup Trial S9321," *Journal of Clinical Oncology*, 24, 929–936.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009), "Improving efficiency and robusstness of the doubly robust estimator for a population mean with incomplete data." *Biometrika*, 96.

Clarke, R., Daly, L., Robinson, K., Naughten, E., Cahalane, S., Fowler, B., and Graham, I. (1991), "Hyperhomocysteinemia: An independent risk factor for vascular disease." *The New England Journal of Medicine*, 324.

Daly, A. K. (2010), "Genome-wide association studies in pharmacogenomics." *Nature Review Genetics*, 11.

de Mendonça, A., Vincent, J.-L., Suter, P. M., Moreno, R., Dearden, N. M., Antonelli,

M., Takala, J., Sprung, C., and Cantraine, F. (2000), "Acute renal failure in the ICU: risk factors and outcome evaluated by the SOFA score." *Intensive Care Medicine*, 26.

Freidlin, B., Jiang, W., and Simon, R. (2010), "The cross-validated adaptive signature design." *Clinical Cancer Research*, 16.

Freidlin, B., McShane, L. M., and Korn, E. L. (2012), "Randomized Clinical Trials With Biomarkers: Design Issues." *Clinical Trials*, 9.

Freidlin, B. and Simon, R. (2005), "Adaptive Signature Design: An Adaptive Clinical Trial Design for Generating and Prospectively Testing A Gene Expression Signature for Sensitive Patients." *Clinical Cancer Research*, 11.

Friedly, J. L., Bresnahan, B. W., Comstock, B., Turner, J. A., Deyo, R. A., Sullivan, S. D., Heagerty, P., Bauer, Z., Nedeljkovic, S. S., Avins, A. L., et al. (2012), "Study Protocol-Lumbar Epidural Steroid Injections for Spinal Stenosis (LESS): a double-blind randomized controlled trial of epidural steroid injections for lumbar spinal stenosis among older adults," *BMC musculoskeletal disorders*, 13, 48.

Gui, J. and Li, H. (2005), "Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data." *Bioinformatics*, 21.

Gunter, L., Zhu, J., and Murphy, S. (2011), "Variable selection for qualitative interactions." *Statistical Methodology*, 4594.

Huang, Y., Gilbert, P. B., and Janes, H. (2012), "Assessing treatment-selection markers using a potential outcomes framework." *Biometrics*, 68(3).

Inci, M. F., Ozkan, F., Ark, B., E., V. U., Ege, M. R., Sincer, I., and Zorlu, A. (2013), "Sonographic Evaluation for Predicting the Presence and Severity of Coronary Artery Disease." *Ultrasound Q.*, (*in print*).

Jacobs Jr, R. D., Kroenke, C., Crow, R., Deshpande, M., Gu, D. F., Gatewood, L., and Blackburn, H. (1999), "PREDICT: A Simple Risk Score for Clinical Severity and Long-Term Prognosis After Hospitalization for Acute Myocardial Infarction or Unstable Angina." *Circulation*, 100.

Janes, H., Pepe, S. M., Bossuyt, P. M., and Barlow, W. E. (2011), "Measuring the performance of markers for guiding treatment decisions." *Annals of Internal Medicine*, 154.

Jiang, W., Freidlin, B., and Simon, R. (2007), "Biomarker-adaptive threshold Design: A Procedure for Evaluating Treatment with possible biomarker-defined subset effect." *Journal of National Cancer Institute*, 99.

Kalbfleisch, J. D. and Prentice, R. L. (1972), "Marginal likelihoods based on Cox's regression and life model." *Biometrika*, 60.

Kamath, P., Wiesner, R., Malinchoc, M., Kremers, W., Therneau, T. M., Kosberg, C. L., D'Amico, G., Dickson, E. R., and Kim, W. (2001), "A model to predict survival in patients with end-stage liver disease." *Hepatology*, 33(2).

Kaplan, E. L. and Meier, P. (1958), "Non-parametric estimation from incomplete observations." *Journal of the American Statistical Association*, 53.

Kooperberg, C., LeBlanc, M., and Obenchain, V. (2010), "Risk prediction using genome-wide association studies." *Genetic Epidemiology*, 34.

Lai, T. L., Lavori, P. W., Shih, M.-C. I., and Sikic, B. I. (2012), "Clinical trial design for testing biomarker-based personalized therapies." *Clinical Trials*, 9.

Liévre, A., Bachet, J.-B., Le Corre, D., Boige, V., Landi, B., Emile, J.-F., Côté, J.-F., Tomasic, G., Penna, C., Ducreux, M., Rougier, P., Penault-Llorca, F., and Laurent-Puig, P. (2006), "KRAS Mutation Status Is Predictive of Response to Cetuximab Therapy in Colorectal Cancer," *Cancer Research*, 66, 3992–3995.

