Semiparametric Regression and Machine Learning Methods for Estimating Optimal Dynamic Treatment Regimes

by

Yebin Tao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2016

Doctoral Committee:

Associate Professor Lu Wang, Chair
Associate Professor Janis M. Miller
Professor Bhramar Mukherjee
Research Associate Professor Matthew J. Schipper

To my parents and my grandma

ACKNOWLEDGEMENTS

This dissertation is an epitome of my five wonderful years at Michigan, full of wisdom, support and encouragement from my instructors, colleagues, friends and family. Those listed below are a subset, and I apologize to anyone left out.

My sincere gratitude goes to my advisor Dr. Lu Wang. I have been fortunate enough to have Lu guide me through my studies and my life at Michigan. I have enjoyed every single meeting with her, no matter when having breakthroughs or being stuck in some research problems. This dissertation could never be finished without her instruction, inspiration and encouragement. Lu's approach to always give the right amount of guidance while leaving enough space for my own development has benefited me greatly. I will continue learning from her as I move on to the next phase of my life.

Special thanks go to my other committee members. Janis is a delight to work with. I am deeply grateful to her for helping me improve my collaboration skills with clinical researchers and fit into the academic life in the US shortly after I came here. I cannot thank Bhramar enough for believing in me, and supporting me both intellectually and financially. Her comprehensive knowledge and exceptional multi-tasking ability have been a great inspiration to me. I would not have gone this far without her guidance and support during the first half of my life at Michigan. I must also thank Matt for supporting me during my PhD studies and it is my great pleasure to work with him. I have learned so much from him given his excellent abilities as an applied statistician

working with medical researchers.

I am grateful to Dr. Wei Huang, my advisor at Peking University, for guiding me into the field of public health and biostatistics. My thanks also go to Dr. Brisa Sánchez for helping me with my research during the MS program. I am thankful to all the knowledgeable faculty members at Michigan for their excellent lectures.

Last but not least, I would like to thank my friends and family, who have made my life so enjoyable. I am especially grateful to my father Weirong Tao, my mother Xueya Huang, and my grandma Aigen Hu for their unconditional love and support, and for making me a better person.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Semiparametric Regression and Machine Learning Methods for Estimating Optimal Dynamic Treatment Regimes

by

Yebin Tao

Chair: Associate Professor Lu Wang

Dynamic treatment regimes (DTRs) are sequential decision rules that focus simultaneously on treatment individualization and adaptation over time. They determine treatment prescriptions based on each individual's specific characteristics (e.g., demographics, clinical outcomes and genetic makeup) and also adapt the prescriptions over time to evolving illness. We develop robust and flexible semiparametric regression and machine learning methods for estimating optimal DTRs.

In the first project, we consider identifying the optimal personalized timing for treatment initiation. Instead of considering multiple fixed decision stages as in most DTR literature, we deal with continuous or multiple random decision points for treatment initiation given each patient's individual disease and treatment history. For a set of predefined candidate DTRs, we propose to fit a flexible survival model with splines of time-varying covariates to estimate patient-specific probabilities of adherence to each DTR. Given the estimated probabilities, an inverse probability weighted estimator for the counterfactual mean utility (prespecified criteria) is employed to assess each DTR

and then the optimal one is identified among all candidates. We conduct simulations to demonstrate the performance of our method and further illustrate the application process with an example of insulin therapy initiation among type 2 diabetic patients.

In the second project, we propose a dynamic statistical learning method, adaptive contrast weighted learning (ACWL), which combines doubly robust semiparametric regression estimators with flexible machine learning methods. Compared to the method in Project 1, ACWL can handle multiple treatments at each stage and does not require prespecifying candidate DTRs, despite being limited to a fixed number of treatment stages. At each stage, we develop robust semiparametric regression-based contrasts with the adaptation of treatment effect ordering for each patient, and the adaptive contrasts simplify the problem of optimization with multiple treatment comparisons to a weighted classification problem that can be solved using existing machine learning techniques. The algorithm is implemented recursively using backward induction. Through simulation studies, we show that the proposed method is robust and efficient for the identification of the optimal DTR. We further illustrate our method using observational data on esophageal cancer.

In the third project, we propose a tree-based reinforcement learning (T-RL) method to directly estimate optimal DTRs in a multi-stage multi-treatment setting. At each stage, T-RL builds an unsupervised decision tree that maintains the nature of batch-mode reinforcement learning. Unlike ACWL, T-RL handles directly the problem of optimization with multiple treatment comparisons, through the purity measure constructed with augmented inverse probability weighted estimators. For the multiple stages, the algorithm is implemented recursively using backward induction. By combining robust semiparametric regression with flexible tree-based learning, we show that T-RL is robust, efficient and easy to interpret for the identification of optimal DTRs. We illustrate our method in a case study to identify dynamic substance abuse

treatment regimes for adolescents.

CHAPTER I

Introduction

As the importance of personalized medicine becomes more and more widely recognized in today's health care, a lot of research efforts are being made in the development of individualized treatment strategies, which are decision rules that dictate what treatment to provide given a patient's specific characteristics (e.g., demographics, clinical outcomes and genetic makeup). Dynamic treatment regimes (DTRs) (*Robins*, 1986, 1997, 2004; *Murphy*, 2003) mathematically generalize personalized medicine to a time-varying treatment setting, and focus simultaneously on treatment individualization and adaptation over time. Identifying optimal DTRs offers an effective vehicle for personalized management of diseases, and helps physicians tailor the treatment strategies dynamically and individually based on clinical evidence, which provides a key foundation for better chronic care (*Wagner et al.*, 2001).

In this dissertation, we consider estimating optimal DTRs, which is difficult due to the complex relationships between the alternating sequences of time-varying treatments and clinical outcomes. Standard regression methods fail without being able to adjust for time-varying confounding. The most popular methods for estimating optimal DTRs include marginal structural models (MSMs) with inverse probability weighting (IPW) (*Robins*, 2000; *Hernán et al.*, 2001; *Wang et al.*, 2012), G-estimation of structural nested mean models (*Robins*, 1986, 1989, 1997), generalized by *Murphy*

(2003) and *Robins* (2004), targeted maximum likelihood estimation (*van der Laan and Rubin*, 2006), and likelihood-based approaches (*Thall et al.*, 2007). Machine learning methods have become popular alternative approaches on estimating optimal DTRs, for example, Q-learning (*Watkins and Dayan*, 1992; *Sutton and Barto*, 1998) and A-learning (*Murphy*, 2003; *Schulte et al.*, 2014), both of which use backward induction (*Bather*, 2000) to first optimize the treatment at the last stage and then sequentially optimize the treatment at each of the earlier stages.

A DTR is essentially a multi-stage decision problem with two or more options at each stage. It has been developed upon the simplest case of single-stage decision-making with binary treatment options, the so-called individualized treatment regime (*Zhang et al.*, 2012b,a; *Zhao et al.*, 2012; *Zhou et al.*, 2015). Other studies have extended the exploration into multiple treatment stages (*Murphy*, 2003; *Zhang et al.*, 2013; *Zhao et al.*, 2015) or multiple treatment options (*Laber and Zhao*, 2015). We aim to continue this endeavor into more decision stages and/or more treatment options, using flexible and robust methods. In this dissertation, we utilize the counterfactual framework for causal inference (*Robins*, 1986) to identify the optimal DTR, which means that the optimal DTR would optimize the expectation of a counterfactual mean outcome/utility.

Our first project as presented in Chapter II takes on the task of handling continuous or multiple random decision points with binary treatment options. It is motivated by the example of type 2 diabetes patients enrolled to initiate insulin therapy. Finding the optimal personalized timing to initiate the therapy is essential to achieve the best balance of treatment effectiveness and risk. In the data example, each type 2 diabetes patient has multiple clinical visits at random time points before treatment initiation and the treatment decisions are made at each clinical visit. With the advance in mobile-health technologies (*Free et al.*, 2013) and wearable biosen-

sor systems for health monitoring (*Gatzoulis and Iakovidis*, 2007; *Pantelopoulos and Bourbakis*, 2010), it is now feasible to obtain biomarker measures, such as blood pressure and heart rate, continuously over time, so that continuous treatment decisions can be made. However, most existing methods for identifying optimal DTRs have only considered multiple fixed treatment decision stages or even a single stage and thus cannot deal with continuous or multiple random decision points. We provide a general framework based on MSMs to identify the optimal personalized time for treatment initiation given random or continuous decision points. To utilize MSMs with IPW, one has to prespecify candidate DTRs, for example, defining the structure of the DTR to depend on a small set of covariates and searching over a grid of thresholds of these covariates. In our study, we consider a set of predefined DTRs with each representing a way of timing the treatment initiation based on patients' up-to-date medical history, such as biomarker trajectories. These candidate DTRs are compared by the expectation of a counterfactual mean utility, which also needs to be predefined.

Instead of focusing on continuous or multiple random stages with binary treatment options, Projects 2 and 3 (Chapters III and IV, respectively) work on multiple treatment options in the context of multiple fixed decision stages. The treatment options can be either multinomial or ordinal. A motivation example is the esophageal cancer data where each patient went through two stages of chemotherapy and radiation therapy. The methods in the second and third projects are fundamentally different from the one in the first project. In Projects 2 and 3, we combine robust semiparametric regression with flexible machine learning methods for multi-stage multi-treatment decision-making. The algorithms are implemented recursively from the last stage using backward induction (*Bather*, 2000). Due to the use of machine learning, there is no need to predefine candidate DTRs.

