# Abstract

FAN WU. Adaptive Projection Confidence Sets and Data Analyses in Personalized Medicine. (Under the direction of Eric B. Laber.)

This dissertation contains three projects. In chapter 2, we proposed a new method for constructing asymptotically valid confidence interval for a group of non-regular estimands. We named it as Adaptive Projection Interval (API). This method works well under both regular and non-regular estimand situations. In particular, we investigate our method under different estimands including: the marginal mean outcomes in Dynamic Treatment Regimes (DTR), the stage 1 coefficients in two-stage $Q$-learning, and the measure of marker performance. In Chapter 3, we explored the efficacy of the possible medications for Bipolar Disorder patients using data from STEP-BD (Systematically Treatment Enhancement Program for Bipolar Disorder) study. Through using the $Q$-learning, and grouped $Q$-learning methods, we tried to construct models to estimate the optimal treatment regimes for bipolar disorder patients from both randomized pathway, and standard pathway in STEP-BD. One interesting finding is that patients with (hypo)manic episode in their early lives are recommended only receiving mood-stabilizers. In Chapter 4, we proposed a sequential dosage assignment method that will give guidance of dosage assignment to patients at each stage based on patients' up-to-date information. This sequential dosage assignment method is motivated by the data collected from phase III clinical trails by Purdue Pharmaceutical company.

# Adaptive Projection Confidence Sets and Data Analyses in Personalized Medicine

by
Fan Wu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2016

APPROVED BY:

| | |
|---|---|
| Leonard Stefanski | Dennis Boos |
| Michael Kosorok | Eric Laber<br>Chair of Advisory Committee |

## Dedication

To my parents, my husband, and my lovely son.

# Biography

Fan Wu has specialized in biostatistics during her extensive career. She is a doctoral student in the Department of Statistics at North Carolina State University, supervised by Dr. Eric Laber. Her dissertation title is: Adaptive Projection Confidence Interval for Non-regular Estimands and Some Data Analysis in Personalized Medicine. She was awarded an M.S. in biostatistics in 2012 from North Carolina State University. In 2006, she was awarded a B.S. in statistics from Sun Yat-Sen University, China. She was an intern in the Center for Statistics in Drug Development (CSDD) at Quintiles, a clinical research organization company in North Carolina research triangle park (RTP), from 2013-2014, and a teaching assistant in the Department of Statistics at North Carolina State University from 2010-2013. In 2013, she was nominated as an outstanding teaching assistant at North Carolina State University. Her research interests include dynamic treatment regimes, non-regular asymptotics, machine learning, and statistical computing. She has written a book chapter on bipolar disorder introducing Sequential Multiple Assignment Randomized Trails (SMART) design.

# Acknowledgements

First, I would like to thank my thesis advisor Dr. Eric Laber for his kind guidance, inspiration, and continued support during my ph.D. study. When I met problems, he always patiently gave me suggestions, and encouraged me to continue my research projects. I greatly appreciate the time and effort that he dedicated to helping me complete the research projects smoothly.

I would like to thank my committee members Dr. Leonard Stefanski, Dr. Michael Kosorok, and Dr. Dennis Boos constructive comments and suggestions for my research. I especially thank Dr. Stefanski for giving me a lot of suggestions during our lab meetings. His suggestions are really helpful.

I am grateful to my intern supervisor Dr. Ilya Lipchovich, and the collaborated professor Dr. Emanuel Severus. Thanks for their help and effort in the STEP-BD project. Without their guidance, I cannot finish the STEP-BD data analysis in my thesis.

I also thank the lab members from Dr. Laber's group. They provide me a lot of suggestions in my research work during the weekly group meeting. I also learned a lot of knowledge from different research area through their presentations and discussion during the group meetings.

Lastly, I would like to thank my family. My parents, and my husband Teng Zhang always give me support and encouragement when I met problems and felt stressful. I thank my lovely son Maxwell. He makes me want to be a better person.

# Table of Contents

# LIST OF TABLES

# LIST OF FIGURES

# Chapter 1

# Introduction

## 1.1 A Group of Non-smooth Estimands

Suppose we observe a dataset: $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^p$ is predictive variable vector, and $Y_i \in \mathbb{R}$ is the response variable. Suppose the parameter of interest is $\theta^*(\beta^*) = \mathrm{E}\{f(X, \beta^*)\}$, where $f(x, \beta^*)$ is non-smooth such that $(x, \beta^*) \in \mathcal{Q}_{(X,\beta^*)}$. Sometimes $f(x, \beta^*)$ can be rewritten as $f(x, \beta^*) = s(x_1, \beta_1^*) \cdot g(x_2, \beta_2^*)$, where $x_1, x_2, \beta_1^*, \beta_2^*$ are subvectors of $x, \beta^*$, $s(\cdot)$ is continuous and differentiable, and $g(\cdot)$ is non-smooth when $(x, \beta^*) \in \mathcal{Q}_{(X,\beta^*)}$. For example, let $\theta^*(\beta^*) = \mathrm{E}\{[X^T\beta^*]_+\}$, then $f(x, \beta^*) = [x^T\beta^*]_+$, $s(x_1, \beta_1^*) = x^T\beta^*$, and $g(x_2, \beta_2^*) = 1_{x^T\beta^* \geq 0}$. Here $1_{(\cdot)}$ is indicator function. Many quantities of interests in history can be written in this form: the marginal mean outcome in Dynamic Treatment Regimes, regression coefficients in two-stage $Q$-learning, coefficients in two-stage outcome weighted learning, and measures of biomarker performance. The non-smooth functions in $\theta^*(\beta^*)$ can lead to non-regular asymptotics, and thus invalidate standard procedures like the bootstrap and the delta method.

The aim of this project is to construct asymptotically valid confidence intervals for this class of estimands. Berger and Boos (1994) proposed a method named projection confidence intervals (PCI) that could provide asymptotically valid CI for this group of estimands. But this method is conservative. For example, the estimated coverage rates from simulations always equal to 1.0. Based on their method, we proposed a method named adaptive projection confidence interval (API). Through this method, the PCI will only be used for points that are near the region $\mathcal{Q}_{(X,\beta^*)}$. This will reduce the conservatism from PCI.

## 1.2 STEP-BD Study

STEP-BD (Systematic Treatment Enhancement Program for Bipolar Disorder) is a long-term study of bipolar disorder funded by the National Institute of Mental Health (NIMH). Its aim was "to generate externally valid answers to treatment effectiveness questions related to bipolar disorder" Sachs et al. (2003). Patients of age older than 15 years fulfilling DSM-IV criteria for any subtype of bipolar disorders could enter the study registry. In total, 4,360 patients from 22 sites in United States enrolled. The study lasted for 7 years (2001-2007).

Antidepressant is one kind of medication that could control the feeling of depression for patients with unipolar or bipolar disorder. But meanwhile, it may introduce some side effects for the patients. For example, it may increase of the probability of suicide for patients with bipolar disorder. Recently, some psychiatrists suggested that antidepressant may be effective for a subgroup of patients. Another common medication that is usually used for bipolar disorder patients is mood-stabilizer.

Nowadays, personalized medicine is a hot topic, and dynamic treatment regime (DTR) is one framework for personalized medicine. A DTR is a sequence of decision rules, which maps patients up-to-date information to the feasible treatments. An optimal DTR is the one that optimized the primary mean outcome across the population of interest. Historically, there are a lot of different methods that could be used to estimate the optimal DTR. These include: $A$-learning (Murphy 2003b), $Q$-learning (Watkins and Dayan 1992; Schulte et al. 2012; Laber et al. 2014c), outcome weighted learning (Zhao et al. 2012), and augmented inverse probability weighted estimator (Zhang et al. 2012a,b). $Q$-learning is one popular method that is easily understand and plug-in.

Our goal for this project is explore the efficacy of antidepressants and mood-stabilizers using $Q$-learning to estimate the optimal treatment regimes for the bipolar disorder patients through the data extracted from STEP-BD study. We first used the data extracted from the randomized sub study of STEP-BD, which named Randomized Acute Depression Pathway (RAD). The method we used is $Q$-learning with linear model estimation for each stage $Q$-functions. We then used a data set named SAD from the observational sub study, Standardized Care Pathway (SCP), of STEP-BD. The method we used is grouped $Q$-learning.

## 1.3 Constrained Sequential Dosage Assignments

When managing a chronic illness, a clinical scientist must decide how adapt treatment both in response to and anticipation of changes in each individual patient's health status. A treatment regime formalizes this decision process as a sequence of functions, one per intervention period, that map current patient information to a recommended treatment (Murphy, 2003a; Robins, 2004; Chakraborty and Moodie, 2013). An optimal treatment regime is defined as maximizing some functional of the outcome distribution, e.g., mean symptom reduction or the probability of surviving disease-free past some time horizon, if the regime were used to select treatments for individuals in a population of interest. We consider the problem of constructing an interpretable treatment regime that adjusts treatment dosage over a potentially large number of intervention periods with the goal of maximizing a cumulative measure while controlling the risk of an adverse event. This work is motivated by a sequence of clinical trials on chronic pain in which the clinical goal was to maximize pain reduction while reducing the risk of constipation.

We use non-parametric Q-learning to form a joint estimator of the marginal mean efficacy and a general measure of risk of an adverse event for any regime in a pre-specified class. An estimator of the optimal regime is obtained by choosing the regime that maximizes expected efficacy among those that satisfy a constraint on risk. The class of regimes is chosen to ensure that the class of regimes is interepretable and easily disseminated among domain experts. We show that non-parametric Q-learning produces estimators that are more stable than (augmented) inverse probability weighting when there are a large number of time points. Furthermore, non-parametric Q-learning can be used to do diagnose severe approximation error in the class of pre-specified regimes.

## 1.4 Outline

The outline of the thesis is as follows:

In Chapter 2, we proposed a method named Adaptive Projection Confidence Interval (API), which can provide asymptotically valid CI for this group of estimands. Section 2.1 will give a detailed introduction. Section 2.2 reviews existing methods for constructing a asymptotically valid CI for non-smooth estimands. The construction and tuning of the API and its theoretical properties are illustrated in Section 2.3. Section 2.4 presents a numerical study of the API. Section 2.5 concludes the chapter.

In Chapter 3, we focused on data analysis of STEP-BD. Section 3.1 introduce the STEP-BD study. In Section 3.2, we analysis the data set from RAD (Randomized Acute Depression Pathway) using $Q$-learning. In Section 3.3, we analysis the data set named SAD (Standardized Acute Depression) using the method of grouped $Q$-learning.

In Chapter 4, we proposed a method giving guidance for dosage assignment with some constrained conditions. Section 4.1 provide motivations for this project. In Section 4.2, we discuss a variant of non-parametric Q-learning that is amenable to constrained policy-search algorithms. In Section 4.3, we discuss computation of the proposed estimator. In Section 4.4, we apply the proposed method to analyze data from a clinical trial on chronic pain. Future work is discussed in Section 4.5.

# Chapter 2

# Adaptive Projection Confidence Interval for Non-smooth Estimands

## 2.1 Introduction

Consider a regression problem in which we observe a dataset, $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^p$ denotes predictive variables, and $Y_i \in \mathbb{R}$ is the response variable. Suppose we are interested in an estimand, say $\theta^*(\beta^*) = \mathrm{E}\{f(X, \beta^*)\}$, where $\beta^*$ is a nuisance parameter, $f(\cdot)$ is a known function, and $X$ is a random vector with the same distribution as predictor $X_i$. Suppose the function $f(x, \beta^*)$ is non-smooth in the region denoted as $\mathcal{Q}_{(X,\beta^*)}$, where we assume $\mathcal{Q}_{(X,\beta^*)}$ is a closed set. The term non-smooth here means that $f(x, \beta^*)$ is not differentiable at any point $(x, \beta^*) \in \mathcal{Q}_{(X,\beta^*)}$. We assume $f(x, \beta^*)$ is continuous and differentiable for all $(x, \beta^*) \notin \mathcal{Q}_{(X,\beta^*)}$. For example, if our target estimand is $\theta^*(\beta^*) = \mathrm{E}\{[X^T\beta^*]_+\}$, where $f(x, \beta^*) = [x^T\beta^*]_+$. The non-smooth region $\mathcal{Q}_{(X,\beta^*)}$ for this example is $\{(x, \beta^*) : x^T\beta^* = 0\}$. In many cases, the function $f(x, \beta^*)$ can be rewritten as the multiplication of two functions: a smooth function $s(x_1, \beta_1^*)$ and a non-smooth function $g(x_2, \beta_2^*)$, where $g(\cdot)$ is an indicator function, and $\beta_1^*$ and $\beta_2^*$ are subsets of $\beta^*$ that can have same elements. For the example above, the corresponding $s(\cdot)$ and $g(\cdot)$ will be $s(x_1, \beta_1^*) = x^T\beta^*$, and $g(x_2, \beta_2^*) = 1_{x^T\beta^* \geq 0}$, where $x_1 = x_2 = x$ and $\beta_1^* = \beta_2^* = \beta^*$. Estimands that can be written in this form include: (i) the marginal mean outcome in Dynamic Treatment Regimes (Robins, 2004); (ii) the stage 1 coefficients in two-stage $Q$-learning (Murphy, 2005a; Schulte et al., 2012); (iii) the coefficients in outcome weighted learning (Zhao et al., 2012); and (iv) the measure of marker performance (Janes et al.,

2014). In this chapter, we focus on the construction of a confidence interval (CI) for the non-smooth estimands of the form $\theta^*(\beta^*) \triangleq \mathrm{E}\{s(X_1, \beta_1^*)g(X_2, \beta_2^*)\}$.

For estimand $\theta^*(\beta^*)$, let $\hat{\beta}_n$ denote an estimator based on the dataset $\mathcal{D}$. A natural estimator of $\theta^*(\beta^*)$ is $\hat{\theta}_n(\hat{\beta}_n) = \mathbb{E}_n\{f(X, \hat{\beta}_n)\} = \mathbb{E}_n\{s(X_1, \hat{\beta}_{1n})g(X_2, \hat{\beta}_{2n})\}$, where $\mathbb{E}_n$ represents empirical expectation, i.e., $\mathbb{E}_n h(X) = \frac{1}{n}\sum_{i=1}^n h(X_i)$. We assume that $\sqrt{n}(\hat{\beta}_n - \beta^*)$ is asymptotically normally distributed with mean 0 and positive definite variance matrix denoted as $\Omega_{\beta^*}$. If $\mathrm{P}\{(x, \beta^*) \in \mathcal{Q}_{(X,\beta^*)}\} = 0$, then the asymptotic distribution of $\sqrt{n}\{\hat{\theta}_n(\hat{\beta}_n) - \theta^*(\beta^*)\}$ can be derived using the delta method. Hence, an asymptotically valid CI can be constructed using this limiting distribution. However, if $\mathrm{P}\{(x, \beta^*) \in \mathcal{Q}_{(X,\beta^*)}\} > 0$, then the limiting distribution of $\sqrt{n}\{\hat{\theta}_n(\hat{\beta}_n) - \theta^*(\beta^*)\}$ may depend abruptly on the value of $\beta^*$ and the distribution of $X$. Therefore, standard methods such as asymptotic approximations or the delta method cannot be applied.

There are some existing approaches constructing confidence sets for $\theta^*(\beta^*)$. One approach is to ignore the non-smoothness and use the non-parametric bootstrap to construct a CI (Tibshirani, 1984; Efron, 1987). The central percentile bootstrap CI works well when $f(x, \beta^*)$ is continuous. But when $f(x, \beta^*)$ is not differentiable, it may yield poor performance because of the estimator's nonregular limiting distribution (Van Der Vaart, 1991; Robins, 2004; Hirano and Porter, 2012; Laber et al., 2014c). Another approach is the projection confidence interval (PCI), which was first proposed by Berger and Boos (1994), and later discussed by Robins (2004). Instead of directly constructing a CI for $\theta^*(\beta^*)$, a CI for nuisance parameter $\beta^*$ is constructed first based on the limiting distribution of $\sqrt{n}(\hat{\beta}_n - \beta^*)$. Then, CIs for the estimand $\theta^*(\beta)$ are constructed corresponding to each fixed point $\beta$ in the CI of nuisance parameter $\beta^*$. It can be shown that the union of these CIs for $\theta^*(\beta)$ will be an asymptotically valid CI for our target estimand $\theta^*(\beta^*)$. However, because the PCI is constructed by taking unions, it may be conservative especially when $f(x, \beta^*)$ is continuous. The $m$-out-of-$n$ subsampling bootstrap is another approach to construct CIs for non-smooth esimands (Shao, 1994; Bickel and Sakov, 2008). The $m$-out-of-$n$ subsampling CI is constructed through drawing bootstrap samples with size $m < n$. The choice of $m$ is connected with a measure of nonregularity, and thus can achieve an asymptotically valid CI for non-smooth estimands. Recently Chakraborty et al. (2013a) applied the $m$-out-of-$n$ subsampling bootstrap method to construct CIs for the first stage coefficients in $Q$-learning. They also extended this $m$-out-of-$n$ method to CI construction of marginal mean outcome in dynamic treatment

regimes (DTR) (Chakraborty et al., 2014). This $m$-out-of-$n$ method can provide valid CI in both regular and nonregular situations. Laber et al. (2014c) proposed the Adaptive Confidence Interval (ACI) for the first stage coefficients in $Q$-learning. They created smooth upper and lower bounds on the non-smooth estimand, and use them to construct confidence intervals. But it is difficult to compute these bounds and to generalize them to new settings. Thus, it is difficult to apply ACI to other estimands.

Based on the PCI, we propose an adaptive method to construct an asymptotically valid CI. We call it Adaptive Projection Interval (API). The API can be used with estimands on the form $\theta^*(\beta^*) = \mathrm{E}\{s(X_1, \beta_1^*)g(X_2, \beta_2^*)\}$ discussed previously. When using the API, standard methods such as the percentile bootstrap will be applied to observations that are not in the region $\mathcal{Q}_{(X,\beta^*)}$, while PCI will be applied to observations that are in the region $\mathcal{Q}_{(X,\beta^*)}$. Therefore the API will gain the benefit from both percentile bootstrap and PCI. We prove that it provides an asymptotically valid CI for both regular and nonregular situations. The outline of this chapter is as follows. Section 2.2 reviews existing methods for constructing a valid CI for non-smooth estimands. The construction and tuning of the API and its theoretical properties are illustrated in Section 2.3. Section 2.4 presents a numerical study of the API. Section 2.5 concludes the chapter. Proofs of theoretical results are referred to an appendix.

## 2.2 Review of Existing methods

In this section, we review the following methods for constructing confidence intervals for non-smooth estimands: the percentile bootstrap confidence interval (CPB), the projection confidence interval (PCI), the $m$-out-of-$n$ subsampling bootstrap confidence interval (MOFN), and the adaptive confidence interval (ACI).

### 2.2.1 Percentile Bootstrap Confidence Interval

The percentile bootstrap method was first proposed by Efron (1979). A $(1-2\alpha)\times 100\%$ percentile bootstrap CI for $\theta^*(\beta^*)$ is constructed as follows.

**Step 1** For $b = 1, 2, ..., B$,

- draw $n$ observations with replacement from $\mathcal{D}$, denote the sampled data $\mathcal{D}^{(b)}$;
- construct the corresponding estimator $\hat{\theta}_n^{(b)}(\hat{\beta}_n^{(b)})$ for $\theta^*(\beta^*)$ using $\mathcal{D}^{(b)}$.

**Step 2** Take the empirical $\alpha \times 100$ and $(1-\alpha) \times 100$ percentiles of bootstrap estimators $\hat{\theta}_n^{(1)}(\hat{\beta}_n^{(1)}), \hat{\theta}_n^{(2)}(\hat{\beta}_n^{(2)}), ..., \hat{\theta}_n^{(B)}(\hat{\beta}_n^{(B)})$. These two empirical percentiles will construct the $(1-2\alpha) \times 100\%$ bootstrap percentile CI for $\theta^*(\beta^*)$.

If we use $\hat{K}_B$ denote the empirical distribution function of these bootstrap values, then the $(1-2\alpha) \times 100\%$ percentile bootstrap confidence interval is :

$$\left(\hat{K}_B^{-1}(\alpha), \hat{K}_B^{-1}(1-\alpha)\right).$$

When $f(x, \beta^*)$ is continuous, or $\beta^*$ is a fixed and known, then under mild regularity condition, the CPB provides an asymptotically valid CI. However, in our situation, $\beta^*$ is a parameter that is unknown, and $f(x, \beta^*)$ is not differentiable in the region $\mathcal{Q}_{(X,\beta^*)}$. Therefore, the CI constructed by percentile bootstrapping through estimator $\hat{\theta}_n(\hat{\beta}_n)$ may not provide nominal coverage.

### 2.2.2 Projection Confidence Interval

The projection confidence interval (PCI) was first introduced by Berger and Boos (1994). Later it was discussed by Robins (2004). A projection region for the target parameter $\theta^*(\beta^*)$ is constructed as follows. First, construct a $(1-\eta) \times 100\%$ confidence region for $\beta^*$, say $\mathcal{C}_{(1-\eta),\beta^*}$. Second, for each $\beta \in \mathcal{C}_{(1-\eta),\beta^*}$, a central bootstrap confidence interval $\mathcal{I}_{(1-\alpha),\theta^*(\beta)}$ for the corresponding estimand $\theta^*(\beta) = \mathrm{E}\{f(X, \beta)\}$ can be formed through the limiting distribution of $\sqrt{n}\{\hat{\theta}_n(\beta) - \theta^*(\beta)\}$, where $\hat{\theta}_n(\beta) = \mathbb{E}_n\{f(X, \beta)\}$. The projection confidence region for $\theta^*(\beta^*)$ is defined as the union of $\mathcal{I}_{(1-\alpha),\theta^*(\beta)}$ over all $\beta \in \mathcal{C}_{(1-\eta),\beta^*}$, which is denoted as $\mathcal{U}_{(1-\alpha-\eta),\theta^*}$. There are two chances to have mistake with this approach: (i) $\beta^*$ may be not belong to $\mathcal{C}_{(1-\eta),\beta^*}$, and this will occur with probability no more than $\eta$; (ii) $\beta^* \in \mathcal{C}_{(1-\eta),\beta^*}$ but $\theta^*(\beta^*)$ may not belong to $\mathcal{I}_{(1-\alpha),\theta^*(\beta)}$, and this will occur with probability no more than $\alpha$. Thus the probability that $\theta^*(\beta^*) \notin \mathcal{U}_{(1-\alpha-\eta),\theta^*}$ satisfies:

$$\begin{aligned}
\Pr\{\theta^*(\beta^*) \notin \mathcal{U}_{(1-\alpha-\eta),\theta^*}\} &= \Pr\{\theta^*(\beta^*) \notin \mathcal{U}_{(1-\alpha-\eta),\theta^*}, \beta^* \in \mathcal{C}_{(1-\eta),\beta^*}\} + \\
&\quad \Pr\{\theta^*(\beta^*) \notin \mathcal{U}_{(1-\alpha-\eta),\theta^*}, \beta^* \notin \mathcal{C}_{(1-\eta),\beta^*}\} \\
&\leq \Pr\{\theta^*(\beta^*) \notin \mathcal{I}_{(1-\alpha),\theta^*(\beta^*)}\} + \Pr\{\beta^* \notin \mathcal{C}_{(1-\eta),\beta^*}\} \\
&\leq \alpha + \eta + o_p(1).
\end{aligned}$$

This means the probability that $\theta^*(\beta^*) \notin \mathcal{U}_{(1-\alpha-\eta),\theta^*}$ is less or equal to $\alpha+\eta$, which satisfies the definition of validity for confidence interval. Hence, the coverage rate of the projection confidence region is at least $(1-\alpha-\eta)\times 100\%$.

In our model set up, the PCI is implemented as follows. Let $\hat{\beta}_n$ denote an consistent estimator of $\beta^*$ such that $\sqrt{n}(\hat{\beta}_n - \beta^*)$ converges in distribution to $\mathrm{N}(\mathbf{0}, \Omega_{\beta^*})$, where $\Omega_{\beta^*}$ is positive semi-definite. Let $\hat{\Omega}_{\hat{\beta}_n}$ denote a consistent estimator of $\Omega_{\beta^*}$, which is smooth in a neighbor of $\beta^*$. A wald-type asymptotic $(1-\eta)\times 100\%$ confidence region for $\beta^*$ is therefore

$$\mathcal{C}_{(1-\eta),\beta^*} \triangleq \left\{\beta \in \mathbb{R}^{\dim(\beta^*)} : n(\hat{\beta}_n - \beta)^T \hat{\Omega}_{\hat{\beta}_n}(\hat{\beta}_n - \beta) \leq \chi^2_{1-\eta,\dim(\beta^*)}\right\},$$

where $\chi^2_{1-\eta,d}$ is the $(1-\eta)\times 100$ percentile of a $\chi^2$-distribution with $d$ degrees of freedom. And for each $\beta \in \mathcal{C}_{(1-\eta),\beta^*}$ fixed, it follows from standard argument that $\sqrt{n}(\hat{\theta}_n(\beta) - \theta^*(\beta))$ is regular, asymptotically normal with mean zero, where $\hat{\theta}_n(\beta) = \mathbb{E}_n\{f(X,\beta)\}$. This is because for $\beta$ fixed and assuming $\mathrm{Var}\{f(X,\beta)\}$ exists, the central limit theorem can be applied to $\sqrt{n}(\hat{\theta}_n(\beta) - \theta^*(\beta))$. Thus, standard methods for constructing confidence interval, e.g. the percentile bootstrap, can be used to form a valid $(1-\alpha)\times 100\%$ confidence interval for $\theta^*(\beta)$, say $\mathcal{I}_{(1-\alpha),\theta^*(\beta)}$. Then, the union

$$\mathcal{U}_{(1-\alpha-\eta),\theta^*} = \bigcup_{\beta \in \mathcal{C}_{(1-\eta),\beta^*}} \mathcal{I}_{(1-\alpha),\theta^*(\beta)}, \tag{2.1}$$

is a $(1-\alpha-\eta)\times 100\%$ projection confidence interval (PCI) for $\theta^*(\beta^*)$.

The PCI is appealing because it is conceptually simple. However, it may be conservative especially when $\mathrm{P}\{(x,\beta^*) \in \mathcal{Q}_{(X,\beta^*)}\} = 0$. For example, let $\theta^*(\beta^*) = \mathrm{E}([X^T\beta^*]_+)$, where the parameter $\beta^*$ is from multiple linear regression model. We know that $[x^T\beta^*]_+$ is non-smooth when $(x,\beta^*) \in \{(x,\beta^*) : x^T\beta^* = 0\}$. Suppose $\mathrm{P}\{X^T\beta^* > 0\} = 1$. Then the estimand can be rewritten as $\mathrm{E}(X^T\beta^*) = \mathrm{E}(X)^T\beta^*$. Thus, standard methods, e.g., the percentile bootstrap, can be used to construct a CI. It can be shown that the CI width from percentile bootstrap is much smaller than the CI width from PCI. For proof details, see Section A.1.1.

### 2.2.3 $m$-out-of-$n$ Subsampling Bootstrap Confidence Interval

The $m$-out-of-$n$ bootstrap is one possible approach to produce valid confidence interval for non-smooth estimands (Bretagnolle 1983; Swanepoel 1986; Shao and Wu 1989; Dümbgen 1993; Shao 1994; Huang et al. 1996; Bickel et al. 1997). Instead of resampling bootstrap samples of size $n$, the $m$-out-of-$n$ bootstrap draws bootstrap samples of size $m$, which is a smaller order than the original sample size $n$. This means $m$ depends on $n$, tends to infinity, and satisfies $\frac{m}{n} \to 0$ (or we say, $m = o(n)$). A $(1-2\alpha) \times 100\%$ $m$-out-of-$n$ bootstrap CI for $\theta^*(\beta^*)$ is constructed as follows:

**Step 1** For $m = 1, 2, ..., B$,

- draw $m$ observations with replacement from $\mathcal{D}$, denote the sampled data $\mathcal{D}_m^{(b)}$;
- construct the corresponding estimator $\hat{\theta}_m^{(b)}(\hat{\beta}_m^{(b)})$ for $\theta^*(\beta^*)$ using $\mathcal{D}_m^{(b)}$.

**Step 2** Take the empirical $\alpha \times 100$ and $(1-\alpha) \times 100$ percentiles of bootstrap estimators $\hat{\theta}_m^{(1)}(\hat{\beta}_m^{(1)}), \hat{\theta}_m^{(2)}(\hat{\beta}_m^{(2)}), ..., \hat{\theta}_m^{(B)}(\hat{\beta}_m^{(B)})$. These two empirical percentiles will construct the $(1-2\alpha) \times 100\%$ $m$-out-of-$n$ bootstrap CI for $\theta^*(\beta^*)$.

One difficulty associated with this method is the choice of $m$ in finite sample. Since the condition $m = o(n)$ is asymptotic, and thus no guidance is provided for finite samples. Intuitively, the choice of the resample size $m$ should reflect the non-regularity of the estimand. Using the idea from Bickel and Sakov (2008), Chakraborty et al. (2013b) proposed a data adaptive choice of $m$ based on a measure of non-regularity. They applied their method to construct CIs for coefficients in $Q$-learning. Chakraborty et al. (2013b) consider a class of resample sizes of the form $m \triangleq n^{f(p)}$, where $f(p)$ satisfies: (i) $f(p)$ is monotone decreasing in $p$, where $p \in (0, 1]$, and $f(0) = 1$; and (ii) $f(p)$ is continuous and its first derivative is bounded. A plug-in estimator of $p$ is defined as $\hat{p} = \mathbb{E}_n 1_{\{\hat{T}_n(X,\hat{\beta}_n) \leq \tau_n\}}$, where $\hat{T}_n(X, \hat{\beta}_n)$ is a test statistic testing $\mathrm{H}_0 : (x, \beta^*) \in \mathcal{Q}_{X,\beta^*}$, reject the test if $\hat{T}_n(x, \hat{\beta}_n) > \tau_n$, and $\tau_n$ is a potential tuning parameter. They used $f(p) = \frac{1+\nu(1-p)}{1+\nu}$, but other choices of $f(p)$ are possible. The corresponding estimator of $m$ is then defined as:

$$\hat{m} = n^{\frac{1+\nu(1-\hat{p})}{1+\nu}},$$

where $\nu > 0$ is a tuning parameter that can be chosen using double bootstrap. After deciding the resample size $m$, the $m$-out-of-$n$ bootstrap confidence interval can be constructed. Chakraborty et al. (2014) also extended this idea of choosing $m$ to the marginal

mean outcome in dynamic treatment regimes (DTRs). The expression for estimator $\hat{m}$ is similar. It can be shown that $m$-out-of-$n$ subsampling confidence interval forming in this way is asymptotically valid.

### 2.2.4 Adaptive Confidence Interval (ACI)

Similar to the adaptive confidence interval for misclassification rate (Laber and Murphy, 2011), Laber et al. (2014c) proposed an adaptive confidence interval (ACI) to construct an asymptotically valid confidence interval for linear combinations of the first stage coefficients in $Q$-learning. Let $\theta^*(\beta^*)$ denote the multiplication of the first stage coefficient vector, and let $\hat{\theta}_n(\hat{\beta}_n)$ denote the estimator of $\theta^*(\beta^*)$. Because of non-regularity, it is impossible to construct a uniformly convergent estimator of the limiting distribution of $\sqrt{n}\{\hat{\theta}_n(\hat{\beta}_n) - \theta^*(\beta^*)\}$ (Van Der Vaart, 1991; Hirano and Porter, 2012). Instead of constructing a CI for $\theta^*(\beta^*)$ directly, Laber et al. (2014c) construct an upper bound and a lower bound for $\sqrt{n}\{\hat{\theta}_n(\hat{\beta}_n) - \theta^*(\beta^*)\}$. These two bounds can be regular and uniformly convergent. Thus a confidence interval for $\theta^*(\beta^*)$ can be formed by bootstrapping these two bounds.

To limit the conservatism, these bounds are defined only based on the non-smooth term $g(X_2, \beta_2^*)$ of the target estimand $\theta^*(\beta^*) = \mathrm{E}\{s(X_1, \beta_1^*)g(X_2, \beta_2^*)\}$, and only applied to subjects in the non-smooth region $\mathcal{Q}_{(X,\beta^*)}$. To achieve this, a "pretest" (Olshen, 1973; Andrews, 2001; Cheng, 2008; Andrews and Guggenberger, 2009) is used to partition the observed data into two groups: (Group 1) observations that are not in the non-smooth region; and (Group 2) observations that are near in the non-smooth region. A test statistic $\hat{T}_n(x, \hat{\beta}_n)$ is used to achieve the partition: assign a subject to Group 1 if $\hat{T}_n(x, \hat{\beta}_n) > \lambda_n$ and Group 2 otherwise. The $\lambda_n$ is a tuning parameter that can be estimated using double bootstrapping.

Let $\mathcal{U}$ denote the upper bound, and let $\mathcal{L}$ denote the lower bound of $\sqrt{n}\{\hat{\theta}_n(\hat{\beta}_n) - \theta^*(\beta^*)\}$ proposed in this procedure. Noting that

$$\hat{\theta}_n(\hat{\beta}_n) - \mathcal{U}/\sqrt{n} \leq \theta^*(\beta^*) \leq \hat{\theta}_n(\hat{\beta}_n) - \mathcal{L}/\sqrt{n},$$

the distribution of the bounds can be approximated by percentile bootstrap. Let $\hat{u}$ denote the $(1-\alpha) \times 100$ percentile of bootstrap distribution of $\mathcal{U}$, and let $\hat{l}$ denote the $\alpha \times 100$ percentile of the bootstrap distribution of $\mathcal{L}$. A $(1-2\alpha) \times 100\%$ adaptive confidence

interval (ACI) for $\theta^*(\beta^*)$ is defined:

$$\left(\hat{\theta}_n(\hat{\beta}_n) - \hat{u}/\sqrt{n}, \hat{\theta}_n(\hat{\beta}_n) - \hat{l}/\sqrt{n}\right).$$

The ACI can provide a asymptotically valid confidence interval for non-regular estimand. But its construction of the upper and lower bounds may be too complicated.