Lumley, T. and Heagerty, P. J. (2000), "Graphical Exploratory Analysis of Survival Data." *Journal of Computational and Graphical Statistic*, 9.

Matsui, S., Simon, R., and Qu, P. (2012), "Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine." *Clinical Cancer Research*, 18.

Murphy, S. A. and Van Der Vaart, A. W. (1997), "Semiparametric likelihood ratio inference." *The Annals of Statistics*, 25.

Peters, U., Jiao, S., Schumacher, F. R., Hutter, C. M., Aragaki, A. K., Baron, J. A., Berndt, S. I., Bézieau, S., Brenner, H., Butterbach, K., Caan, B. J., Campbell, P. T., Carlson, C. S., Casey, G., Chan, A. T., Chang-Claude, J., Chanock, S. J., Chen, L. S., Coetzee, G. A., Coetzee, S. G., Conti, D. V., Curtis, K. R., Duggan, D., Edwards, T., Fuchs, C. S., Gallinger, S., Giovannucci, E. L., Gogarten, S. M., Gruber, S. B., Haile, R. W., Harrison, T. A., Hayes, R. B., Henderson, B. E., Hoffmeister, M., Hopper,

J. L., Hudson, T. J., Hunter, D. J., Jackson, R. D., Jee, S. H., Jenkins, M. A., Jia, W. H., Kolonel, L. N., Kooperberg, C., Küry, S., Lacroix, A. Z., Laurie, C. C., Laurie, C. A., Le Marchand, L., Lemire, M., Levine, D., Lindor, N. M., Liu, Y., Ma, J., Makar, K. W., Matsuo, K., Newcomb, P. A., Potter, J. D., Prentice, R. L., Qu, C., Rohan, T., Rosse, S. A., Schoen, R. E., Seminara, D., Shrubsole, M., Shu, X. O., Slattery, M. L., Taverna, D., Thibodeau, S. N., Ulrich, C. M., White, E., Xiang, Y., Zanke, B. W., Zeng, Y. X., Zhang, B., Zheng, W., Hsu, L., Registry, C. C. F., the Genetics, and of Colorectal Cancer Consortium., E. (2013), "Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis." *Gastroenterology*, 144(4).

Qian, M. and Murphy, S. A. (2011), "Performance guarantees for individualized treatment rules." *Annals of Statistics*, 39.

Rafiq, S., Tapper, W., Collins, A., Khan, S., Politopoulos, I., Gerty, S., Blomqvist, C., Couch, F. J., Nevanlinna, H., Liu, J., and Eccles, D. (2013), "Identification of Inherited Genetic Variations Influencing Prognosis in Early-Onset Breast Cancer." *Cancer Research*, 73(6).

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American Statistical Association*, 89.

Simon, R. (2008), "Development and validation of biomarker classifiers for treatment selection." *Journal of Statistical Planning and Inference*, 138.

Simon, R. and Maitournam, A. (2004), "Evaluating the Efficiency of Targeted Designs for Randomized Clinical Trials." *Clinical Cancer Research*, 10, 6759–6763.

Sitlani, C. M. and Heagerty, P. J. (2014), "Using longitudinal structrural mixed models for characterizing accuracy of markers used to select treatment." *Statistics in Medicine*.

Sitlani, C. M., Heagerty, P. J., Blood, E. A., and Tosteson, T. D. (2012), "Longitudinal structural mixed models for the analysis of surgical trials with noncompliance." *Statistics in Medicine*.

Song, X. and Pepe, M. S. (2004), "Evaluating Markers for Selecting a Patient's Treatment." *Biometrics*, 60.

Spence, J. D., Howard, V. J., Chambless, L. E., Malinow, M. R., Pettigrew, L. C., Stampfer, M., and Toole, J. F. (2001), "Vitamin Intervention for Stroke Prevention (VISP) Trial: Rationale and Design." *Neuroepidemiology*, 20.

Tibsirani, R. (1996), "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society*, 58.

— (1997), "The LASSO method for variable selection in the Cox model." *Statistics in Medicine*, 16.

Urabe, Y., Ochi, H., Kato, N., Kumar, V., Takahashi, A., Muroyama, R., Hosono, N., Otsuka, M., Tateishi, R., Lo, P., Tanikawa, C., Omata, M., Koike, K., Miki, D., Abe, H., Kamatani, N., Toyota, J., Kumada, H., Kubo, M., Chayama, K., Nakamura, Y., and Matsuda, K. (1997), "A genome-wide association study of HCV-induced liver

cirrhosis in the Japanese population identifies novel susceptibility loci at the MHC region." *Circulation*, 95.

van der Laan, M. J. and Rose, S. (2011), *Targeted Learning*, New York, NY: Springer Series in Statistics.

Wilks, S. S. (1938), "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *The Annals of Mathematical Statistics*, 9.

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