The problem of multi-stage decision-making has strong resemblance to reinforcement learning (RL), which is a branch of machine learning (*Chakraborty and Moodie*, 2013). Unlike supervised learning (SL) (e.g., regression and classification), the desired output value or the optimal decision, known as *label*, is not observed in RL, and the learning agent has to keep interacting with the environment to learn the best policy for decision-making. In a DTR problem, the optimal treatment for each patient at each stage is also not observed and can only be inferred based on observed treatments and outcomes from all subjects. In Chapter III, we develop a dynamic statistical learning method, adaptive contrast weighted learning (ACWL), to directly estimate the optimal DTR through a sequence of weighted classifications. Basically, ACWL transforms RL into SL by obtaining *label* from a working semiparametric regression model which estimates the treatment effect ordering for each patient at each stage. ACWL can deal with more than two treatments at each stage due to the use of contrasts with the adaptation of treatment effect ordering. The proposed adaptive contrasts stand for the minimum or maximum expected loss in the outcome given any sub-optimal treatment for each patient, and simplify the problem of optimization with multiple treatment comparisons to a weighted classification problem at each stage.

We show that ACWL is robust and efficient for the identification of the optimal DTR and can be easily implemented using existing regression and classification methods. However, it requires the extra step of transforming RL into SL, which may induce additional uncertainty through the identification of *label*. It also may not the most efficient method by avoiding multiple treatment comparisons. Therefore in Chapter IV, we propose a tree-based reinforcement learning (T-RL) method to directly handle the problem of optimization with multiple treatment comparisons while maintaining the RL nature of the DTR problem. At each stage, T-RL builds an unsupervised decision trees using a purity measure constructed with augmented inverse probability weighted estimators for all treatment options. T-RL enjoys the advantages of typical

tree-based methods as being straightforward to understand and interpret, and capable of handling various types of data without distributional assumptions. It is also robust and efficient by combining robust semiparametric regression with flexible tree-based learning. However, for non-tree-type underlying DTRs, ACWL may have better performance with the ability of incorporating non-tree-based classification methods.

CHAPTER II

Optimizing the Personalized Timing for Treatment Initiation with Continuous or Multiple Random Decision Points

2.1 Introduction

Many chronic diseases such as cancer and diabetes are of long duration and progressive nature. Therefore, long-term health monitoring and dynamic treatment processes with sequential intensification are necessary for patients with such diseases. An important but challenging problem is to find the optimal personalized timing to initiate a treatment for the next stage of disease condition. For example, patients diagnosed with type 2 diabetes usually start with oral anti-diabetic medications, such as metformin (Glucophage), and their glycated hemoglobin (HbA1c) levels are constantly checked during their regular clinical visits. According to the American Diabetes Association, a reasonable HbA1c goal for many non-pregnant adults is $< 7\%$ (*American Diabetes Association*, 2014). As the disease progresses, most patients eventually require and benefit from insulin therapy (*Turner et al.*, 1999). Delayed insulin therapy has been found related to reduced life expectancy and increased risk of microvascular and macrovascular complications (*Goodall et al.*, 2009). However, intensive

glucose control (e.g., targeting HbA1c $< 6\%$) is rarely effective in achieving tight glycemic control (*Hayward et al.*, 1997) and is associated with adverse effects such as hypoglycemia and weight gain (*Gerstein et al.*, 2008; *Patel et al.*, 2008). Therefore, finding the optimal timing to initiate insulin therapy is essential to achieve the best balance of treatment effectiveness and risk. In most observational data, patients with chronic diseases have their own schedules for examinations of clinical biomarkers (e.g., HbA1c) and their physicians make treatment decisions each time the biomarkers are measured. The frequency of clinical visits can be considered as a random variable which likely depends on a patient's disease progression, physical status and the physician's personal judgment, and the clinical visits are in fact multiple random decision points for treatment decisions. Moreover, with the advance in mobile-health technologies (*Free et al.*, 2013) and wearable biosensor systems for health monitoring (*Gatzoulis and Iakovidis*, 2007; *Pantelopoulos and Bourbakis*, 2010), it is now feasible to obtain some biomarker measures (e.g., blood pressure and heart rate) continuously over time, so that continuous treatment decisions, which are more timely and precise, can be made.

Motivated by these examples, we consider a situation where key biomarkers of disease severity are monitored at continuous or multiple random time points during a follow-up period and each time the biomarkers are measured, a decision on treatment initiation is made based on the patient's up-to-date biomarker and treatment history. Hence, personalized decisions on treatment initiation are made dynamically over time. Regardless of continuous or multiple random visits, the time for each clinical visit can be considered as a continuous random variable from a population perspective. The difference is that with multiple random visits, the biomarker trajectories are not fully observed and we may have to extrapolate using parametric or nonparametric methods. Our goal is to find the optimal timing for treatment initiation given a patient's biomarker and treatment history. This can be framed as a specific type of dynamic

treatment regimes (DTRs) (*Robins*, 1986, 1997, 2004; *Murphy*, 2003). Instead of having multiple fixed stages of treatment decisions as in most DTR literature, we consider continuous or multiple random decision points for treatment initiation according to a patient's up-to-date medical history. Identifying such optimal DTRs offers an effective vehicle for personalized management of diseases, and helps physicians tailor the treatment strategies dynamically and individually based on clinical evidence, which provides a key foundation for better chronic care (*Wagner et al.*, 2001).

Many recent studies have explored the designs of sequential multiple assignment randomized trials (SMARTs) (*Murphy*, 2005) that aim at evaluating DTRs, as well as analytic tools to estimate the effects of DTRs using longitudinal observational or experimental data, as reviewed in *Wang et al.* (2012). Although SMARTs are desirable for causal inference, the more common source of data for constructing DTRs is observational studies. A lot of statistical research has focused on dealing with observational data (*Murphy*, 2003; *Robins*, 2004; *Henderson et al.*, 2010), where careful thoughts and assumptions are required in order to make valid inference (*Robins and Hernán*, 2009). Diverse statistical methods have been developed including G-estimation of structural nested mean models (SNMMs) (*Robins*, 1986, 1989, 1997), generalized by *Murphy* (2003) and *Robins* (2004), marginal structural models (MSMs) with inverse probability weighting (IPW) (*Robins*, 2000; *Hernán et al.*, 2001), targeted maximum likelihood estimators (*van der Laan and Rubin*, 2006), Q- and A-learning (*Watkins and Dayan*, 1992; *Robins*, 2004; *Murphy*, 2005; *Huang and Ning*, 2012; *Moodie et al.*, 2012; *Chakraborty and Moodie*, 2013; *Schulte et al.*, 2014), outcome weighted learning and classification-based methods (*Zhang et al.*, 2012a; *Zhao et al.*, 2012, 2015). However, susceptibility to model misspecification remains as a major limitation of many methods in this field and the computational burden would increase as the number of decision stages increases. Moreover, for methods that rely on backward induction (*Bather*, 2000), one needs to line up the treatment stages to a finite number. Thus

most existing studies have only considered multiple fixed treatment decision stages or even a single stage for decision making. Very few attempts exist to handle continuous decision points. For example, *Lok* (2008) provides a conceptual framework and mathematical formalization of SNMs in continuous time. However, real data application of their SNMMs based method is still limited to a finite number of stages (*Lok and DeGruttola*, 2012). *Johnson and Tsiatis* (2005) consider the duration-response relationship with treatment duration being a continuous random variable, and they apply MSMs to estimate the optimal regimes that are determined solely by the treatment duration.

We aim to provide a general framework based on MSMs to identify the optimal personalized time for treatment initiation given random or continuous decision points, which also applies to dynamic decisions on binary and monotonic treatment switch. We compare a set of predefined DTRs with each representing a way of timing the treatment initiation based on patients' up-to-date medical history, such as biomarker trajectories. The cause of continuous or random decision points under this framework is that the biomarkers are measured continuously over time or at multiple randomly time points and decisions are made each time the biomarkers are measured. In the type 2 diabetes example, if we consider the timing by the HbA1c level, a possible DTR could be that we initiate insulin therapy only when a patient's HbA1c level is between $6 \sim 6.5\%$. In this case, decisions for treatment initiation are made continuously based on the HbA1c trajectories. Furthermore, under each DTR, instead of modeling directly the time from enrollment to treatment initiation, we focus on the duration when a patient adheres to a given regime. The adherence duration, as a function of the biomarker trajectories and the definition of the DTR, is a continuous random variable specific to a given regime. Most MSMs based methods are limited to a finite number of aligned stages so that one can apply pooled logistic regression to estimate probabilities of adherence to a given DTR at all stages (*Robins et al.*, 2000; *Hernán*

et al., 2001). It is nontrivial to extend this problem to accommodate continuous or multiple random decision points, considering that as the number of stages goes to infinity, the probabilities estimated by pooled logistic regression may go to zero and no longer work in IPW.

The remainder of this paper is organized as follows. We define the DTRs of interest within the framework of causal inference in Section 2.2 and establish our estimation procedure in Section 2.3. We propose to build a survival model with splines of time-varying covariates to calculate the probability of adherence to a given DTR for all patients, given their own covariate history. This model allows much flexibility on how the the risk of failure to follow a specific regime depends on time-varying biomarkers. Then we use the estimated probability to construct IPW estimators for the counterfactual mean of the utility of interest (e.g., a prespecified measure balancing treatment efficacy and toxicity). The simulation studies (Section 2.4) bring out the salient features of our proposed method. We further illustrate the application process in Section 2.5, using the example of insulin therapy initiation among type 2 diabetic patients, where we consider a class of DTRs that the patients initiate insulin therapy the first time their HbA1c levels reach a certain threshold. Finally, we conclude with some discussions and suggestions for future studies in Section 2.6.