## 2.3 Adaptive Projection Interval

In this section, we propose the adaptive projection interval, which is a asymptotically valid confidence interval for the estimands we have defined before. We first introduce the construction of API. We then discuss the algorithm of tuning the parameter in API.

### 2.3.1 Construction of API

We propose a method of constructing a confidence interval that is consistent in non-regular framework. We refer to this method as the Adaptive Projection Interval (API). The API is inspired on the idea of projection confidence interval (PCI), which is proposed by Robins (2004). The details for PCI is introduced in 2.2.2. Note that the estimand of interest has the term: $\theta^*(\beta^*) = \mathrm{E}\{s(X_1, \beta_1^*)g(X_2, \beta_2^*)\}$, where $s(\cdot)$ represents smooth part and $g(\cdot)$ represents non-smooth part (e.g. indicator function) in a region $\mathcal{Q}_{(X,\beta^*)}$. Let $\hat{\beta}_n$ denote a consistent estimator for $\beta^*$. And define estimator $\hat{\theta}_n(\hat{\beta}_n) = \mathbb{E}_n\{s(X_1, \hat{\beta}_{1n})g(X_2, \hat{\beta}_{2n})\}$. Recall that generally it is impossible to construct a uniformly convergent estimator of the limiting distribution of $\sqrt{n}(\hat{\theta}_n(\hat{\beta}_n) - \theta^*(\beta^*))$. Projection confidence interval is one possible approach to construct confidence interval for $\theta^*(\beta^*)$. But it may be too conservative. Our approach is applying the idea of projection confidence interval to an adaptive estimator for $\theta^*(\beta^*)$.

First, note that the observed dataset $\mathcal{D}$ can be partitioned into two groups: (Group 1) points that are not in the non-smooth region $\mathcal{Q}_{(X,\beta^*)}$; and (Group 2) points that are in the non-smooth region $\mathcal{Q}_{(X,\beta^*)}$. Then by the group division, the estimator $\hat{\theta}_n(\hat{\beta}_n)$ can

be decomposed as:

$$\begin{aligned}\hat{\theta}_n(\hat{\beta}_n) &= \mathbb{E}_n\{s(X_1,\hat{\beta}_{1n})g(X_2,\hat{\beta}_{2n})\} \\ &= \mathbb{E}_n\{s(X_1,\hat{\beta}_{1n})g(X_2,\hat{\beta}_{2n})1_{(X,\beta^*)\notin\mathcal{Q}_{(X,\beta^*)}}\} \\ &\quad +\mathbb{E}_n\{s(X_1,\hat{\beta}_{1n})g(X_2,\hat{\beta}_{2n})1_{(X,\beta^*)\in\mathcal{Q}_{(X,\beta^*)}}\},\end{aligned}$$

where 1. is the indicator function. The first term on the right-hand side corresponds to points that are not in the non-smooth region (Group 1), and the second term corresponds to points that are in the non-smooth region (Group 2). Similarly, the estimand $\theta^*(\beta^*)$ can also be written in this style:

$$\begin{aligned}\theta^*(\beta^*) &= \mathrm{E}\{s(X_1,\beta_1^*)g(X_2,\beta_2^*)\} \\ &= \mathrm{E}\{s(X_1,\beta_1^*)g(X_2,\beta_2^*)1_{(X,\beta^*)\notin\mathcal{Q}_{(X,\beta^*)}}\} \\ &\quad +\mathrm{E}\{s(X_1,\beta_1^*)g(X_2,\beta_2^*)1_{(X,\beta^*)\in\mathcal{Q}_{(X,\beta^*)}}\}.\end{aligned}$$

Since $\beta^*$ is unknown, to achieve the partition, we use a "pretest" (Olshen, 1973; Andrews, 2001; Cheng, 2008; Andrews and Guggenberger, 2009; Andrews and Soares, 2010). The pretest is based on $\hat{T}_n(X,\hat{\beta}_n)$, which is a test statistic that diverges to $+\infty$ when $(x,\beta^*)\notin\mathcal{Q}_{(X,\beta^*)}$ but is bounded in probability when $(x,\beta^*)\in\mathcal{Q}_{(X,\beta^*)}$. For each $X=x$, this test statistic testing: $H_0:(x,\beta^*)\in\mathcal{Q}_{(X,\beta^*)}$ against the alternative. Reject the test if $\hat{T}_n(x,\hat{\beta}_n)>\lambda_n$, where $\lambda_n$ is a potential tuning parameter. Then using the pretest, $\hat{\theta}_n(\hat{\beta}_n)$ can be rewritten as:

$$\begin{aligned}\hat{\theta}_n(\hat{\beta}_n) &= \mathbb{E}_n\{s(X_1,\hat{\beta}_{1n})g(X_2,\hat{\beta}_{2n})1_{\hat{T}_n(X,\hat{\beta}_n)>\lambda_n}\} \\ &\quad +\mathbb{E}_n\{s(X_1,\hat{\beta}_{1n})g(X_2,\hat{\beta}_{2n})1_{\hat{T}_n(X,\hat{\beta}_n)\leq\lambda_n}\}.\end{aligned}$$

Let $\mathcal{C}_{(1-\eta),\beta^*}$ denote the $(1-\eta)\times 100\%$ confidence region for nuisance parameter $\beta^*$ from limiting distribution of $\sqrt{n}(\hat{\beta}_n-\beta^*)$. In projection confidence interval, for each $\beta\in\mathcal{C}_{(1-\eta),\beta^*}$, a confidence interval is constructed based on the limiting distribution of $\sqrt{n}\{\hat{\theta}_n(\beta)-\theta^*(\beta)\}$. But in adaptive projection interval, we consider an adaptive estimator $\hat{\theta}_n(\beta,\hat{\beta}_n)$:

$$\begin{aligned}\hat{\theta}_n(\beta,\hat{\beta}_n) &= \mathbb{E}_n\{s(X_1,\hat{\beta}_{1n})g(X_2,\hat{\beta}_{2n})1_{T_n(X,\hat{\beta}_n)>\lambda_n}\} \\ &\quad +\mathbb{E}_n\{s(X_1,\hat{\beta}_{1n})g(X_2,\beta_2)1_{T_n(X,\hat{\beta}_n)\leq\lambda_n}\},\end{aligned} \tag{2.2}$$

where $\beta_2$ is the corresponding subset of $\beta$. We also define the estimand $\theta^*(\beta, \beta^*)$:

$$\begin{aligned}\theta^*(\beta,\beta^*) &= \mathrm{E}\{s(X_1,\beta_1^*)g(X_2,\beta_2^*)1_{(X,\beta^*)\notin\mathcal{Q}_{(X,\beta^*)}}\} \\ &\quad +\mathrm{E}\{s(X_1,\beta_1^*)g(X_2,\beta_2)1_{(X,\beta^*)\in\mathcal{Q}_{(X,\beta^*)}}\}. \end{aligned} \tag{2.3}$$

It can be shown that $\sqrt{n}\{\hat{\theta}_n(\beta,\hat{\beta}_n) - \theta^*(\beta,\beta^*)\}$ is asymptotically normal with finite variance (details see in Section A.1). Hence for each $\beta \in \mathcal{C}_{(1-\eta),\beta^*}$, a $(1-\alpha)\times 100\%$ confidence interval based on this limiting distribution can be constructed using non-parametric bootstrapping. We denote this confidence interval as $\mathcal{I}_{(1-\alpha),\theta^*(\beta,\beta^*)}$. Then the union:

$$\widetilde{\mathcal{U}}_{(1-\alpha-\eta),\theta^*(\beta^*)} = \bigcup_{\beta\in\mathcal{C}_{(1-\eta),\beta^*}} \mathcal{I}_{(1-\alpha),\theta^*(\beta,\beta^*)} \tag{2.4}$$

is a $(1-\alpha-\eta)\times 100\%$ adaptive projection interval (API) for $\theta^*(\beta^*)$. The API is adaptive in two ways. First, it provides an asymptotically valid confidence interval regardless of the non-smoothness. Specifically, this is achieved by applying PCI to points that are in the region $\mathcal{Q}_{(X,\beta^*)}$ through a test statistic. Secondly, it restricts the scale of non-smooth part. Because instead of using $f(X,\beta) = s(X_1,\beta_1)g(X_2,\beta_2)$ to construct the PCI, it uses $f(X,\beta,\hat{\beta}_n) = s(X_1,\hat{\beta}_n)g(X_2,\beta_2)$ to form the PCI.

### 2.3.2 Choice of Tuning Parameter

In the construction of adaptive projection interval (API), it is important to decide the value of tuning parameter $\lambda_n$. Recall that $\sqrt{n}(\hat{\beta}_n - \beta^*)$ is asymptotic normal with mean 0 and variance matrix $\Omega_{\beta^*}$. Then the test statistic $\hat{T}_n(x,\hat{\beta}_n)$ can be $\frac{n(x^T\hat{\beta}_n)^2}{x^T\hat{\Omega}_{\beta^*}x}$, where $\hat{\Omega}_{\beta^*}$ is a plug-in estimator for $\Omega_{\beta^*}$. Then a potential cutting value can be defined as $\chi^2_{1-\alpha,\dim(\beta^*)}$, which is the $(1-\alpha)\times 100$ percentile of a $\chi^2$-distribution with $\dim(\beta^*)$ degrees of freedom. We then consider a range of values of $\lambda_n$ of the form $\lambda_n = \tau\chi^2_{1-\alpha,\dim(\beta^*)}$, where $\tau \in [m, M]$ with $0 < m < M < +\infty$. A double bootstrap procedure (see Davison and Hinkley 1997 for details) is used here for choosing the tuning parameter $\lambda_n$ (or $\tau$) in a data-driven manner. This procedure appears to reduce conservatism in simulations.

Consider a grid of candidate values for $\tau$; for example, $\{0.025, 0.05, 0.075, ..., 1\}$. The algorithm is as follows:

- Calculate the estimator $\hat{\theta}_n(\hat{\beta}_n)$ for data $\mathcal{D}$.

- Draw $B_1$ bootstrap samples from the data $\mathcal{D}$. For each bootstrap sample data $\mathcal{D}^{(b_1)}$, $b_1 = 1, ..., B_1$, and for each $\tau^i \in \{0.025, 0.05, 0.075, ..., 1\}$ do the following:

    1. calculate the corresponding confidence region of $\beta^*$, $\mathcal{C}^{(b_1)}_{(1-\eta),\beta^*}$, based on the limiting distribution of $\sqrt{n}(\hat{\beta}^{(b_1)}_n - \beta^*)$.
    2. Conditional on sample $\mathcal{D}^{(b_1)}$, draw $B_2$ bootstrap samples: $\mathcal{D}^{(b_1,1)}, ..., \mathcal{D}^{(b_1,B_2)}$.
    3. For each double bootstrap sample $\mathcal{D}^{(b_1,b_2)}$, $b_2 = 1, ..., B_2$, and each point $\beta \in \mathcal{C}^{(b_1)}_{(1-\eta),\beta^*}$, compute the adaptive estimator $\hat{\theta}_n(\beta, \hat{\beta}^{(b_1,b_2)}_n)$ with tuning parameter $\lambda_n = \tau^i \chi^2_{1-\alpha,\dim(\beta^*)}$.
    4. For each fixed $\beta$, calculate the non-parametric bootstrap $(1-\alpha)\times 100\%$ confidence interval $\mathcal{I}^{(b_1)}_{(1-\alpha),\theta^*(\beta,\beta^*)}$ for $\theta^*(\beta, \beta^*)$ based on these $\hat{\theta}_n(\beta, \hat{\beta}^{(b_1,b_2)}_n)$.
    5. Take unions of $\mathcal{I}^{(b_1)}_{(1-\alpha),\theta^*(\beta,\beta^*)}$ for all $\beta \in \mathcal{C}^{(b_1)}_{(1-\eta),\beta^*}$ to get the adaptive confidence interval $\widetilde{\mathcal{U}}^{(b_1)}_{(1-\alpha-\eta),\theta^*(\beta^*)}$.

- Estimate the coverage rate of the double bootstrap CIs $\widetilde{\mathcal{U}}^{(b_1)}_{(1-\alpha-\eta),\theta^*(\beta^*)}$ from first-stage bootstrap data $\mathcal{D}^{(b_1)}$ for each $\tau^i$

$$\frac{1}{B_1}\sum_{b_1=1}^{B_1} 1_{\hat{\theta}_n(\hat{\beta}_n)\in\widetilde{\mathcal{U}}^{(b_1)}_{(1-\alpha-\eta),\theta^*(\beta^*)}}.$$

- Pick the value of $\tau^i$ that has the coverage rate nearest to the significant level $(1-\alpha-\eta)$.

## 2.4 Empirical Study

One main advantage of the API is that it provides valid coverage rate for the estimand with expression $\theta^*(\beta^*) = \mathrm{E}\{s(X_1, \beta^*_1)g(X_2, \beta^*_2)\}$ under both smooth and non-smooth situations. It can be used for numerous estimands. Moreover, it is easily understand and applied. In this section, we use simulated examples to demonstrate the numerical performance of API in various scenarios. We consider four monte carlo examples. The first example is a toy example, which is related to a stochastic process. The second example shows a case that API presents asymptotically valid confidence interval for the marginal mean outcome of one stage dynamic treatment regime. The third example demonstrates

the performance of API for inference of multi-stage $Q$-learning. The forth example shows the performance of API for the estimand related to biomarker evaluation. In general, we consider the construction of 95% confidence intervals for these four examples. We compare the empirical performance of the API with the following methods: the centered percentile bootstrap (CPB) as described in Section 2.2.1; the projection confidence interval (PCI) as described in Section 2.2.2 with parameter fixed at $\eta = 0.01, \alpha = 0.05$; the adaptive $m$-out-of-$n$ (MOFN) bootstrap with data-drive tuning as described in Section 2.2.3.

### 2.4.1 Toy Example

The target estimand for this example is $\theta^*(\beta^*) = \mathrm{E}[X^T\beta^*]_+$. The non-smooth region of $(X, \beta^*)$ for this example satisfying: $X^T\beta^* = 0$. Five generative models are used in these evaluations. Each of the models are generated from:

- $X \sim \mathrm{N}_p(\mathbf{0}, \mathrm{AR}(0.5))$
- $\beta^* = (\frac{i}{\sqrt{np}}, ..., \frac{i}{\sqrt{np}})_p^T$
- $Y = X^T\beta^* + \epsilon$, $\epsilon \sim \mathrm{N}(0, 1)$

where $X$ and $\epsilon$ are independent, and variance matrix $\mathrm{AR}(d)$ is a symmetric matrix with value $d^{|i-j|}$ for the $i$th row the $j$th column element. Sample size is fixed at $n = 20$; dimension of $\beta^*$ is also fixed at $p = 3$; and $i$ has values: 0.5, 1, 2, 3, 4. As $i$ increases, $\theta^*(\beta^*)$ will be far away from the non-smooth point 0.

Table 2.1 shows the coverage probabilities of the simulations. Table 2.2 shows the mean width of confidence intervals. As expected, CIs constructed via the CPB method have the smallest average width; however, these are associated with under-coverage under non-regular situation. On the other hand, PCI method always gives the widest CIs, with coverage rates often over the nominal level. Our proposed API method with data-driven $\lambda_n$ always offers valid coverage rates. By fine-tuning $\lambda_n$, it reduce the conservatism in regular cases.

### 2.4.2 Marginal Mean Outcome for Dynamic Treatment Regimes

Chronic illness is a long term disease. Many kinds of common diseases are belonged to chronic illness. For example, alcohol and drug abuse (Lei et al., 2012), cancer (Thall

Table 2.1: Monte Carlo estimates of coverage probabilities of confidence intervals for the toy example at 95% nominal level. The rows represent different methods of constructing CIs: (i) the $n$-out-of-$n$ centered percentile bootstrap (CPB); (ii) the projection confidence interval (PCI); (iii) the proposed adaptive projection interval (API). Estimates are constructed using 200 datasets, and 500 bootstraps drawn from each dataset. Coverage rate significantly different from 0.95 at the 0.05 level are in bold. The ones that significantly below 0.95 are marked with $*$.

| Toy Example | $i = 0.5$ | $i = 1$ | $i = 2$ | $i = 3$ | $i = 4$ |
|---|---|---|---|---|---|
| CPB | **0.88**$^*$ | 0.93 | 0.93 | 0.94 | 0.94 |
| PCI | 1.0 | 0.98 | 0.98 | 0.99 | 1.0 |
| API | 0.93 | 0.93 | 0.95 | 0.93 | 0.96 |

Table 2.2: Monte Carlo estimates of the mean width of confidence intervals for the toy example at 95% nominal level. The rows represent different methods of constructing CIs: (i) the $n$-out-of-$n$ centered percentile bootstrap (CPB); (ii) the projection confidence interval (PCI); (iii) the proposed adaptive projection interval (API). estimates are constructed using 200 datasets, and 500 bootstraps from each dataset. Widths with coverage rate significantly different from 0.95 at the 0.05 level are in bold. The ones with coverage rate significantly below 0.95 are marked with $*$.

| Toy Example | $i = 0.5$ | $i = 1$ | $i = 2$ | $i = 3$ | $i = 4$ |
|---|---|---|---|---|---|
| CPB | **0.49**$^*$ | 0.56 | 0.67 | 0.73 | 0.85 |
| PCI | 0.51 | 0.59 | 0.77 | 0.86 | 1.08 |
| API | 0.45 | 0.56 | 0.67 | 0.82 | 0.85 |

et al., 2007; Zhao et al., 2011), HIV infection (Khalili and Armaou, 2008; Robins et al., 2008), and mental illnesses (Dawson and Lavori, 2004; Shortreed and Moodie, 2012). A dynamic treatment regime (DTR) is a sequence of decision rules, one per stage, which dictates individualized treatment assignment evolving patients' up-to-date information. These regimes can be widely used for chronic illness. Usually, we use $Value$ to evaluate the quality of a DTR. The $Value$ of a DTR is the average primary outcome obtained when the DTR is applied to the entire population of interest. We say a DTR is *optimal* if its corresponding $Value$ is highest.

Here, we focus on one-stage DTR. Define data $\mathcal{D} = \{(X_i, A_i, Y_i) : i = 1, ..., n; X_i \in \mathbb{R}^p; Y_i \in \mathbb{R}; A_i \in \{-1, 1\}\}$ where: $X_i$ is the $i$th subject information, $Y_i = r(X_i, A_i)$ is the continuous outcome for $i$th subject with a known function $r(\cdot)$, and $A_i$ denotes treatment assigned for the $i$th subject. The outcome $Y_i$ is the larger the better. We define treatment regime $\pi$ as a map: $\mathbb{R}^p \to \{-1, 1\}$ such that patient with information $X = x$ is assigned to treatment $\pi(x)$. Then for a fixed DTR $\pi$, its $Value$ is

$$V^\pi = \mathrm{E}_\pi [Y] = \mathrm{E}_\pi [r(X, A)],$$

where $\mathrm{E}_\pi$ denotes expectation with respect to the distribution of entire data. Usually for a feasible regime $\pi$, one can use inverse probability weighting (IPW) to express the $Value$ of the regime in terms of the generative model (Robins et al., 2000; Murphy et al., 2001). This IPW estimator is

$$V^\pi = \mathrm{E}_\pi [Y] = \mathrm{E}_\pi \left[ \frac{1_{\pi(X)=A}}{\Pr(A|X)} Y \right], \tag{2.5}$$

where $1_{(\cdot)}$ is indicator function, and $\Pr(A|X)$ is the treatment allocation probability. Then a plug-in estimator for $V^\pi$ will be:

$$\hat{V}^\pi = \mathbb{E}_n \left[ \frac{1_{\pi(X)=A}}{\Pr(A|X)} Y \right], \tag{2.6}$$

where $\mathbb{E}_n$ denotes the empirical expectation. From (2.5), we know that the $Value$ of a DTR is non-regular because of the indicator function.

In this simulation, we suppose the plug-in model for $r(X_i, A_i)$ is $Q$-learning (Murphy, 2005a). And the data $\mathcal{D}$ is generated from SMART (sequential multiple assignment randomized trial, for details see Murphy 2005b). Sixteen generative models are considered.

Each of the models are generated from:

- $X = W1_{u \geq \rho} + \mathbb{E}^{\perp}_{\beta^*} W 1_{u<\rho}$
- $W \sim \mathrm{N}_p(0, I_p)$, $\mathbb{P}^{\perp}_{\beta^*} W = W - \frac{\langle W, \beta^* \rangle}{\langle \beta^*, \beta^* \rangle} {\beta^*}^T$
- $A \in \{-1, 1\}$, $A \sim$ Bernoulli(0.5)
- $Y = X^T \psi^* + A X^T \beta^* + e$, $e \sim \mathrm{N}(0, 1)$

where parameter $u$ is uniform distributed in $(0, 1)$, and $\rho$ is a runing parameter taking values $\{0.1, 0.4, 0.7, 0.9\}$. We fix $\psi^* = \mathbf{1}_p$, and define $\beta^* = c\mathbf{1}_p$, where $c$ is a tuning parameter taking values $\{0.1, 0.4, 0.7, 0.9\}$. We also fix dimension $p = 10$.

Figure 2.1 shows the estimated coverage rate for different methods: CPB (central percentile bootstrap), PCI (projection confidence interval), MOFN ($m$-out-of-$n$ adaptive bootstrap), and API (adaptive projection interval). Figure 2.2 shows the corresponding average width for these confidence intervals. CIs constructed via CPB method have the smallest average width; however, under non-regular situations CPB often under-coverage. On the contrast, CIs formed by PCI method always have the largest average width, but with coverage rate often significantly above the 95% nominal level. It is surprising that in some non-regular cases, coverage rates from confidence intervals constructed by MOFN are significantly lower than the 95% nominal level. But our proposed method, API, always have coverage rates around the 95% nominal level, with considerable average CI width.

### 2.4.3 *Q*-learning First Stage Covariates

In section 2.4.2, we mentioned the idea of dynamic treatment regime (DTR). An optimal dynamic treatment regime is the DTR that optimizes the expectation of a cumulative outcome over a population of interest. Recently, there are numerous methods proposed to estimate the optimal DTR. These include $Q$-learning (Murphy, 2005a; Chakraborty and Moodie, 2013; Laber et al., 2014a), $A$-learning (Murphy, 2003a; Robins, 2004), outcome weighted learning (Zhao et al., 2012, 2014), and augmented value maximization (Zhang et al., 2012b, 2013). Among these methods, $Q$-learning is one of the most important one. So in this subsection, we discuss the simulation results from $Q$-learning.

We consider two stage two treatment $Q$-learning. The observed data set is defined as $\mathcal{D} = \{(X_{1i}, A_{1i}, X_{2i}, A_{2i}, Y_i)\}_{i=1}^n$, where $X_1 \in \mathbb{R}^{p_1}$ denotes baseline (pre-randomization)

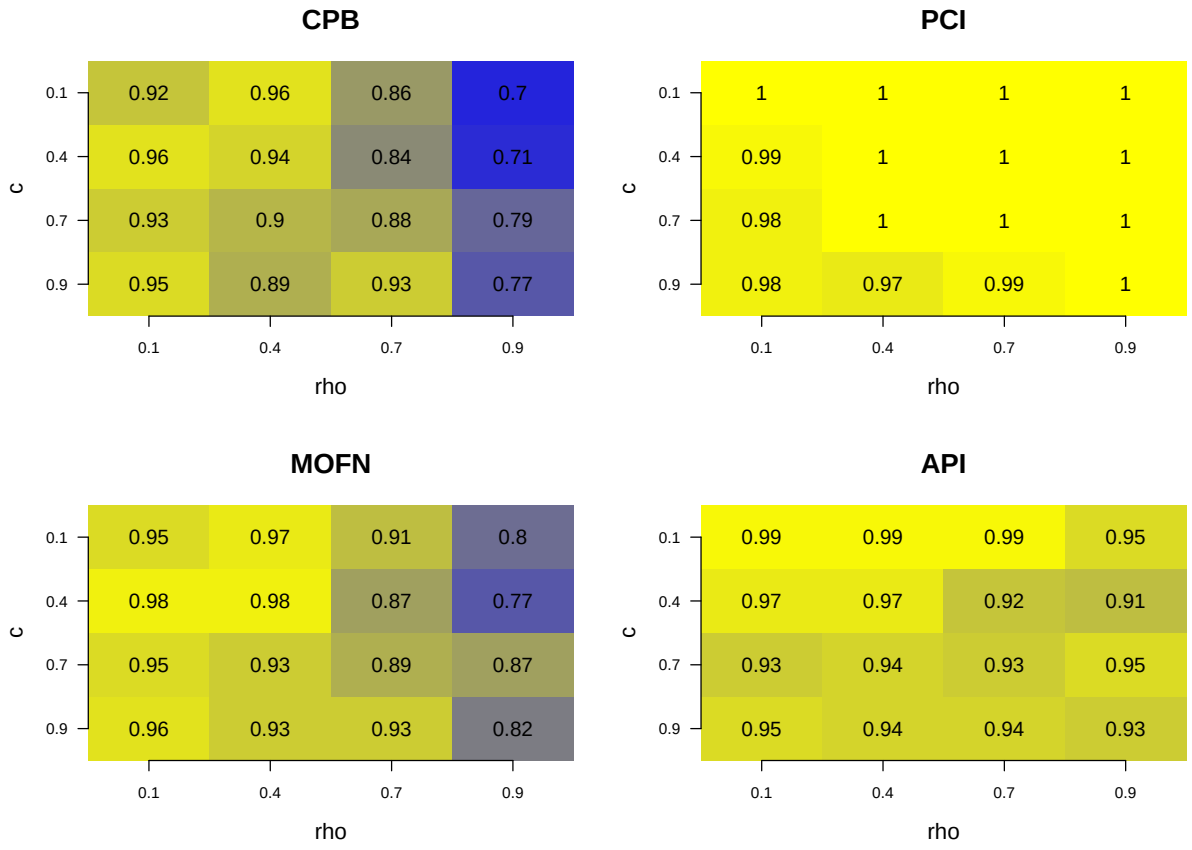


Figure 2.1: Monte Carlo estimates of coverage rates of confidence intervals for the $Value$ at 95% nominal level. The texts in the plot represent the coverage rate. $x$-axis and $y$-axis are values of the two parameters $c, \rho$ for generated models. Sample size $n = 150$, Monte Carlo replication $N_{rep} = 100$, and Bootstrap sample size $B = 1000$.

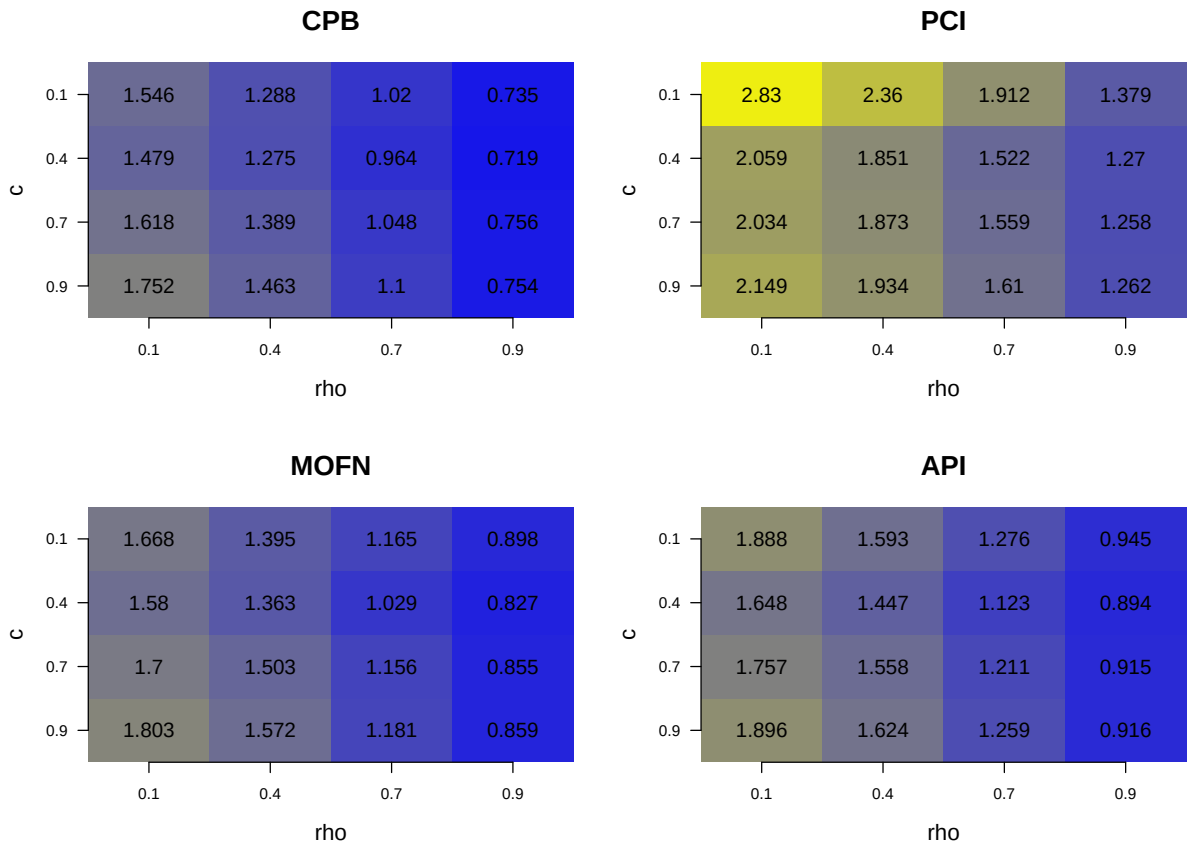


Figure 2.2: Monte Carlo estimates of the mean width of confidence intervals for the $Value$ at 95% nominal level. The texts in the plot represent the average width of the confidence intervals. $x$-axis and $y$-axis are values of the two parameters $c, \rho$ for generated models. Sample size $n = 150$, Monte Carlo replication $N_{rep} = 100$, and Bootstrap sample size $B = 1000$.

subject information; $A_1 \in \mathcal{A}_1$ denotes the first stage treatment assignment; $X_2 \in \mathbb{R}^{p_2}$ denotes information collected during the course of the first stage treatment including information dictating first-stage responder status; $A_2 \in \mathcal{A}_2$ denotes the second stage treatment assignment; and $Y \in \mathbb{R}$ denotes a continuous outcome measured at the end of the study coded so that larger values are better. Let $H_t$, $t = 1, 2$ denote the information collected before time $t$. Then $H_1 = X_1$, and $H_2 = (X_1^T, A_1, X_2^T)$. Also, we define $A_t \in \{-1, 1\}$ The optimal DTR, $\pi^{\text{opt}}$, is the one satisfying $\mathrm{E}^{\pi^{\text{opt}}} Y = \sup_\pi \mathrm{E}^\pi Y$. In two-stage setting, a DTR is defined as a pair of decision rules $\pi = (\pi_1, \pi_2)$, where $\pi_t : \text{dom}(H_t) \to \text{dom}(A_t)$. The optimal DTR can be characterized by $Q$-functions. In the two-stage setting, we define the $Q$-functions as (Sutton and Barto, 1998; Murphy, 2005a)

$$\begin{aligned} Q_2(h_2, a_2) &\triangleq \mathrm{E}(Y | H_2 = h_2, A_2 = a_2), \\ Q_1(h_1, a_1) &\triangleq \mathrm{E}\left\{\max_{a_2} Q_2(H_2, a_2) | H_1 = h_1, A_1 = a_1\right\}. \end{aligned}$$

In this setting, the optimal DTR can be calculated by dynamic programming (Bellman, 1957), and the solution is $\pi_t^{\text{dp}}(h_t) = \text{argmax}_{a_t} Q_t(h_t, a_t)$.