2.2 Notation and Dynamic Treatment Regimes

Suppose that N patients, a random sample from a large target population, are followed up for a maximum duration of T^* since enrollment. For patient i at time t, where $i = 1, ..., N$ and $0 \leq t \leq T^*$, let $\mathbf{X}_i(t)$ denote the time-varying biomarkers and we allow $\mathbf{X}_i(0)$ to further include all other baseline covariates. Let $A_i(t)$ denote the treatment prescription that takes the value 1 for starting treatment and 0 oth-

erwise. Our interest is in the optimal timing to initiate treatment and we assume that all patients have the same fixed post-initiation treatment plan during the study period. Let S_i denote patient i's treatment initiation time and S_i is considered as administratively censored if no treatment is given during the follow-up period, i.e., $S_i > T^*$. Therefore, patient i's observed follow-up duration for treatment initiation is $min(S_i, T^*)$, which we denote as T_i. Notably, T_i is also the patient's duration of continuous random decisions on treatment initiation. In some cases, a patient could terminate study participation before T^* without being treated due to uncontrollable factors such as death or simply loss to follow-up. These types of censoring can be incorporated in our method if they are non-informative conditional on the observed data. To simplify the problem, we only consider administrative censoring herein after.

For brevity, we suppress the patient index i in the following text when no confusion exists. Following the convention in the literature, we use overbars to denote the history of variables up to the indexed time, underbars to denote the future of variables from the indexed time, capital letters for random variables or vectors, and small letters for observations of the corresponding random variables. For example, the treatment history up to time t is denoted as $\overline{A}(t) = \{A(s) : 0 \leq s \leq t\}$ and a possible observed treatment history is denoted as $\overline{a}(t)$ with value in the range of $\overline{A}(t)$, where $a(t) \in \mathcal{A}(t)$. We denote the observational data up to time t as $\overline{\mathbf{O}}(t) = \{\overline{A}(t^-), \overline{\mathbf{X}}(t)\}$, where $\overline{A}(t^-)$ is the treatment history up to, but not including, time t and $\overline{\mathbf{X}}(t)$ denotes the covariate history up to time t including baseline covariates $\mathbf{X}(0)$.

Since the treatment decisions are made continuously and dynamically based on patients' own medical history, $A(t)$ depends on $\overline{A}(t^-)$ and $\overline{\mathbf{X}}(t)$. Recursively, the time-dependent biomarkers (e.g., HbA1c and morbidity) are also affected by previous treatment decisions and past covariate history, i.e., $\mathbf{X}(t)$ also depends on $\overline{A}(t^-)$ and $\overline{\mathbf{X}}(t^-)$. A feasible DTR $g = \left\{g\{t; \overline{A}(t^-), \overline{\mathbf{X}}(t)\} : 0 \leq t \leq T^*\right\}$ is a sequential rule for deter-

mining the next treatment prescription $A(t)$ at time t. For convention, when $t = 0$, we let $\overline{A}(0^-) = \varnothing$ and $\overline{\mathbf{X}}(0) = \mathbf{X}(0)$. We denote the collection of all DTRs g of interest as $\mathcal{G}$ and $g\{t; \overline{A}(t^-), \overline{\mathbf{X}}(t)\}$ is a map

$$[\overline{A}(t^-), \overline{\mathbf{X}}(t)] \to g\{t; \overline{A}(t^-), \overline{\mathbf{X}}(t)\} \in \mathcal{A}(t) \text{ for any } t \in [0, T^*],$$

which may depend on a patient's part or all of the recorded treatment history before time t and biomarker information up to time t.

In our study, we consider a set of predefined g's. The optimal DTR is the one that optimizes the expected utility function if all patients in the population follow this rule, which may be contrary to fact. To assess the causal effect of a specific DTR, we consider the counterfactual framework for causal inference (*Robins*, 1986). Let $\overline{\mathbf{x}}^g(t)$ denote the counterfactual $\mathbf{X}$-history up to time t (e.g., biomarkers or clinical outcomes) that would be observed in the world where the patient had followed regime g. Similarly, let $\overline{a}^g(t)$ denote the counterfactual A-history (e.g., treatment, action or intervention) up to time t that would be observed in the world in which the patient had followed regime g. To assess the treatment effects of g, we define a utility function U at the end of the study that depends on both $\overline{\mathbf{X}}(T^*)$ and $\overline{A}(T^*)$. Without loss of generality, we assume that smaller values of U is preferred. Let U^g denote the counterfactual utility function if all subjects have followed regime g. The optimal DTR is the one that minimizes the expected counterfactual utility function, i.e.,

$$g^{opt} = \text{argmin}_{g \in \mathcal{G}} E(U^g). \tag{2.1}$$

Note that the expected utility depends on regime g, while the treatment decision at time t according to regime g recursively depends on the patient's up-to-date biomarker history $\overline{\mathbf{X}}(t)$. Thus, minimizing $E(U^g)$ provides the optimal personalized treatment

decisions. Obviously, the utility function is the key to assess and compare various regimes of interest, and one can define the utility function based on the goal of the study. Our method can be applied to any study with a target utility function well defined for every subject, but cannot deal with cases where the utility function may be censored, for example, a time-to-event utility function. In that case, an alternative is to redefine the utility function, such as restricted mean survival time, so as to remove the censoring. Future research on time-to-event utility functions may be of great interest, especially considering that the ultimate goal in many medical studies is to prolong patients' survival time.

2.3 Estimation of the Optimal Dynamic Treatment Regime

2.3.1 Identifiability of the Counterfactual Mean Utility

Note that in the ideal case, to solve the optimization problem (2.1), we need the counterfactual data under all regimes $g \in \mathcal{G}$, which is impractical since each patient can only experience one treatment history. Therefore, instead of using data collected from the counterfactual world as if everyone had followed g, we need to estimate $E(U^g)$ using the observed data $\overline{\mathbf{O}}(T^*)$. In order to do that, we make the following three assumptions suggested by previous studies (*Murphy et al.*, 2001; *Robins and Hernán*, 2009; *Orellana et al.*, 2010a).

- Consistency assumption: For any regime g, if a given patient has $A(t) = g\{t; \overline{A}(t^-), \overline{\mathbf{X}}(t)\}$ for any $t \in [0, \tilde{t}]$, then $\overline{\mathbf{X}}_g(\tilde{t}) = \overline{\mathbf{X}}(\tilde{t})$ for any $\tilde{t} \in [0, T^*]$; and if $A(t) = g\{t; \overline{A}(t^-), \overline{\mathbf{X}}(t)\}$ for any $t \in [0, T^*]$, then $U^g = U$ for that patient. That is to say, if the actual treatment history observed for a patient is compatible with DTR g, his or her observed biomarker and utility function are the same

as the counterfactual ones under g.

- No unmeasured confounder assumption (NUCA): NUCA implies that at any $t \in [0, T^*]$,

$$A(t) \perp \{U^g, \underline{\mathbf{X}}(t)\} | \{\overline{A}(t^-), \overline{\mathbf{X}}(t)\}$$

for regime g, where $\underline{\mathbf{X}}(t)$ is the future of variables $\mathbf{X}$ starting from time t. In other words, the treatment decision at time t is independent of future observations and the counterfactual outcomes, conditional on $\{\overline{A}(t^-), \overline{\mathbf{X}}(t)\}$ that are recorded until just prior to assigning $A(t)$.

- Positivity assumption: We assume that at any time $t \in [0, T^*]$,

$$P^g \left(Pr \left[A(t) = g\{t; \overline{A}(t^-), \overline{\mathbf{X}}(t)\} | \overline{A}(t^-), \overline{\mathbf{X}}(t) \right] > 0 \right) = 1,$$

where P^g is the law of $\{\overline{A}(t^-), \overline{\mathbf{X}}(t)\}$ in the counterfactual world where regime g were enforced for the entire population. This assumption basically guarantees that if in the counterfactual world where everyone followed regime g, there were patients with history $\overline{\mathbf{x}}(t)$ and $\overline{a}(t^-)$, then in the observational data among the subjects with the same covariates history $\overline{\mathbf{x}}(t)$ who actually followed regime g up to time t^-, there is a subset who also follow regime g at time t.

With these assumptions, we are able to make inference about $E(U^g)$ using only the observed data (*Robins and Hernán*, 2009; *Orellana et al.*, 2010a,b). Specifically, for a DTR g at any $t \in [0, T^*]$, we define $C^g(t) = I[A(t) = g\{t; \overline{A}(t^-), \overline{\mathbf{X}}(t)\}]$ as the indicator of adherence that takes the value 1 if a patient adheres to g at time t and 0 otherwise. Therefore, if $\overline{C}^g(t) = \overline{1}(t)$, where $\overline{1}(t)$ is a function with constant value of 1 up to t, it means that this patient follows g up to t. As defined before, a patient's decision process for treatment initiation is random till $T = min(S, T^*)$, and thus we assess the patient's adherence history up to T, i.e., $\overline{C}^g(T)$. Let F^g

denote the time to failure of adherence to DTR g and we have $F^g = min_t\{t \in [0,T] : C^g(t) = 0\}$ for a patient whose failure of adherence is observed within T. On the other hand, for a patient who follows g all the way to T, whether treated (i.e., $T = S$) or administratively censored (i.e., $T = T^*$), F^g is not observed and the only information is $F^g > T$. In other words, the patient's time to failure of adherence to g is censored at T.

Using our notation, only patients with $\overline{C}^g(T) = \overline{1}(T)$ actually follow g for treatment initiation during the study and have U^g observed. The observed utility U of other patients who have $C^g(t) = 0$ for some $t \in [0,T]$ are actually affected by other treatment assignments on and after F^g, and thus cannot be used to estimate $E(U^g)$. If we denote the counterfactual law under DTR g by P^g and the observational law by P, then under Assumptions (1) - (3) we have

$$\frac{dP^g}{dP}\left\{\overline{A}(T), \overline{\mathbf{X}}(T)\right\} = \frac{I\left[\overline{C}^g(T) = \overline{1}(T)\right]}{Pr\left[\overline{C}^g(T) = \overline{1}(T)|\overline{\mathbf{X}}(T)\right]},$$

where $I(\cdot)$ is the indicator function that takes the value 1 if $\cdot$ is true and 0 otherwise.

In our study, the patients compatible with g have their time to failure of adherence to g censored at T. Therefore,

$$Pr\left[\overline{C}^g(T) = \overline{1}(T)|\overline{\mathbf{X}}(T)\right] = Pr\left[F^g > T|\overline{\mathbf{X}}(T)\right].$$

If we define

$$\omega^g = \frac{I\left[\overline{C}(T) = \overline{1}(T)\right]}{Pr\left[F^g > T|\overline{\mathbf{X}}(T)\right]},$$

then under Assumptions (1) - (3),

$$E(U^g) = E\left[U\frac{dP^g}{dP}\left\{\overline{A}(T), \overline{\mathbf{X}}(T)\right\}\right] = E(U\omega^g). \tag{2.2}$$

Therefore, the bias induced from using the subset of patients who actually follow g in the observational data can be corrected for by weighting each patient with the corresponding inverse probability of adherence. We apply the IPW method (*Murphy et al.*, 2001; *Wang et al.*, 2012) and estimate $E(U^g)$ as the weighted mean of the observed U among patients adherent to DTR g throughout the study period.

2.3.2 Estimation of the Weights and Counterfactual Mean Utility

We denote $\pi^g = Pr[F^g > T|\overline{\mathbf{X}}(T)]$, the adherence probability. Then the weight ω^g for an adherent patient is $1/\pi^g$, while that of a non-adherent patient is 0. According to (2.2), the estimation of $E(U^g)$ depends on ω^g and equivalently, π^g. Therefore, a robust estimate of π^g is essential to guarantee the validity of $\widehat{E}(U^g)$. We propose a flexible time to failure of adherence model with time-varying biomarkers, which allows both linear terms and flexible spline functions of covariates. Specifically, from the biomarker set $\mathbf{X}$, we select biomarkers $X_l^L(l = 1, ..., m)$ as linear terms (e.g., categorical variables) and $X_j^S(j = 1, ..., k)$ to be fit in spline terms (e.g., continuous variables with unknown effects on adherence), which can be determined by scientific knowledge related to the study and regime definition. Then the flexible Cox model for the hazard of failing to adhere to DTR g at time t is

$$\lambda^g\{t|\overline{\mathbf{X}}(t) = \overline{\mathbf{x}}(t)\} = \lambda_0^g(t)exp\left[\sum_{l=1}^{m}\beta_l^g x_l^L(t) + \sum_{j=1}^{k} f_j^g\left\{x_j^S(t)\right\}\right] \tag{2.3}$$

where $\lambda_0^g(t)$ is an unspecified baseline hazard function, β_l^g is an unknown parameter for X_l^L, and $f_j^g(\cdot)$ is a spline function for X_j^S with details given below. Note that we model the hazard at t with only the current biomarker observations, which can be easily extended to include summary variables indicating certain aspects of the covariate history up to t (e.g., percent of time with HbA1c over 8% from enrollment

to t). For convenience, we write all covariates in the time-varying form and simply let baseline covariates to be constant over time. The parameterization used for the splines is

$$f_j^g(x) = \sum_{h=1}^{M+3} \theta_{jh}^g B_{jh}(x)$$

where B_{jh}'s are standard cubic B-spline basis functions (*De Boor*, 1978), θ_{jh}^g is the parameter corresponding to B_{jh} and M is the number of knots. The constant term is absorbed in the unknown baseline hazard $\lambda_0^g(t)$. The function $f_j^g(\cdot)$ is only required to be smooth enough to have continuous second derivatives, and is not restricted to any specific parametric form. Therefore, it allows enough flexibility in (2.3) to get a robust estimate of π^g, the adherence probability. Then we can obtain $\widehat{\omega}^g$, and estimate $E(U^g)$ consistently with the IPW estimator (*Wang et al.*, 2012)

$$\widehat{E}(U^g) = \frac{\sum_{i=1}^n \widehat{\omega}_i^g U_i}{\sum_{i=1}^n \widehat{\omega}_i^g}. \tag{2.4}$$

According to *Robins* (2000), the IPW estimator (2.4) will be consistent if the models for estimating π^g, the denominator of ω^g, are correctly specified, and furthermore, $\widehat{E}(U^g)$ will be $\sqrt{n}$ consistent if $\widehat{\pi}^g$ converges at a rate of $n^{1/4}$ or faster. It implies that our IPW estimator can perform well as long as the adherence probability estimated from model (2.3) is not exceedingly variable. To ensure this, we follow *Gray* (1992) to use fixed knot splines with a modest number of knots and use penalized partial likelihood to estimate the parameters of model (2.3). To maximize the penalized log-partial likelihood, the smoothing parameters for the penalty terms are solved for by first specifying degrees of freedom for the spline smoothers (*Buja et al.*, 1989; *Gray*, 1992). We choose the optimal degrees of freedom according to the corrected Akaike information criteria (AIC) of Hurvich *et al.* (*Hurvich et al.*, 1998). For the baseline hazard, we use the Breslow estimator (*Breslow*, 1972, 1974), which converges at a

rate of $n^{1/2}$ (*Tsiatis*, 1981).

2.4 Simulations

We conduct simulation studies to evaluate the performance of the proposed method given access to the simulated counterfactual data. For each patient under each defined regime, we simulate the counterfactual biomarkers and utility functions, and calculate the real causal outcome using these data. The proposed method is applied to each simulation scenario and then compared to the truth and several competing methods. Our estimation is based on 500 replicates each with sample size $N = 500$.

2.4.1 Simulation Settings

For simplicity, we simulate one biomarker X that is linearly increasing before treatment initiation over the study period $[0, T^* = 120]$. For patient i at visiting time t (before treatment initiation), we generate the biomarker observation from $X_i(t) = \beta_{0i} + \beta_{1i}t + \epsilon_{ti}$, where patient-specific intercept β_{0i} and slope β_{1i} are independently drawn from $N(2.5, 0.5^2)$ and $N(0.07, 0.02^2)$, respectively, and measurement error $\epsilon_{ti} \sim N(0, 0.1^2)$. We also simulate a binary baseline covariate Z_i with success rate of 0.5. We define the DTR g to be initiating treatment when X falls into an interval $[\tau, \tau + 0.5)$ and basically, a patient fails to follow the regime g if he or she starts treatment with $X < \tau$ (too early) or $X \geq \tau$ (too late). We consider ten regimes $g_1, g_2, \ldots, g_{10}$ with $\tau_1 = 5.0, \tau_2 = 5.5, \ldots, \tau_{10} = 9.5$, respectively. Given $X_i(t)$, we calculate the counterfactual treatment initiation time $S_i^{g_j}$ $(j = 1, 2, \ldots, 10)$ had the subject followed g_j. We randomly assigned one of the ten regimes to each subject as the true underlying DTR, which is not observed. The observed adherence depends on the biomarker trajectory up to the time of treatment initiation. The observed

decision period for patient i with underlying DTR g is $T_i^g = min\{S_i^g, T^*\}$. Apparently, whether or not a patient is administratively censoring is jointly determined by T^*, X and g. An administratively censored patient may be compatible with multiple regimes. For example, a patient with X increasing from 4.0 at baseline to 6.2 at T^* without treatment is compatible with regimes that have $\tau > 6.2$, i.e., $g_4, \ldots, g_{10}$.

In the ideal case, we need all patients' biomarkers fully observed throughout the study to apply the survival model (2.3). However, in practice, we oftentimes only observe X at a limited number of clinical visits. Moreover, patients with worse conditions tend to have more clinical visits. To investigate how this can affect our selection of the optimal DTR, we create two biomarker scenarios:

- In Scenario 1, we have access to biomarker observations at each event time (i.e., the time one or more patients fail to follow g) for patients at risk (i.e., adherent to g and untreated before).

- In Scenario 2, patient i has X observed at enrollment and T_i (i.e., time of treatment initiation or end of study), and the number of visits in-between follows $binomial(\lceil 0.1T_i \rceil, \rho_i)$, where $\lceil \cdot \rceil$ means taking the smallest integer not less than $\cdot$ and $\rho_i = 1/[1 + exp(1 - 0.3\beta_{0i} - 10\beta_{1i} + 0.2Z_i)]$. The time of each visit between enrollment and T_i is uniformly sampled from $(0, T_i)$.

Scenario 1 provides the whole biomarker history needed for the estimation of the adherence model (2.3) (e.g., the case of mobile medicine). Scenario 2 is to mimic a more common situation with longitudinal biomarker measurements at random clinical visits. We let ρ depend on X and Z in a way that subjects with larger X and smaller Z are expected to have more clinical visits. We extrapolate the values of X between visits for model (2.3). For example, one can fit a linear mixed model for the trajectories using polynomial terms for t and make patient-specific predictions for

X. However, in the case of informative observation times, further adjustments may be necessary to correct for bias (*Sun et al.*, 2005, 2007). Another simpler method is to conduct patient-specific extrapolation without borrowing information from other subjects. In our simulations, we use linear extrapolation to estimate the biomarker value at a given event time for each patient still adherent to g.