$Q$-learning estimates the optimal DTR by plug-in regression models for the $Q$-functions. Usually, we consider linear regression models, which have the form $Q_t(h_t, a_t; \beta_t) = h_{t,0}^T \beta_{t,0} + a_t h_{t,1}^T \beta_{t,1}$, $h_{t,0}$ and $h_{t,1}$ are known sub vectors of $h_t$ and $\beta_t = (\beta_{t,0}^T, \beta_{t,1}^T)^T$. Then by $Q$-learning algorithm, the estimated coefficients from stage 1 regression is :

$$\hat{\beta}_1 = \arg\min_{\beta_1} \mathbb{E}_n \left(\tilde{Y} - Q_1(H_1, A_1; \beta_1)\right)^2,$$

where $\tilde{Y}$ represents the predicted second stage outcome: $\tilde{Y} = \max_{a_2} Q_2(H_2, a_2; \hat{\beta}_2)$ with $\hat{\beta}_2 = \arg\min_{\beta_2} \mathbb{E}_n \left(Y - Q_2(H_2, A_2; \beta_2)\right)^2$. We then define the following population analogs of the estimators in $Q$-learning:

$$\beta_1^* = \arg\min_{\beta_1} \mathrm{E} \left(\tilde{Y}^* - Q_1(H_1, A_1; \beta_1)\right)^2,$$

where $\tilde{Y}^* = \max_{a_2} Q_2(H_2, a_2; \beta_2^*) = H_{2,0}^T \beta_{2,0}^* + |H_{2,1}^T \beta_{2,1}^*|$, and $\beta_2^* = \arg\min_{\beta_2} \mathrm{E}\left(Y - Q_2(H_2, A_2; \beta_2)\right)^2$. Our target here is to construct 95% confidence interval for $c^T \beta_1^*$, where $c$ is a constant vector. It can be found that this is a non-regular estimand because of the absolute operator. In this simulation, we compare the empirical performance of the API with the following methods: the centered percentile bootstrap (CPB); the adaptive $m$-out-of-$n$

Table 2.3: Parameters indexing the example models. Examples are designated NR=nonregular, NNR=near-nonregular, R=regular.

| EX. | $\gamma$ | $\delta$ | Type | Regularity Measures | |
|---|---|---|---|---|---|
| 1 | $(0,0,0,0,0,0,0)^T$ | $(0.5,0.5)^T$ | NR | $p=1$ | $\phi=0/0$ |
| 2 | $(0,0,0,0,0.01,0,0)^T$ | $(0.5,0.5)^T$ | NNR | $p=0$ | $\phi=\infty$ |
| 3 | $(0,0,-0.5,0,0.5,0,0.5)^T$ | $(0.5,0.5)^T$ | NR | $p=1/2$ | $\phi=1.0$ |
| 4 | $(0,0,-0.5,0,0.5,0,0.49)^T$ | $(0.5,0.5)^T$ | NNR | $p=0$ | $\phi=1.02$ |
| 5 | $(0,0,-0.5,0,1.0,0.5,0.5)^T$ | $(1.0,0.0)^T$ | NR | $p=1/4$ | $\phi=1.41$ |
| 6 | $(0,0,-0.5,0,0.25,0.5,0.5)^T$ | $(0.1,0.1)^T$ | R | $p=0$ | $\phi=0.35$ |
| A | $(0,0,-0.25,0,0.75,0.5,0.5)^T$ | $(0.1,0.1)^T$ | R | $p=0$ | $\phi=1.035$ |
| B | $(0,0,0,0,0.25,0,0.25)^T$ | $(0,0)^T$ | NR | $p=1/2$ | $\phi=1.00$ |
| C | $(0,0,0,0,0.25,0,0.24)^T$ | $(0,0)^T$ | NNR | $p=0$ | $\phi=1.03$ |

(MOFN) bootstrap with data-driven tuning (Chakraborty et al., 2013a). Nine generative models are used in these evaluations (Chakraborty et al., 2009). Each of the models can be expressed as:

- $X_t \in \{-1,1\}$, $A_t \in \{-1,1\}$ for $t \in \{1,2\}$
- $P(A_1=1)=P(A_1=-1)=0.5$, $P(A_2=1)=P(A_2=-1)=0.5$
- $X_1 \sim$ Bernoulli(0.5), $X_2|X_1,A_1 \sim$ Bernoulli(expit$(\delta_1 X_1+\delta_2 A_1)$)
- $Y_2=\gamma_1+\gamma_2 X_1+\gamma_3 A_1+\gamma_4 X_1 A_1+\gamma_5 A_2+\gamma_6 X_2 A_2+\gamma_7 A_1 A_2+\epsilon$, $\epsilon \sim N(0,1)$

where expit$(x)=e^x/(1+e^x)$. The nine generative models are generated by different $\delta,\gamma$ combinations. Table 2.3 shows the information of nine generative models. The patients' history is defined as

$$H_{2,0}=(1,X_1,A_1,X_1A_1,X_2)^T, \quad H_{2,1}=(1,X_2,A_1)^T,$$
$$H_{1,0}=(1,X_1)^T, \quad H_{1,1}=(1,X_1)^T.$$

Then the simulation models are given by $Q_2(H_2,A_2;\beta_2)=H_{2,0}^T\beta_{2,0}+H_{2,1}^T\beta_{2,1}A_2$, and $Q_1(H_1,A_1;\beta_1)=H_{1,0}^T\beta_{1,0}+H_{1,1}^T\beta_{1,1}A_1$. The regularity is defined as $p=P(\gamma_5 A_2+\gamma_6 X_2 A_2+\gamma_7 A_1 A_2=0)$; and the standardized effect size is defined as $\phi=\mathrm{E}[\gamma_5+\gamma_6 X_2+\gamma_7 A_1]/\sqrt{\mathrm{Var}(\gamma_5+\gamma_6 X_2+\gamma_7 A_1)}$. These two quantities can be treated as measures of non-regularity (Chakraborty et al., 2009).

Table 2.4: Monte Carlo estimates of coverage probabilities of confidence intervals for the main effect of treatment at the 95% nominal level. The rows represent different methods of constructing CIs: (i) the $n$-out-of-$n$ centered percentile bootstrap (CPB); (ii) the projection confidence interval (PCI); (iii) the $m$-out-of-$n$ bootstrap; (iv) the proposed adaptive projection interval (API). Estimates are constructed using 200 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Examples are designated NR=nonregular, NNR=near-nonregular, R=regular.

| $Q$-learning | Ex.1 | Ex.2 | Ex.3 | Ex.4 | Ex.5 | Ex.6 | Ex.A | Ex.B | Ex.C |
|---|---|---|---|---|---|---|---|---|---|
| $A_1$ coefficient | NR | NNR | NR | NNR | NR | R | R | NR | NNR |
| CPB | 0.97 | 0.99 | **0.91*** | **0.91*** | 0.93 | 0.95 | 0.92 | **0.90*** | **0.87*** |
| PCI | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| MOFN | 0.98 | 0.97 | **0.91*** | **0.91*** | 0.93 | 0.94 | 0.94 | 0.95 | 0.98 |
| API | 0.98 | 0.96 | 0.92 | 0.97 | 0.95 | 0.95 | 0.94 | 0.93 | 0.93 |

We provide confidence intervals for the coefficienst of $A_1$, say $\beta^*_{1,1}$. Table 2.4 shows the estimated coverage for $\beta^*_{1,1}$. The sample size of this simulation is 150; the number of Monte Carlo replications is 200; the number of bootstrap samples is 1000. The target coverage is 0.95. Table 2.5 presents the average CI width for $\beta^*_{1,1}$. CIs constructed via the CPB method have the smallest average width; however, these are associated with under-coverage under some non-regular or near non-regular situations. But PCI method often gives the widest CIs, with coverage rates often be 1.0, which are conservative. The CIs from MOFN are often around 0.95, but sometimes the average CI width is too wide. Our proposed API method with data-driven $\lambda_n$ always offers valid coverage rates, with average CI width dramatically smaller than average CI width from PCI.

### 2.4.4 Estimands Related to Biomarker Evaluation

Recently, numerous amount of research effort are devoted to evaluating markers that can predict a patient's chance of responding to a treatment. Usually we name these kinds of markers as biomarker. It is known that the treatment selection markers, may be called "predictive" (Simon, 2008) or "prescriptive" (Gunter et al., 2007) markers, have the possibility to improve patient outcomes and reduce medical costs by restricting the treatment used to subjects that seems most likely to have benefit. Since these biomarkers are useful, it is important to provide methods to evaluate the biomarkers.

In the medical literature, there are some existing approaches for marker evaluation.

Table 2.5: Monte Carlo estimates of the mean width of confidence intervals for the main effect of treatment at the 95% nominal level. The rows represent different methods of constructing CIs: (i) the $n$-out-of-$n$ centered percentile bootstrap (CPB); (ii) the projection confidence interval (PCI); (iii) the $m$-out-of-$n$ bootstrap; (iv) the proposed adaptive projection interval (API). Estimates are constructed using 200 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Examples are designated NR=nonregular, NNR=near-nonregular, R=regular.

| $Q$-learning | Ex.1 | Ex.2 | Ex.3 | Ex.4 | Ex.5 | Ex.6 | Ex.A | Ex.B | Ex.C |
|---|---|---|---|---|---|---|---|---|---|
| $A_1$ coefficient | NR | NNR | NR | NNR | NR | R | R | NR | NNR |
| CPB | 0.38 | 0.39 | **0.42*** | **0.42*** | 0.45 | 0.44 | 0.45 | **0.43*** | **0.42*** |
| PCI | 0.79 | 0.79 | 0.87 | 0.87 | 0.92 | 0.82 | 0.90 | 0.87 | 0.87 |
| MOFN | 0.77 | 0.79 | **0.71*** | **0.72*** | 0.95 | 0.44 | 1.35 | 1.30 | 1.32 |
| API | 0.41 | 0.42 | 0.49 | 0.48 | 0.46 | 0.44 | 0.53 | 0.54 | 0.48 |

One comprehensive approach to evaluating markers for treatment selection is proposed by Janes et al. (2014). Suppose we have two treatment options, which are denoted as *treatment* ($T = 1$) and *no treatment* ($T = 0$). Define $D \in \{0, 1\}$, a binary indicator of an "adverse event", which is the clinical outcome of interest within a specific time-frame following treatment assignment. For example, $D$ may be chosen to represent an indicator of treatment-associated toxicity or death. We use $Y$ to denote a marker that is measured prior to treatment provision. The question here is whether $Y$ useful for identifying a group of subjects who can avoid treatment. Define the absolute treatment effect given marker value $Y$ as:

$$\Delta(Y) = Pr(D = 1|T = 0, Y) - Pr(D = 1|T = 1, Y).$$

The treatment rule is defined as following: do not treat if $\Delta(Y) < 0$. Refer to subjects with $\Delta(T) < 0$ as $marker - negatives$, and $\Delta(Y) > 0$ as $marker - positives$. Following the notations from Janes et al. (2014), the useful measures are listed below:

- Average benefit of no treatment among marker-negatives,

$$\begin{aligned} B_{neg} &= Pr(D = 1|T = 1, \Delta(Y) < 0) - Pr(D = 1|T = 0, \Delta(Y) < 0) \\ &= \mathrm{E}(-\Delta(Y)|\Delta(Y) < 0) \end{aligned}$$

- Average benefit of treatment among marker-positives,

$$\begin{aligned} B_{pos} &= Pr(D=1|T=1,\Delta(Y)>0) - Pr(D=1|T=0,\Delta(Y)>0) \\ &= \mathrm{E}(-\Delta(Y)|\Delta(Y)>0) \end{aligned}$$

- Proportion marker-negative, $P_{neg} = Pr(\Delta(Y)<0)$

- Decrease in population event rate under marker-based treatment,

$$\begin{aligned} \Theta &= Pr(D=1|T=1) - [Pr(D=1|T=1,\Delta(Y)>0)Pr(\Delta(Y)>0) \\ &\quad + Pr(D=1|T=0,\Delta(Y)<0)Pr(\Delta(Y)<0)] \\ &= \mathrm{E}(-\Delta(Y)|\Delta(Y)<0)\cdot Pr(\Delta(Y)<0) \\ &= B_{neg}\cdot P_{neg} \end{aligned}$$

In fact, the measure $\Theta$, or its variation, has been advocated as a global measure of marker performance in many papers (Song and Pepe, 2004; Gunter et al., 2007; Janes et al., 2011; McKeague and Qian, 2011; Zhang et al., 2012b). Our target here is to construct a $(1-\alpha-\eta)\times 100\%$ CI for $\Theta$. Given data consisting of observations $(Y_i, T_i, D_i), i = 1, ..., n$, a general linear regression risk model (e.g. logistic regression) with an interaction between $T$ and $Y$ is used:

$$g(Pr(D=1|T,Y)) = \beta_0 + \beta_1 T + \psi_1 Y + \psi_2 TY.$$

The corresponding estimator of $\Delta(Y)$ then can be defined as:

$$\hat{\Delta}(Y) = g^{-1}(\hat{\beta}_0 + \hat{\psi}_1 Y) - g^{-1}(\hat{\beta}_0 + \hat{\psi}_1 Y + \hat{\beta}_1 T + \hat{\psi}_2 TY).$$

Hence, the estimator for $\Theta$ is

$$\begin{aligned} \hat{\Theta} &= \hat{B}_{neg}\cdot\hat{P}_{neg} \\ &= \mathbb{E}_n\{-\hat{\Delta}(Y)|\hat{\Delta}(Y)<0\}\mathbb{E}_n(1_{\hat{\Delta}(Y)<0}), \end{aligned}$$

where $\mathbb{E}_n$ denotes the empirical expectation. The non-smooth region here satisfies $\beta_1 T + \psi_2 TY = 0$.

In this simulation study, we assume logistic regression model for $g(Pr(D=1|T,Y))$. Here is the simulation setting: Treatment $T$ is from bernoulli distribution with probability equals to 0.5. Marker $Y$ is from mixed normal distribution with p.d.f defined as:

$$\mathrm{f}(y) = p \cdot \phi(y; 0, c_1) + (1-p) \cdot \phi(y; 0, c_2),$$

where $\phi$ represents normal density and $p$ is the proportion has values between $(0, 1)$. We fix the variance of $Y$ at 1, and define the variance ratio $k = \frac{c_2}{c_1}$. So during simulation, we vary the distribution of $Y$ through tuning the two parameters, $k$ and $p$. The data generation for the adverse event $D$ follows the logistic regression model, where the parameters $\psi_1 = \psi_2 = 1$, and $\beta_0, \beta_1$ is defined by two tuning parameters $q_0, q_1$ satisfying that:

$$\begin{aligned} q_0 &= \frac{1}{1+e^{-\beta_0}}, \\ q_1 &= \frac{1}{1+e^{-(\beta_0+\beta_1)}}, \end{aligned}$$

where the values of $q_0$ and $q_1$ fall in $(0, 1)$. Because $q_0$ is related to the probability of the adverse event for people with treatment and $q_1$ is the probability of adverse event for people without treatment, we define $q_1 \leq q_0$ such that our setting is making sense. In the simulation, let $q_0 = 0.1, 0.3, 0.5, 0.7$, $q_1 = 0.1, 0.3, 0.5, 0.7$ with $q_1 \leq q_0$. Parameters for marker $Y$ are defined by: $k = 4$ and $p = 0.8$. The sample size will be fixed at $n = 100$ with bootstrap replication $B = 500$ ,and monte carlo replication $Nrep = 200$.

We plug in the three CI methods to construct 95% confidence intervals: central percentile bootstrap confidence interval (CPB), projection bootstrap confidence interval (PCI), and our proposed adaptive projection confidence interval (API). Figure 2.3 shows the estimated coverage rates from these CI methods. Figure 2.4 shows the average width for these CI methods. Similar to other simulation results, CIs constructed via the CPB method have the smallest average width; however, these are associated with under-coverage under non-regular situation. But PCI method always gives the widest CIs, with coverage rates often be 1.0, which are conservative. Our proposed API method with data-driven $\lambda_n$ always offers valid coverage rates.

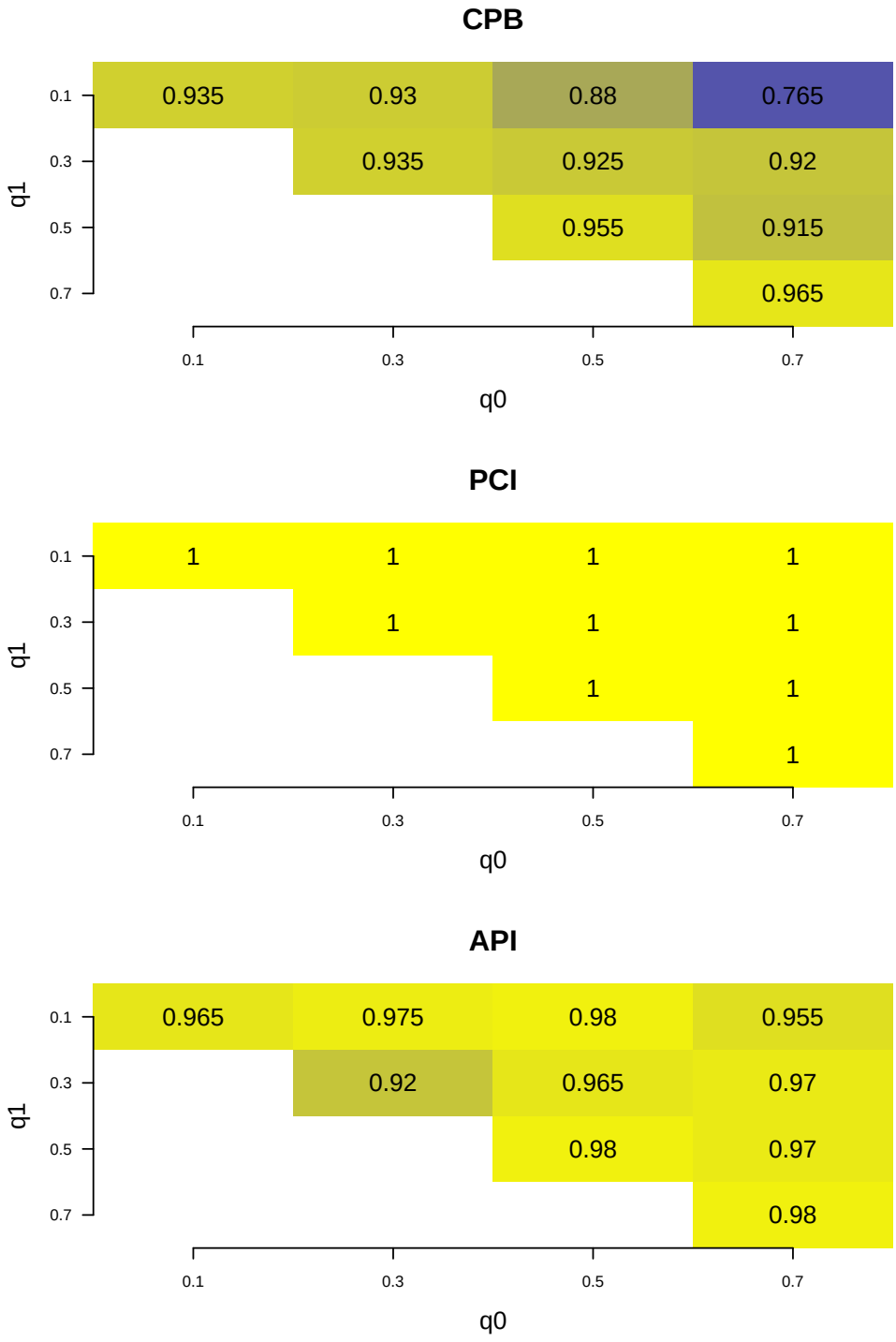


Figure 2.3: Monte Carlo estimates of coverage rates of confidence intervals for the measure related to biomarker evaluation at 95% nominal level. The texts in the plot represent the coverage rate. $x$-axis and $y$-axis are values of the two parameters $q_0, q_1$ for generated models. Sample size $n = 100$, Monte Carlo replication $N_{rep} = 200$, and Bootstrap sample size $B = 1000$.

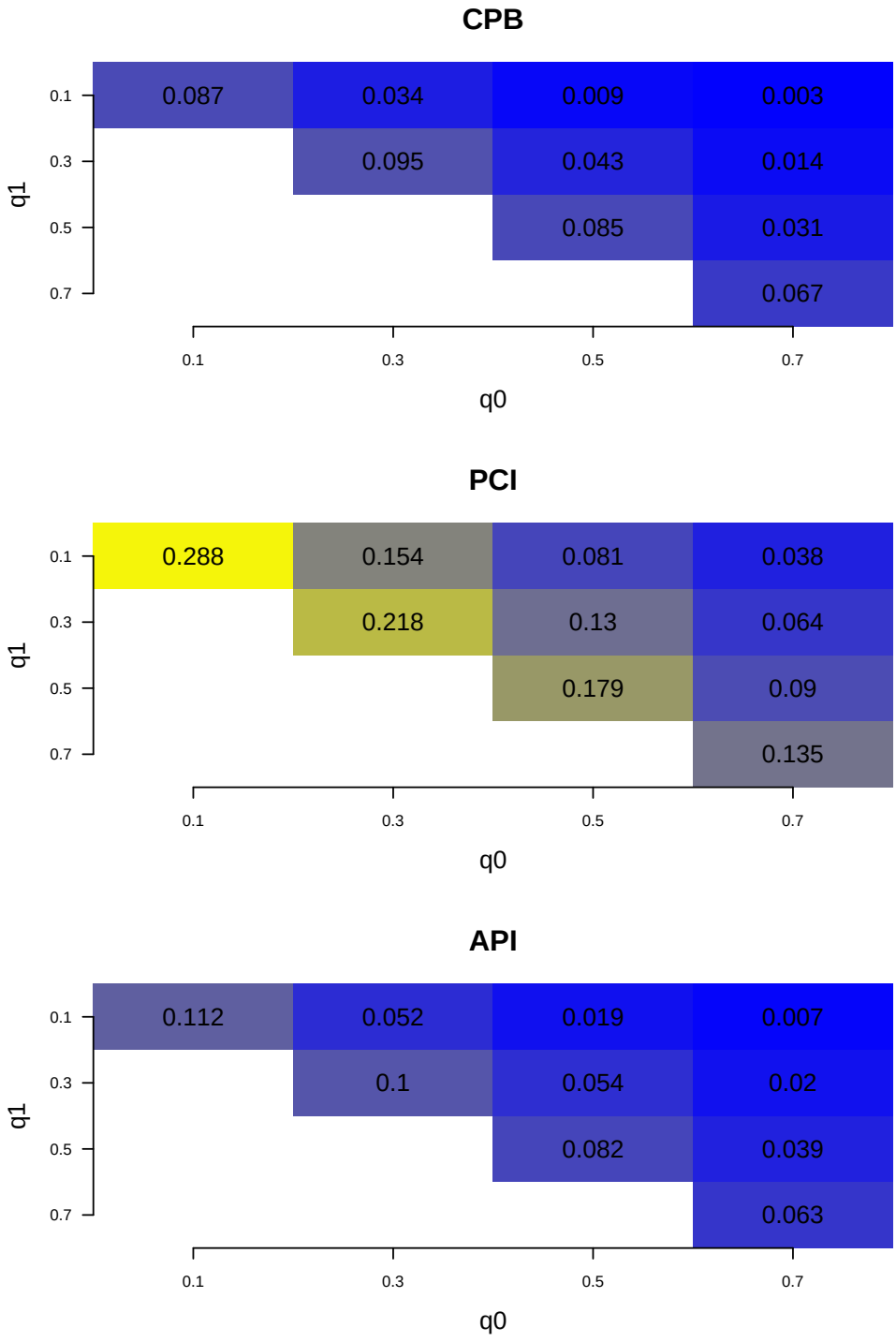


Figure 2.4: Monte Carlo estimates of the mean width of confidence intervals for the measure related to biomarker evaluation at 95% nominal level. The texts in the plot represent the coverage rate. $x$-axis and $y$-axis are values of the two parameters $q_0, q_1$ for generated models. Sample size $n = 100$, Monte Carlo replication $N_{rep} = 200$, and Bootstrap sample size $B = 1000$.

## 2.5 Discussion

In this section, we have illustrated the problem of non-regularity in a class of estimands defined as $\theta^*(\beta^*) = \mathrm{E}s(X_1, \beta_1^*)g(X_2, \beta_2^*)$, where $s(\cdot)$ is smooth, and $g(\cdot)$ is non-smooth in a region $\mathcal{Q}_{(X,\beta^*)}$. We reviewed the existing approaches to construct CIs for this class of estimands, and discussed the limitation of these current methods. We proposed an adaptive projection interval (API) method to construct a asymptotic valid $(1-\alpha-\eta)\times 100\%$ confidence interval for these estimands. The API is an adaptive method that is motivated from projection confidence interval proposed by Robins (2004). The double bootstrap procedure is used to tun the important parameter $\lambda_n$ in API. The API has the advantage of being easy to understand, and simple to program. For illustration, we gave four different simulations: toy example, one-stage value function, first stage coefficients in multi-stage $Q$-learning, and biomarker evaluation. These empirical studies suggest that API always provide valid coverage rate for the estimands.

Even though the empirical results suggest promising performance of the API, theoretical properties are needed. We have derived partial theoretical properties for the proposed API method. These theoretical results are shown in the appendix. Besides, the evaluation of different divisions for $s(\cdot)$ and $g(\cdot)$ may be needed for further discussion.

# Chapter 3

# Case Study for STEP-BD (Systematically Treatment Enhancement Program for Bipolar Disorder)

## 3.1 Introduction of STEP-BD Study

Bipolar disorders are a group of chronic lifelong recurrent psychiatric disorders characterized by episodic shifts in mood, energy, social and vocational functioning, and activity levels (Phillips and Kupfer, 2013). Worldwide, bipolar disorders are a leading cause of disability (Vos et al., 2013) and associated with a substantial economic burden on society (Kleine-Budde et al., 2013). Standard antidepressant medications have been proved to be effective for acute and long-term treatment of unipolar depression (Bauer et al., 2013); however, supporting evidence for the inclusion of standard antidepressants in the acute and long-term treatment of bipolar depression is more limited and controversial (Grunze et al., 2010; Pacchiarotti et al., 2013).

STEP-BD (Systematic Treatment Enhancement Program for Bipolar Disorder) is a long-term study of bipolar disorder funded by the National Institute of Mental Health (NIMH). Its aim was "to generate externally valid answers to treatment effectiveness questions related to bipolar disorder" (Sachs et al., 2003). Figure 3.1 shows the STEP-BD patient registration procedure. Patients of age older than 15 years fulfilling DSM-IV

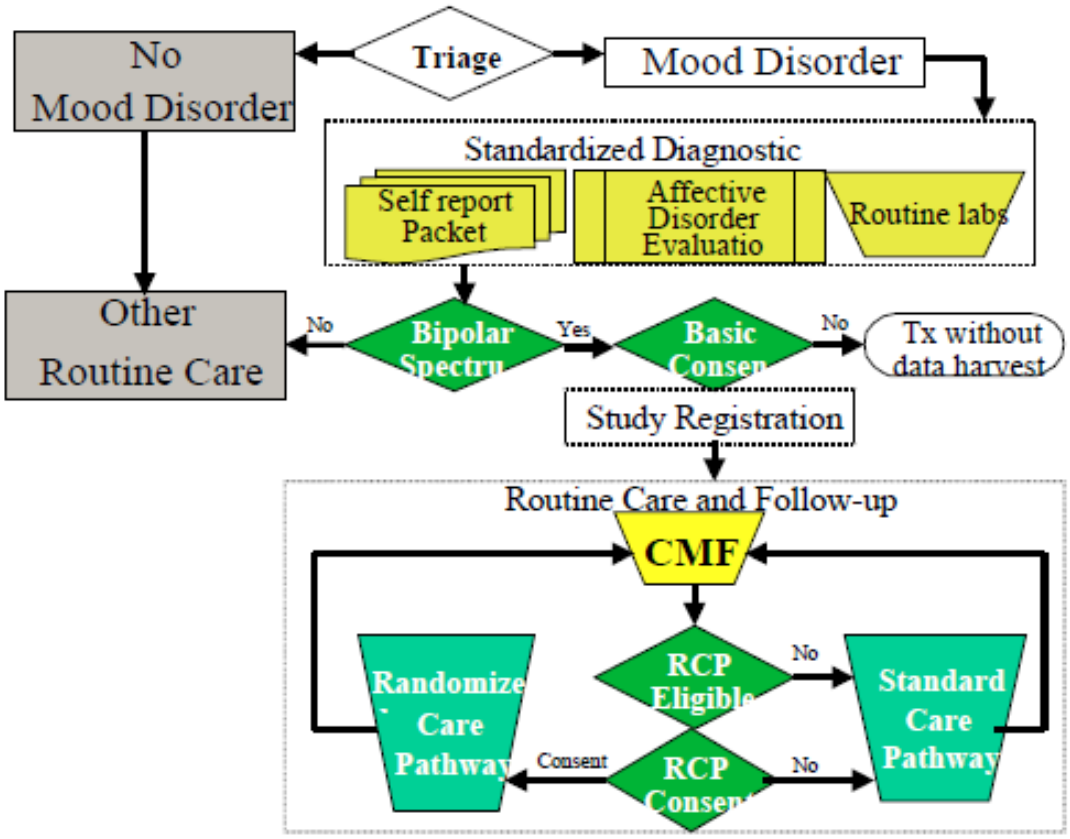


Figure 3.1: Registration procedure for STEP-BD study.

criteria for any subtype of bipolar disorders could enter the study registry. In total, 4,360 patients from 22 sites in United States enrolled. The study lasted for 7 years (2001-2007). In STEP-BD, there are two different treatment pathways: Standard Care Pathway (SCP) and Randomized Care Pathway (RCP). SCP is open to all participants with a diagnosis of bipolar disorders. Each treatment delivered is open and will follow treatment guidelines. Decisions are made on the basis of shared decision making. After patients signed informed consent for entry into the STEP-BD study registry all patients enter the SCP. If a patient's status meets the eligibility criteria at one of the follow-up visits during the SCP for a study within the RCP, additional consent is requested for entry into that RCP. The RCP utilizes methods appropriate for efficacy studies and random assignment is needed to provide answers to clinical questions. In the RCP, there are three different pathways each addressing unmet needs in the treatment of bipolar dis-

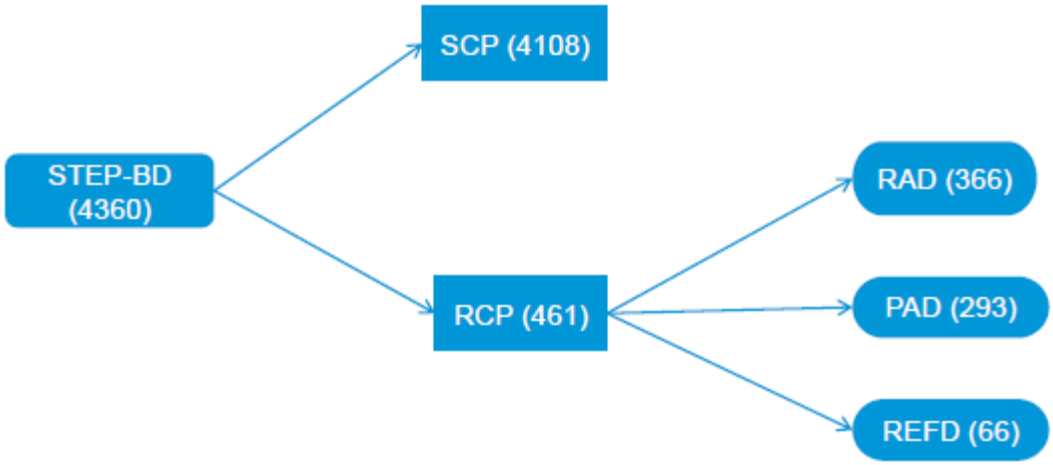


Figure 3.2: Different pathways in STEP-BD study.

order: Acute Depression Randomized Pathway (RAD); Acute Depression Psychosocial Intervention Pathway (PAD); and Refractory Depression Pathway (REFD). If patients are unwilling to consent to one of the RCPs they remain in the SCP. In general, the decision of pathway (SCP versus RCP) is based on both the doctor's and patient's opinion. In STEP-BD, patients could switch pathways based on doctor's or their own preference as well as inclusion and exclusion criteria. Figure 3.2 shows the diagram of STEP-BD study.

In this chapter, we analysis datasets from both randomized trails and observational trails. Section 3.2 is a case study for acute depression randomized pathway (RAD). Section 3.3 is the data analysis for observational acute depression pathway (SAD). Section 3.4 ends with discussion.

## 3.2 A Reanalysis of RAD using $Q$-learning

In this section, our analysis utilizes patients enrolled in RAD. We will first introduce the Acute Depression Randomized Pathway (RAD). We then review the knowledge about dynamic treatment regimes (DTRs) and two-stage $Q$-learning. Data analysis results using $Q$-learning is shown in section 3.2.3.

### 3.2.1 Acute Depression Randomized Pathway (RAD)

As mentioned above, the acute Depression Randomized Pathway (RAD) is one of three RCPs (Randomized Care Pathways) in STEP-BD. In addition to satisfying the general entry criteria for STEP-BD study registry, patients had to be at least 18 years old and fulfill the DSM-IV criteria for a major depressive episode in the context of bipolar I or bipolar II disorder. The goal of RAD was to explore the effectiveness of adjunctive antidepressant treatment. All patients with a history of intolerance or nonresponse to both bupropion and paroxetine were excluded, as well as those requiring current short-term treatment for a coexisting substance-abuse disorder or requiring the addition of antipsychotic medication or a change in the dose of a long-term antipsychotic medication (Sachs et al., 2007). In addition, patients had to take a mood stabilizer at the time of randomization or agree to begin treatment with a mood stabilizer. Moreover they had to agree to have all non-study antidepressants tapered after initiation of study drug, with the antidepressant discontinued by the end of week 2. The purpose of RAD was to explore the effectiveness of adjunctive antidepressant treatment, in addition to a mood stabilizer. Initially the mood stabilizers were limited to lithium, valproate, the combination of lithium and valproate, or carbamazepine. However, later on, any FDA-approved antimanic agent could be used as mood stabilizers. At week 0, patients were randomly assigned to one antidepressant (150 mg of a sustained release formulation of bupropion or 10 mg of paroxetine to begin with) or placebo. After 6 weeks, patients with non-response on the placebo were randomized to either paroxetine or bupropipon; patients with non-response on the antidepressant were assigned to either openly increase the dose of their current antidepressant or add another antidepressant. At week 8, 10, or 12 clinicians will make final decision for patients based on their clinical status collected from CMF (clinical monitoring form). During the study patients need to visit their doctors every week to fill in the Clinical Monitoring Form (CMF, Sachs et al. 2002). Patients were allowed to switch to SCP (opt out) at any time by their preference or doctor's opinion. Patients who had severe adverse effects or met criteria for hypomania or mania discontinued the antidepressant or placebo and received open treatment while remaining in STEP-BD. Since after 6 weeks, only one patient with non-response on the antidepressant was assigned to add another antidepressant, we ignored this one observation and supposed patients with non-response to antidepressant after 6 weeks were only assigned to increase the dose of their current antidepressant. Figure 3.3 shows the RAD diagram.

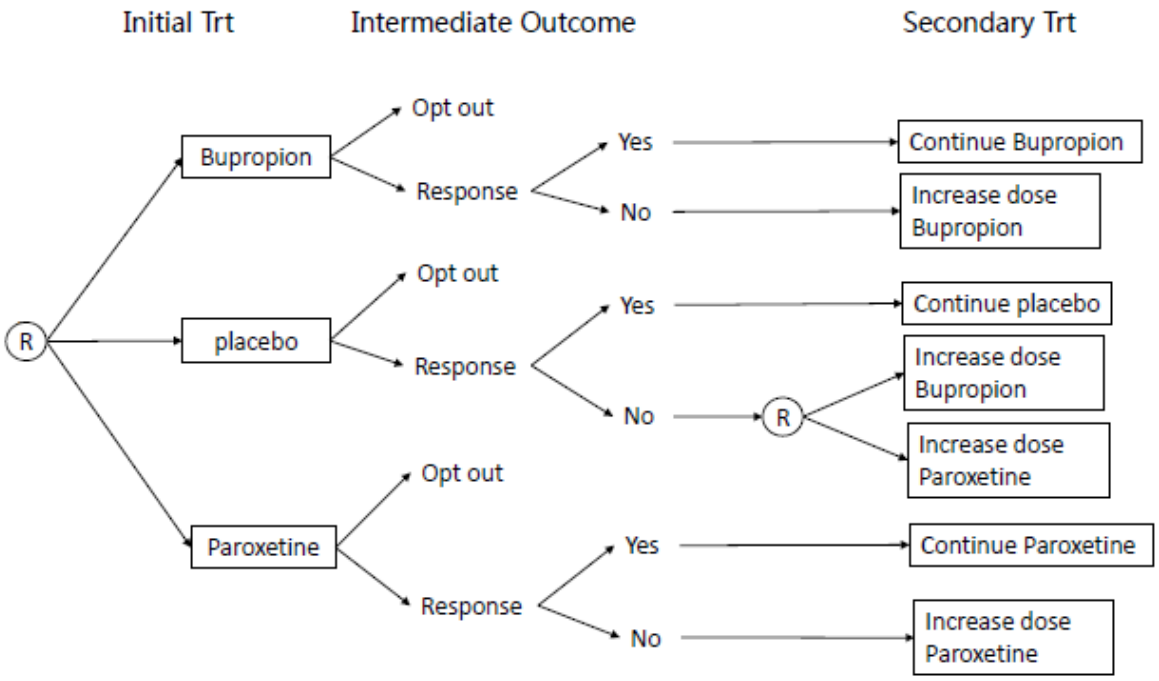


Figure 3.3: At the beginning (stage 1), there are 365 patients in total. 85 patients take Bupropion, 93 patients take Paroxetine and 187 patients take placebo. After 6 weeks, 104 patients' information are lost. Only 78 patients are tracked with non-response at the end of stage 1. At stage 2, patients with non-response are assigned to secondary treatment intervention. Patients taking Bupropion or Paroxetine at stage 1 will increase current doses. But Patients taking placebo at stage 1 will be assigned Bupropion or Paroxetine.

Response for a given subject was defined as at least 50% improvement over their initial SUM-D score and not meeting the DSM-IV criteria for hypomania or mania. Scores on the continuous symptom subscale for depression (SUMD) range from 0 to 22, with higher scores indicating more severe symptoms. Both SUM-D and SUMM (symptom subscale for mood elevation, SUMM scores range from 0 to 16) are part of the modified Clinical Monitoring Form for mood disorders (Sachs et al., 2002).

Because subjects in RAD are potentially randomized multiple times with randomizations occurring at crucial points of the disease process, RAD is an example of a Sequential Multiple Assignment Randomized Trial (Murphy, 2003a; Lavori and Dawson, 2004, SMART). Data collected in SMARTs can be used to efficiently estimate and

evaluate DTRs (dynamic treatment regimes). In the next section we will formalize the notion of an optimal DTR and introduce a regression-based approach called $Q$-learning for estimating an optimal DTR from a SMART.

### 3.2.2 Dynamic Treatment Regimes and $Q$-learning

The effective management of a chronic illness requires ongoing personalized treatment (Wagner et al., 2001). Dynamic treatment regimes (DTRs) formalize clinical decision making as sequence of *decision rules*, one per treatment decision, that map patient information to a recommended treatment. An optimal DTR yields the minimal mean outcome when applied to assign treatment to a population of interest. One method for estimating an optimal DTR from observational or randomized study data is $Q$-learning (Murphy, 2005a; Schulte et al., 2012). $Q$-learning is an approximate dynamic programming algorithm that can be viewed as an extension of regression to multi-stage decision problems (Nahum-Shani et al., 2012). As our focus is the application of $Q$-learning to the RAD study within the RCP pathway, we focus on data from a two-stage randomized trial with a terminal continuous outcome; however, $Q$-learning applies in much more general settings (Watkins and Dayan, 1992; Sutton and Barto, 1998; Goldberg and Kosorok, 2012; Schulte et al., 2012; Laber et al., 2014c; Moodie et al., 2013).

$Q$-learning estimates an optimal regime using backward induction. For simplicity we assume that the entire treatment period contains two stages with a distal outcome measured after completion of the second stage; treatment decisions are made in the beginning of each stage. $Q$-learning proceeds in two steps. In the first step it estimates an optimal treatment rule for the second stage of treatment given patient-level data accumulated up to and immediately preceding this second treatment assignment. This information includes each patient's baseline information, stage 1 treatment assignment and intermediate, i.e., proximal, outcomes measured during the course of the first stage of treatment. These inputs to the second stage rule are treated as "independent variables" with no attempt to infer what decision at stage 1 would be optimal for a given patient. This first step is achieved by regressing the distal outcome on patient information up to decision stage 2 and manipulating the obtained analytic expression to find for each patient which treatment at stage 2 optimizes the expected distal outcome.

At the second step, $Q$-learning looks for treatment assignment at stage 1 that would result in optimal distal outcome, assuming that subsequent stage 2 treatment will be

determined by the rule constructed in step 1 of the procedure. Such backward reasoning allows $Q$-learning to factor in future decisions when making treatment decisions at earlier stages. This can be contrasted with a myopic strategy that only looks at intermediate (proximal) outcomes of a current treatment assignment. For example, treatments at stage 1 may lead to temporary alleviation of symptoms and therefore appear beneficial for a patient, however, the long-term benefits may become questionable after the later (e.g., second) stage decisions are factored in.

We now present formal mathematical description of $Q$-learning. We assume that data available to estimate a DTR are in the form of $n$ independent, identically distributed trajectories $\{(X_{1i}, A_{1i}, X_{2i}, A_{2i}, Y_i)\}_{i=1}^n$, one for each subject where: $X_1 \in \mathbb{R}^{p_1}$ denotes baseline (pre-randomization) subject information; $A_1 \in \mathcal{A}_1$ denotes the first stage treatment assignment; $X_2 \in \mathbb{R}^{p_2}$ denotes information collected during the course of the first stage treatment including information dictating first-stage responder status; $A_2 \in \mathcal{A}_2$ denotes the second stage treatment assignment; and $Y \in \mathbb{R}$ denotes a continuous outcome measured at the end of the study coded so that lower values are better. To match the RAD study, we assume that responders are not re-randomized. In the RAD study, $X_1$ contains a subject's age, race, gender, marital status, annual household income, employment status, education level, nine different side effect measures, medical insurance type, as well as baseline measures of bipolar type, clinical status prior to depressive episode, scale scores for mood elevation (SUMM), and scale scores for depression (SUMD); $A_1$ denotes low-dose Bupropion, low-dose Paroxetine, or placebo; $X_2$ contains responder status at the end of stage 1, as well as SUMM and SUMD at the end of stage 1; $A_2$ denotes either high-dose Bupropion or high-dose Paroxetine; $Y$ is SUMD measured at the end of stage 2.

Define $H_1 = X_1$ and $H_2 = (X_1^T, A_1, X_2^T)^T$, so that $H_j$ denotes information available to the decision maker at stage $j = 1, 2$. A DTR is pair of functions $\pi = (\pi_1, \pi_2)$ where $\pi_j : \operatorname{dom} H_j \to \operatorname{dom} A_j$ so that a patient presenting with $H_j = h_j$ at stage $j$ is assigned treatment $\pi_j(h_j)$. For any $h_j \in \operatorname{dom} H_j$, let $\mathcal{F}_j(h_j)$ denote the set of feasible treatments for a patient presenting at stage $j$ with $H_j = h_j$. In the RAD study $\mathcal{F}_1(h_1) = \{\text{Bupropion, Paroxetine, placebo}\}$. At the second stage responders are not

re-randomized; feasible second stage treatments for non-responders are

$$\mathcal{F}_2(h_2) = \begin{cases} \{\text{high–dose Buproprion}\} & \text{if } A_1 = \text{ Buproprion,} \\ \{\text{high–dose Paroxetine}\} & \text{if } A_1 = \text{ Paroxetine,} \\ \{\text{high–dose Buproprion, high–dose Parotexine,}\} & \text{if } A_1 = \text{placebo.} \end{cases}$$

Let $\Pi = \{\pi = (\pi_1, \pi_2),: \ \pi_j(h_j) \in \mathcal{F}_j(h_j), \forall h_j \in \text{dom}\, H_j\}$ denote the class of *feasible* DTRs (for a more formal discussion of feasibility see (Schulte et al., 2012)). An optimal DTR, say $\pi^{\text{opt}}$, satisfies $\text{E}^{\pi^{\text{opt}}} Y \geq \text{E}^{\pi} Y$ for all $\pi \in \Pi$, where $\text{E}^{\pi}$ denotes expectation under the restriction that $A_j = \pi_j(H_j)$. Define $Q_2(h_2, a_2) = \text{E}(Y|H_2 = h_2, A_2 = a_2)$ and $Q_1(h_1, a_1) = \text{E}\min_{a_2} Q_2(H_2, a_2)|H_1 = h_1, A_1 = a_1)$. The function $Q_2(h_2, a_2)$ measures the 'quality' of assigning treatment $a_2$ to a patient presenting at stage two with $H_2 = h_2$; the function $Q_1(h_1, a_1)$ measures the quality of assigning treatment $a_1$ to a patient presenting at stage one with $H_1 = h_1$ assuming optimal subsequent treatment. It follows from dynamic programming (Bellman, 1957) that $\pi_j^{\text{opt}}(h_j) = \arg\min_{a_j \in \mathcal{F}_j(h_j)} Q_j(h_j, a_j)$. In practice, dynamic programming cannot be applied because the true $Q$-functions are not known; instead, estimation of $\pi^{\text{opt}}$ must rely on the observed data. $Q$-learning is an approximate dynamic programming algorithm which mimics the dynamic programming solution by replacing the conditional expectations required by dynamic programming with regression models fit to the observed data. Let $Q_j(h_j, a_j; \theta_j)$ denote a postulated working model for $Q_j(h_j, a_j)$ indexed by unknown parameter $\theta_j$.

In RAD, only patients who receive placebo as their first stage treatment and failed to respond are randomized at the second stage. Thus, we only use these subjects to estimate $\pi_2$ Let $R$ denote a subjects first-stage responder status so that $R = 1$ for responders and $R = 0$ for non-responders. A version of the $Q$-learning algorithm that applies to data from RAD is:

**Algorithm 3.1: Q-learning for RAD**

**(Q1)** Compute $\widehat{\theta}_2 = \arg\min_{\theta_2} \sum_{i=1}^{n} \{Y_i - Q_2(H_{2i}, A_{2i}; \theta_2)\}^2 1_{A_{1i}=\text{placebo}}(1 - R_i)$; and subsequently estimator $Q_2(h_2, a_2; \widehat{\theta}_2)$ of $Q_2(h_2, a_2)$.

**(Q2)** Define $\widehat{Y}_i = 1_{A_{1i}=\text{placebo}}(1-R_i) \min_{a_2 \in \mathcal{F}_2(H_{2i})} Q_2(H_{2i}, a_2; \widehat{\theta}_2) + (1_{A_{1i} \neq \text{placebo}} + R_i 1_{A_{1i}=\text{placebo}}) Y_i$.

**(Q3)** Compute $\widehat{\theta}_1 = \arg\min_{\theta_1} \sum_{i=1}^{n} \left\{\widehat{Y}_i - Q_1(H_{1i}, A_{1i}; \theta_1)\right\}^2$ and subsequently estimator $Q_1(h_1, a_1; \widehat{\theta}_1)$ of $Q_1(h_1, a_1)$.

The $Q$-learning estimated optimal regime is $\widehat{\pi}_j(h_j) = \arg\min_{a_j \in \mathcal{F}_j(h_j)} Q_j(h_j, a_j; \widehat{\theta}_j)$. To estimate $\pi^{\text{opt}}$ using data from the RAD study, we assume linear working models for the $Q$-functions. For $Q_1(h_1, a_1)$ we posit a model of the form $Q_1(h_1, a_1; \theta_1) = h_{10}^T\beta_{10} + a_{11}h_{11}^T\beta_{11} + a_{12}h_{12}^T\beta_{12}$, where $\theta_1 = (\beta_{10}^T, \beta_{11}^T, \beta_{12}^T)^T$, $h_{1k}$, $k = 0, 1, 2$ are known summary vectors of $h_1$, and $a_{1k}, k = 1, 2$ are dummy variables coding two of the three possible treatments at the first stage. For $Q_2(h_2, a_2)$ we posit a model of the form $Q_2(h_2, a_2; \theta_2) = h_{20}^T\beta_{20} + a_2h_{21}^T\beta_{21}$, where $\theta_2 = (\beta_{20}^T, \beta_{21}^T)^T$, $h_{2k}$, $k = 0, 1$ are known summary vectors of $h_2$, and $a_2$ is a dummy variable coding one of the two possible treatments at the second stage.

### 3.2.3 Data Analysis for RAD

The $Q$-learning algorithm stated in the preceding section assumes: (i) complete data; and (ii) working models for the $Q$-functions. However, in RAD, as in most clinical trials, a non-trivial amount of covariate and outcome information are missing. Furthermore, there is not strong theory to suggest working models for the $Q$-functions so we must use the data to assist in the choice of these models. We combine multiple imputation with stepwise variable selection to estimate the $Q$-functions and subsequently the optimal treatment regime.

#### Missing data

Figure 3.4 shows the fraction of missing data for the variables under consideration in our analysis of the RAD data. There is a significant amount of missing covariate information at both stages; thus, discarding subjects with missing information is inefficient and may introduce bias (Little and Rubin, 2002).

One approach to dealing with missing data is multiple imputation(Rubin, 2004, MI). MI creates multiple complete datasets and is thereby suited for conducting a series of exploratory and secondary analyses including estimation of an optimal treatment regime (Shortreed et al., 2011). We use Bayesian MI to "fill in" the missing values with draws from the posterior predictive distribution of the missing values given the observed data (for details and underlying assumptions see Little and Rubin, 2002; Van Buuren, 2012). Implementation of Bayesian MI requires specification of a prior and likelihood for the observed data. We specify the joint likelihood through the conditional distribution of each variable on all other variables (for discussion of this approach see Raghunathan et al., 2001; Van Buuren et al., 2006; Van Buuren, 2007). Thus, the likelihood is determined

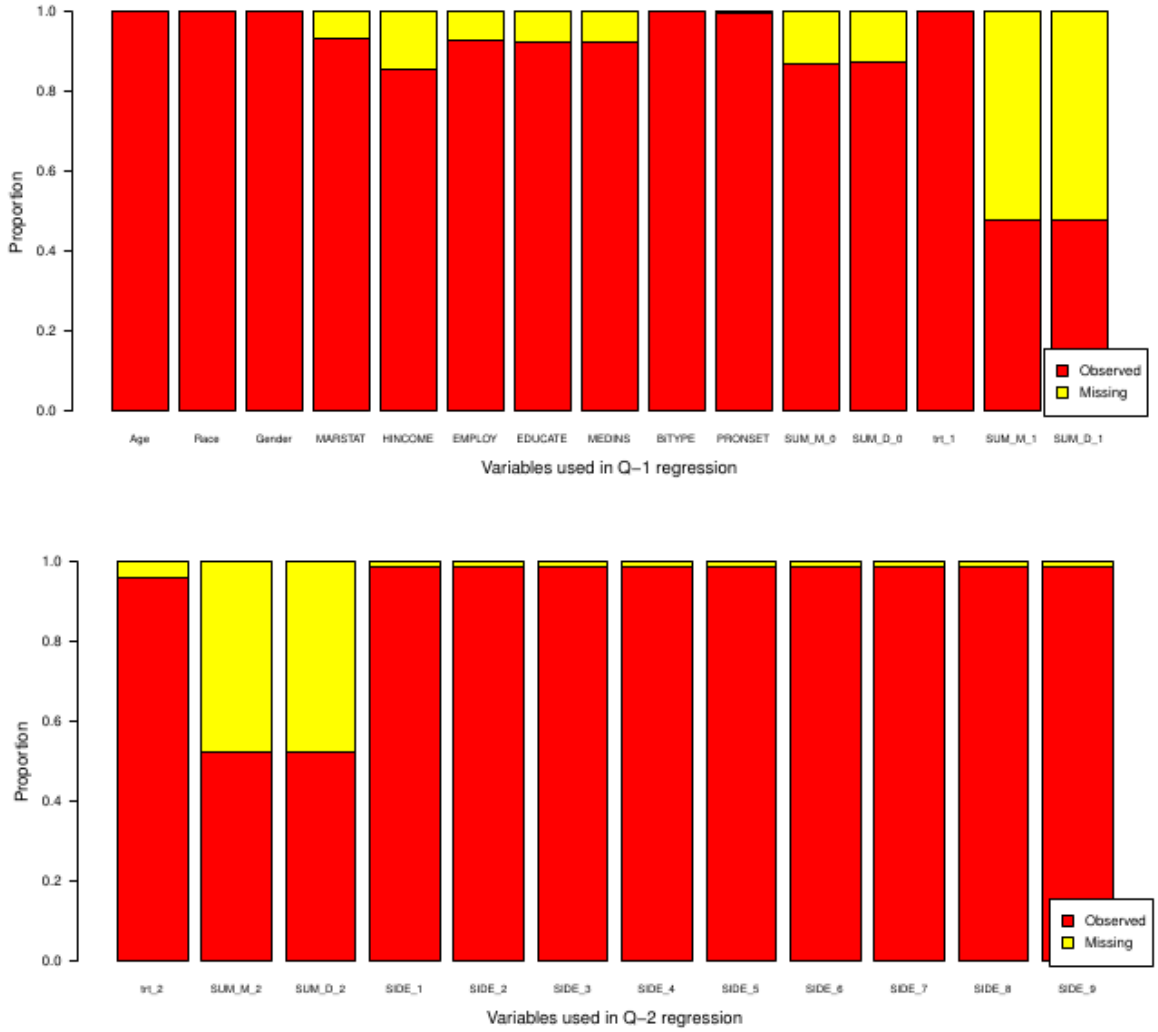


Figure 3.4: Variables with missing data are listed. The SUMM$i$ and SUMD$i$ denote continuous symptom subscales for depression and mood elevation at i th stage. The Trt$i$ denotes current treatment at stage i. The response$i$ denotes patients' clinical status at the end of stage i. The SIDE$j$ represents different side effects. PRONSET denotes patients' prior to onset clinical status. EDUCATE, EMPLOY, MARSTAT, MEDINS and HINCOME are the indicators for patients' education level, employment status, marriage status, medical insurance and annual home income respectively.

implicitly through a series of regression models, one for each variable that contains missing information. For continuous variables we use predictive mean matching and for binary variables we use logistic regression models. To reduce variance, we use forward stepwise variable selection applied to the complete data to select predictors for each conditional model. Flat improper priors were used for all parameters. Imputations were carried out using the freely available and open source `mice` package with the default settings (http://cran.r-project.org/web/packages/mice/index.html).

Using the procedure described above we impute $m$ complete datasets. For a given choice of $h_{1,k},\ k = 0, 1, 2$ and $h_{2,k},\ k = 0, 1$, we can apply the $Q$-learning algorithm to each imputed dataset to obtain estimated $Q$-functions $Q_j(h_j, a_j; \widehat{\theta}_j^{(\ell)})$, $j = 1, 2$, $\ell = 1, \ldots, m$. The final estimated optimal decision rule is obtained as the minimizer of the averaged imputed $Q$-functions $\widehat{\pi}_j(h_j) = \arg\min_{a_j \in \mathcal{F}_j(h_j)} m^{-1} \sum_{\ell=1}^{m} Q_j\left(h_j, a_j; \widehat{\theta}_j^{(\ell)}\right)$.

**Variable Selection and estimation of the optimal treatment regime**

In order to estimate an optimal treatment regime using $Q$-learning we need to select which covariates to include the models for the $Q$-functions. Recall that the 12 week depression SUMD score was used as the outcome ($Y$). We identified 23 potential predictors; these predictors are listed in Table 3.1. We select a subset of these predictors for each $Q$-function using stepwise variable selection to minimize the Bayes Information Criterion (BIC, Schwarz, 1978) averaged over the multiply imputed data sets. Let $\mathcal{M}_2$ denote a subset of predictors dictating the features $h_{2,k},\ k = 0, 1$ and let $\widehat{\theta}_2^{(\ell)}(\mathcal{M}_2)$ denote the coefficients obtained by applying step (Q1) of the $Q$-learning algorithm with predictors $\mathcal{M}_2$ to the $\ell$th imputed data set. Define $\widehat{Y}_i^{(\ell)}(\mathcal{M}_2)$, $i = 1, \ldots, n$ to be the predicted outcomes computed in step (Q2) of the $Q$-learning algorithm using the $\ell$th imputed dataset and parameter $\widehat{\theta}_2^{(\ell)}(\mathcal{M}_2)$. Let $\mathcal{M}_1$ denote a subset of predictors dictating $h_{1,k},\ k = 0, 1, 2$ and let $\widehat{\theta}_1^{(\ell)}(\mathcal{M}_1, \mathcal{M}_2)$ denote the coefficients estimated in step (Q3) of the $Q$-learning algorithm using predictors $\mathcal{M}_1$ and predicted outcomes $\widehat{Y}_i^{(\ell)}(\mathcal{M}_2)$, $i = 1, \ldots, n$. In addition, let $\text{BIC}^{(\ell)}(\mathcal{M}_2)$ denote the BIC for a second stage model $Q(h_2, a_2; \widehat{\theta}_2^{(\ell)}(\mathcal{M}_2))$ calculated on the $\ell$th imputed dataset. Similarly, and let $\text{BIC}^{(\ell)}(\mathcal{M}_1, \mathcal{M}_2)$ denote the BIC for the first stage model $Q(h_1, a_1; \widehat{\theta}_1^{(\ell)}(\mathcal{M}_1, \mathcal{M}_2))$ calculated on the $\ell$th imputed dataset. This procedure that we use to construct models for the $Q$-learning algorithm is:

Table 3.1: Candidate predictors for regression models in $Q$-learning. Those that are only available for the second stage regression model are starred.

| Variable | Description | Type | Values (range or level) | Mean (SD) or Frequency (%) |
|---|---|---|---|---|
| AGE | Age at entry (years) | Numerical | 18-77 | 40.59 (11.74) |
| RACE | Race | Binary | white or causasian, non white | 90.4%, 9.6% |
| GENDER | Gender | Trinary | male, female, transgender | 43%, 56%, 1% |
| MARSTAT | Marital status | Trinary | never married, married, separated | 35.6%, 33.8%, 30.6% |
| HINCOME | Annual household income ($\times$\$1000) | Binary | $< 40$, $\geq 40$ | 58.5%, 41.5% |
| EMPLOY | Employment status | Binary | employed, unemployed | 46.9%, 53.1% |
| EDUCATE | Education level | Binary | college or below, technical school or above | 53%, 47% |
| MEDINS | Indicator of medical insurance | Binary | yes, no | 72.8%, 27.2% |
| BITYPE | Bipolar type at entry | Binary | type I, type II | 70.4%, 29.6% |
| PRONSET | Clinical status before depressive episode | Trinary | remission, (hypo)manic, mixed | 45.9%, 33.2%, 20.9% |
| SUMD0 | Scaled depression at entry | Numerical | 0.75-18 | 7.47 (2.30) |
| SUMD1* | Scaled depression at the end of stage 1 | Numerical | 0-14 | 4.49 (3.07) |
| SUMM0 | Scaled mood elevation at entry | Numerical | 0-7 | 1.19 (1.09) |
| SUMM1* | Scaled mood elevation at the end of stage 1 | Numerical | 0-6.75 | 0.95 (1.30) |
| Trt1* | Treatment received at stage 1 | Trinary | Bupropion, Paroxetine, placebo | 23.3%, 25.5%, 51.2% |
| SIDE1 | Tremor | Binary | yes, no | 26.9%, 73.1% |
| SIDE2 | Dry mouth | Binary | yes, no | 21.1%, 78.9% |
| SIDE3 | Sedation | Binary | yes, no | 17.1%, 82.9% |
| SIDE4 | Constipation | Binary | yes, no | 5.7%, 94.3% |
| SIDE5 | Diarrhea | Binary | yes, no | 12%, 88% |
| SIDE6 | Headache | Binary | yes, no | 13.7%, 86.3% |
| SIDE7 | Poor memory | Binary | yes, no | 14.3%, 85.7% |
| SIDE8 | Sexual dysfunction | Binary | yes, no | 9.7%, 90.3% |
| SIDE9 | Increased appetite | Binary | yes, no | 12.6%, 87.4% |

Table 3.2: Point estimates and confidence intervals for the coefficients indexing the second stage $Q$-function.

| Variable | Coefficient Estimate | 90% Confidence Interval |
|---|---|---|
| SUMM1 | 0.18 | (-0.14, 0.90) |
| SUMD1 | 0.50 | (0.48, 0.83) |
| SIDE3 | -0.41 | (-2.65, 0.42) |
| Intercept | 1.21 | (0.05, 2.00) |
| $A_2 \times$ SUMM1 | 0.77 | (-0.16, 1.18) |
| $A_2 \times$ SIDE3 | 1.82 | (-1.05, 3.18) |
| $A_2$ | -1.18 | (-1.98, 0.00) |

(S1) Using forward variable selection compute

$$\widehat{\mathcal{M}}_2 = \arg\min_{\mathcal{M}_2} \frac{1}{m} \sum_{\ell=1}^{m} \text{BIC}^{(\ell)}(\mathcal{M}_2);$$

(S2) Using forward variable selection compute

$$\widehat{\mathcal{M}}_1 = \arg\min_{\mathcal{M}_1} \frac{1}{m} \sum_{\ell=1}^{m} \text{BIC}^{(\ell)}(\mathcal{M}_1, \widehat{\mathcal{M}}_2);$$

(S3) Let $Q_2(h_2, a_2; \widehat{\theta}_2(\widehat{\mathcal{M}}_2))$ and $Q_1(h_1, a_2; \widehat{\theta}_1(\widehat{\mathcal{M}}_1, \widehat{\mathcal{M}}_2))$ denote the second and first stage estimated $Q$-functions respectively.

The variables included in the model $\widehat{\mathcal{M}}_2$ are: SIDE3, SUMD1 and SUMM1. The variables included in the model $\widehat{\mathcal{M}}_1$ are: AGE, PRONSET, SUMD0 and SUMM0. Thus, the second stage $Q$-functions has the form $Q_2(h_2, a_2; \theta_2) = \beta_{20}^T h_{20} + a_2 \beta_{21}^T h_{21}$, where $h_{20} = (1, SUMM1, SUMD1, SIDE3)^T$, $h_{21} = (1, SUMM1, SIDE3)^T$, and $A_2$ is indicator variable for stage 2 treatment coded so that $A_2 = 1$ denotes high-dose Bupropion and $A_2 = 0$ denotes high-dose Paroxetine. The estimated coefficients $\widehat{\beta}_{20}, \widehat{\beta}_{21}$ along with 90% adaptive projection bootstrap confidence intervals (API) are shown in Table 3.2. The table shows a significant main effect of $A_2$ as well as significant interaction between second $A_2$ and SUMM1. To visualize the optimal decision rule we approximated $\widehat{\pi}_2(h_2)$ with a decision tree; this tree was fit by applying the classification and regression tree (CART) algorithm (Breiman et al., 1984) to $\{(H_{2i}, \widehat{\pi}_2(H_{2i}))\}_{i:A_{1i}=placebo \text{ and } R_i=1}$. This

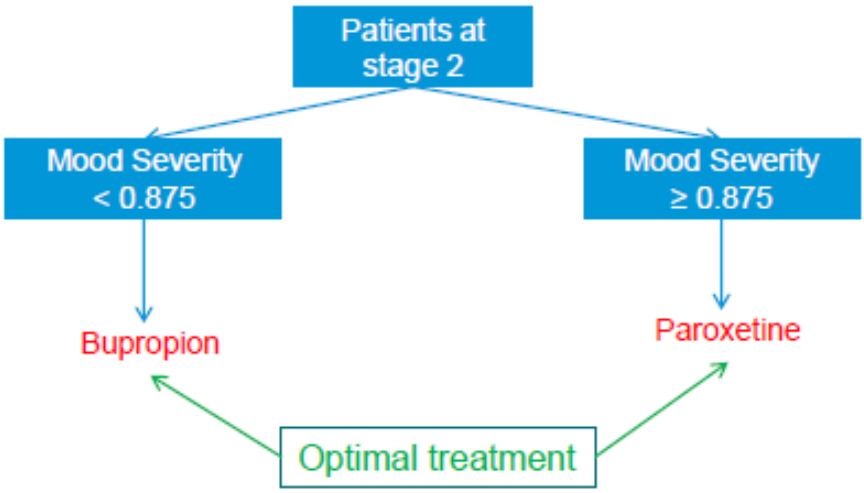


Figure 3.5: Estimated optimal second stage decision rule. The $Q$-learning estimated optimal second stage decision rule represented as a tree. As anticipated by the estimated second stage $Q$-function, SUMM1 (scale score for mood elevation) is used to dictate treatment. The tree was fit using the CART algorithm (Breiman et al., 1984) to the data $\{(H_{2i}, \widehat{\pi}_2(H_{2i}))\}_{i:A_{1i}=placebo \text{ and } R_i=1}$.

tree is displayed in Figure 3.5. As anticipated by estimated second stage $Q$-function, SUMM1 (mood severity) dictates treatment selection; subjects with low mood severity are recommended to high-dose Bupropion and those with high mood severity are recommended to high-dose Paroxetine.

The first stage $Q$-function has the form $Q_1(h_1, a_1; \theta_1) = \beta_{10}^T h_{10} + a_{11}\beta_{11}^T h_{11} + a_{12}\beta 12^T h_{12}$, where:

$$\begin{aligned} h_{10} &= (1, \text{AGE}, \text{SUMM0}, \text{SUMD0}, \text{PRONSET1}, \text{PRONSET2})^T; \\ h_{11} &= (1, \text{SUMM0}, \text{PRONSET1}, \text{PRONSET2})^T; \\ h_{12} &= (1, \text{SUMM0}, \text{PRONSET1}, \text{PRONSET2})^T; \end{aligned}$$

$a_{11} = 1$ if $a_1 =$ Bupriopion otherwise $a_{11} = 0$; $a_{12} = 1$ if $a_1 =$ Paroxetine otherwise $a_{12} = 0$; PRONSET1 = 1 if PRONSET = remission otherwise PRONSET1 = 0; and PRONSET2 = 1 if PRONSET = manic or hypo–manic otherwise PRONSET2 = 0. The estimated coefficients and 90% confidence intervals are listed in Table 3.3. Figure 3.6

Table 3.3: Point estimates and confidence intervals for the coefficients indexing the first stage $Q$-function.

| Variable | Coefficient Estimate | 90% Confidence Interval |
|---|---|---|
| AGE | 0.02 | (-0.01, 0.04) |
| SUMM0 | 0.48 | (0.35, 0.70) |
| SUMD0 | 0.20 | (0.15, 0.36) |
| PRONSET1 | -0.42 | (-1.07, 0.42) |
| PRONSET2 | -0.86 | (-1.49, -0.05) |
| Intercept | 1.57 | (-0.49, 2.50) |
| $A_{11} \times$ AGE | 0.01 | (-0.04, 0.07) |
| $A_{11} \times$ PRONSET1 | 0.66 | (-0.80, 2.29) |
| $A_{11} \times$ PRONSET2 | 1.13 | (-0.55, 2.82) |
| $A_{11}$ | -1.55 | (-4.07, 1.25) |
| $A_{12} \times$ AGE | -0.03 | (-0.08, 0.02) |
| $A_{12} \times$ PRONSET1 | 0.79 | (-0.51, 1.92) |
| $A_{12} \times$ PRONSET2 | 1.62 | (0.05, 2.90) |
| $A_{12}$ | 0.73 | (-1.48, 3.08) |

shows the first stage optimal decision rule implied by the estimated $Q$-function approximated as a decision tree. An interesting feature of the first stage decision rule is that subjects with a prior (hypo) manic episode are recommended to receive placebo. This supports the hypothesis that subjects with a prior (hypo) manic episodes might not benefit from an adjutant antidepressant.

Recall that the optimal treatment regime minimizes the expected depression score SUMD measured at week 12. Thus, it is of interest to compare the estimated expected 12 week SUMD under the estimated optimal treatment regime and other potential treatment regimes of interest. Table 3.4 shows the estimated depression score under the estimated regime and four static treatment regimes. Estimates were computed using the inverse-probability weighted estimator (IPWE Zhang et al., 2013) and confidence intervals using the $n$-out-of-$n$ non-parametric bootstrapping method. The estimated optimal regime performs significantly better than any fixed regime under consideration.