Our utility function applies to both treated and administratively censored patients. Specifically, for subject i following g with threshold τ, we have

$$U_i^g = \begin{cases} 70\beta_{1i} + e_i & \text{if administratively censored} \\ 0.5\tau + 70\beta_{1i} \cdot \phi(\tau) + e_i & \text{if treated} \end{cases}$$

where $e_i \sim N(0, 0.1^2)$. Note that the utility function depends on both the slope of the biomarker trajectory and the DTR. We consider two utility functions with different $\phi(\tau)$: $U1$ with $\phi(\tau) = 0.6 + 0.3cos(0.8\tau - 2)$, which is minimized at $\tau = 6.0$, and $U2$ with $\phi(\tau) = 0.6 + 0.3cos(0.8\tau - 3.6)$, which is minimized at $\tau = 8.0$, as shown in Figure 2.1. These two functions correspond to two situations where treatments are more effective at earlier and later stages of disease, respectively. Furthermore, in our simulation setting, all DTRs have a similar number of treated compatible patients (~ 50). However, regimes with smaller thresholds have a much lower percentage of administrative censoring among their compatible patients (range: $\sim 5\%$ for τ_1 to $\sim 60\%$ for τ_{10}), and thus we are also able to investigate how various percentages of censoring would affect the detection of the optimal regime. We use $U1$ and $U2$ in both biomarker scenarios.

For comparison, we consider the unweighted method, simply averaging the observed utility of all patients compatible with the regime of interest, as well as the weighted method with weights from two types of pooled logistic regression (*Robins et al.*, 2000; *Hernán et al.*, 2001). The first type uses polynomial terms up to a degree of 3 for

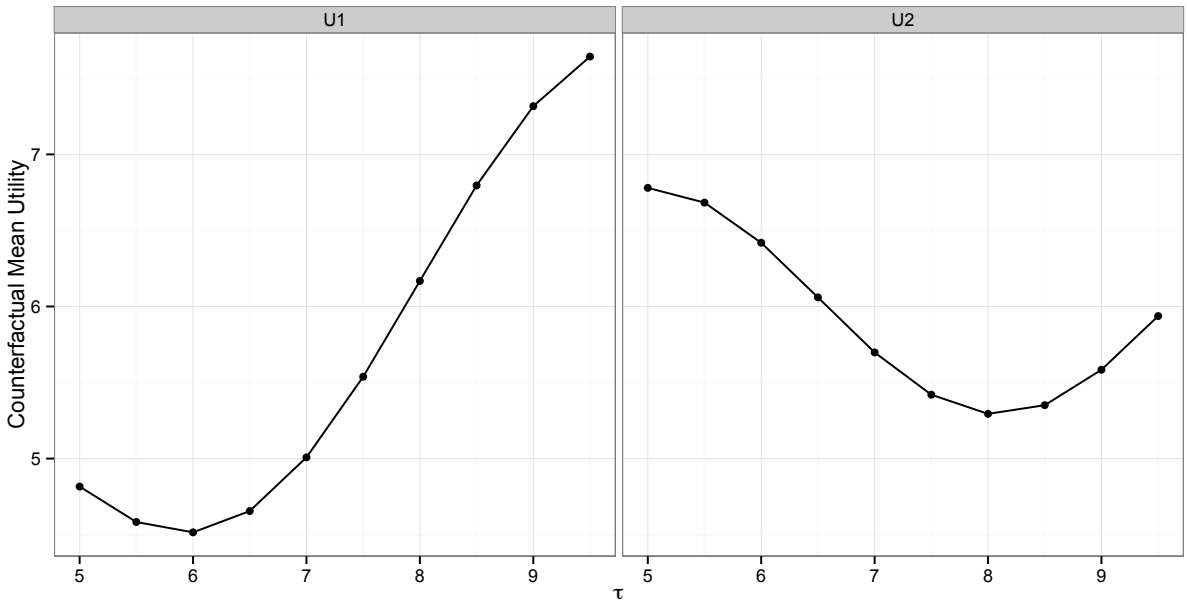

Figure 2.1: Counterfactual mean utility for the 10 regimes in simulations with various thresholds τ given utility function U1 (left) or U2 (right). Each point in the plots is calculated from 1000 Monte Carlo samples.

X, denoted as PPL, while the second one applies generalized additive model (GAM) with smoothing splines for X, denoted as NPL. Since time to failure of adherence is continuous, we discretize the period into five intervals by quintiles of event time and apply logistic regression within each interval. We use the biomarker observations in each interval as the one at event time for a patient who fails in that interval, or the one linearly extrapolated at the center of that interval for a patients who is still adherent.

2.4.2 Simulation Results

Table 2.1 summarizes the performance of all methods in evaluating the ten regimes in biomarker Scenario 1, in term of estimated counterfactual mean utility function $\widehat{E}(U^g)$ and the percentage of replicates selecting g as the optimal DTR (opt%). Generally, the weighted methods are far superior to the unweighted one. With $U1$ where the

optimal DTR has $\tau = 6$, all weighted methods have very accurate selection of the optimal regime ($> 95\%$) and our proposed method is only slightly better. The two pooled logistic methods are almost identical in selection of g^{opt} and NPL has slightly smaller bias and variance in $\widehat{E}(U^g)$. Comparing the results with the utility function $U1$ to those with $U2$, all methods have worse performance. PPL has the largest drop in opt% (from 95.2% to 67.6%) and a wide range of mis-selected regimes. NPL, as a more flexible method than PPL, still has an acceptable performance with 77% of the time selecting the correct optimal DTR and only three different regimes selected. Our proposed method maintains very satisfactory performance with opt% of 94.4%, and much smaller bias and variance in $\widehat{E}(U^g)$. From Scenario 1, we can see that our proposed method can well handle the increased bias due to administrative censoring.

In Table 2.2, we present the results for biomarker Scenario 2, where we extrapolate the biomarker observations from longitudinal visits. Compared to Scenario 1, all weighted methods have worse performance with the biomarker trajectories only partially observed. Given more uncertainty in the biomarker extrapolation, the bias and variance in $\widehat{E}(U^g)$ increases and thus more mis-selection has occurred. Note that results from the full adherent (counterfactual) and the unweighted methods are the same as in Table 2.1 since we use the same simulated samples. In terms of opt%, PPL drops over 20% with $U1$ from Scenario 1 to Scenario 2, and has only 60.2% correctness with $U2$ in Scenario 2. NPL is still slightly better than PPL. Our proposed method works much better than the other methods, with about 20% higher chance to select g^{opt}, smaller bias and variance, and fewer mis-selected regimes. From Scenario 1 to Scenario 2, it only drops less than 10% in opt% and has over 80% correctness with $U2$ in Scenario 2, which is the most noisy situation. Through the two biomarker scenarios each combined with two different utility functions, we can see that the performance of the proposed method is robust.

Table 2.1: Estimated counterfactual mean utility $\widehat{E}(U^g)$ (empirical SD in parentheses) and percentage of replicates selecting g as the optimal DTR (opt%) for regimes in biomarker Scenario 1 of Chapter II.

	Regime	Full Adherent		Unweighted		PPL[1]		NPL[2]		Proposed Method	
		$\widehat{E}(U^g)$	opt%	$\widehat{E}(U^g)$	opt%	$\widehat{E}(U^g)$	opt%	$\widehat{E}(U^g)$	opt%	$\widehat{E}(U^g)$	opt%
$U1$	$\tau_1 = 5.0$	4.816 (0.032)	0	4.524 (0.154)	0	4.921 (0.221)	0	4.845 (0.133)	0	4.793 (0.143)	0
	$\tau_2 = 5.5$	4.583 (0.028)	0	4.206 (0.162)	0.2	4.492 (0.099)	0.4	4.501 (0.081)	0.2	4.555 (0.071)	0
	$\tau_3 = 6.0$	4.515 (0.027)	100	4.015 (0.160)	47.4	4.373 (0.092)	95.2	4.394 (0.073)	95.4	4.492 (0.062)	97.3
	$\tau_4 = 6.5$	4.655 (0.030)	0	4.014 (0.161)	44.4	4.518 (0.058)	4.4	4.534 (0.069)	4.8	4.594 (0.057)	2.7
	$\tau_5 = 7.0$	5.008 (0.037)	0	4.160 (0.175)	7.6	4.845 (0.076)	0	4.863 (0.072)	0	4.923 (0.068)	0
	$\tau_6 = 7.5$	5.538 (0.047)	0	4.426 (0.211)	0.4	5.334 (0.084)	0	5.352 (0.084)	0	5.417 (0.073)	0
	$\tau_7 = 8.0$	6.168 (0.066)	0	4.735 (0.266)	0	5.912 (0.119)	0	5.919 (0.110)	0	6.036 (0.089)	0
	$\tau_8 = 8.5$	6.796 (0.088)	0	5.027 (0.274)	0	6.538 (0.268)	0	6.509 (0.139)	0	6.720 (0.125)	0
	$\tau_9 = 9.0$	7.317 (0.110)	0	5.307 (0.360)	0	7.031 (0.203)	0	7.065 (0.169)	0	7.218 (0.173)	0
	$\tau_{10} = 9.5$	7.643 (0.130)	0	5.524 (0.375)	0	7.450 (0.247)	0	7.511 (0.240)	0	7.775 (0.238)	0
$U2$	$\tau_1 = 5.0$	6.780 (0.057)	0	6.320 (0.253)	0	7.022 (0.329)	0	6.845 (0.244)	0	6.648 (0.252)	0
	$\tau_2 = 5.5$	6.683 (0.055)	0	6.050 (0.283)	0	6.497 (0.213)	0	6.516 (0.172)	0	6.624 (0.127)	0
	$\tau_3 = 6.0$	6.419 (0.052)	0	5.572 (0.278)	0	6.087 (0.379)	1.0	6.165 (0.161)	0	6.235 (0.118)	0
	$\tau_4 = 6.5$	6.060 (0.049)	0	5.084 (0.246)	0	5.804 (0.113)	0.4	5.839 (0.120)	0	5.925 (0.083)	0
	$\tau_5 = 7.0$	5.698 (0.048)	0	4.642 (0.216)	0.4	5.471 (0.104)	0.2	5.499 (0.099)	0	5.654 (0.071)	0
	$\tau_6 = 7.5$	5.420 (0.045)	0	4.349 (0.203)	8.0	5.223 (0.080)	2.4	5.239 (0.080)	2.0	5.387 (0.060)	0.8
	$\tau_7 = 8.0$	5.294 (0.049)	100	4.204 (0.203)	37.2	5.120 (0.091)	67.6	5.122 (0.083)	77.2	5.231 (0.063)	92.2
	$\tau_8 = 8.5$	5.351 (0.055)	0	4.197 (0.180)	39.0	5.203 (0.178)	28.0	5.182 (0.087)	20.8	5.293 (0.079)	7.0
	$\tau_9 = 9.0$	5.584 (0.065)	0	4.331 (0.227)	12.8	5.425 (0.120)	0.4	5.445 (0.103)	0	5.496 (0.096)	0
	$\tau_{10} = 9.5$	5.937 (0.080)	0	4.555 (0.242)	2.6	5.826 (0.158)	0	5.853 (0.128)	0	6.023 (0.129)	0