### 3.2.4 Discussion

We estimated an optimal DTR for patients presenting with bipolar depression using data from the RAD pathway in the STEP-BD study. The estimated treatment regime

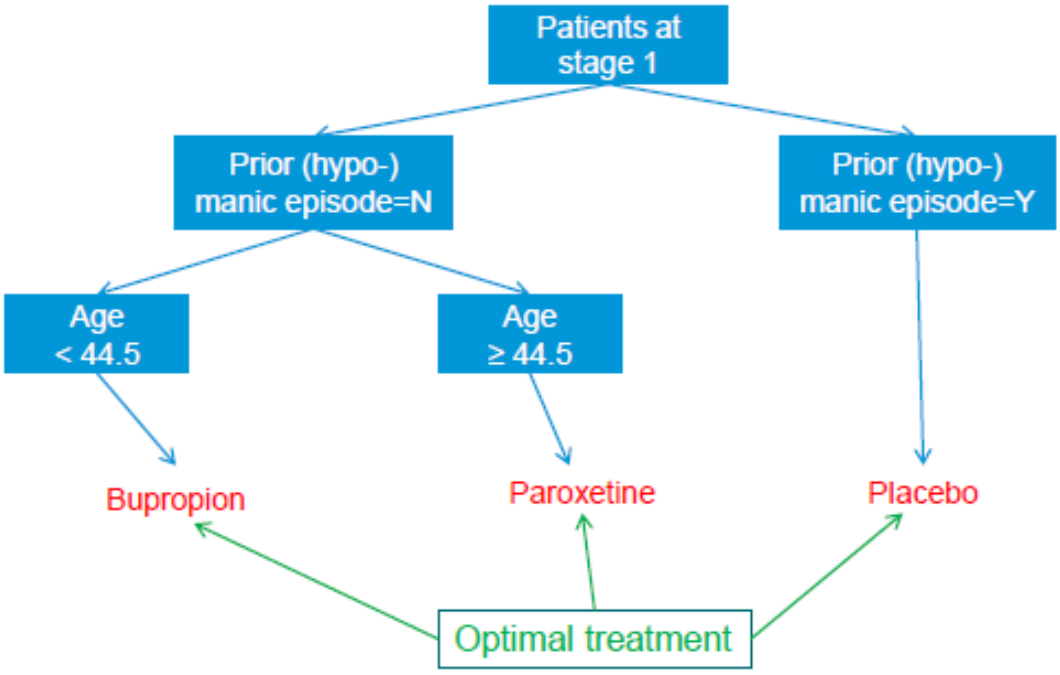


Figure 3.6: Estimated optimal first stage decision rule. The $Q$-learning estimated optimal first stage decision rule represented as a tree. Note that subjects with a prior (hypo) manic episodes are recommended to receive placebo. The tree was fit using the CART algorithm (Breiman et al., 1984) to the data $\{H_{1i}, \widehat{\pi}_1(H_{1i})\}_{i=1}^n$.

Table 3.4: Point estimates and confidence intervals for the expected depression score SUMD at week 12 under static regimes(first line treatment, second line treatment) and estimated DTR.

| Regime $(\pi_1, \pi_2)$ | Estimated SUMD | 90% Confidence Interval |
|---|---|---|
| Estimated DTR | 2.13 | (1.34, 2.86) |
| (Bupropion, high-dose Bupropion) | 6.91 | (6.27, 7.71) |
| (Paroxetine, high-dose Paroxetine) | 8.25 | (7.39, 9.07) |
| (placebo, high-dose Bupropion) | 3.71 | (3.38, 4.04) |
| (placebo, high-dose Paroxetine) | 4.51 | (4.10, 4.90) |

suggested that bipolar-depression patients with (hypo) mania immediately preceding a major depressive episode should not receive adjunctive antidepressant treatment with either paroxetine or bupropion, whereas the opposite is true for who were in remission or experienced a mixed episode before the current major depressive episode. This is a novel finding, which has not been explored so far. At present there is a consensus that antidepressants in the acute treatment of bipolar depression may be used when there is a history of previous positive response to antidepressants, while they should be avoided in patients with an acute bipolar depressive episode with two or more concomitant core manic symptoms in the presence of psychomotor agitation, in patients with a high number of previous episodes or with a history of rapid cycling and during depressive episodes with mixed features (Pacchiarotti et al., 2013). Furthermore the use of antidepressants is discouraged if there is a history of past mania, hypomania, or mixed episodes emerging during antidepressant treatment (Pacchiarotti et al., 2013). In our study the scale scores for measuring symptoms of depression as well as mania were available for baseline and stage 1 to model the $Q$-functions but did not turn out to be helpful in building an optimal DTR.

So far there are no reliable data on the differential efficacy of paroxetine and bupropion in adult patients older or younger than 44.5 years with bipolar depression. In unipolar depression, a recent meta-analysis suggests that the efficacy of antidepressants in general may be reduced in trials involving patients aged 65 years or older (Tedeschini et al., 2011). Similarly there havent been any reliable data suggesting that patients with higher scores on mood elevation scales do better on paroxetine than bupropion and vice versa (Pacchiarotti et al., 2013). What is well known on the other hand is that paroxetine 20mg/day does not seem to be associated with an increased risk of switch into (hypo)mania in patients with bipolar depression, even in monotherapy (McElroy et al., 2010). The data for our analyses stem from a double-blind, randomized, placebo-controlled trial (Sachs et al., 2007). Consequently we do not know whether in clinical practice not adding any medication or intervention to a mood stabilizer is of comparable benefit for those who do best on placebo in our analyses (Severus et al., 2012).

Estimation of an optimal DTR is typically done as a secondary, exploratory analysis and viewed as a method of generating hypotheses for follow-up confirmatory experiments. The latter are just about to start, using patients with bipolar depression being openly treated within the SCP pathway of STEP-BD using the same rating forms, in particular

the clinical monitoring form for mood disorders.

## 3.3 A Follow up Observational Data Analysis of STEP-BD

In last section, we use $Q$-learning to analysis the data set from Randomized Acute Depression Pathway. Through the analysis, we estimated the optimal treatment regime for patients with acute depression (see Figure 3.6, and 3.5). One important finding from this analysis is that patients with (hypo)manic episode during their early lives are suggested not to receive antidepressants. In this section, we will follow up to do an observational data analysis using data from standardized care pathway (SCP). The analysis will utilize patients enrolled in SCP and with acute depression. We name the extracted data set as SAD (Standardized Acute Depression). This section is constructed as following: We will first introduce the Standardized Acute Depression (SAD) dataset. We then give the detail of the method proposed to analysis SAD. Data analysis results are discussed at the end of this section.

### 3.3.1 Standardized Acute Depression Dataset (SAD)

The standardized acute depression dataset is derived from the standardized care pathway (SCP) in STEP-BD. Because our purpose is to validate what we find in RAD using data from SCP, we want to extract observations that have similar information as observations in RAD. In order to have similar information, patients in SAD should satisfy entering criteria of RAD pathway: (i) Patients should be over 18 years old; (ii) Paitients should be diagnosed as Bipolar type I Disorder or Bipolar type II Disorder. The starting point of SAD for each patient is defined as the time point that the patient's current clinical status is depression. In SAD, at week 0, in addition to receiving moodstabilizers, some patients were assigned to one or some combination of antidepressants. Both the medications and the dosages are decided by the doctors. After the initial assignment, the following treatment for each patient in SAD is also decided by the doctor. The final response for each patient in SAD is collected at week 6, 7, or 8. Figure 3.7 shows the SAD diagram. In RAD, there are only two kinds of antidepressants to be considered (Bupropion and Paroxetine). Differed to RAD, in SAD, there are 10 potential antidepressants

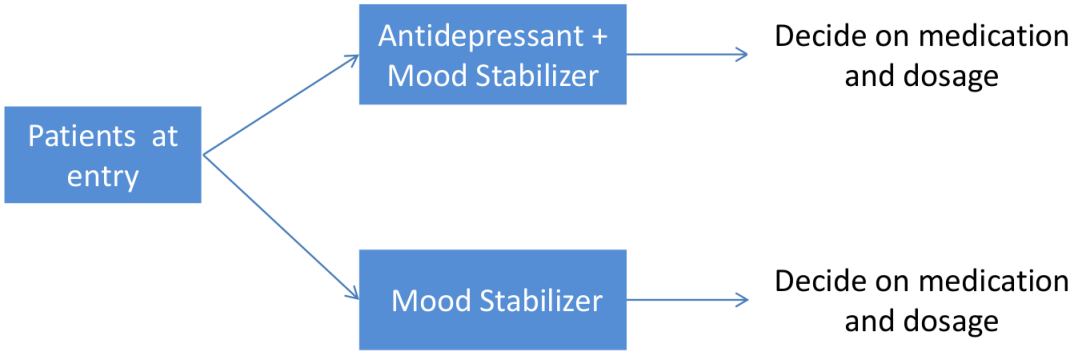


Figure 3.7: SAD is an observational study. The dosages of antidepressant and mood stabilizer are decided by doctors.

to be considered. These include: Deseryl, Serzone, Citalopram, Escitalopram Oxalate, Prozac, Fluvoxamine, Paroxetine, Zoloft, Venlafaxine, and Bupropion. The dosages of these antidepressants a specific patient should receive are decided by the doctors. Based on the suggestions from psychiatrists, we divide these antidepressants into 4 groups. And the dosage for each medication are also divided into 3 levels: high, median, and low. Table 3.5 shows the grouping definitions for the antidepressants. In SAD, besides antidepressants, 9 different mood-stabilizers are being used, which includes: Tegertol, Valproate, Olanzapine, Quetiapine, Clozapine, Lithium, Risperdal, Geodon, and Abilify. These mood-stabilizers are divided into 5 groups with 3 dosage levels for each medication. Table 3.6 shows the grouping definitions for the mood-stabilizers. In SAD, there are 4 group of antidepressants. And the dosage in each group is divided into 3 levels. Since the $Q$-learning method we used for RAD cannot be applied to SAD directly. We decide to use Grouped $Q$-learning to analysis the SAD. In the next subsection, we will describe the data structure, and introduce the idea of grouped $Q$-learning.

### 3.3.2 $Q$-learning with grouped treatment

Ongoing personalized treatment is required by the effective management of a chronic illness (Wagner et al., 2001). Dynamic treatment regimes (DTRs) formalize clinical decision making as sequence of *decision rules*, one per treatment decision, that map patient information to a recommended treatment. An optimal DTR yields the minimal (maxi-

Table 3.5: Antidepressants are divided into 4 groups, and the dosage for each medication is divided into 3 levels: high, median, and low.

| Group Number | Medication Name | Definition of Dosage Level (Low, Median, High) (mg) |
|---|---|---|
| Anti 1 | Deseryl | $< 200, 200 - 400, > 400$ |
| | Serzone | $< 200, 200 - 400, > 400$ |
| Anti 2 | Citalopram | $< 20, 20 - 40, > 40$ |
| | Escitalopram Oxalate | $< 10, 10 - 20, > 20$ |
| | Prozac | $< 20, 20 - 40, > 40$ |
| | Fluvoxamine | $< 100, 100 - 200, > 200$ |
| | Paroxetine | $< 20, 20 - 40, > 40$ |
| | Zoloft | $< 50, 50 - 100, > 100$ |
| Anti 3 | Venlafaxine | $< 75, 75 - 150, > 150$ |
| Anti 4 | Bupropion | $< 150, 150 - 300, > 300$ |

Table 3.6: Mood-stabilizers are divided into 5 groups, and the dosage for each medication is divided into 3 levels: high, median, and low.

| Group Number | Medication Name | Definition of Dosage Level (Low, Median, High) (mg) |
|---|---|---|
| Mood 1 | Tegertol | $< 400, 400 - 800, > 800$ |
| | Valproate | $< 1000, 1000 - 2000, > 2000$ |
| Mood 2 | Olanzapine | $< 10, 10 - 20, > 20$ |
| | Quetiapine | $< 400, 400 - 800, > 800$ |
| Mood 3 | Clozapine | $< 200, 200 - 400, > 400$ |
| Mood 4 | Lithium | $< 900, 900 - 1800, > 1800$ |
| Mood 5 | Risperdal | $< 2, 2 - 4, > 4$ |
| | Geodon | $< 80, 80 - 160, > 160$ |
| | Abilify | $< 15, 15 - 30, > 30$ |

mal) mean outcome when applied treatment assignments to the population of interest. One method for estimating an optimal DTR from observational or randomized study data is $Q$-learning (Murphy, 2005a; Schulte et al., 2012). $Q$-learning is an approximate dynamic programming algorithm that can be viewed as an extension of regression to multi-stage decision problems (Nahum-Shani et al., 2012). As our purpose is the application of $Q$-learning to the SAD data within SCP pathway, we focus on data from a one-stage clinical trial with a terminal continuous outcome. Also, since treatments in SAD contain the idea of groups and dosage levels, we will focus on introducing $Q$-learning with grouped treatments here.

We now present formal mathematical description of $Q$-learning with grouped treatments. We assume that data available to estimate a DTR are in the form of $n$ independent, identically distributed trajectories $\{(X_i, A_i, Y_i)\}_{i=1}^n$, one for each subject where: $X \in \mathbb{R}^p$ denotes baseline subject information; $A \in \mathcal{A}$ denotes the treatment assignment; and $Y \in \mathbb{R}$ denotes a continuous outcome measured at the end of the study coded so that lower values are better. To match the SAD study, we assume that treatments are assigned based on the doctors' suggestions. In the SAD study, $X$ contains a subject's baseline age, race, gender, marital status, annual home income, employment status, education level, as well as baseline measures of bipolar type, clinical status prior to depressive episode, scale scores for mood elevation (SUMM), and scale scores for depression (SUMD); $Y$ is SUMD measured at the end of the study; and $A$ is composed of one treatment for $\mathcal{T}_1$, and a second treatment for $\mathcal{T}_2$ such that $a \in \mathcal{A}$ has the form

$$a = (t_1, d_1, t_2, d_2),$$

where $t_j \in \mathcal{T}_j$ and $d_j$ is the dosage level at $t_j$ $(j = 1, 2)$. In SAD, $\mathcal{T}_1$ denotes the treatment of mood-stabilizer, and $\mathcal{T}_2$ denotes the treatment of antidepressant. If a subject is receiving a single treatment, then $t_2, d_2$ are set to $NA$. Throughout, we assume that $d_j$ is an ordinal variable with three levels coded as $\{1, 2, 3\}$ that represent low, medium, and high dosage correspondingly. Furthermore, assume that the treatments in $\mathcal{T}_j$ can be treated as groups $\mathcal{T}_j = \bigcup_{k=1}^{\mu_j} \mathcal{G}_{kj}$ based on some scientific theories.

Define $H = X$, so that $H$ denotes information available to the decision maker at during treatment assignment time point. A DTR is a function $\pi$ where $\pi : \operatorname{dom} H \to \operatorname{dom} A$ so that a patient presenting with $H = h$ is assigned treatment $\pi(h)$. For any $h \in \operatorname{dom} H$, let $\mathcal{F}(h)$ denote the set of feasible treatments for a patient with $H = h$. In

the SAD study, feasible treatments regarding Mood 1 in $\mathcal{F}(h)$ are shown in Table 3.7. In total, there are 195 feasible treatments in $\mathcal{F}(h)$. Let $\Pi = \{\pi : \pi(h) \in \mathcal{F}(h), \forall h \in \text{dom}\, H\}$ denote the class of *feasible* DTRs (for a more formal discussion of feasibility see (Schulte et al., 2012)). An optimal DTR, say $\pi^{\text{opt}}$, satisfies $\text{E}^{\pi^{\text{opt}}} Y \geq \text{E}^{\pi} Y$ for all $\pi \in \Pi$, where $\text{E}^{\pi}$ denotes expectation under the restriction that $A = \pi(H)$. Define $Q(h, a) = \text{E}(Y|H = h, A = a)$. The function $Q(h, a)$ measures the 'quality' of assigning treatment $a$ to a patient with $H = h$. It follows from dynamic programming (Bellman, 1957) that $\pi^{\text{opt}}(h) = \arg\min_{a \in \mathcal{F}(h)} Q(h, a)$. In practice, dynamic programming cannot be applied because the true $Q$-functions are not known; instead, estimation of $\pi^{\text{opt}}$ must rely on the observed data. $Q$-learning is an approximate dynamic programming algorithm which mimics the dynamic programming solution by replacing the conditional expectations required by dynamic programming with regression models fit to the observed data. Let $Q(h, a; \theta)$ denote a postulated working model for $Q(h, a)$ indexed by unknown parameter $\theta$. A version of the $Q$-learning algorithm that applies to the data from SAD is:

**Algorithm 3.2: Q-learning for RAD**

**(Q1)** Compute $\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \{Y_i - Q(H_i, A_i; \theta)\}^2$; and subsequently estimator $Q(h, a; \widehat{\theta})$ of $Q(h, a)$.

**(Q2)** The estimated optimal regime is $\widehat{\pi}(h) = \arg\min_{a \in \mathcal{F}(h)} Q(h, a; \widehat{\theta})$.

To estimate $\pi^{\text{opt}}$ using data from the SAD study, we assume linear working models for the $Q$-functions with grouped effects. For $Q(h, a)$ we posit a model of the form:

$$Q(h, a; \theta) \;=\; \sum_{a' \in \mathcal{A}} 1_{a'=a} X^T \beta_{a'}$$

where

$$\begin{aligned} \theta &= \beta_{a'} \\ &= \sum_{k=1}^{\mu_1} 1_{t_1 \in \mathcal{G}_{k1}} \left( \alpha_{0k} + \sum_{l=2}^{3} \alpha_{1l} 1_{d_1=l} + \delta_{t_1} \right) \\ &\quad + \sum_{k=1}^{\mu_2} 1_{t_2 \in \mathcal{G}_{k2}} \left( \eta_{0k} + \sum_{l=2}^{3} \eta_{1l} 1_{d_2=l} + \nu_{t_2} \right). \end{aligned}$$

The parameters $\alpha_{0k}, \alpha_{1l}, \delta_{t_1}, \eta_{0k}, \eta_{1l}$, and $\nu_{t_2}$ can be multi-dimensional. This model has

Table 3.7: Part of feasible treatments $\mathcal{F}(h)$ in SAD study. It is a combination of treatment $\mathcal{T}_1$ and $\mathcal{T}_2$. Here the possible combination treatments regarding Mood 1 are listed.

| $t_1$ | $d_1$ | $t_2$ | $d_2$ |
|---|---|---|---|
| Mood 1 | 1 | NA | NA |
| | 2 | NA | NA |
| | 3 | NA | NA |
| Mood 1 | 1 | Anti 1 | 1 |
| | 2 | | 1 |
| | 3 | | 1 |
| Mood 1 | 1 | Anti 1 | 2 |
| | 2 | | 2 |
| | 3 | | 2 |
| Mood 1 | 1 | Anti 1 | 3 |
| | 2 | | 3 |
| | 3 | | 3 |
| Mood 1 | 1 | Anti 2 | 1 |
| | 2 | | 1 |
| | 3 | | 1 |
| Mood 1 | 1 | Anti 2 | 2 |
| | 2 | | 2 |
| | 3 | | 2 |
| Mood 1 | 1 | Anti 2 | 3 |
| | 2 | | 3 |
| | 3 | | 3 |
| Mood 1 | 1 | Anti 3 | 1 |
| | 2 | | 1 |
| | 3 | | 1 |
| Mood 1 | 1 | Anti 3 | 2 |
| | 2 | | 2 |
| | 3 | | 2 |
| Mood 1 | 1 | Anti 3 | 3 |
| | 2 | | 3 |
| | 3 | | 3 |
| Mood 1 | 1 | Anti 4 | 1 |
| | 2 | | 1 |
| | 3 | | 1 |
| Mood 1 | 1 | Anti 4 | 2 |
| | 2 | | 2 |
| | 3 | | 2 |
| Mood 1 | 1 | Anti 4 | 3 |
| | 2 | | 3 |
| | 3 | | 3 |

main effects for groups, dose, and a treatment specific deviation. When we fit the model, we will shrink the $\delta_{t_1}$ and $\nu_{t_2}$ values toward 0, hence the estimator for $\theta$ will be defined as:

$$\begin{aligned}\hat{\theta}_n^{\lambda,\zeta} &= \arg\min_\theta \mathbb{E}_n \left( Y - \sum_{a'} 1_{a=a'} X^T \beta_{a'} \right)^2 \\ &\quad + \lambda \sum_{t_1} ||\delta_{t_1}|| + \zeta \sum_{t_2} ||\nu_{t_2}||,\end{aligned}$$

where $\lambda, \zeta \geq 0$ are tuning parameters, and $\mathbb{E}_n$ is empirical expectation such that $\mathbb{E}_n X = \frac{1}{n}\sum_{i=1}^n X_i$.

### 3.3.3 Data Analysis for SAD

The $Q$-learning algorithm stated in the preceding section assumes: (i) complete data; and (ii) working models for the $Q$-functions. However, in SAD, as in most clinical trials, a non-trivial amount of covariates and outcome information are missing. Furthermore, there is not strong theory to suggest working models for the $Q$-functions so we must use the data to assist in the choice of these models. We combine multiple imputation with stepwise variable selection to estimate the $Q$-function and the optimal treatment regime.

#### Missing data

Figure 3.8 shows the fraction of missing data for the variables under consideration in our analysis of the SAD data. There is a significant amount of missing covariate information in SAD; thus, discarding subjects with missing information is inefficient and may introduce bias (Little and Rubin, 2002).

Similar to what we did in RAD data analysis, here we use multiple imputation (Rubin, 2004, MI) to deal with missing data. MI creates multiple complete datasets and is thereby suited for conducting a series of exploratory and secondary analyses including estimation of an optimal treatment regime (Shortreed et al., 2011). We use Bayesian MI to "fill in" the missing values with draws from the posterior predictive distribution of the missing values given the observed data (for details and underlying assumptions see Little and Rubin, 2002; Van Buuren, 2012). Implementation of Bayesian MI requires specification of a prior and likelihood for the observed data. We specify the joint likelihood through the

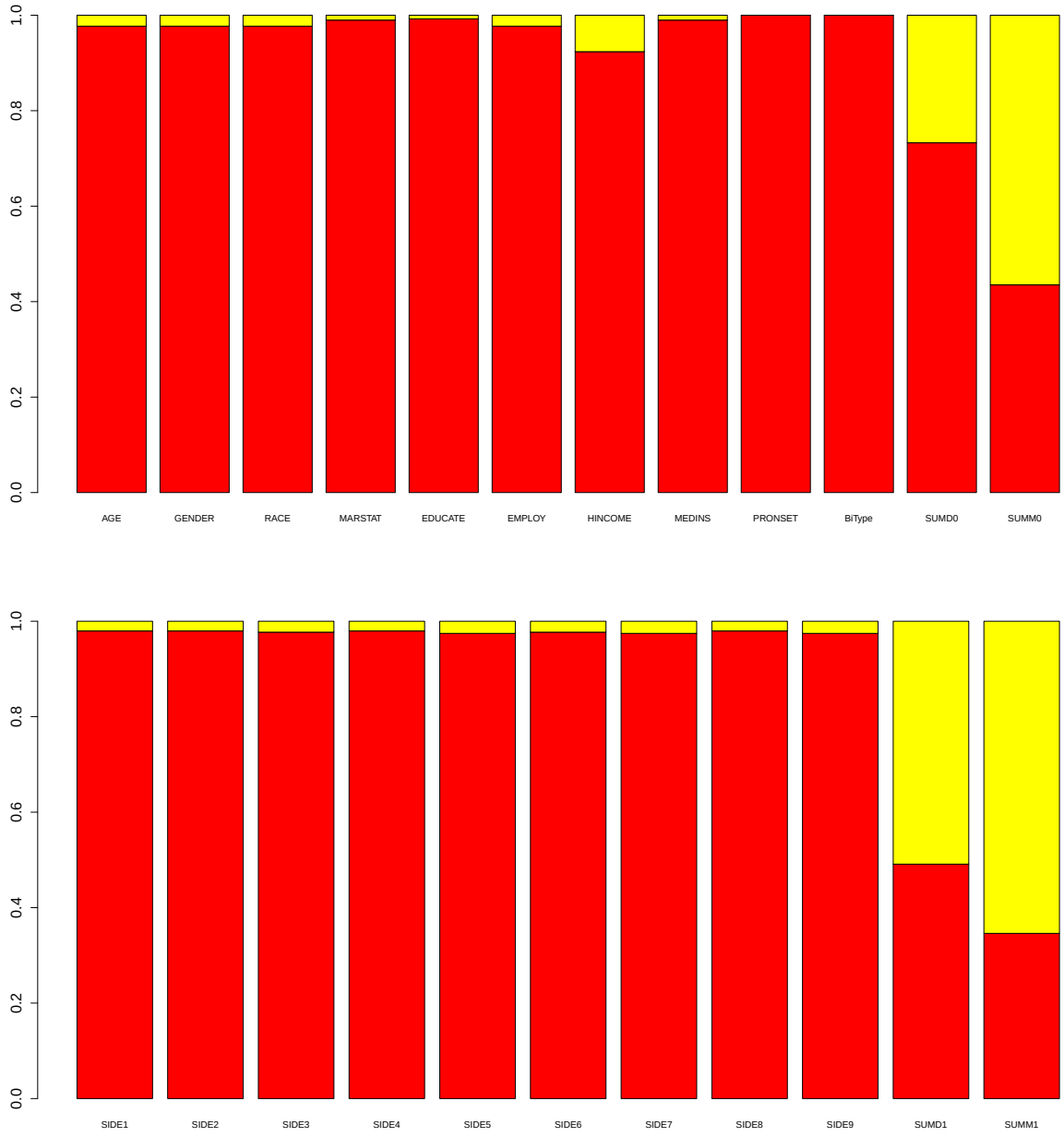


Figure 3.8: Variables with missing data are listed. The SUMM$i$ and SUMD$i$ denote continuous symptom subscales for depression and mood elevation at i th time point. The SIDE$j$ represents different side effects. PRONSET denotes patients' prior to onset clinical status. EDUCATE, EMPLOY, MARSTAT, MEDINS and HINCOME are the indicators for patients' education level, employment status, marriage status, medical insurance and annual home income respectively.

conditional distribution of each variable on all other variables (for discussion of this approach see Raghunathan et al., 2001; Van Buuren et al., 2006; Van Buuren, 2007). Thus, the likelihood is determined implicitly through a series of regression models, one for each variable that contains missing information. For continuous variables we use predictive mean matching and for binary variables we use logistic regression models. To reduce variance, we use forward stepwise variable selection applied to the complete data to select predictors for each conditional model. Flat improper priors were used for all parameters. Imputations were carried out using the freely available and open source `mice` package with the default settings (http://cran.r-project.org/web/packages/mice/index.html).

Using the procedure described above we impute $m$ complete datasets. For a given choice of $h$, we can apply the $Q$-learning algorithm to each imputed dataset to obtain estimated $Q$-functions $Q(h, a; \widehat{\theta}^{(\ell)})$, $\ell = 1, \ldots, m$. The final estimated optimal decision rule is obtained as the minimizer of the averaged imputed $Q$-functions $\widehat{\pi}(h) = \arg\min_{a \in \mathcal{F}(h)} m^{-1} \sum_{\ell=1}^{m} Q\left(h, a; \widehat{\theta}^{(\ell)}\right)$.

**Variable Selection and estimation of the optimal treatment regime**

In order to estimate an optimal treatment regime using $Q$-learning with grouped treatments, we need to select which covariates to include the models for the $Q$-functions. Recall that the average of 6-8 week depression SUMD score was used as the outcome ($Y$). We identified 21 potential predictors; these predictors are listed in Table 3.8. We select a subset of these predictors for the $Q$-function using stepwise variable selection to minimize the mean square error (MSE) using cross validation method averaged over the multiply imputed data sets (Shao, 1993). Let $\mathcal{M}$ denote a subset of predictors dictating $h$ and let $\widehat{\theta}^{(\ell)}(\mathcal{M})$ denote the coefficients estimated in step (Q1) of the $Q$-learning algorithm using predictors $\mathcal{M}$ to the $\ell$th imputed data set. In addition, let $\text{MSE}^{(\ell)}(\mathcal{M})$ denote the MSE for the model $Q(h, a; \widehat{\theta}^{(\ell)}(\mathcal{M}))$ calculated on the $\ell$th imputed dataset. This procedure that we use to construct models for the $Q$-learning algorithm is:

(S1) Using forward variable selection compute

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}} \frac{1}{m} \sum_{\ell=1}^{m} \text{MSE}^{(\ell)}(\mathcal{M});$$

(S2) Let $Q(h, a; \widehat{\theta}(\widehat{\mathcal{M}}))$ denote the estimated $Q$-function.

Table 3.8: Candidate predictors for regression models in $Q$-learning.

| Variable | Description | Type | Values (range or level) |
|---|---|---|---|
| AGE | Age at entry (years) | Numerical | 16-77 |
| RACE | Race | Binary | white or causasian, non white |
| GENDER | Gender | Trinary | male, female, transgender |
| MARSTAT | Marital status | Trinary | never married, married, separated |
| HINCOME | Annual household income ($\times$\$1000) | Binary | $< 40$, $\geq 40$ |
| EMPLOY | Employment status | Binary | employed, unemployed |
| EDUCATE | Education level | Binary | college or below, technical school or above |
| MEDINS | Indicator of medical insurance | Binary | yes, no |
| BITYPE | Bipolar type at entry | Binary | type I, type II |
| PRONSET | Clinical status before depressive episode | Trinary | remission, (hypo)manic, mixed |
| SUMD0 | Scaled depression at entry | Numerical | 0-22 |
| SUMM0 | Scaled mood elevation at entry | Numerical | 0-12 |
| SIDE1 | Tremor | Binary | yes, no |
| SIDE2 | Dry mouth | Binary | yes, no |
| SIDE3 | Sedation | Binary | yes, no |
| SIDE4 | Constipation | Binary | yes, no |
| SIDE5 | Diarrhea | Binary | yes, no |
| SIDE6 | Headache | Binary | yes, no |
| SIDE7 | Poor memory | Binary | yes, no |
| SIDE8 | Sexual dysfunction | Binary | yes, no |
| SIDE9 | Increased appetite | Binary | yes, no |

Table 3.9: Estimated optimal regime with RACE $= 1, \text{MEDINS} = 1$. M$i$, A$j$ represent group $i$ mood-stabilizer and group $j$ antidepressant respectively. "Low", "Medium", and "High" denote mood-stabilizer and antidepressants dosage levels.

| $SUMD_0$ | Optimal Treatment |
|---|---|
| $[0, 6)$ | Low M2, Medium A2 |
| $[6, 8)$ | Low M2, High A1 |
| $[8, 10)$ | High M1, High A1 |
| $[10, 12)$ | Medium M1, Medium A2 |
| $[12, 14)$ | Medium M1, Low A2 |
| $[14, 16]$ | Medium M2, High A3 |

Table 3.10: Estimated optimal regime with RACE $= 1, \text{MEDINS} = 0$. M$i$, A$j$ represent group $i$ mood-stabilizer and group $j$ antidepressant respectively. "Low", "Medium", and "High" denote mood-stabilizer and antidepressants dosage levels.

| $SUMD_0$ | Optimal Treatment |
|---|---|
| $[0, 8)$ | Low M2, Medium A2 |
| $[8, 10)$ | Low M2, High A1 |
| $[10, 12)$ | Medium M1, Low A2 |
| $[12, 16)$ | Medium M2, Low A2 |

The variables included in the model $\widehat{\mathcal{M}}$ are: SUMD0, MEDINS, and RACE.

The $Q$-function has the form $Q(h, a; \theta) = \sum_{a' \in \mathcal{A}} 1_{a'=a} h^T \beta_{a'}$, where:

$$\begin{aligned} \theta &= \beta_{a'} \\ &= \sum_{k=1}^{\mu_1} 1_{t_1 \in \mathcal{G}_{k1}} \left( \alpha_{0k} + \sum_{l=2}^{3} \alpha_{1l} 1_{d_1=l} + \delta_{t_1} \right) \\ &\quad + \sum_{k=1}^{\mu_2} 1_{t_2 \in \mathcal{G}_{k2}} \left( \eta_{0k} + \sum_{l=2}^{3} \eta_{1l} 1_{d_2=l} + \nu_{t_2} \right). \end{aligned}$$

$t_1$ denotes treatment from mood-stabilizers, and $t_2$ denotes treatment from antidepressants. From Table 3.5 and Table 3.6, $\mu_1 = 5$ and $\mu_2 = 4$. Point estimates for the coefficients in the $Q$-function above is shown in the Appendix. Table 3.9, 3.10, 3.11, and 3.12 show the estimated optimal treatment regime when the SUMD0, MEDINS, and RACE have different values.

Table 3.11: Estimated optimal regime with RACE $= 0$, MEDINS $= 1$. M$i$, A$j$ represent group $i$ mood-stabilizer and group $j$ antidepressant respectively. "Low", "Medium", and "High" denote mood-stabilizer and antidepressants dosage levels.

| $SUMD_0$ | Optimal Treatment |
|---|---|
| $[0, 6)$ | Low M2, Medium A2 |
| $[6, 8)$ | Low M2, High A1 |
| $[8, 10)$ | High M2, Medium A2 |
| $[10, 16]$ | High M2, High A3 |

Table 3.12: Estimated optimal regime with RACE $= 0$, MEDINS $= 0$. M$i$, A$j$ represent group $i$ mood-stabilizer and group $j$ antidepressant respectively. "Low", "Medium", and "High" denote mood-stabilizer and antidepressants dosage levels.

| $SUMD_0$ | Optimal Treatment |
|---|---|
| $[0, 8)$ | Low M2, Medium A2 |
| $[8, 16)$ | High M2, High A3 |

### 3.3.4 Discussion and Future Work

We estimated an optimal DTR for patients presenting with bipolar depression using data SAD from the standardized care pathway (SCP) that have similar characteristic as RAD in the STEP-BD study. The estimated treatment regime suggested that bipolar-depression patients with different medical insurance, race, and baseline SUMD score should receive different combination of antidepressant and moodstabilizer group as well as different dosage levels. This is a novel finding, which has not been explored so far.

Estimation of an optimal DTR is typically done as a secondary, exploratory analysis and viewed as a method of generating hypotheses for follow-up confirmatory experiments. Grouped $Q$-learning gave advice to the choice of treatment and dosage. The confirmation of findings in SAD is needed.