[1]PPL: pooled logistic method with parametric terms for biomarkers.
[2]NPL: pooled logistic method with smoothing splines (GAM) for biomarkers.

Table 2.2: Estimated counterfactual mean utility $\widehat{E}(U^g)$ (empirical SD in parentheses) and percentage of replicates selecting g as the optimal DTR (opt%) for regimes in biomarker Scenario 2 of Chapter II.

	Regime	Full Adherent		Unweighted		PPL[1]		NPL[2]		Proposed Method	
		$\widehat{E}(U^g)$	opt%	$\widehat{E}(U^g)$	opt%	$\widehat{E}(U^g)$	opt%	$\widehat{E}(U^g)$	opt%	$\widehat{E}(U^g)$	opt%
$U1$	$\tau_1 = 5.0$	4.816 (0.032)	0	4.524 (0.154)	0	4.774 (0.186)	1.2	4.778 (0.123)	0	4.780 (0.144)	0
	$\tau_2 = 5.5$	4.583 (0.028)	0	4.206 (0.162)	0.2	4.625 (0.191)	8.9	4.610 (0.125)	2.3	4.690 (0.108)	1.0
	$\tau_3 = 6.0$	4.515 (0.027)	100	4.015 (0.160)	47.4	4.361 (0.225)	70.2	4.401 (0.106)	75.3	4.492 (0.084)	90.2
	$\tau_4 = 6.5$	4.655 (0.030)	0	4.014 (0.161)	44.4	4.482 (0.262)	18.8	4.510 (0.093)	21.0	4.582 (0.071)	8.8
	$\tau_5 = 7.0$	5.008 (0.037)	0	4.160 (0.175)	7.6	4.800 (0.110)	0.9	4.795 (0.101)	0.4	4.901 (0.079)	0
	$\tau_6 = 7.5$	5.538 (0.047)	0	4.426 (0.211)	0.4	5.203 (0.118)	0	5.200 (0.107)	0	5.348 (0.092)	0
	$\tau_7 = 8.0$	6.168 (0.066)	0	4.735 (0.266)	0	5.680 (0.152)	0	5.677 (0.139)	0	6.032 (0.101)	0
	$\tau_8 = 8.5$	6.796 (0.088)	0	5.027 (0.274)	0	6.142 (0.184)	0	6.163 (0.167)	0	6.700 (0.143)	0
	$\tau_9 = 9.0$	7.317 (0.110)	0	5.307 (0.360)	0	6.612 (0.231)	0	6.670 (0.215)	0	7.220 (0.221)	0
	$\tau_{10} = 9.5$	7.643 (0.130)	0	5.524 (0.375)	0	7.030 (0.245)	0	7.120 (0.223)	0	7.785 (0.245)	0
$U2$	$\tau_1 = 5.0$	6.780 (0.057)	0	6.320 (0.253)	0	6.470 (0.346)	0	6.491 (0.300)	0	6.452 (0.254)	0
	$\tau_2 = 5.5$	6.683 (0.055)	0	6.050 (0.283)	0	6.335 (0.375)	0	6.467 (0.290)	0	6.816 (0.137)	0
	$\tau_3 = 6.0$	6.419 (0.052)	0	5.572 (0.278)	0	6.045 (0.400)	2.4	6.101 (0.158)	0	6.246 (0.125)	0
	$\tau_4 = 6.5$	6.060 (0.049)	0	5.084 (0.246)	0	5.761 (0.292)	2.0	5.765 (0.138)	0	5.910 (0.107)	0
	$\tau_5 = 7.0$	5.698 (0.048)	0	4.642 (0.216)	0.4	5.431 (0.166)	3.4	5.424 (0.121)	0.5	5.589 (0.094)	0
	$\tau_6 = 7.5$	5.420 (0.045)	0	4.349 (0.203)	8.0	5.110 (0.117)	5.8	5.116 (0.113)	7.1	5.310 (0.080)	1.8
	$\tau_7 = 8.0$	5.294 (0.049)	100	4.204 (0.203)	37.2	4.982 (0.107)	60.2	4.983 (0.100)	64.9	5.200 (0.085)	83.0
	$\tau_8 = 8.5$	5.351 (0.055)	0	4.197 (0.180)	39.0	5.016 (0.117)	25.7	5.029 (0.107)	27.0	5.266 (0.090)	15.2
	$\tau_9 = 9.0$	5.584 (0.065)	0	4.331 (0.227)	12.8	5.244 (0.128)	0.5	5.286 (0.118)	0.5	5.473 (0.107)	0
	$\tau_{10} = 9.5$	5.937 (0.080)	0	4.555 (0.242)	2.6	5.660 (0.155)	0	5.703 (0.139)	0	5.913 (0.144)	0

[1]PPL: pooled logistic method with parametric terms for biomarkers.
[2]NPL: pooled logistic method with smoothing splines (GAM) for biomarkers.

2.5 Application to Diabetes Example

In this section, we applied the proposed method to Eli Lilly's diabetes electronic medical record (EMR) database from Humedica. We considered patients with only type 2 diabetes, without diagnosis of type 1 or gestational diabetes, and our data was recorded from January 1, 2005 to December 31, 2010. All patients had baseline age and body mass index (BMI) measured at their first clinical visit, and during the follow-up period, their HbA1c level, use of oral anti-diabetic drugs, hospitalizations and comorbidities were recorded at several random visits (range of visiting times: $4 \sim 68$). HbA1c is the primary biomarker tested for diabetes management and research, as well as for initiating insulin therapy (*Hayward et al.*, 1997). Our goal is to use DTRs to identify the best timing to initiate insulin therapy based on patients' HbA1c levels.

We defined viable DTRs as initiating insulin therapy when a patient's HbA1c level is observed to be in an interval $[\tau, \tau + 0.5)$, similar to the definition in simulations. In other words, a patient fails to follow a given DTR if he or she starts insulin therapy with HbA1c measurement $< \tau$ during the same clinical visit or does not start with HbA1c measurement $\geq \tau + 0.5$. We considered the threshold τ from 5.5 to 8.5% with an increment of 0.5%. To be included in our analysis, each patient must have at least two clinical visits before insulin treatment and baseline HbA1c level lower than 10%, resulting in a cohort of 1220 patients. The patients are mostly seniors with obesity (BMI > 30). Table 2.3 compares patients with HbA1c $< 6.5\%$ and those with HbA1c $\geq 6.5\%$ at treatment initiation. The two groups are very similar in all aspects except HbA1c. The group with lower HbA1c at treatment initiation also has lower HbA1c at enrollment and at the end of study. However, its average HbA1c increases from 6.3 to 6.8% while in the other group, the average HbA1c decreases from 7.8 to 7.7%. Given the variables available in our data, we defined a utility function for each patient as

Table 2.3: Summary statistics (mean ± SD) for patients with HbA1c $< 6.5\%$ at initiation of insulin therapy and those with HbA1c $\geq 6.5\%$

Variable	HbA1c at Initiation of Insulin Therapy	
	$< 6.5\%$ ($n = 162$)	$\geq 6.5\%$ ($n = 1058$)
Age (yrs)	63.5 ± 8.6	63.0 ± 8.4
BMI	34.6 ± 6.7	34.7 ± 7.5
Anti-diabetic drugs before insulin (frequency)	0.6 ± 0.9	0.7 ± 1.1
New complications at end of study (type)	4.1 ± 2.4	4.0 ± 2.3
Hospitalizations at end of study (frequency)	1.5 ± 5.6	1.4 ± 3.8
HbA1c at baseline (%)	6.3 ± 0.7	7.8 ± 1.0
HbA1c at end of study (%)	6.8 ± 1.1	7.7 ± 1.4

the sum of two relative levels. The first one is relative HbA1c level calculated as the mean HbA1c level during follow-up divided by the diagnosis threshold of 6.5%. The second one is relative morbidity level calculated as the total number of new diseases developed during follow-up divided by four which is the median in this database. With this utility function, we aim to find the optimal DTR that controls both glycemic level and comorbidities that come either as side effects of insulin therapy or from diabetes progression. We applied the survival model (2.3) to estimate the weights. As in simulation Scenario 2, we used linear extrapolation to estimate the biomarker value at the event time for patients at risk. We used cubic B-splines for time-varying HbA1c observations and linear terms for all other covariates.