# Chapter 4

# Dosing regimes with adverse events

## 4.1 Introduction

In course of managing a chronic illness, a clinical scientist must decide how to adapt treatment both in response to and anticipation of changes in each individual patients health status. A treatment regime formalizes this decision process as a sequence of functions, one per intervention period, that map current patient information to a recommended treatment (Murphy, 2003a; Robins, 2004; Chakraborty and Moodie, 2013). An optimal treatment regime is defined as maximizing some functional of the outcome distribution, e.g., mean symptom reduction or the probability of surviving disease-free past some time horizon, if the regime were used to select treatments for individuals in a population of interest. We consider the problem of constructing an interpetable treatment regime that adjusts treatment dosage over a potentially large number of intervention periods with the goal of maximizing a cumulative measure while controlling the risk of an adverse event. This work is motivated by a sequence of clinical trials on chronic pain in which the clinical goal was to maximize pain reduction while reducing the risk of constipation.

When there are multiple intervention periods, estimating a regime that is both interpretable and efficacious is difficult because the optimal regime is generally a complex, nonsmooth functional of the generative model (Laber et al., 2014a,c; Zhao et al., 2014). One approach to estimating a high-quality regime is to use a regression-based approximate dynamic programming algorithm, e.g., Q-learning, with flexible regression models (Zhao et al., 2011; Moodie et al., 2013); however, this leads to estimated regimes that are difficult to interpret. An alternative is to use policy-search algorithms wherein one

first constructs an estimator of the targeted functional of the outcome distribution for every regime within a pre-specifed class and then chooses the maximizer as the estimated optimal regime (Robins et al., 2008; Orellana et al., 2010; Zhang et al., 2012b,a, 2013, 2015). In this approach, it is possible to enforce parsimony and intepretability through the pre-specified class of regimes; however, these estimators use either: (i) an augmented inverse probability weighted estimator of the mean outcome which becomes unstable as the number of intervention periods increases; or (ii) a marginal structural model which is best suited for low-dimensional problems.

In addition to the foregoing problems, the work on estimating an optimal regime under risk constraints is limited. Set-valued treatment regimes allow competing outcomes, e.g., efficacy and risk of adverse events, but have not been developed to handle constraints (Laber et al., 2014b; Lizotte and Laber, 2015). Constrained interactive Q-learning allows constraints but only applies to settings with two intervention periods and the form of the estimated regime is generally non-linear and therefore difficult to interpret (Linn et al., 2015). Luedtke and van der Laan (2015) derived estimators of the optimal regime under constraints on the proportion of subjects that can be prescribed an expensive treatment, thus, the constraints are on treatment assignments rather than outcomes.

We use non-parametric Q-learning to form a joint estimator of the marginal mean efficacy and a general measure of risk of an adverse event for any regime in a pre-specified class. An estimator of the optimal regime is obtained by choosing the regime that maximizes expected efficacy among those that satisfy a constraint on risk. The class of regimes is chosen to ensure that the class of regimes is interepretable and easily disseminated among domain experts. We show that non-parametric Q-learning produces estimators that are more stable than (augmented) inverse probability weighting when there are a large number of time points. Furthermore, non-parametric Q-learning can be used to do diagnose severe approximation error in the class of pre-specified regimes.

In Section 4.2, we discuss a variant of non-parametric Q-learning that is amenable to constrained policy-search algorithms. In Section 4.3, we discuss computation of the proposed estimator. In Section 4.4, we apply the proposed method to analyze data from a clinical trial on chronic pain. Future work is discussed in Section 4.5.

## 4.2 Policy-search through non-parametric $Q$-learning

### 4.2.1 Notation Set-up

The available data to estimate an optimal treatment regime are of the form $\{ (X_{0,i}, A_{0,i}, Y_{0,i}, Z_{0,i}, X_{1,i}, A_{1,i}, Y_{1,i}, Z_{1,i}, \ldots, X_{T_i,i}, A_{T_i,i}, Y_{T_i,i}, Z_{T_i,i})\}_{i=1}^{n}$, where: $T \in \{0, 1, 2, ..., \mathcal{T}\}$ denotes the number of treatment periods; $X_t$ denotes patient covariates measured during treatment period $t$; $A_0 \in [0, 1]$ denotes the treatment (dosage) received at baseline; $A_t \in \{0, 1, ..., K\}$ denotes the treatment received at during treatment period $t$ for $t = 1, 2, ..., \mathcal{T}$; $Y_t \in \mathbb{R}$ denotes the primary outcome, e.g., efficacy, observed at the end of treatment period $t$; and $Z_t \in \mathbb{R}$ denotes a secondary outcome, e.g., side-effect burden, observed at the end of treatment period $t$. Define $H_0 = X_0$ and for $t \geq 1$ recursively define $H_t = (H_{t-1}, A_{t-1}, Y_{t-1}, Z_{t-1}, X_t)$; thus, $H_t$ is all the information collected before treatment assignment in stage $t$. Define a state functions $\phi_t : \text{dom}H_t \to \mathbb{R}^p$, so that $S_t = \phi_t(H_t)$ is a summary of the information accumulated by time $t$. Furthermore, for each state $s \in \mathbb{R}^p$, define $A_0(s) \in [0, 1]$, and $A_1(s) \subseteq \{0, 1, ..., K\}$ to be the set of feasible treatments options for a patient presenting with state $s$. We define two stationary treatment regimes as maps $\pi_0 : \mathbb{R}^p \to [0, 1]$, and $\pi_1 : \mathbb{R}^p \to \{0, 1, ..., K\}$ that satisfy $\pi_0(s) \in A_0(s)$, and $\pi_1(s) \in A_1(s)$ for all $s \in \mathbb{R}^p$; under $\pi_0, \pi_1$, a patient presenting with state $S_t = s_t$ at time $t = 0$ is recommended to treatment $\pi_0(s_t)$; a patient presenting with state $S_t = s_t$ at time $t \geq 1$ is recommended to treatment $\pi_1(s_t)$. We define $\pi = (\pi_0, \pi_1)$

The foregoing framework includes, as a special, the classic definition of a dynamic treatment regime as a sequence of functions, one per intervention period, that map current patient history to a recommended treatment (Schulte et al., 2014); to see this, choose $p = \text{dim}H_{\mathcal{T}} + 1$, define $S_t = (H_t, t, 0, ..., 0)^T$, and define $\pi_1(s) = \pi_t(h_t)$ with $t \geq 1$, $\pi_0(s) = \pi_0(h_0)$ where $\pi_t : \text{dom}H_t \to \{0, 1, ..., K\}$ with $t \geq 1$, and $\pi_0 : \text{dom}H_0 \to [0, 1]$. The reason for defining the regime as a function of $S_t$, rather than $H_t$, is that in settings where the same treatment decisions are being made at each time point, e.g., increase, decrease, or do not change current dose of pain medication, it is desirable to construct a summary and treatment regime that is qualitatively fixed over time. In the context of managing chronic pain we might consider a regime of the form: increase dose if current pain score is above 5 and side-effect burden is below 3; decrease dose if current pain score is below 2 or side-effect burden has been above 8 for two or more time points; and leave dose unchanged otherwise. A regime of this form is easier to disseminate to clinical

scientists and evaluate using domain knowledge than a regime that is changing across time points. Nevertheless, as described above, our framework allows the regime and the summary function to change over time and could therefore be used in more general settings than the motivating application we consider here.

To define an optimal regime we use the language of potential outcomes (Rubin, 1978; Splawa-Neyman et al., 1990). For each $t = 0, 1, 2, ..., T$ we use an overline to denote the history up to time $t$, e.g., $\overline{a}_t = (a_0, a_1, ..., a_t)$, and we use underline notation to indicate history from $t$ until $\mathcal{T}$ , e.g., $\underline{a}_t = (a_t, ..., a_\mathcal{T})$. Define $Y_t^*(\underline{a}_t)$ and $Z_t^*(\underline{a}_t)$ to be the potential outcomes under treatment sequence at and let $X_t^*(\overline{a}_t)$ denote the set of covariates that would arise at time $t$ under $\overline{a}_t$ ; similarly, the potential states are defined as $S_t^*(\overline{a}_t) = \phi_t(H_t^*(\overline{a}_t))$. Define $\Delta T_t^*(\overline{a}_t)$ to be the potential indicator that a subject leaves the study at time $t$ under treatment sequence at and define the follow-up time under sequence $\overline{a}_\mathcal{T}$ as $\mathcal{T}^*(\overline{a}_\mathcal{T}) = \arg\min\{0 \leq t \leq \mathcal{T} : \Delta T_t^*(\overline{a}_t) = 1\}$; if $T^*(\overline{a}_\mathcal{T}) = t$ then $T^*(\overline{a}_t, \underline{a}'_{t+1}) = t$ for any sequence of treatments $\underline{a}'_{t+1}$, hence we will write the event $T^*(\overline{a}_t) = t$ as shorthand for the event $T^*(\overline{a}_\mathcal{T})$ where $\underline{a}'_{t+1}$ is defined arbitrarily. For each $t = 0, 1, 2, ..., T$ let $u_t : \mathbb{R}^{t+1} \to \mathbb{R}$ be a utility function coded so that higher values are better and let $c_t : \mathbb{R}^{t+1} \to \mathbb{R}$ denote a cost function coded so that lower values are better. For any regime, $\pi$, define the expected utility under $\pi$ as

$$V_u(\pi) = \mathbb{E}\left(\sum_{t=0}^{\mathcal{T}}\left[\sum_{\overline{a}_t:T^*(\overline{a}_t)=t} u_t\{Y_0^*(a_0), Y_1^*(\overline{a}_1), ..., Y_t^*(\overline{a}_t)\}\prod_{v=1}^{t} 1_{\pi\{S_v^*(\overline{a}_{v-1})\}=a_v}\right]\right),$$

where $1_v$ denotes the indicator function of event $v$. The expected cost under $\pi$, denoted $V_c(\pi)$, is defined analogously by substituting $c_t(\cdot)$ for $u_t(\cdot)$ and $Z_t^*(\overline{a}_t)$ for $Y_t^*(\overline{a}_t)$ in the above equation. That is:

$$V_c(\pi) = \mathbb{E}\left(\sum_{t=0}^{\mathcal{T}}\left[\sum_{\overline{a}_t:T^*(\overline{a}_t)=t} c_t\{Z_0^*(a_0), Z_1^*(\overline{a}_1), ..., Z_t^*(\overline{a}_t)\}\prod_{v=1}^{t} 1_{\pi\{S_v^*(\overline{a}_{v-1})\}=a_v}\right]\right).$$

Let $\Pi$ denote a class of regimes of interest, and let $\tau \in \mathbb{R}$ denote a threshold on cost. We define the $\tau-$optimal regime, $\pi_\tau^{opt}$, as a solution to $\text{Sup}_{\pi\in\Pi} V_u(\pi)$ such that $V_c(\pi) \leq \tau$, provided such a solution exists. The canonical problem of this form is to maximize average efficacy, e.g., $u_t(y_0, ..., y_t) = t^{-1}\sum_{j=0}^{t} y_j$, where $y_t$ denotes a measure of efficacy at time $t$, subject to constraint on harm, e.g., $c_t(z_0, ..., z_t) = \max_{0\leq j\leq t} z_j$, where $z_t$ denotes an

adverse event indicator. Another common example is to maximize average efficacy but subject to the constraint that the average time to all-cause treatment discontinuation exceeds some threshold, e.g., $c_t(z_0, ..., z_t) = t$. In the settings that are interested in, patients leave the study because adverse events prevent them from staying on treatment; thus, it would not be feasible (ethical) to keep a patient on treatment after they leave the study. This is in contrast to settings, where the primary interest is what would be the best treatment regime if all patients remained in the study and on treatment (e.g., Shortreed et al., 2014).

### 4.2.2 *Q*-learning for Policy Search

Let $\pi$ denote a fixed regime. We assume that an indicator of whether or not a patient has left the study is in $H_t$ so that the events $T^*(\overline{a}_t) = t$ and $T^*(\overline{a}_t) > t$ are in $\sigma\{H_t^*(\overline{a}_t)\}$. Define the $\mathcal{T}-$stage $Q$-function

$$Q_{\mathcal{T}}^Y(h_{\mathcal{T}}, a_{\mathcal{T}}) = \mathbb{E}\left[1_{T^*(\overline{a}_{\mathcal{T}})=\mathcal{T}} u_{\mathcal{T}}\{Y_0^*(a_0), Y_1^*(\overline{a}_1), ..., Y_{\mathcal{T}}^*(\overline{a}_{\mathcal{T}})\} | H_{\mathcal{T}} = h_{\mathcal{T}}\right], \tag{4.1}$$

where $\overline{a}_{\mathcal{T}-1}$ is contained in $h_{\mathcal{T}}$. The function $Q_{\mathcal{T}}(h_{\mathcal{T}}, a_{\mathcal{T}})$ measures the expected outcome for patient presenting with history $H_{\mathcal{T}} = h_{\mathcal{T}}$ assigned treatment $a_{\mathcal{T}}$ (Sutton and Barto, 1998; Schulte et al., 2014). Define $v_{\mathcal{T}}(h_{\mathcal{T}}; \pi) = Q_{\mathcal{T}}[h_{\mathcal{T}}, \pi\{\phi_{\mathcal{T}}(h_{\mathcal{T}})\}]$ so that $v_{\mathcal{T}}^{\pi}(h_{\mathcal{T}})$ denotes the expected outcome for a patient presenting with $H_{\mathcal{T}} = h_{\mathcal{T}}$ and treated according to $\pi$. For $t = \mathcal{T} - 1, ..., 1, 0$ define

$$Q_t^Y(h_t, a_t; \pi) = \mathbb{E}\Bigg[1_{T^*(\overline{a}_t)=t} u_t\{Y_0^*(a_0), Y_1^*(\overline{a}_1), ..., Y_t^*(\overline{a}_t)\} + 1_{T^*(\overline{a}_t)>t} v_{t+1}\{H_{t+1}^*(\overline{a}_{t+1}); \pi\} | H_t = h_t\Bigg], \tag{4.2}$$

and subsequently $v_t(h_t; \pi) = Q_t^Y[h_t, \pi\{\phi_t(h_t)\}; \pi]$. Define $Q_{\mathcal{T}}^Z(h_{\mathcal{T}}, a_{\mathcal{T}})$ and $Q_t^Z(h_t, a_t; \pi)$ analogously by replacing $Y_t(\cdot)$ with $Z_t(\cdot)$, and $u_t(\cdot)$ with $c_t(\cdot)$, $t = 1, ..., \mathcal{T}$, in 4.1 and 4.2. Then $Q_{\mathcal{T}}^Z(h_{\mathcal{T}}, a_{\mathcal{T}})$ is defined as

$$Q_{\mathcal{T}}^Z(h_{\mathcal{T}}, a_{\mathcal{T}}) = \mathbb{E}\left[1_{T^*(\overline{a}_{\mathcal{T}})=\mathcal{T}} c_{\mathcal{T}}\{Z_0^*(a_0), Z_1^*(\overline{a}_1), ..., Z_{\mathcal{T}}^*(\overline{a}_{\mathcal{T}})\} | H_{\mathcal{T}} = h_{\mathcal{T}}\right],$$

and $Q_t^Z(h_t, a_t; \pi)$ is defined as

$$Q_t^Z(h_t, a_t; \pi) = \mathbb{E}\Big[1_{T^*(\overline{a}_t)=t} c_t\{Z_0^*(a_0), Z_1^*(\overline{a}_1), ..., Z_t^*(\overline{a}_t)\} + 1_{T^*(\overline{a}_t)>t} v_{t+1}\{H_{t+1}^*(\overline{a}_{t+1}); \pi\} | H_t = h_t\Big].$$

The following results characterize the $\tau$-optimal regime in terms of the $Q$- and $v$-functions; proofs of these results follow immediately from expanding the function definitions.

**Lemma 4.1.** *Let $\pi$ be a fixed regime and assume the expectations in 4.1 and 4.2 are well defined. Then, $V_u(\pi) = \mathbb{E}\{v_0^T(H_0; \pi)\}$ and $V_c(\pi) = \mathbb{E}\{v_0^Z(H_0; \pi)\}$.*

**Corollary 4.2.** *Let $\Pi$ denote a class of regimes and assume the expectations in 4.1 and 4.2 are well defined. Any solution of $Sup_{\pi \in \Pi} \mathbb{E}\{v_0^Y(H_0; \pi)\}$ such that $\mathbb{E}\{v_0^Z(H_0; \pi)\} \leq \tau$, is a $\tau$-optimal regime.*

The above characterization of a $\tau$-optimal regime facilitates estimation an optimal regime by first estimating the $Q$-functions for each $\pi \in \Pi$ and then checking that the conditions of Corollary 4.2 are satisfied. Of course, in implementation the class $\Pi$ may be infinite so an iterative solver must by used. In order to estimate the $Q$-functions we must connect them with the underlying generative model.

## 4.3 Computation Algorithm

### 4.3.1 Estimating the $Q$-functions

To estimate a $\tau$-optimal regime from the observed data, we make a series of assumptions about the data-generating mechanism. The set of potential outcomes is

$$W^* = \Big\{X_0, Y_0^*(a_0), Z_0^*(a_0), X_1^*(\overline{a}_1), ..., X_{T^*(\overline{a}_\mathcal{T})}^*, Y_{T^*(\overline{a}_\mathcal{T})}^*, Z_{T^*(\overline{a}_\mathcal{T})}^*, T^*(\overline{a}_\mathcal{T}) : \overline{a}_\mathcal{T} \in \{0, 1, ..., K\}^{\otimes \mathcal{T}}\Big\},$$

where we have included $X_0$ in $W^*$ for convenience. We make the following assumptions: (A1) sequential ignorability, $A_t \perp W|H^t$; (A2) consistency, $T = T^*(\overline{A}_T, \underline{a}_{T+1})$ for every sequence $\underline{a}_{T+1}$, $X_t = X_t^*(\overline{A}_t)$, $Y_t = Y_t^*(\overline{A}_t)$, and $Z_t = Z_t^*(\overline{A}_t)$; and (A3) positivity, $P(A_t = a_t|H_t = h - t) \geq \epsilon$ for some $\epsilon > 0$, for all $a_t \in \mathcal{A}\{\phi_t(h_t)\}$, and for all $h_t$ outside of a null set. Assumptions (A1)-(A3) are standard in the treatment regimes literature (Robins, 2004; Chakraborty and Moodie, 2013; Schulte et al., 2014); assumptions (A1) and (A3) are satisfied by design in a randomized study but must be considered carefully in an observational study.

Under (A1)-(A3), it follows from standard arguments (Schulte et al., 2014) that $Q_{\mathcal{T}}^Y(h_{\mathcal{T}}, a_{\mathcal{T}}) = \mathbb{E}\{1_{T=\mathcal{T}} u_{\mathcal{T}}(Y_0, ..., Y_{\mathcal{T}})|H_{\mathcal{T}} = h_{\mathcal{T}}, A_{\mathcal{T}} = a_{\mathcal{T}}\}$ and, recursively for $t = \mathcal{T} - 1, \mathcal{T}-2, ..., 1, 0$, $Q_t^Y(h_t, a_t; \pi) = \mathbb{E}\{1_{T=t} u_t(Y_0, ..., Y_t) + 1_{T>t} v_{t+1}(H_{t+1}; \pi)|H_t = h_t, A_t = a_t\}$. The Q-functions related to cost can be similarly expressed in terms of the underlying generative model. To fix ideas, we first describe regression-based estimation in terms of parametric models fit by least squares; generalizations are discussed subsequently. A regression-based procedure for estimating $V_u(\pi)$ and $V_c(\pi)$ is as follows.

1. Postulate a working model $Q_{\mathcal{T}}^Y(\mathbf{h}_{\mathcal{T}}, a_{\mathcal{T}}; \beta_{\mathcal{T}})$ for $Q_{\mathcal{T}}^Y(\mathbf{h}_{\mathcal{T}}, a_{\mathcal{T}})$ indexed by $\beta_{\mathcal{T}} \in \mathbb{R}^{p_{\mathcal{T}}}$ and compute $\widehat{\beta}_{\mathcal{T}} = \arg\min_{\beta \in \mathbb{R}^{p_{\mathcal{T}}}} \sum_{i\,:\,T_i = \mathcal{T}} \left\{u_{\mathcal{T}}(Y_{0,i}, Y_{1,i}, \ldots, Y_{\mathcal{T},i}) - Q_{\mathcal{T}}^Y(\mathbf{H}_{\mathcal{T},i}, A_{\mathcal{T},i}; \beta)\right\}^2$. Define $\widehat{\nu}_{\mathcal{T}}(\mathbf{h}_{\mathcal{T}}; \pi) = Q_{\mathcal{T}}^Y\left\{\mathbf{h}_{\mathcal{T}}, \pi(\mathbf{h}_{\mathcal{T}}); \widehat{\beta}_{\mathcal{T}}\right\}$.

2. For $t = \mathcal{T} - 1, \ldots, 1, 0$ postulate a working model $Q_t^Y(\mathbf{h}_t, a_t; \beta_t^\pi)$ for $Q_t^Y(\mathbf{h}_t, a_t; \pi)$ indexed by $\beta_t^\pi \in \mathbb{R}^{p_t}$ and compute

$$\widehat{\beta}_t^\pi = \arg\min_{\beta \in \mathbb{R}^{p_t}} \sum_{i\,:\,T_i \geq t} \left\{1_{T_i = t} u_t(Y_{0,i}, Y_{1,i}, \ldots, Y_{t,i}) + 1_{T_i \geq t} \widehat{\nu}_{t+1}(\mathbf{H}_{t+1,i}, \pi) - Q_t^Y(\mathbf{H}_{t,i}, A_{t,i}; \beta)\right\}^2.$$

   Define $\widehat{\nu}_t(\mathbf{h}_t, \pi) = Q_t^Y\left\{\mathbf{h}_t, \pi(\mathbf{h}_t); \widehat{\beta}_t^\pi\right\}$.

Define $\widehat{V}_u(\pi) = n^{-1} \sum_{i=1}^n \widehat{\nu}_0^Y(\mathbf{H}_{0,i}, \pi)$ and $\widehat{V}_c(\pi) = n^{-1} \sum_{i=1}^n \widehat{\nu}_0^Z(\mathbf{H}_{0,i}, \pi)$ The preceding algorithm defines the map $\pi \mapsto \left\{\widehat{V}_u(\pi), \widehat{V}_c(\pi)\right\}$ from which we can construct $\widehat{\pi}_\tau = \arg\max_{\pi \in \Pi} \widehat{V}_u(\pi)$ such that $\widehat{V}_c(\pi) \leq \tau$ provided such a solution exists. To compute an $\widehat{\pi}_\tau$ for a parametric class of policies, e.g., $\Pi = \{\pi(\mathbf{s}; \rho) = 1_{\mathbf{s}^\intercal \rho > 0} : \rho \in \mathbb{R}^p\}$, we use grid search. Code implementing the proposed methods and to replicate the simulation study presented in Section 4.4 is in the Supplemental Materials.

For the purpose of exposition, the above algorithm describes $Q$-learning for policy-search using parametric models fit using least squares. However, to safeguard against

potential model-misspecification, it maybe be desirable to use more flexible models for the $Q$-functions (Zhao et al., 2011; Moodie et al., 2013; Laber et al., 2014a). Indeed, any regression algorithm, e.g., trees, ensemble methods, support vector regression, etc., may be incorporated into the $Q$-learning for policy-search algorithm as follows. Let $\mathcal{Q}_\mathcal{T}$ denote a postulated class of functions for $Q^Y_\mathcal{T}(\mathbf{h}_\mathcal{T}, a_\mathcal{T})$ and and $L_\mathcal{T}\mathbb{R}^2 \rightarrow \mathbb{R}_+$ denote a loss function used to define an optimal model in $\mathcal{Q}_\mathcal{T}$. Define $\widehat{Q}^Y_\mathcal{T} = \arg\min_{Q\in\mathcal{Q}_\mathcal{T}} \sum_{i:T_i=\mathcal{T}} L_\mathcal{T}\left\{u_\mathcal{T}(Y_{0,i}, Y_{1,i}, \ldots, Y_{\mathcal{T},i}), Q(\mathbf{H}_{\mathcal{T},i}, A_{\mathcal{T},i})\right\}$. Subsequently, define $\widehat{\nu}^Y_\mathcal{T}(\mathbf{h}_\mathcal{T}; \pi) = \widehat{Q}^Y_\mathcal{T}\left\{\mathbf{h}_\mathcal{T}, \pi(\mathbf{h}_\mathcal{T})\right\}$. For $t = \mathcal{T}-1, \ldots, 1, 0$ postulate class of models $\mathcal{Q}_t$ and loss function $L_t$ and compute

$$\widehat{Q}^Y_t = \arg\min_{Q\in\mathcal{Q}_t} \sum_{i:T_i\geq t} L_t\left\{1_{T_i=t} u_t\left(Y_{0,i,}, Y_{1,i}, \ldots, Y_{t,i}\right) + \widehat{\nu}_{t+1}\left(\mathbf{H}_{t+1,i}, \pi\right), Q\left(\mathbf{H}_{t,i}, A_{t,i}\right)\right\},$$

and subsequently $\widehat{\nu}^Y_t(\mathbf{h}_t, \pi) = \widehat{Q}_t\left\{\mathbf{h}_t, \pi(\mathbf{h}_t)\right\}$. The estimated means under $\pi$ are defined as before, i.e., $\widehat{V}_u(\pi) = n^{-1}\sum_{i=1}^n \widehat{\nu}^Y_0\left(\mathbf{H}_{0,i}, \pi\right)$ and $\widehat{V}_c(\pi) = n^{-1}\sum_{i=1}^n \widehat{\nu}^Z_0\left(\mathbf{H}_{0,i}, \pi\right)$.

## 4.4 Case Study

In this section, we will apply our proposed methods to two studies. Both of the data sets are from Purdue Pharmaceutical Company phase III clinical trails. We name these two studies OXN and BUP. The participants in both studies are patients with Osteoarthritis pain. Patients in these studies received the pain killer medication with predefined initial dosage. At each follow up visit, doctors will decide the change of current dosage (increase, decrease, or stay the same dosage) for each patient.

The potential outcome (efficacy) $Y$ is defined as the average pain score over each visit, and the secondary outcome (safety) $Z$ is defined as indicator of severe adverse event through the whole study. The baseline variables used including: patients age ($age$) and baseline weight ($blweight$).

### 4.4.1 Estimation of $Q$-functions in OXN and BUP

As we mentioned before, in order to plug in our proposed method, we need to estimate the $Q$-functions for each stage. For this case study, we decided to use Random Forest (Breiman 2001) estimate $Q$-functions at time point $t \geq 1$. And for $t = 0$, we decided to

use unimodal regression using Bernstein-Schoenberg Splines and Penalties to estimate $Q_0$ (Köllmann et al. 2014). Next, we will review Random Forest, and unimodal regression using Bernstein-Schoenberg splines and penalties.

### Review of Random Forest

Motivated by the idea of random selection (Amit, 1997), Breiman (2001) proposed a new tree-based ensemble method, random forest (RF). The idea of RF is a substantial modification of bagging. It uses un-pruned decision tree learners with a randomized feature at each split (Hastie et al., 2009).

Suppose we have training data set $\mathcal{D} = \{(x_i, y_i) : x_i = (x_{i1}, x_{i2}, ..., x_{ip});\ i = 1, ..., n\}$, the definition of RF for classification is proposed by Breiman (2001) (Def. 5.1). The algorithm of RF is shown in Algorithm 4.1 (Hastie et al., 2009).

**Def. 4.1** Random Forest for Regression
A random forest is a model consisting of a collection of tree-structured models $\{T(\boldsymbol{x}, \Theta_b), b = 1, 2, ...\}$ where $\{\Theta_b\}$ are independent identically distributed random vectors and each tree casts the average value for the assigned subgroup at input $x$.

**Algorithm 4.1** Random Forest

**(I)** For $b = 1$ to $B$:

**(a)** Draw a bootstrap sample $\mathcal{D}^b$ of size $n$ from the training data $\mathcal{D}$.

**(b)** Grow a random-forest tree predictor $T_b(x) = T(x, \Theta_b)$ from the bootstrapped data $\mathcal{D}^b$, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{min}$ is reached without pruning.
**i.** Select $m$ variables randomly from $p$ variables, where $m \leq p$ is a tuning parameter.
**ii.** Pick the best split (feature variable and split point) among the $m$ variables.
**iii.** Split the node into two children nodes.

**(II)** Output the ensemble of trees $\{T_b(x)\}_1^B$ and define the final RF predictor $T(x)$:

- For regression, $T(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$;

- For classification, $T(x) = \underset{k}{\text{argmax}} \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}(T_b(x) = k)$, where $k = 1, ..., K$ are the possible class labels.

Where minimum node size $n_{min}$ is a tuning parameter with default value 3 (in R).

In RF, we have two randomization procedures. The first randomization procedure is for bootstrap: we generate samples $\mathcal{D}^b$ of size $n$ by random sampling the training data $\mathcal{D}$ with replacement. The second randomization procedure is for the predictor subsets: we select a random subset inputs of size $m \leq p$ for splitting. The second randomization procedure improves computational efficiency especially when $p$ is large. Empirical studies show that the suggested $m$ is $\lfloor \sqrt{p} \rfloor$ for classification and $\lfloor p/3 \rfloor$ for regression (Hastie et al., 2009).

**Review of Unimodal Regression using Bernstein-Schoenberg Splines and Penalties**

Let's consider regression problem, with one continuous predictive variable $X$ and one continuous response $Y$. In a variety of applications, $Y$ increases with higher values of $X$ up to a maximum, and then decreases again. This means the functional relationship between $X$ and $Y$ is unimodal. Dose-response analysis is one prominent example, where $Y$ here is some beneficial effect of a substance that increases with increasing dose $X$ up to a saturation point, after which the effect begin to decrease again because of the toxic effects from the substance. Köllmann et al. (2014) proposed a non-parametric method to estimate the unimodal relationship using Bernstein-Schoenberg Splines with penalties.

Bernstein-Schoenberg spline is based on B-spline. Let $N_{j,k+1}(x)$ be the normalized B-spline basis function of degree $k \geq 1$ with knots $\tau_j, ..., \tau_{j+k+1}$, which can be defined by: $N_{j,1}(x) = 1_{\tau_j \leq x < \tau_{j+1}}$, $N_{j,k+1}(x) = \frac{x-\tau_j}{\tau_{j+k}-\tau_j} N_{j,k}(x) + \frac{\tau_{j+k+1}-x}{\tau_{j+k+1}-\tau_{j+1}} N_{j+1,k}(x)$ for $j = -k, ..., g$, where $g \geq 0$ is the number of inner knots (Dierckx 1995). The Bernstein-Schoenberg (B-S) operator or Bernstein-Schoenberg spline of a function $f$ has specific coefficients, and is defined for $x \in [0, 1]$ as $\mathcal{V}f(x) = \sum_{j=-k}^{g} f(\tau_j^*) N_{j,k+1}(x)$, where $\tau_j^* = \frac{1}{k} \sum_{i=1}^{k} \tau_{j+i}$, $j = -k, ..., g$ are so-called knot averages. If $f$ is a unimodal function, then it is shown that it can be approximated by B-S splines with uniform convergence properties (Köllmann et al. 2014).

In real life, the functional relationship $f$ between $X$ and $Y$ is unknown. Thus it is impossible to choose $\beta_j = f(\tau_j^*)$. We have to estimate the coefficients. One possible approach is to use a penalized spline to estimate the coefficients. The corresponding

Table 4.1: Dosage assignment situation at each time point. There are 852 patients in total. At each time point, patients may drop off due to the severe adverse events.

| Dose Change | t=1 | t=2 | t=3 | t=4 | t=5 |
|---|---|---|---|---|---|
| $A_t = 1$ | 30 | 17 | 14 | 5 | 8 |
| $A_t = 0$ | 747 | 696 | 646 | 610 | 545 |
| $A_t = -1$ | 10 | 11 | 4 | 5 | 10 |

objective function is described below:

$$||\frac{1}{\sigma}(y - B\beta)||_2^2 + \lambda \sum_{j=-k+q}^{g} (\Delta^q \beta_j)^2,$$

where $\Delta\beta_j = \beta_j - \beta_{j-1}$, $\Delta^2\beta_j = \Delta(\Delta\beta_j) = \beta_j - 2\beta_{j-1} + \beta_{j-2}$, and so on. The estimator of the coefficients are those that minimized the objective function, and the parameter $\lambda > 0$ enables tuning of the penalization (Eilers and Marx 1996). Köllmann et al. (2014) wrote a package in R named *uniReg* that contains functions solve this unimodal regression model using the B-S splines with penalties. We will use the function *unireg* in the package to estimate $Q_0$ function in BUP and OXN studies.