Figure 2.2 presents the estimated counterfactual mean utility for each of the regimes of interest. The number of compatible patients ranges from 32 to 251 for different regimes. The regime with HbA1c threshold of 6.0% has the minimal estimated counterfactual mean utility and thus is identified as g^{opt} by our proposed method. It means that initiation of insulin treatment for diabetic patients with HbA1c level of $6 \sim 6.5\%$ may lead to the best overall control of diabetes severity and related comorbidities. However, this result might be sensitive to the choice of the utility function.

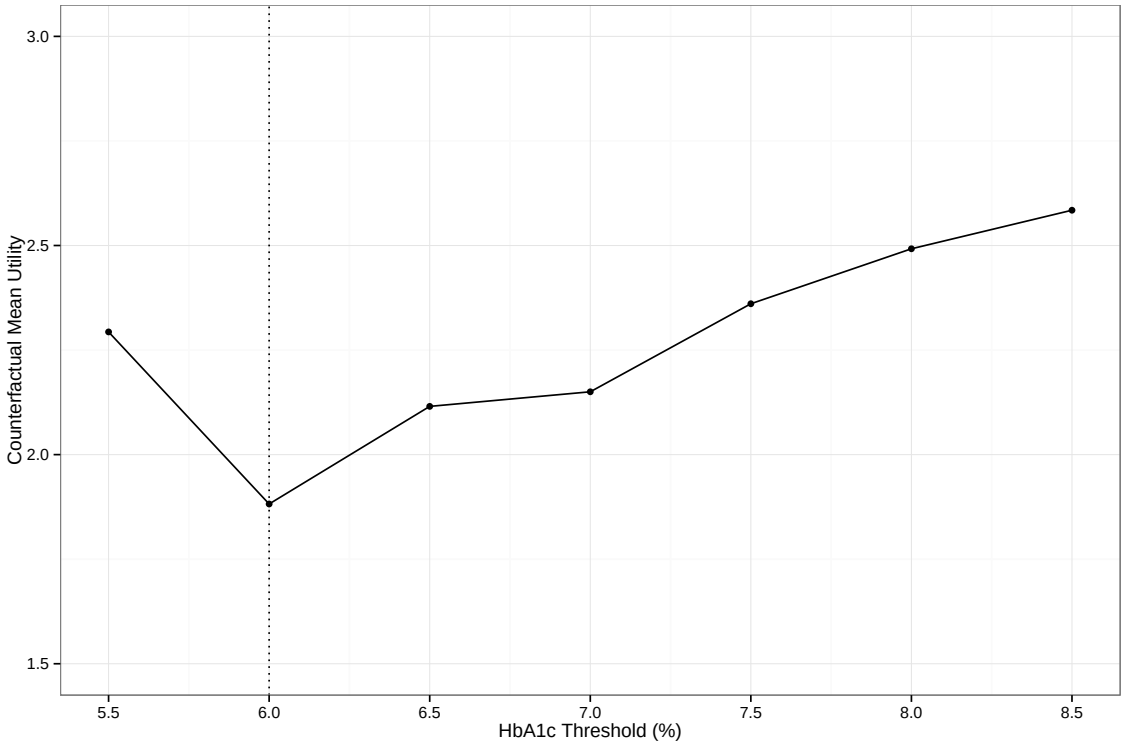

Figure 2.2: The estimated counterfactual mean utility for dynamic treatment regimes with various HbA1c thresholds to initiate insulin therapy for the diabetes example.

2.6 Discussion

In this article, we address the challenge of estimating the optimal DTR for treatment initiation given continuous random decision points based on patients' up-to-date clinical records. Our proposed method successfully fills the gap in DTR literature where most existing studies have only considered several fixed stages of treatment decisions. Compared to existing methods, the proposed method has multiple strengths. First of all, it makes minimal model assumptions and is more robust to model misspecification. Our assumption is primarily made on estimating π^g through adherence model (2.3), and the use of splines allows much flexibility on how the risk of failure to follow a specific regime depends on time-varying biomarkers. In contrast, the pooled logistic regression requires assumptions about the treatment models at every stage, and Q- and A-learning methods both require assumptions about the structure on the outcome and treatment models at every stage, resulting in high susceptibility to model misspecification. Furthermore, with one adherence model (2.3) for each DTR g regardless of the number of decision points, our method is more stable, compared to pooled logistic regression, and more computationally feasible, compared to Q- and A-learning methods. As the number of stages increase, the weights derived from pooled logistic regression may become very unstable if, for example, there are very few events at a certain stage, leading to both bias and imprecision (*Robins and Hernán*, 2009). Q- and A-learning methods use backward recursive fitting procedure (*Bather*, 2000), and as the number of stages increases, the computational burden may increase dramatically.

The proposed method can be used for many types of DTRs with multiple random decision points, not limited to the problem of treatment initiation. The key is to prespecify a class of meaningful candidate DTRs and build an appropriate model to estimate the probability of adherence to each DTR. For a time to failure of adherence

model like model (2.3), there should be enough biomarker observations at different decision points and the time to failure of adherence should be a continuous random variable. In order to assess adherence, the treatment decision rules should be clearly defined and consistent throughout the study period.

Several improvements and extensions can be explored in future studies given the restrictions in our method. First, more flexible methods can be considered for robust weight estimation, for example, the random survival forest (*Ishwaran et al.*, 2008). *Bou-Hamad et al.* (2011) has extended the method to allow for time-varying covariates. However, a more flexible method may have a lower convergence rate and one should proceed carefully considering the sample size. Second, we have only considered a limited number of well defined DTRs. In practice, there could be a larger or even infinite number of candidate regimes. For example, if we treat the HbA1c thresholds as continuous, we will have an infinite number of regimes. Furthermore, if we have many variables that may affect treatment decisions, we may also end up with a large number of DTRs that deal with various scenarios. Then with a limited sample size, it is impossible to find enough compatible patients for each regime to make inference separately. A feasible solution is to fit a nonsaturated marginal structural model for the utility conditional on the variables that define the DTRs (*Robins*, 2000; *Hernán et al.*, 2001). One may also apply classification or machine learning methods to select a subset or some combinations of variables for regime definition so as to reduce the number of DTRs of interest. Third, joint modeling of a multivariate utility function (e.g., toxicity, efficacy and management cost) may be of interest for future research. A univariate utility function is easier for comparison of various DTRs, but its definition may be arbitrary. However, with a multivariate utility function, one needs to search over a multi-dimensional plane to find the optimal DTR that achieves the best joint payoff, which can be computationally complex.

CHAPTER III

Adaptive Contrast Weighted Learning for Multi-Stage Multi-Treatment Decision-Making

3.1 Introduction

Individualized treatment strategies (ITS) are decision rules that dictate treatment prescriptions based on a patient's specific characteristics (e.g., demographics, clinical outcomes and genetic makeup). Given the increasingly popular theme of personalized medicine, many clinical and intervention scientists have now become interested in the development of ITS. Treatment individualization is important due to the fact that many diseases, such as cancer and diabetes, have complex causes by the interplay among genetic, physiological and environmental factors that vary from person to person. The effectiveness of a given treatment is usually determined not only by a patient's current disease status but also by his/her past treatment and disease history and perhaps other concurrent medical conditions. Moreover, due to the progressive nature of many chronic diseases, treatment adaptation over time is also crucial to optimize treatment effects.

Dynamic treatment regimes (DTRs) (*Robins*, 1986, 1997, 2004; *Murphy*, 2003) mathematically generalize personalized medicine to a time-varying treatment setting. They

are sequential decision rules that focus simultaneously on treatment individualization and adaptation over time. Identifying the optimal DTRs offers an effective vehicle for personalized management of diseases, and helps physicians tailor the treatment strategies dynamically and individually based on clinical evidence, which provides a key foundation for better chronic care (*Wagner et al.*, 2001). However, it is challenging to identify optimal DTRs in a multi-stage treatment setting due to the complex relationships between the alternating sequences of time-varying treatments and clinical outcomes. Recent research on estimating optimal DTRs has focused on sequential multiple assignment randomized trials (SMARTs) (*Murphy*, 2005), which are desirable for causal inference, as well as longitudinal observational studies (*Murphy*, 2003; *Robins*, 2004), which are the more common source of data. The observational data may restrict the set of DTRs that can be assessed due to possible violation of key causal assumptions and thus require careful thoughts and formulations in order to make valid inference (*Robins and Hernán*, 2009). Diverse statistical methods have been developed including marginal structural models with inverse probability weighting (IPW) (*Robins*, 2000; *Hernán et al.*, 2001; *Wang et al.*, 2012), G-estimation of structural nested mean models (*Robins*, 1986, 1989, 1997), generalized by *Murphy* (2003) and *Robins* (2004), targeted maximum likelihood estimation (*van der Laan and Rubin*, 2006), and likelihood-based approaches (*Thall et al.*, 2007). However, susceptibility to model misspecification remains as a major limitation of many methods in this field due to the inherent difficulty of modeling high-dimensional information in a time-varying setting.