### 4.4.2 BUP Study Result

In BUP, the data are collected from the extension study of phase III clinical trail. During the extension procedure, all the patients received the same medication denoted as BTDS, which is one kind of pain killer. There are 6 visits during the study. At baseline, $t = 0$ patients were received the initial dosages. Thus $A_0$ is continuous in $[0, 1]$. At each following stage, $t = 1, 2, ..., 5$, doctors made decisions of dosage change for each patient. At time point $t = 1, 2, ..., 5$: $A_t = 1$ for the decision of increasing the dosage, $A_t = 0$ for the decision of no change of the dosage, and $A_t = -1$ for the decision of decreasing the dosage. Table 4.1 shows the number of patients receiving treatment $A_t$ at each time point $t$. The numbers in the table represent how many patients receiving treatment $A_t$ at time point $t$. Since after $t = 2$, the number of patients receiving increase or decrease dosage decision are very small. We only consider time points at $t = 1, t = 2$. The baseline variables including: patients' age ($age$), baseline weight ($blweight$). The efficacy score defined here is the average cumulative BPI pain score. Let $e_t$ denote the efficacy score

Table 4.2: The estimated optimal treatment regime with different thresholds $\tau$. $\beta_1$ and $\gamma$ are parameters for regime $\pi_1$. $\beta_0$ is the parameter vector for regime $\pi_0$.

| $\tau$ | $\beta_1$ $(age, blweight, e_t, s_t)$ | $\gamma$ | $\beta_0$ $(age, blweight)$ | Estimated Average Pain Score |
|---|---|---|---|---|
| $[0.005, 0.01)$ | (-0.1, -0.3, -0.6, -0.1) | 0.01 | (-0.3, 0.9) | 0.245 |
| $[0.01, 0.02)$ | (0.0, -0.1, -0.6, 0.0) | 0.01 | (0.3, -0.6) | 0.243 |
| $[0.02, 0.03)$ | (0.0, -0.1, -0.3, -0.1) | 0.01 | (-0.3, -0.3) | 0.232 |
| $[0.030, 0.06)$ | (0.0, 0.0,-0.6, -0.3) | 0.10 | (0.3, 0.0) | 0.214 |
| $[0.060, 0.35)$ | (0.0, 0.0, -0.9, -0.6) | 0.10 | (-0.3, -0.9) | 0.213 |

at each time point $t$, which is the BPI pain score recorded at the end of each stage. Let $s_t$ denote the cumulative adverse event indicator. The primary outcome $Y$ equals to $\frac{1}{T}\sum_{t=1}^{T} e_t$. The secondary outcome $Z$ equals to $s_T$. We want to minimize $\mathrm{E}Y$ subject to $\mathrm{E}Z < \tau$, where $\tau \in (0, 1)$ is a predefined value (threshold). Let $\pi_1$ denote the treatment regime that maps patients' up-to-date information to a treatment at time point $t = 1, 2$. Let $X_t = c(age, blweight, e_t, s_t)^T, t = 1, 2$. Then the policy $\pi_1$ could have the expression:

$$\pi_1 = \begin{cases} 1: & x_t^T\beta_1 > \gamma \\ 0: & -\gamma \leq x_t^T\beta_1 \geq \gamma \\ -1: & x_t^T\beta_1 < -\gamma \end{cases}$$

Where $\beta$ and $\gamma$ are tuning parameters, and $t = 1, 2$. We also consider the initial dosage assignment, the initial dosage is in the range $[0, 20]$, and we rescale it into $[0, 1]$. The policy $\pi_0$ for initial treatment assignment has the expression:

$$\mathrm{logit}\pi_0 = X_0^T\beta_0,$$

where $X_0 = c(age, blweight)^T$.

Figure 4.1 shows the changes of the estimated average cumulative pain score over different thresholds $\tau$. As $\tau$ increases, the efficacy score decreases. Table 4.2 shows the estimated average cumulative pain score corresponding to the optimal treatment with different thresholds $\tau$.

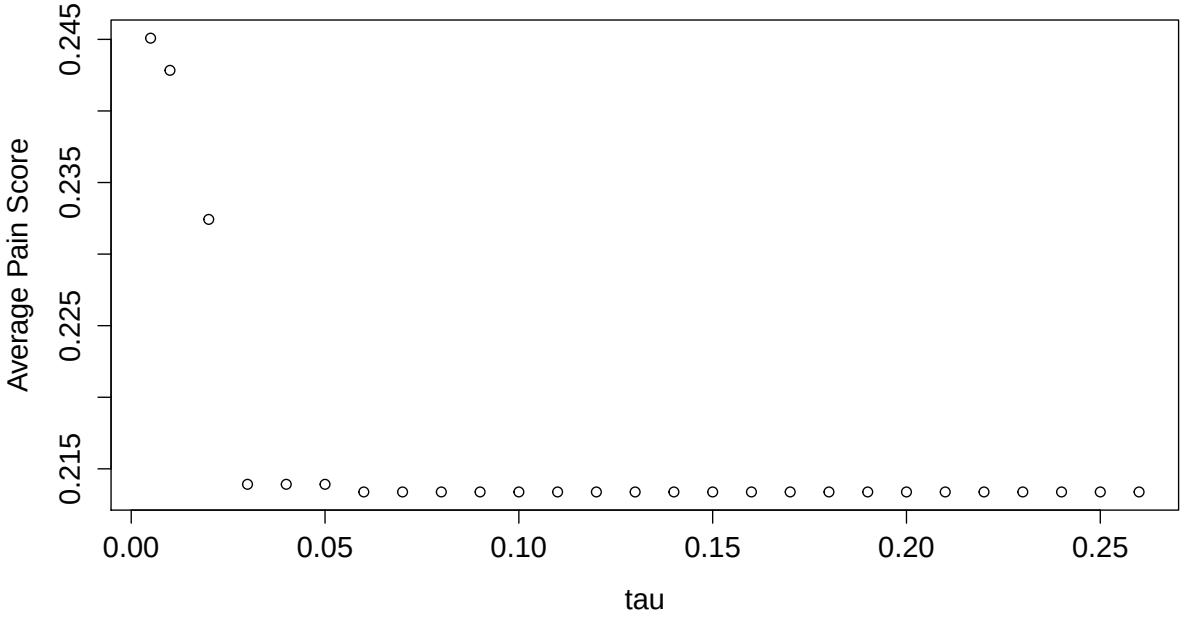


Figure 4.1: Estimated average pain score with different thresholds. $x$-axis represents the different values of threshold $\tau$ for the constraint function. $y$-axis represents the estimated efficacy score corresponding to different thresholds.

Table 4.3: Dosage assignment situation at each time point. There are 460 patients in total. At each time point, patients may drop off due to the severe adverse events.

| Dose Change | t=1 | t=2 | t=3 | t=4 | t=5 |
|---|---|---|---|---|---|
| $A_t = 1$ | 55 | 50 | 61 | 35 | 10 |
| $A_t = 0$ | 395 | 397 | 388 | 407 | 118 |
| $A_t = -1$ | 10 | 12 | 8 | 12 | 324 |

### 4.4.3 OXN Study Result

OXN is another study that is similar to BUP. Patients in this study received dosage change during extension procedure with medication OXY, which is another kind of pain killer. The dataset has 5 stages (time points). Table 4.3 shows the dosage assignment situation at each time point. The numbers in the table represent how many patients receiving treatment $A_t$ at time point $t$. At each time point $t$, two potential outcomes are recorded: average pain score last 24 hours, and adverse event indicator. The baseline variables including: age, baseline weight (blweight). Let $e_t$ denote the average pain score last 24 hours at each time point $t$. The primary outcome $Y$ here equals to $\frac{1}{T}\sum_{t=1}^{T} e_t$. Let $s_t$ denote the cumulative adverse event indicator until time point $t$. The secondary outcome $Z$ is the defined as $s_T$. We want to minimize E$Y$ subject to E$Z < \tau$, where $\tau \in (0, 1)$ is a predefined value (threshold). Similar to the model in the BUP study, we define $X_t = c(age, blweight, e_t, s_t)^T$ for $t = 1, 2, ..., 5$, and $X_0 = c(age, blweight)^T$. The policy of treatment regimes is defined the same as the BUP study.

Table 4.4 shows the estimated average cumulative pain score corresponding to the optimal treatment regime with different thresholds $\tau$. Figure 4.2 shows the changes of the estimated average cumulative pain score over different thresholds $\tau$.

## 4.5 Discussion

In this section, we proposed a method that gives guidance to dosage assignment for patients with consideration of adverse events or side effects. We addressed a continuous initial dosage assignment, and a sequence of decision rules for the following time visits. The proposed method used the non-parametric $Q$-learning framework, but with constraints. We suggested using non-parametric regression models estimate each stage $Q$-functions, and search the best treatment regime from the region of all feasible treatments. The

Table 4.4: The estimated optimal treatment regime with different thresholds $\tau$. $\beta_1$ and $\gamma$ are parameters for regime $\pi_1$. $\beta_0$ is the parameter vector for regime $\pi_0$.

| $\tau$ | $\beta_1$ $(age, blweight, e_t, s_t)$ | $\gamma$ | $\beta_0$ $(age, blweight)$ | Estimated Average Pain Score |
|---|---|---|---|---|
| $[0.005, 0.01)$ | (0.3, -0.6, 0.1, 0.3) | 0.01 | (-0.9, -0.1) | 0.216 |
| $[0.01, 0.02)$ | (0.1, -0.9, 0.1, 0.6) | 0.10 | (0.1, -0.9) | 0.178 |
| $[0.02, 0.03)$ | (-0.1, 0.0, 0.0, 0.6) | 0.30 | (0.1, -0.9) | 0.091 |
| $[0.03, 0.04)$ | (0.0, -0.1, 0.0, 0.9) | 0.40 | (0.9, -0.3) | 0.081 |
| $[0.04, 0.05)$ | ( 0.1, 0.0, 0.3, 0.9) | 0.70 | (0.6, 0.6) | 0.058 |
| $[0.05, 0.06)$ | (0.0, 0.0, 0.3, 0.9) | 0.70 | (0.9, 0.3) | 0.055 |
| $[0.06, 0.07)$ | (0.0, -0.1, 0.3, 0.9) | 0.80 | (0.6, 0.6) | 0.048 |
| $[0.07, 0.08)$ | (0.0, 0.1, 0.3, 0.9) | 0.80 | (-0.1, 0.6) | 0.044 |
| $[0.08, 0.09)$ | (0.0, 0.0, 0.1, 0.3) | 0.30 | (0.9, -0.3) | 0.037 |
| $[0.09, 0.14)$ | (0.0, 0.0, 0.1, 0.6) | 0.60 | (0.3, -0.1) | 0.036 |
| $[0.14, 0.28)$ | (0.3, 0.1, -0.3, -0.6) | 0.80 | (0.3, -0.1) | 0.035 |

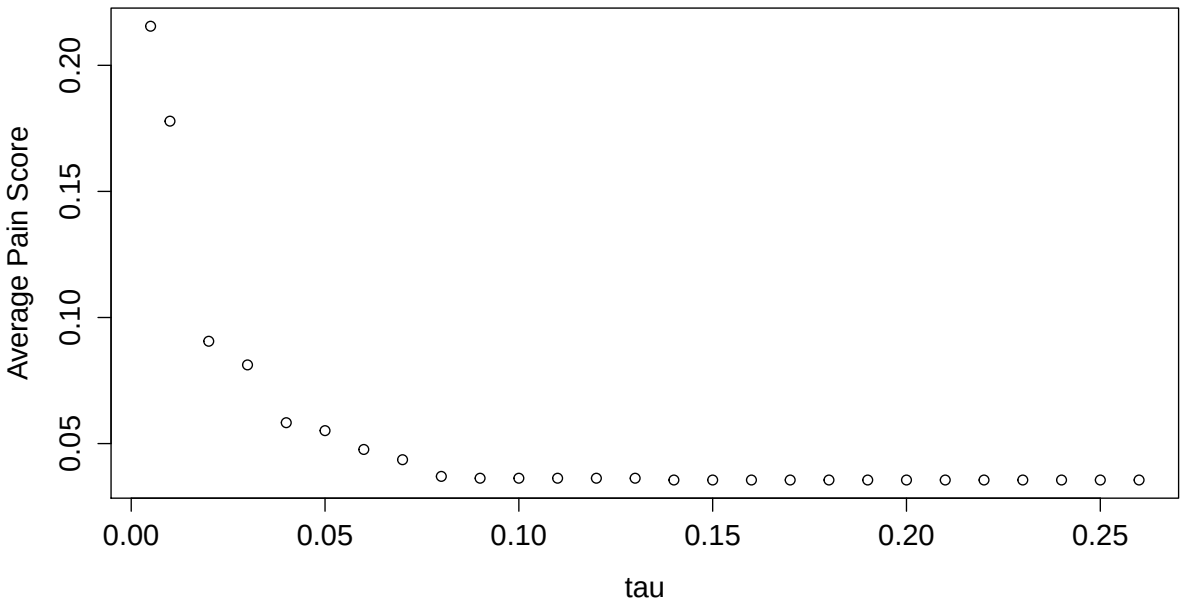


Figure 4.2: Estimated average pain score with different thresholds. $y$-axis represents the estimated efficacy score corresponding to different thresholds.

"best" treatment regime is defined as the one that optimizes the mean outcome across the population of interest while still satisfies some predefined constraints.

When applying our methods to the two real dose-response data sets from Purdue Pharmaceutical company, we chose using shape-constrained spline regression techniques (Köllmann et al., 2014) for estimating initial dosage assignment $Q$-function, and using the random forest method (Breiman, 2001) to estimate the following sequential $Q$-functions. The constraint we considered is the probability of an severe adverse event happened during the treatment.

An interesting aspect for further research are confidence or credible regions. The estimator's distribution under constraint is not known. Thus, bootstrapping may be one potential way to construct confidence intervals for the corresponding estimands. But there is still need to investigate its theoretical properties in the presence of model with constraints.

# References


Yali Amit. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, October 1997.

Donald WK Andrews. Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69(3):683–734, 2001.

Donald WK Andrews and Patrik Guggenberger. Incorrect asymptotic size of subsampling procedures based on post-consistent model selection estimators. *Journal of Econometrics*, 152(1):19–27, 2009.

Donald WK Andrews and Gustavo Soares. Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157, 2010.

Michael Bauer, Andrea Pfennig, Emanuel Severus, Peter C Whybrow, Jules Angst, and Hans-Jürgen Möller. World federation of societies of biological psychiatry (wfsbp) guidelines for biological treatment of unipolar depressive disorders, part 1: update 2013 on the acute and continuation treatment of unipolar depressive disorders. *The World Journal of Biological Psychiatry*, 14(5):334–385, 2013.

R E Bellman. *Dynamic programming*. Princeton University Press, Princeton, NY, 1957.

Roger L Berger and Dennis D Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

P J Bickel and A Sakov. On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, 18:967–985, 2008.

PJ Bickel, F Götze, and WR van Zwet. Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica*, 7:1–31, 1997.

L Breiman, J Friedman, R Olshen, and C Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

J Bretagnolle. Lois limites du bootstrap de certaines fonctionnelles. In *Annales de l'IHP Probabilités et statistiques*, volume 19, pages 281–296. Elsevier, 1983.

B Chakraborty, E B Laber, and Y Zhao. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723, 2013a.

Bibhas Chakraborty and Erica EM Moodie. Statistical reinforcement learning. In *Statistical Methods for Dynamic Treatment Regimes*, pages 31–52. Springer, 2013.

Bibhas Chakraborty, Susan Murphy, and Victor Strecher. Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 2009.

Bibhas Chakraborty, Eric B Laber, and Yingqi Zhao. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3): 714–723, 2013b.

Bibhas Chakraborty, Eric B Laber, and Ying-Qi Zhao. Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials*, 11(4):408–417, 2014.

Xu Cheng. Robust confidence intervals in nonlinear regression under weak identification. *Unpublished working paper, Department of Economics, Yale University*, 2008.

Anthony C Davison and David V Hinkley. Bootstrap methods and their applications, cambridge series in statistical and probabilistic mathematics, 1997.

Ree Dawson and Philip W Lavori. Placebo-free designs for evaluating new mental health treatments: the use of adaptive treatment strategies. *Statistics in medicine*, 23(21): 3249–3262, 2004.

Paul Dierckx. *Curve and surface fitting with splines*. Oxford University Press, 1995.

Lutz Dümbgen. On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95(1):125–140, 1993.

Bradley Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185, 1987.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

Yair Goldberg and Michael R Kosorok. Q-learning with censored data. *Annals of statistics*, 40(1):529, 2012.

HEINZ Grunze, Eduard Vieta, Guy M Goodwin, CHARLES Bowden, Rasmus W Licht, Hans-Juergen Möller, and Siegfried Kasper. The world federation of societies of biological psychiatry (wfsbp) guidelines for the biological treatment of bipolar disorders: update 2010 on the treatment of acute bipolar depression. *World J Biol Psychiatry*, 11(2):81–109, 2010.

Lacey Gunter, Ji Zhu, and Susan Murphy. Variable selection for optimal decision making. In *Artificial Intelligence in Medicine*, pages 149–154. Springer, 2007.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning. Data mining, inference, and prediction. 2nd ed.* Springer Series in Statistics. New York, NY: Springer., 2009.

Keisuke Hirano and Jack R Porter. Impossibility results for nondifferentiable functionals. *Econometrica*, 80(4):1769–1790, 2012.

JS Huang, PK Sen, and J Shao. Bootstrapping a sample quantile when the density has a jump. *Statistica Sinica*, 6(1):299–309, 1996.

Holly Janes, Margaret S Pepe, Patrick M Bossuyt, and William E Barlow. Measuring the performance of markers for guiding treatment decisions. *Annals of internal medicine*, 154(4):253–259, 2011.

Holly Janes, Marshall D Brown, Ying Huang, and Margaret S Pepe. An approach to evaluating and comparing biomarkers for patient treatment selection. *The international journal of biostatistics*, 10(1):99–121, 2014.

Samira Khalili and Antonios Armaou. An extracellular stochastic model of early hiv infection and the formulation of optimal treatment policy. *Chemical Engineering Science*, 63(17):4361–4372, 2008.

Katja Kleine-Budde, Elina Touil, Jörn Moock, Anke Bramesfeld, Wolfram Kawohl, and Wulf Rössler. Cost of illness for bipolar disorder: a systematic review of the economic burden. *Bipolar disorders*, 2013.

Claudia Köllmann, Björn Bornkamp, and Katja Ickstadt. Unimodal regression using bernstein–schoenberg splines and penalties. *Biometrics*, 70(4):783–793, 2014.

Eric B Laber and Susan A Murphy. Adaptive confidence intervals for the test error in classification. *Journal of the American Statistical Association*, 106(495), 2011.

Eric B Laber, Kristin A Linn, and Leonard A Stefanski. Interactive model building for q-learning. *Biometrika*, page asu043, 2014a.

Eric B Laber, Daniel J Lizotte, and Bradley Ferguson. Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, 70(1):53–61, 2014b.

Eric B Laber, Daniel J Lizotte, Min Qian, William E Pelham, Susan A Murphy, et al. Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics*, 8:1225–1272, 2014c.

Philip W Lavori and Ree Dawson. Dynamic treatment regimes: practical design considerations. *Clinical trials*, 1(1):9–20, 2004.

H Lei, I Nahum-Shani, K Lynch, D Oslin, and SA Murphy. A smart design for building individualized treatment sequences. *Annual review of clinical psychology*, 8, 2012.

K.A. Linn, E.B. Laber, and L.A. Stefanski. Constrained estimation for competing outcomes. In E.E.M. Moodie and M.R. Kosorok, editors, *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. CRC Press, Boca Raton, FL, 2015.

R J A Little and D B Rubin. *Statistical analysis with missing data (second edition)*. Chichester: Wiley, 2002.

D.J. Lizotte and E.B. Laber. Multi-objective markov decision support systems. *Under review*, 2015.

Alexander R Luedtke and Mark J van der Laan. Optimal dynamic treatments in resource-limited settings. 2015.

Susan L McElroy, Richard H Weisler, William Chang, Bengt Olausson, Björn Paulsson, Martin Brecher, Vasavan Agambaram, Charles Merideth, Arvid Nordenhem, and Allan H Young. A double-blind, placebo-controlled study of quetiapine and paroxetine as monotherapy in adults with bipolar depression (embolden ii). *Journal of Clinical Psychiatry*, 71(2):163–174, 2010.

IW McKeague and M Qian. Evaluation of treatment policies via sparse functional linear regression. *Biometrics*, 2011.

Erica EM Moodie, Nema Dean, and Yue Ru Sun. Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, pages 1–21, 2013.

S A Murphy. Optimal dynamic treatment regimes (with discussion). *Journal of the Royal Statistical Society*, 65(2):331–366, 2003a.

S A Murphy, M J Van Der Laan, and J M Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003b.

Susan A Murphy. A generalization error for q-learning. *Journal of machine learning research: JMLR*, 6:1073, 2005a.

Susan A Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481, 2005b.

I Nahum-Shani, M Qian, D Almirall, W E Pelham, B Gnagy, G A Fabiano, J G Waxmonsky, J Yu, and S A Murphy. Q-learning: A data analysis method for constructing adaptive interventions. *Psychological Methods*, 17(4):478–494, December 2012.

Richard A Olshen. The conditional level of the ftest. *Journal of the American Statistical Association*, 68(343):692–698, 1973.

L. Orellana, A. Rotnitzky, and J. Robins. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content. *Int. Jrn. of Biostatistics*, 6(2), 2010.

Isabella Pacchiarotti, David J Bond, Ross J Baldessarini, Willem A Nolen, Heinz Grunze, Rasmus W Licht, Robert M Post, Michael Berk, Guy M Goodwin, Gary S Sachs, et al. The international society for bipolar disorders (isbd) task force report on antidepressant use in bipolar disorders. *American Journal of Psychiatry*, 170(11):1249–1262, 2013.

Mary L Phillips and David J Kupfer. Bipolar disorder diagnosis: challenges and future directions. *The Lancet*, 381(9878):1663–1671, 2013.

Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, and Peter Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology*, 27(1):85–96, 2001.

J M Robins, M A Hernan, and B Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

James Robins, Liliana Orellana, and Andrea Rotnitzky. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in medicine*, 27(23):4678–4721, 2008.

James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

D B Rubin. *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: John Wiley & Sons, 2004.

D.B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.

Gary S Sachs, Constance Guille, and Stephanie L McMurrich. A clinical monitoring form for mood disorders. *Bipolar disorders*, 4(5):323–327, 2002.

Gary S Sachs, Michael E Thase, Michael W Otto, Mark Bauer, David Miklowitz, Stephen R Wisniewski, Philip Lavori, Barry Lebowitz, Mathew Rudorfer, Ellen Frank, et al. Rationale, design, and methods of the systematic treatment enhancement program for bipolar disorder (step-bd). *Biological psychiatry*, 53(11):1028–1042, 2003.

Gary S Sachs, Andrew A Nierenberg, Joseph R Calabrese, Lauren B Marangell, Stephen R Wisniewski, Laszlo Gyulai, Edward S Friedman, Charles L Bowden, Mark D Fossey, Michael J Ostacher, et al. Effectiveness of adjunctive antidepressant treatment for bipolar depression. *New England Journal of Medicine*, 356(17):1711–1722, 2007.

Phillip J Schulte, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Q- and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*, In Press, 2012.

G Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

Emanuel Severus, Florian Seemüller, Michael Berger, Sandra Dittmann, Michael Obermeier, Andrea Pfennig, Michael Riedel, Sophia Frangou, Hans-Jürgen Möller, and Michael Bauer. Mirroring everyday clinical practice in clinical trial design: a new concept to improve the external validity of randomized double-blind placebo-controlled trials in the pharmacological treatment of major depression. *BMC medicine*, 10(1):67, 2012.

J Shao. Bootstrap sample size in nonregular cases. *Proceedings of the American Mathematical Society*, 122(4):1251–1262, 1994.

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

Jun Shao and CF Jeff Wu. A general theory for jackknife variance estimation. *The Annals of Statistics*, pages 1176–1197, 1989.

Susan M Shortreed and Erica EM Moodie. Estimating the optimal dynamic antipsychotic treatment regime: evidence from the sequential multiple-assignment randomized clinical antipsychotic trials of intervention and effectiveness schizophrenia study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):577–599, 2012.

Susan M Shortreed, Eric Laber, Daniel J Lizotte, T Scott Stroup, Joelle Pineau, and Susan A Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, 84(1-2):109–136, 2011.

Richard Simon. Lost in translation: problems and pitfalls in translating laboratory observations to clinical utility. *European Journal of Cancer*, 44(18):2707–2713, 2008.

Xiao Song and Margaret Sullivan Pepe. Evaluating markers for selecting a patient's treatment. *Biometrics*, 60(4):874–883, 2004.

Jerzy Splawa-Neyman, DM Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. Cambridge Univ Press, 1998.

Jan WH Swanepoel. A note on proving that the (modified) bootstrap works. *Communications in Statistics-Theory and Methods*, 15(11):3193–3203, 1986.

Enrico Tedeschini, Yeciel Levkovitz, Nadia Iovieno, Victoria E Ameral, J Craig Nelson, and George I Papakostas. Efficacy of antidepressants for late-life depression: a meta-analysis and meta-regression of placebo-controlled randomized trials. *The Journal of clinical psychiatry*, 72(12):1660–1668, 2011.

Peter F Thall, Christopher Logothetis, Lance C Pagliaro, Sijin Wen, Melissa A Brown, Dallas Williams, and Randall E Millikan. Adaptive therapy for androgen-independent prostate cancer: A randomized selection trial of four regimens. *Journal of the National Cancer Institute*, 99(21):1613–1622, 2007.

Robert J Tibshirani. Bootstrap confidence intervals. Technical report, DTIC Document, 1984.

Stef Van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242, 2007.

Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2012.

Stef Van Buuren, Jaap PL Brand, CGM Groothuis-Oudshoorn, and Donald B Rubin. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064, 2006.

Aad Van Der Vaart. On differentiable functionals. *The Annals of Statistics*, pages 178–204, 1991.

Theo Vos, Abraham D Flaxman, Mohsen Naghavi, Rafael Lozano, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, Victor Aboyans, et al. Years lived with disability (ylds) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. *The Lancet*, 380(9859):2163–2196, 2013.

E H Wagner, B T Austin, C Davis, M Hindmarsh, J Schaefer, and A Bonomi. Improving chronic illness care: Translating evidence into action. *Health Aff*, 20(6):64–78, November 2001.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Baqun Zhang, Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012a.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012b.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694, 2013.

Yichi Zhang, Eric B Laber, Anastasios Tsiatis, and Marie Davidian. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, In press, 2015.

Ying-Qi Zhao, Donglin Zeng, Eric B Laber, and Michael R Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, (just-accepted):00–00, 2014.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Yufan Zhao, Donglin Zeng, Mark A Socinski, and Michael R Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4): 1422–1433, 2011.

# Appendix

# Appendix A

# Proofs and Additional Results

## A.1 Some Proofs in Chapter 2

### A.1.1 PCI width for Toy Example

Recall that the target estimand is $\theta^*(\beta^*) = \mathrm{E}[X^T\beta^*]_+$, where $\mathrm{P}\{X^T\beta^* > 0\} = 1$. Hence, this esimand can be rewritten as $\theta^*(\beta^*) = \mathrm{E}(X)^T\beta^*$. We assume that estimand is based on linear regression model: $Y = X^T\beta^* + \epsilon$, where random variable $X$ and $\epsilon$ are independent. Let $\mathrm{E}(X) = \mu_x$, $\mathrm{Var}(X) = \Omega_x$, $\mathrm{E}(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2$. Let $\hat{\beta}_n = (\mathbb{E}_n XX^T)^{-1}\mathbb{E}_n XY$ denote the least square estimator of $\beta^*$.

#### Standard CI Width

Then we can derive:

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta^*) =& (\mathbb{E}_n XX^T)^{-1}\sqrt{n}(\mathbb{E}_n - \mathrm{E})X(Y - X^T\beta^*) \\ =& (\mathbb{E}_n XX^T)^{-1}\sqrt{n}(\mathbb{E}_n - \mathrm{E})X\epsilon \\ =& (\mathrm{E}XX^T)^{-1}\sqrt{n}(\mathbb{E}_n - \mathrm{E})X\epsilon + o_p(1)\end{aligned} \tag{A.1}$$

By Slutsky theorem, $\sqrt{n}(\hat{\beta}_n - \beta^*) \xrightarrow{D} \mathrm{N}\left(0, (\mathrm{E}XX^T)^{-1}\sigma_\epsilon^2\right)$. Let $\hat{\theta}_n(\hat{\beta}_n) = \mathbb{E}_n X^T \hat{\beta}_n$ denote the corresponding estimator for $\theta^*(\beta^*)$. We can derive:

$$\begin{aligned}\sqrt{n}\{\hat{\theta}_n(\hat{\beta}_n) - \theta^*(\beta^*)\} =& \sqrt{n}(\mathbb{E}_n X^T \hat{\beta}_n - \mathrm{E}X^T\beta^*) \\ =& \sqrt{n}\mathbb{E}_n X^T(\hat{\beta}_n - \beta^*) + \sqrt{n}(\mathbb{E}_n - \mathrm{E})X^T\beta^* \\ =& \mathrm{E}X^T(\mathrm{E}XX^T)^{-1}\sqrt{n}(\mathbb{E}_n - \mathrm{E})X\epsilon + \beta^{*^T}\sqrt{n}(\mathbb{E}_n - \mathrm{E})X + o_p(1)\end{aligned} \tag{A.2}$$

Noting that $X$ and $\epsilon$ are independent. We have:

$$\begin{aligned}\mathrm{E}X\epsilon &= 0 \\ \mathrm{E}(X - \mu_x)\epsilon &= 0 \\ \mathrm{E}XX^T\epsilon &= 0\end{aligned} \tag{A.3}$$

By CLT (central limit theorem), we can get:

$$\sqrt{n}(\mathbb{E}_n - \mathrm{E})\begin{pmatrix}(\mathrm{E}XX^T)^{-1}X\epsilon \\ X\end{pmatrix} \xrightarrow{D} \mathrm{MVN}\left(0, \begin{pmatrix}(\mathrm{E}XX^T)^{-1}\sigma_\epsilon^2 & 0 \\ 0 & \Omega_x\end{pmatrix}\right). \tag{A.4}$$

Hence the limiting distribution of $\sqrt{n}\{\hat{\theta}_n(\hat{\beta}_n) - \theta^*(\beta^*)\}$ will be:

$$\sqrt{n}\{\hat{\theta}_n(\hat{\beta}_n) - \theta^*(\beta^*)\} \xrightarrow{D} \mathrm{N}\left(0, \mu_x^T(\mathrm{E}XX^T)^{-1}\mu_x\sigma_\epsilon^2 + \beta^{*^T}\Omega_x\beta^*\right). \tag{A.5}$$

Hence a $(1-\alpha-\eta)\times 100\%$ confidence interval (CI) of $\theta^*(\beta^*)$ will be:

$$\hat{\theta}_n(\hat{\beta}_n) \pm z_{1-\frac{\alpha+\eta}{2}}\sqrt{\frac{\mu_x^T(\mathrm{E}XX^T)^{-1}\mu_x\sigma_\epsilon^2 + \beta^{*^T}\Omega_x\beta^*}{n}}, \tag{A.6}$$

where $z_c$ represents the $c$th quantile value from standard normal distribution.

**Projection CI Width**

Based on limiting distribution of $\beta^*$, the $(1-\eta)\times 100\%$ Wald CI of $\beta^*$ can be written as:

$$\mathcal{C}_{(1-\eta),\beta^*} \triangleq \{\beta:\ n(\beta - \hat{\beta}_n)^T\frac{\mathrm{E}XX^T}{\sigma_\epsilon^2}(\beta - \hat{\beta}_n) \le C\}, \tag{A.7}$$

where $C = \chi^2_{p,1-\eta}$. If we define $\nu \in \mathrm{R}^p$ as any unit vector and we define:

$$\gamma = \hat{\beta}_n + \frac{(\mathrm{E}XX^T)^{-\frac{1}{2}}\nu\sqrt{C}\sigma_\epsilon}{\sqrt{n}}, \tag{A.8}$$

then $\gamma \in \mathcal{C}_{(1-\eta,\beta^*)}$. Thus, for $\forall\gamma$, by the CLT, we can derive:

$$\sqrt{n}(\mathbb{E}_n - \mathrm{E})(X^T\gamma) \xrightarrow{D} \mathrm{N}\left(\gamma^T\Omega_x\gamma\right). \tag{A.9}$$

A $(1-\alpha)\times 100\%$ CI of $\mathrm{E}X^T\gamma$ will be:

$$\mathcal{I}_{(1-\alpha),\theta^*(\gamma)} \triangleq \mathbb{E}_n X^T\gamma \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\gamma^T\Omega_x\gamma}{n}}. \tag{A.10}$$

Further, by the projection interval method, a $(1-\alpha-\eta)\times 100\%$ CI of $\theta^*(\beta^*)$ will be:

$$\bigcup_{\gamma\in\mathcal{C}_{(1-\eta),\beta^*}} \mathcal{I}_{(1-\alpha),\theta^*(\gamma)}. \tag{A.11}$$

Noting that by the definition of $\gamma$, we can derive:

$$\gamma^T\Omega_x\gamma = \hat{\beta}_n^T\Omega_x\hat{\beta}_n + O_p(\frac{1}{\sqrt{n}}). \tag{A.12}$$

Then the upper bound of (A.11) is at least as big as:

$$\mathbb{E}_n X^T\left\{\hat{\beta}_n + \frac{(\mathrm{E}XX^T)^{-\frac{1}{2}}\nu\sqrt{C}\sigma_\epsilon}{\sqrt{n}}\right\} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\beta}_n^T\Omega_x\hat{\beta}_n}{n}} + O_p(\frac{1}{n}) \tag{A.13}$$

Similarly, using $-\nu$, we can get the lower bound:

$$\mathbb{E}_n X^T\left\{\hat{\beta}_n - \frac{(\mathrm{E}XX^T)^{-\frac{1}{2}}\nu\sqrt{C}\sigma_\epsilon}{\sqrt{n}}\right\} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\beta}_n^T\Omega_x\hat{\beta}_n}{n}} + O_p(\frac{1}{n}) \tag{A.14}$$

Therefore the width of the projection CI will be at least:

$$\frac{2}{\sqrt{n}}\mathbb{E}_n X^T(\mathrm{E}XX^T)^{-\frac{1}{2}}\nu\sqrt{C}\sigma_\epsilon + \frac{2}{\sqrt{n}}z_{1-\frac{\alpha}{2}}\sqrt{\hat{\beta}_n^T\Omega_x\hat{\beta}_n} \tag{A.15}$$

**The Comparison of Width of Projection CI and Standard CI**

By the definition, we can derive the following:

$$\begin{aligned}\sqrt{C} =&\chi^2_{p,1-\eta} \geq z_{1-\frac{\eta}{2}}\\ z_{1-\frac{\eta}{2}} \geq& z_{1-\frac{\eta+\alpha}{2}}\\ z_{1-\frac{\alpha}{2}} \geq& z_{1-\frac{\eta+\alpha}{2}}\end{aligned}$$

Define $\nu \triangleq \frac{(\mathrm{E}XX^T)^{-\frac{1}{2}}\hat{\mu}_x}{\|(\mathrm{E}XX^T)^{-\frac{1}{2}}\hat{\mu}_x\|}$. Then the width of projection CI can be written as:

$$\begin{aligned}&\frac{2}{\sqrt{n}}\left(\hat{\mu}_x(\mathrm{E}XX^T)^{-\frac{1}{2}}\nu\sqrt{C}\sigma_\epsilon + z_{1-\frac{\alpha}{2}}\sqrt{\hat{\beta}_n^T\Omega_x\hat{\beta}_n}\right)\\ =&\frac{2}{\sqrt{n}}\left(\sqrt{C}\sigma_\epsilon\frac{\hat{\mu}_x^T(\mathrm{E}XX^T)^{-1}\hat{\mu}_x}{\parallel(\mathrm{E}XX^T)^{-\frac{1}{2}}\hat{\mu}_x\parallel} + z_{1-\frac{\alpha}{2}}\sqrt{\hat{\beta}_n^T\Omega_x\hat{\beta}_n}\right)\\ =&\frac{2}{\sqrt{n}}\left(\sqrt{C}\sigma_\epsilon\parallel(\mathrm{E}XX^T)^{-\frac{1}{2}}\hat{\mu}_x\parallel + z_{1-\frac{\alpha}{2}}\sqrt{\hat{\beta}_n^T\Omega_x\hat{\beta}_n}\right)\\ \geqslant&\frac{2}{\sqrt{n}}\sqrt{C\sigma_\epsilon^2\hat{\mu}_x^T(\mathrm{E}XX^T)^{-1}\hat{\mu}_x + z^2_{1-\frac{\alpha}{2}}\hat{\beta}_n^T\Omega_x\hat{\beta}_n},\end{aligned} \tag{A.16}$$

where the last term is derived by Minkowski Inequality. Using $\mu_x$ and $\beta^*$ instead of $\hat{\mu}_x$ and $\hat{\beta}_n$, the (A.16) can be written as:

$$\frac{2}{\sqrt{n}}\sqrt{C\sigma_\epsilon^2\mu_x^T(\mathrm{E}XX^T)^{-1}\mu_x + z^2_{1-\frac{\alpha}{2}}\beta^{*T}\Omega_x\beta^*}. \tag{A.17}$$

Noting that the width of standard CI is:

$$\frac{2}{\sqrt{n}}\sqrt{z^2_{1-\frac{\eta+\alpha}{2}}\sigma_\epsilon^2\mu_x^T(\mathrm{E}XX^T)^{-1}\mu_x + z^2_{1-\frac{\eta+\alpha}{2}}\beta^{*T}\Omega_x\beta^*}. \tag{A.18}$$

By (A.1.1), (A.17) $\geq$ (A.18). Therefore, the width of PCI is larger than the width of standard CI.

## A.1.2 Proof of API Properties

Suppose we observe a dataset, $\mathbf{D} = \{(X^i, Y^i)\}_{i=1}^n$, where $X^i \in \mathbb{R}^p$ represents predictive variable vector with an unknown distribution, and $Y^i$ is the response variable. Suppose

we are interested in an estimand, say $\theta^*(\beta^*) = \mathrm{E}\{f(X, \beta^*)\}$, where $\beta^*$ is a nuisance parameter, which is related to the distribution of the dataset. $f(\cdot)$ is a function, which is not differentiable (non-smooth) at any points that satisfies $(x, \beta^*) \in \mathcal{Q}$. The region $\mathcal{Q}$ is a closed set, and $X$ represents a random vector with the same distribution as predictor $X^i$. We assume that $f(x, \beta^*)$ is first and second order differentiable (smooth) at any point that satisfies $(x, \beta^*) \notin \mathcal{Q}$. Our goal is to construct a confidence interval for $\theta^*(\beta^*)$. In many cases, the function $f(X, \beta^*)$ can be rewritten as $s(X_1, \beta_1^*)g(X_2, \beta_2^*)$, where $X_1, X_2$ are subsets of $X$, $\beta_1^*, \beta_2^*$ are subsets of $\beta^*$, $s(x_1, \beta_1^*)$ is smooth, and $g(x_2, \beta_2^*)$ is non-smooth in the region $\mathcal{Q}$. Define:

$$\theta^*(r, \beta^*) \;=\; \mathrm{E}\{s(X_1, \beta_1^*)g(X_2, r)1_{(X,\beta^*)\in\mathcal{Q}} + s(X_1, \beta_1^*)g(X_2, \beta_2^*)1_{(X,\beta^*)\notin\mathcal{Q}}\},$$

where $r$ is a fixed vector. Let $\hat{\beta}_n$ denote a consistent estimator of $\beta^*$, where $\sqrt{n}(\hat{\beta}_n - \beta^*) \to N_p(0, \Omega_{\beta^*})$ as $n \to \infty$ and $\Omega_{\beta^*}$ denotes the variance matrix. Let $T_n(x)$ denote a test statistic testing $H_0 : (x, \beta^*) \in \mathcal{Q}$ against the alternative. Reject $H_0$ if $T_n(x) > \lambda_n(x, \mathbf{D})$, where $\lambda_n(x, \mathbf{D})$ is a potential tuning parameter related to the observed dataset $\mathbf{D}$. Then we can define an estimator:

$$\hat{\theta}_n(r, \hat{\beta}_n) \;=\; \mathbb{E}_n\{s(X_1, \hat{\beta}_{1n})g(X_2, r)1_{T_n(X)\le\lambda_n} + s(X_1, \hat{\beta}_{1n})g(X_2, \hat{\beta}_{2n})1_{T_n(X)>\lambda_n}\},$$

where $\hat{\beta}_{1n}, \hat{\beta}_{2n}$ are subsets of $\hat{\beta}_n$, and $\mathbb{E}_n$ represents empirical expectation with respect to $X$. Let $\mathcal{C}_{\beta^*,1-\eta}$ denote the $(1-\eta)\times 100\%$ CI for $\beta^*$. This CI is derived from the limiting distribution of $\sqrt{n}(\hat{\beta}_n - \beta^*)$. Let $\mathcal{I}_{\theta(r,\beta^*),1-\alpha}$ denote the $(1-\alpha)\times 100\%$ CI for $\theta^*(r, \beta^*)$. We suppose $\mathcal{I}_{\theta^*(r,\beta^*),1-\alpha}$ is constructed through the limiting distribution of $\sqrt{n}\{\hat{\theta}_n(r, \hat{\beta}_n) - \theta^*(r, \beta^*)\}$. We argue that a $(1-\alpha-\eta)\times 100\%$ CI for $\theta^*(\beta^*)$ is:

$$\bigcup_{r\in\mathcal{C}_{\beta^*,1-\eta}} \mathcal{I}_{\theta^*(r,\beta^*),1-\alpha}.$$

We want to use the limiting distribution of $\sqrt{n}\{\hat{\theta}_n(r, \hat{\beta}_n) - \theta^*(r, \beta^*)\}$ to construct a CI for $\theta^*(r, \beta^*)$. We first prove that $\sqrt{n}\{\hat{\theta}_n(r, \hat{\beta}_n) - \theta^*(r, \beta^*)\}$ converges to a normal distribution with mean 0 and finite variance.

Because we use the percentile bootstrap to construct a confidence interval for $\theta^*(r, \beta^*)$, we also prove that the limiting distribution of $\sqrt{n}\{\hat{\theta}_n^{(b)}(r, \hat{\beta}_n) - \hat{\theta}_n(r, \hat{\beta}_n)\}$ has the same limiting distribution as $\sqrt{n}\{\hat{\theta}_n(r, \hat{\beta}_n) - \theta^*(r, \beta^*)\}$.

Then, we prove that the CI we construct satisfies:

$$\Pr\left(\theta^*(\beta^*) \notin \bigcup_{r \in \mathcal{C}_{\beta^*,1-\eta}} \mathcal{I}_{\theta^*(r,\beta^*),1-\alpha}\right) \leq \alpha + \eta + o(1).$$

**Proof limiting distribution of $\sqrt{n}\{\hat{\theta}_n(r,\hat{\beta}_n) - \theta^*(r,\beta^*)\}$**

We want to prove $\sqrt{n}\{\hat{\theta}_n(r,\hat{\beta}_n) - \theta^*(r,\beta^*)\}$ is asymptotically normal with mean 0 and a finite variance. We first define two functions $\psi(\cdot), \phi(\cdot)$:

$$\begin{aligned}
\psi(X, r, \hat{\beta}_n, \lambda_n) &= s(X_1, \hat{\beta}_{1n}) g(X_2, r) 1_{T_n(X) \leq \lambda_n} + s(X_1, \hat{\beta}_{1n}) g(X_2, \hat{\beta}_{2n}) 1_{T_n(X) > \lambda_n} \\
\phi(X, r, \hat{\beta}_n, \beta^*) &= s(X_1, \hat{\beta}_{1n}) g(X_2, r) 1_{(X,\beta^*) \in \mathcal{Q}} + s(X_1, \hat{\beta}_{1n}) g(X_2, \hat{\beta}_{2n}) 1_{(X,\beta^*) \notin \mathcal{Q}}.
\end{aligned}$$

Here we assume that: $\sqrt{n}\mathbb{E}_n\{\psi(X, r, \hat{\beta}_n, \lambda_n) - \phi(X, r, \hat{\beta}_n, \beta^*)\} \xrightarrow{p} 0$. Our target $\sqrt{n}\{\hat{\theta}(r,\hat{\beta}_n) - \theta^*(r,\beta^*)\}$ can be rewritten as:

$$\begin{aligned}
& \sqrt{n}\left[\mathbb{E}_n\{\psi(X, r, \hat{\beta}_n, \lambda_n)\} - \mathrm{E}\{\phi(X, r, \beta^*, \beta^*)\}\right] \\
= \; & \sqrt{n}\left[\mathbb{E}_n\{\psi(X, r, \hat{\beta}_n, \lambda_n)\} - \mathbb{E}_n\{\phi(X, r, \hat{\beta}_n, \beta^*)\}\right] + \sqrt{n}\left[\mathbb{E}_n\{\phi(X, r, \hat{\beta}_n, \beta^*)\} - \mathrm{E}\{\phi(X, r, \hat{\beta}_n, \beta^*)\}\right] \\
& + \sqrt{n}\left[\mathrm{E}\{\phi(X, r, \hat{\beta}_n, \beta^*)\} - \mathrm{E}\{\phi(X, r, \beta^*, \beta^*)\}\right] \\
= \; & \sqrt{n}\{(1) + (2) + (3)\}.
\end{aligned}$$

By assumption, (1) converges in probability to 0. We then derive the limiting distribution of (3), which can be written as:

$$\begin{aligned}
(3) &= \sqrt{n}\left[\mathrm{E}\{\phi(X, r, \hat{\beta}_n, \beta^*)\} - \mathrm{E}\{\phi(X, r, \beta^*, \beta^*)\}\right] \\
&= \sqrt{n}\,\mathrm{E}\left\{s(X_1, \hat{\beta}_{1n}) g(X_2, \hat{\beta}_{2n}) 1_{(X,\beta^*) \notin \mathcal{Q}} + s(X_1, \hat{\beta}_{1n}) g(X_2, r) 1_{(X,\beta^*) \in \mathcal{Q}}\right\} \\
&\quad - \sqrt{n}\,\mathrm{E}\left\{s(X_1, \beta_1^*) g(X_2, \beta_2^*) 1_{(X,\beta^*) \notin \mathcal{Q}} + s(X_1, \beta_1^*) g(X_2, r) 1_{(X,\beta^*) \in \mathcal{Q}}\right\} \\
&= \sqrt{n}\mathrm{E}\left[\left\{s(X_1, \hat{\beta}_{1n}) g(X_2, \hat{\beta}_{2n}) - s(X_1, \beta_1^*) g(X_2, \beta_2^*)\right\} 1_{(X,\beta^*) \notin \mathcal{Q}}\right] \\
&\quad + \sqrt{n}\mathrm{E}\left[\left\{s(X_1, \hat{\beta}_{1n}) g(X_2, r) - s(X_1, \beta_1^*) g(X_2, r)\right\} 1_{(X,\beta^*) \in \mathcal{Q}}\right] \\
&= (3.1) + (3.2),
\end{aligned}$$

In (3.1), $s(X_1,\beta_1)g(X_2,\beta_2)$ is differentiable with respect to $\beta$, which is in the neighborhood of $\beta^*$ such that $(X,\beta)\notin\mathcal{Q}$. By Taylor expansion and mean-value theorem, there exists a via point $\tilde{\beta}_n$ that is on the linear segment of $\beta^*$ and $\hat{\beta}_n$, and such that (3.1) can be written as:

$$\begin{aligned}(3.1) &= \sqrt{n}(\hat{\beta}_n-\beta^*)^T\mathrm{E}\left(\frac{\mathrm{d}s(X_1,\beta_1)g(X_2,\beta_2)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\notin\mathcal{Q}}\right)\\ &\quad+\sqrt{n}\frac{1}{2}(\hat{\beta}_n-\beta^*)^T\mathrm{E}\left(\frac{\mathrm{d}^2s(X_1,\beta_1)g(X_2,\beta_2)}{\mathrm{d}\beta^T\mathrm{d}\beta}|_{\beta=\tilde{\beta}_n}1_{(X,\beta^*)\notin\mathcal{Q}}\right)(\hat{\beta}_n-\beta^*)\\ &= (3.1.1)+(3.1.2).\end{aligned}$$

We know that as $n\to\infty$, $\hat{\beta}_n\xrightarrow{p}\beta^*$. This implies $\tilde{\beta}_n\xrightarrow{p}\beta^*$. Therefore, as $n$ becomes large enough, $s(\cdot)g(\cdot)$ is differentiable at $\tilde{\beta}_n$. We assume that $\|\mathrm{E}\left(\frac{\mathrm{d}s(X_1,\beta_1)g(X_2,\beta_2)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\notin\mathcal{Q}}\right)\|$ and $|\mathrm{E}\left(\frac{\mathrm{d}^2s(X_1,\beta_1)g(X_2,\beta_2)}{\mathrm{d}\beta^T\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\notin\mathcal{Q}}\right)|$ are finite. In (3.1.1), $\sqrt{n}(\hat{\beta}_n-\beta^*)\to N_p(0,\Omega_{\beta^*})$. And by assumption, each element in $\mathrm{E}\left(\frac{\mathrm{d}s(X_1,\beta_1)g(X_2,\beta_2)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\notin\mathcal{Q}}\right)$ is finite. By Slutsky theorem, (3.1.1) converges to a normal distribution with mean 0 and variance:

$$\sigma_1^2=\mathrm{E}\left(\frac{\mathrm{d}s(X_1,\beta_1)g(X_2,\beta_2)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\notin\mathcal{Q}}\right)^T\Omega_{\beta^*}\mathrm{E}\left(\frac{\mathrm{d}s(X_1,\beta_1)g(X_2,\beta_2)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\notin\mathcal{Q}}\right).$$

In (3.1.2), let $d_{ij}(X,\tilde{\beta}_n)$ denote the $i$th row, $j$th column element in $\frac{\mathrm{d}^2s(X_1,\beta_1)g(X_2,\beta_2)}{\mathrm{d}\beta^T\mathrm{d}\beta}|_{\beta=\tilde{\beta}_n}1_{(X,\beta^*)\notin\mathcal{Q}}$. Then:

$$\begin{aligned}|\mathrm{E}d_{ij}(X,\tilde{\beta}_n)| &\le \mathrm{E}|d_{ij}(X,\tilde{\beta}_n)|\\ &\le \mathrm{E}\sup_{\beta\in N_{\beta^*}}|d_{ij}(X,\beta)|+\mathrm{E}|d_{ij}(X,\tilde{\beta}_n)|1_{\tilde{\beta}_n\notin N_{\beta^*}}\\ &< \infty\end{aligned}$$

where $N_{\beta^*}$ is the neighborhood of $\beta^*$ such that point $\beta$ in $N_{\beta^*}$ satisfies $(X,\beta)\notin\mathcal{Q}$. And as $n\to\infty$, $\tilde{\beta}_n\xrightarrow{p}\beta^*$. Therefore, $\mathrm{E}|d_{ij}(X,\tilde{\beta}_n)|1_{\tilde{\beta}_n\notin N_{\beta^*}}=o_p(1)$. Hence by Slutsky theorem, (3.1.2) converges in probability to 0.

Similarly, by Taylor expansion and mean-value theorem, (3.2) can be written as:

$$\begin{aligned}(3.2) &= \sqrt{n}(\hat{\beta}_n - \beta^*)^T \mathrm{E}\left(\frac{\mathrm{d}s(X_1,\beta_1)g(X_2,r)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\in\mathcal{Q}}\right) \\ &\quad + \sqrt{n}\frac{1}{2}(\hat{\beta}_n - \beta^*)^T \mathrm{E}\left(\frac{\mathrm{d}^2 s(X_1,\beta_1)g(X_2,r)}{\mathrm{d}\beta^T\mathrm{d}\beta}|_{\beta=\tilde{\beta}_n}1_{(X,\beta^*)\in\mathcal{Q}}\right)(\hat{\beta}_n - \beta^*) \\ &= (3.2.1) + (3.2.2).\end{aligned}$$

Similar to (3.1.1) and (3.1.2), we assume that $\|\mathrm{E}\left(\frac{\mathrm{d}s(X_1,\beta_1)g(X_2,r)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\in\mathcal{Q}}\right)\|$ and $|\mathrm{E}\left(\frac{\mathrm{d}^2 s(X_1,\beta_1)g(X_2,r)}{\mathrm{d}\beta^T\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\in\mathcal{Q}}\right)|$ are finite. Similarly by Slutsky theorem, (3.2.1) converges to normal distribution with mean 0 and variance:

$$\sigma_2^2 = \mathrm{E}\left(\frac{\mathrm{d}s(X_1,\beta_1)g(X_2,r)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\in\mathcal{Q}}\right)^T \Omega_{\beta^*}\mathrm{E}\left(\frac{\mathrm{d}s(X_1,\beta_1)g(X_2,r)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\in\mathcal{Q}}\right);$$

(3.2.2) converges in probability to 0. Combining (3.1) and (3.2) together, (3) converges to normal distribution with mean 0 and variance:

$$\sigma^2 = u_{\beta^*}^T \Omega_{\beta^*} u_{\beta^*},$$

where $u_{\beta^*} = \mathrm{E}\left(\frac{\mathrm{d}s(X_1,\beta_1)g(X_2,\beta_2)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\notin\mathcal{Q}} + \frac{\mathrm{d}s(X_1,\beta_1)g(X_2,r)}{\mathrm{d}\beta}|_{\beta=\beta^*}1_{(X,\beta^*)\in\mathcal{Q}}\right)$.

(2) can be decomposed as:

$$\begin{aligned}(2) &= \sqrt{n}(\mathbb{E}_n - \mathrm{E})\{\phi(X,r,\hat{\beta}_n,\beta^*)\} \\ &= \sqrt{n}(\mathbb{E}_n - \mathrm{E})\{\phi(X,r,\hat{\beta}_n,\beta^*) - \phi(X,r,\beta^*,\beta^*)\} + \sqrt{n}(\mathbb{E}_n - \mathrm{E})\phi(X,r,\beta^*,\beta^*) \\ &= (2.1) + (2.2).\end{aligned}$$

By Taylor Expansion and mean-value theorem, there exists a vector $\tilde{\beta}_n$ such that (2.1) can be written as:

$$\begin{aligned}(2.1) &= \sqrt{n}(\mathbb{E}_n - \mathrm{E})(\hat{\beta}_n - \beta^*)^T \frac{\mathrm{d}\phi(X,r,\beta,\beta^*)}{\mathrm{d}\beta}|_{\beta=\beta^*} \\ &\quad + \sqrt{n}(\mathbb{E}_n - \mathrm{E})\frac{1}{2}(\hat{\beta}_n - \beta^*)^T \frac{\mathrm{d}^2\phi(X,r,\beta,\beta^*)}{\mathrm{d}\beta^T\mathrm{d}\beta}|_{\beta=\tilde{\beta}_n}(\hat{\beta}_n - \beta^*) \\ &= (2.1.1) + (2.1.2).\end{aligned}$$

The first term (2.1.1) can be written as:

$$(2.1.1) = (\hat{\beta}_n - \beta^*)^T \sqrt{n}(\mathbb{E}_n - \mathrm{E}) \frac{\mathrm{d}\phi(X, r, \beta, \beta^*)}{\mathrm{d}\beta}|_{\beta=\beta^*}.$$

Similar to the proof for (3), we assume that $\frac{\mathrm{d}\phi(X,r,\beta,\beta^*)}{\mathrm{d}\beta}|_{\beta=\beta^*}$ exists, and $\mathrm{Var}(\frac{\mathrm{d}\phi(X,r,\beta)}{\mathrm{d}\beta}|_{\beta=\beta^*})$ is positive semi-definite. Thus, using Central Limit Theorem (CLT), we can derive that $\sqrt{n}(\mathbb{E}_n - \mathrm{E})\frac{\mathrm{d}\phi(X,r,\beta)}{\mathrm{d}\beta}|_{\beta=\beta^*}$ converges to a normal distribution with mean $\mathbf{0}$, and a positive semi-definite covariance matrix. Hence, by slutsky theorem: (2.1.1) $\xrightarrow{p} 0$. (2.1.2) can be written as:

$$(2.1.2) = \frac{\sqrt{n}}{2}(\hat{\beta}_n - \beta^*)^T(\mathbb{E}_n - \mathrm{E})\frac{\mathrm{d}^2\phi(X, r, \beta, \beta^*)}{\mathrm{d}\beta^T \mathrm{d}\beta}|_{\beta=\tilde{\beta}_n}(\hat{\beta}_n - \beta^*),$$

We know that $\frac{\mathrm{d}^2\phi(X,r,\beta,\beta^*)}{\mathrm{d}\beta^T\mathrm{d}\beta}|_{\beta=\tilde{\beta}_n}$ is a Hessian matrix. If we use $d_{ij}(X, r, \tilde{\beta}_n)$ denote the element in $i$th row, $j$th column, then:

$$\begin{aligned}
|(\mathbb{E}_n - \mathrm{E})d_{ij}(X, r, \tilde{\beta}_n)| &\leq (\mathbb{E}_n + \mathrm{E})|d_{ij}(X, r, \tilde{\beta}_n)| \\
&\leq (\mathbb{E}_n + \mathrm{E}) \sup_{\beta \in N_{\beta^*}} |d_{ij}(X, r, \beta)| + (\mathbb{E}_n + \mathrm{E})|d_{ij}(X, r, \tilde{\beta}_n)|1_{\tilde{\beta}_n \notin N_{\beta^*}} \\
&= (\mathbb{E}_n - \mathrm{E}) \sup_{\beta \in N_{\beta^*}} |d_{ij}(X, r, \beta)| + 2\mathrm{E} \sup_{\beta \in N_{\beta^*}} |d_{ij}(X, r, \beta)| + o_p(1) \\
&< \infty,
\end{aligned}$$

where $N_{\beta^*}$ is the neighborhood of $\beta^*$. Therefore, by slutsky theorem, (2.1.2) $\xrightarrow{p} 0$.

For (2.2), by assumption and delta method, it converges to a normal distribution: $N(0, \sigma_3^2)$, where $\sigma_3^2 = \mathrm{Var}(\frac{\mathrm{d}\phi(X,r,\beta)}{\mathrm{d}\beta}|_{\beta=\beta^*})$. Hence, combining (2.1.1), (2.1.2), and (2.2), by slutsky theorem, we can derive:

$$(2) = \sqrt{n}\left(\mathbb{E}_n\{\phi(X, r, \hat{\beta}_n, \beta^*)\} - \mathrm{E}\{\phi(X, r, \hat{\beta}_n, \beta^*)\}\right) \xrightarrow{d} N(0, \sigma_3^2).$$

Now, combing (1), (2), (3). We have proved that (1) $\xrightarrow{p} 0$; (2) and (3) converge to a normal distribution. Since $\hat{\beta}_n$ is the MLE, thus by taylor expansion

$$\begin{aligned}
0 &= \bigtriangledown H(\hat{\beta}_n) \\
&= \bigtriangledown H(\beta^*) + \bigtriangledown^2 H(\tilde{\beta}_n)(\hat{\beta}_n - \beta^*),
\end{aligned}$$

where $H(\beta)$ is the log-likelihood with parameter $\beta$ for data set $\mathcal{D}$, and $\tilde{\beta}_n$ is a via point between $\hat{\beta}_n$ and $\beta^*$. Then by assumptions from MLE,

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta^*) &= -\sqrt{n}(\frac{1}{n}\bigtriangledown^2 H(\tilde{\beta}_n))^{-1}\frac{1}{n}\bigtriangledown H(\beta^*) \\ &= \sqrt{n}I(\beta^*)^{-1}\left(\mathbb{E}_n - \mathrm{E}\right)\tilde{\phi}(Y;\beta^*) + o_p(1),\end{aligned}$$

where $\tilde{\phi}(Y;\beta^*) = \log p(Y,\beta^*)$,which is the log pdf from dataset $\mathcal{D}$. Then by CLT, the joint limiting distribution of

$$\begin{aligned}&\sqrt{n}\{(\hat{\beta}_n, \mathbb{E}_n\phi(X,r,\beta^*,\beta^*)) - (\beta^*, \mathrm{E}\phi(X,r,\beta^*,\beta^*))\} \\ =\ &\sqrt{n}\{\mathbb{E}_n - \mathrm{E}\}\left(\{I(\beta^*)^{-1}\tilde{\phi}(Y;\beta^*)\}^T, \phi(X,r,\beta^*,\beta^*)\right)^T\end{aligned}$$

is a normal distribution. If we denote its variance matrix as $\Omega^*$, then the limiting distribution of (2)+(3) is still normal with mean 0 and finite variance $\tilde{u}_{\beta^*}^T\Omega^*\tilde{u}_{\beta^*}$, where $\tilde{u}_{\beta^*}$ is defined as a vector that makes $\tilde{u}_{\beta^*}^T\left(\{I(\beta^*)^{-1}\tilde{\phi}(Y;\beta^*)\}^T, \phi(X,r,\beta^*,\beta^*)\right)^T$ equals to (2)+(3). Hence, by Slutsky theorem, $\sqrt{n}\{\hat{\theta}_n(r,\hat{\beta}_n) - \theta^*(r,\beta^*)\}$ converges to a normal distribution with mean 0 and finite variance $\tilde{u}_{\beta^*}^T\Omega^*\tilde{u}_{\beta^*}$.

**Proof for $\sqrt{n}\{\hat{\theta}_n^{(b)}(r,\hat{\beta}_n^{(b)}) - \hat{\theta}_n(r,\hat{\beta}_n)\}$**

We want to prove $\sqrt{n}\{\hat{\theta}_n^{(b)}(r,\hat{\beta}_n^{(b)}) - \hat{\theta}_n(r,\hat{\beta}_n)\}$ has the same limiting distribution as $\sqrt{n}\{\hat{\theta}_n(r,\hat{\beta}_n) - \theta^*(r,\beta^*)\}$, which is asymptotically normal with mean 0 and a finite variance. This can be rewritten as:

$$\begin{aligned}&\sqrt{n}\left[\hat{\mathbb{E}}_n^{(b)}\{\psi(X,r,\hat{\beta}_n^{(b)},\lambda_n)\} - \mathbb{E}_n\{\psi(X,r,\hat{\beta}_n,\lambda_n)\}\right] \\ =\ &\sqrt{n}\left[\hat{\mathbb{E}}_n^{(b)}\{\psi(X,r,\hat{\beta}_n^{(b)},\lambda_n) - \phi(X,r,\hat{\beta}_n^{(b)},\beta^*)\} - \mathbb{E}_n\{\psi(X,r,\hat{\beta}_n,\lambda_n) - \phi(X,r,\hat{\beta}_n,\beta^*)\}\right] \\ &+\sqrt{n}\left[\hat{\mathbb{E}}_n^{(b)}\{\phi(X,r,\hat{\beta}_n^{(b)},\beta^*)\} - \mathbb{E}_n\{\phi(X,r,\hat{\beta}_n,\beta^*)\}\right] \\ =\ &\sqrt{n}\{(1)+(2)\}.\end{aligned}$$

(1) converges in probability to 0, we focus on (2).

Table A.1: The estimated coefficients of mood-stabilizer grouped effect in SAD. $\alpha_{0k}$ represents the $k$th group defined in Table 3.6.

| Variable Names | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{03}$ | $\alpha_{04}$ | $\alpha_{05}$ |
|---|---|---|---|---|---|
| SUMD0 | 0.92 | 3.69 | 0.47 | -0.37 | -1.83 |
| MEDINS | -0.24 | -1.84 | -1.56 | -0.37 | 3.57 |
| RACE | 0.23 | -0.15 | 0.19 | 0.15 | -0.07 |

Table A.2: The estimated coefficients of antidepressants grouped effect in SAD. $\eta_{0k}$ represents the $k$th group defined in Table 3.5.

| Variable Names | $\eta_{01}$ | $\eta_{02}$ | $\eta_{03}$ | $\eta_{04}$ |
|---|---|---|---|---|
| SUMD0 | 0.10 | -1.56 | 0.41 | -6.45 |
| MEDINS | -1.92 | 4.09 | 2.62 | 3.96 |
| RACE | 0.35 | -0.29 | -0.65 | 0.39 |

## A.2 Complete Details of Point estimators for Coefficients in 3.3

Table A.1, A.2 shows the estimated coefficients of grouped effect for mood-stabilizer $\alpha_{0k}$ ($k = 1, 2, 3, 4, 5$) and antidepressants $\eta_{0k}$ ($k = 1, 2, 3, 4$) respectively. Table A.3, A.4 show the estimated coefficients of mood-stabilizer dose effect with dosage level medium and high ($\alpha_{12}$, $\alpha_{13}$). Table A.5 shows the estimated coefficients of treatment effect with each mood-stabilizer group ($\delta_{t_1}$). Table A.6, A.7 show the estimated coefficients of antidepressants dose effect with dosage level medium and high ($\eta_{12}$, $\eta_{13}$). Table A.8 shows the estimated coefficients of treatment effect with each antidepressant group ($\nu_{t_2}$).

Table A.3: The estimated coefficients of mood-stabilizer dose effect with level medium in SAD. Note, only Mood 1 and Mood 2 are considered.

| Variable Names | $\alpha_{12}$ Mood 1 | $\alpha_{12}$ Mood 2 |
|---|---|---|
| SUMD0 | -0.30 | -5.35 |
| MEDINS | 0.13 | 3.38 |
| RACE | -0.008 | 0.53 |

Table A.4: The estimated coefficients of mood-stabilizer dose effect with level high in SAD. Note, only Mood 1 and Mood 2 are considered.

| Variable Names | $\alpha_{13}$ Mood 1 | $\alpha_{13}$ Mood 2 |
|---|---|---|
| SUMD0 | 0.91 | -5.29 |
| MEDINS | 1.21 | 4.14 |
| RACE | -0.22 | 0.42 |

Table A.5: The estimated coefficients of treatment effect with each mood-stabilizer group ($\delta_{t_1}$) based on Table 3.6.

| Variable Names | $\delta_{t_1}$ Mood 1 | $\delta_{t_1}$ Mood 2 | $\delta_{t_1}$ Mood 5 | $\delta_{t_1}$ Mood 5 |
|---|---|---|---|---|
| SUMD0 | 0.67 | 7.47 | -0.12 | 0 |
| MEDINS | -2.60 | -0.10 | 2.05 | 0 |
| RACE | 0.50 | 0.47 | 2.30 | 0 |

Table A.6: The estimated coefficients of antidepressants dose effect with level medium in SAD.

| Variable Names | $\alpha_{12}$ Anti 1 | $\alpha_{12}$ Anti 2 | $\alpha_{12}$ Anti 3 | $\alpha_{12}$ Anti 4 |
|---|---|---|---|---|
| SUMD0 | -2.85 | 2.37 | -1.24 | 7.64 |
| MEDINS | 6.29 | -3.86 | 2.47 | -7.20 |
| RACE | 0.05 | 0.35 | 0.41 | 0.12 |

Table A.7: The estimated coefficients of antidepressants dose effect with level high in SAD.

| Variable Names | $\alpha_{13}$ Anti 1 | $\alpha_{13}$ Anti 2 | $\alpha_{13}$ Anti 3 | $\alpha_{13}$ Anti 4 |
|---|---|---|---|---|
| SUMD0 | 1.83 | 2.74 | -0.63 | 6.49 |
| MEDINS | 1.83 | -5.36 | -2.13 | -4.04 |
| RACE | -0.39 | 0.52 | 0.86 | -0.08 |

Table A.8: The estimated coefficients of treatment effect with each antidepressant group ($\nu_{t_2}$) based on Table 3.5.

| Variable Names | $\nu_{t_2}$ Anti 1 | $\nu_{t_2}$ Anti 2 | $\nu_{t_2}$ Anti 2 | $\nu_{t_2}$ Anti 2 | $\nu_{t_2}$ Anti 2 | $\nu_{t_2}$ Anti 2 |
|---|---|---|---|---|---|---|
| SUMD0 | -0.37 | 1.51 | 0.79 | -1.49 | 0.45 | 1.04 |
| MEDINS | -0.37 | 1.69 | 1.85 | -1.49 | 2.05 | -0.57 |
| RACE | 0.15 | -0.35 | -0.10 | 0.31 | -0.21 | -0.001 |