Machine learning methods have become popular alternative approaches on estimating optimal DTRs. The commonly employed methods include Q-learning (*Watkins and Dayan*, 1992; *Sutton and Barto*, 1998) and A-learning (*Murphy*, 2003; *Schulte et al.*, 2014), both of which use backward induction (*Bather*, 2000) to first optimize the treatment at the last stage and then sequentially optimize the treatment at each

of the earlier stages. Q- and A- learning are both indirect approaches as they rely on maximizing or minimizing an objective function to infer the optimal DTRs and thus emphasize prediction accuracy of the clinical response model instead of directly optimizing the decision rule (*Zhao et al.*, 2012). *Zhang et al.* (2012a) propose a framework to transform the problem of estimating the optimal treatment regime into a weighted classification problem, and then directly estimate the optimal regime. Their proposed method is robust and efficient due to a combination of semiparametric regression estimators and nonparametric classification methods. However, their approach is limited to a single decision point with binary treatment options. For multi-stage decisions, *Zhao et al.* (2015) propose outcome weighted learning (OWL) to convert the optimal DTR problem into an either sequential or simultaneous classification problem. OWL utilizes existing machine learning techniques, such as support vector machines (SVM) (*Cortes and Vapnik*, 1995), to directly estimate the optimal DTR, which is flexible without the specification of outcome regression models. However, it is also not as efficient as model-based approaches if the models can be well approximated. As reviewed by *Zhou et al.* (2015), OWL is susceptible to trying to retain the actually observed treatments and is also unstable in general since its estimated individualized treatment rule is affected by a simple shift of the outcome. Moreover, OWL is susceptible to the misspecification of propensity score models since it is based on IPW. To our knowledge, few research attempts exist that deal with more than two discrete treatment options at each stage and estimate the optimal DTR in a robust and efficient way.

In this article, we develop a dynamic statistical learning method, adaptive contrast weighted learning (ACWL), to directly estimate the optimal DTR through a sequence of weighted classification for multi-stage multi-treatment decision-making in observational studies. The algorithm is implemented recursively using backward induction. Our method has multiple strengths and novelties compared to existing methods. First

of all, it can handle more than two treatments at each stage. Extending from two treatment options to more than two is nontrivial since one must account for multiple treatment comparisons without sacrificing too much on efficiency, especially when the number of treatment options is large. We achieve this by using contrasts with the adaptation of treatment effect ordering for each patient at each stage. The proposed adaptive contrasts stand for the minimum or maximum expected loss in the outcome given any sub-optimal treatment for each patient, and simplify the problem of optimization with multiple treatment comparisons to a weighted classification problem at each stage. Second, ACWL is robust and efficient by combining semiparametric regression estimators with machine learning methods. Following *Zhang et al.* (2012a), we employ the doubly robust augmented inverse probability weighted (AIPW) estimator (*Robins et al.*, 1994; *Scharfstein et al.*, 1999) to estimate the treatment effect ordering and adaptive contrasts at each stage. Last but not least, ACWL can be easily implemented using existing regression and classification methods, and is also flexible given the capability of incorporating various modeling and machine learning techniques.

The remainder of this paper is organized as follows. In Section 3.2, we formalize the problem of estimating the optimal DTR in a multi-stage multi-treatment setting using the counterfactual framework and transform it to a sequence of weighted classification using adaptive contrasts. The performance of our proposed method in various scenarios is evaluated by simulation studies in Section 3.3. We further illustrate our method in Section 3.4 using esophageal cancer data. Finally, we conclude with some discussions and suggestions for future research in Section 3.5.

3.2 Adaptive Contrast Weighted Learning (ACWL)

3.2.1 Notation

Consider a clinical trial or observational study with n subjects from a population of interest and T treatment stages. For brevity, we suppress the patient index i ($i = 1, \ldots, n$) in the following text when no confusion exists. For $j = 1, \ldots, T$, let A_j denote the multi-categorical treatment indicator at the j^{th} stage with observed value $a_j \in \mathcal{A}_j = \{1, \ldots, K_j\}$ ($K_j \geq 2$). Let $\mathbf{X}_j$ denote the vector of patient characteristics history just prior to treatment assignment A_j, containing both baseline and time-varying covariates, and $\mathbf{X}_{T+1}$ denote the entire characteristics history up to the end of stage T. Let R_j be the clinical outcome following A_j, also known as rewards, which depends on the precedent covariate history $\mathbf{X}_j$ and treatment history $A_1, \ldots, A_j$, and is also a part of the covariate history $\mathbf{X}_{j+1}$. We consider the overall outcome of interest to be $Y = f(R_1, \ldots, R_T)$, where $f(\cdot)$ is a prespecified function (e.g., sum), and assume that Y is bounded and preferable with larger values.

A DTR $\mathbf{g} = (g_1, \ldots, g_T)$ is a set of rules for personalized treatment decisions at all T stages, where g_j is a map from the domain of covariate and treatment history $\mathbf{H}_j = (A_1, \ldots, A_{j-1}, \mathbf{X}_j^\top)^\top$ to the domain of treatment assignment A_j, and we set $A_0 = \emptyset$. The optimal DTR is the one that maximizes the expectation of Y if used to assign treatments to all patients in the population of interest.

3.2.2 ACWL with $T = 1$

To facilitate the presentation of our method, we start with optimizing the treatment regime for a single stage and $K(\geq 2)$ treatment options. The method is essentially the same for optimizing the regime for the last stage in a multi-stage decision prob-

lem. We suppress the stage index in this section for brevity. To define and identify the optimal treatment regime, we consider the counterfactual framework for causal inference (*Robins*, 1986). Let $Y^*(a), a = 1, \ldots, K$, denote the counterfactual outcome had a subject received treatment a. We make the following three assumptions in order to estimate $E\{Y^*(a)\}$. First, we assume that the observed outcome is the same as the counterfactual outcome under the treatment a patient is actually given, i.e., $Y = \sum_{a=1}^{K} Y^*(a) I(A = a)$, where $I(\cdot)$ is the indicator function that takes the value 1 if $\cdot$ is true and 0 otherwise. This is referred to as the consistency assumption, which also implies that there is no interference between subjects. Second, we make the no unmeasured confounding assumption (NUCA); treatment A is randomly assigned with probability possibly dependent on $\mathbf{H}$, i.e., $\{Y^*(1), \ldots, Y^*(K)\} \perp\!\!\!\perp A|\mathbf{H}$, where $\perp\!\!\!\perp$ denotes statistical independence. Third, we assume that with probability one, the propensity score $\pi_a(\mathbf{H}) = Pr(A = a|\mathbf{H})$ is bounded away from zero, which is known as the positivity assumption.

We define the counterfactual outcome for a patient following regime g as

$$Y^*(g) = \sum_{a=1}^{K} Y^*(a) I\{g(\mathbf{H}) = a\},$$

and thus conditioning on $\mathbf{H}$, we have

$$E\{Y^*(g)\} = E_{\mathbf{H}} \left[\sum_{a=1}^{K} E\left\{Y^*(a)|\mathbf{H}\right\} I\{g(\mathbf{H}) = a\} \right],$$

where $E_{\mathbf{H}}(\cdot)$ denotes expectation with respect to the marginal joint distribution of $\mathbf{H}$. Under NUCA, we can further show that

$$E\{Y^*(g)\} = E_{\mathbf{H}} \left[\sum_{a=1}^{K} E\{Y^*(a)|A = a, \mathbf{H}\} I\{g(\mathbf{H}) = a\} \right],$$

and given the consistency assumption, we have

$$E\{Y^*(g)\} = E_{\mathbf{H}}\left[\sum_{a=1}^{K} E(Y|A=a,\mathbf{H})I\{g(\mathbf{H})=a\}\right].$$

The positivity assumption assures the identifiability of $E(Y|A=a,\mathbf{H})$.

The optimal regime, g^{opt}, is the one that maximizes the expected counterfactual outcome among the class of all potential regimes, $\mathcal{G}$. If we denote the conditional mean $E(Y|A=a,\mathbf{H})$ as $\mu_a(\mathbf{H})$, we have

$$g^{opt} = arg\max_{g\in\mathcal{G}} E_{\mathbf{H}}\left[\sum_{a=1}^{K} \mu_a(\mathbf{H})I\{g(\mathbf{H})=a\}\right].$$

Let $\mu_{(1)}(\mathbf{H}) \leq \ldots \leq \mu_{(K)}(\mathbf{H})$ denote the order statistics of $\mu_1(\mathbf{H}),\ldots,\mu_K(\mathbf{H})$, and l_a denote the treatment effect order with $\mu_{(a)}(\mathbf{H}) = \mu_{l_a}(\mathbf{H})$. Note that l_a depends on $\mathbf{H}$. Therefore, we get

$$g^{opt} = arg\max_{g\in\mathcal{G}} E_{\mathbf{H}}\left[\sum_{a=1}^{K} \mu_{(a)}(\mathbf{H})I\{g(\mathbf{H})=l_a(\mathbf{H})\}\right].$$

By subtracting $\mu_{(K)}(\mathbf{H})$ and reversing the sign, we have

$$g^{opt} = arg\min_{g\in\mathcal{G}} E_{\mathbf{H}}\left[\sum_{a=1}^{K-1} \{\mu_{(K)}(\mathbf{H}) - \mu_{(a)}(\mathbf{H})\}I\{g(\mathbf{H})=l_a(\mathbf{H})\}\right]. \tag{3.1}$$

According to (3.1), g^{opt} minimizes the expected loss in the outcome due to sub-optimal treatments in the entire population of interest. It would classify as many patients as possible to their corresponding treatment l_K (i.e., letting $I\{g(\mathbf{H})=l_a(\mathbf{H})\}=0, a=1,\ldots,K-1$) while putting more emphasis on patients with larger contrasts (i.e., larger values of $\mu_{(K)}(\mathbf{H})-\mu_{(a)}(\mathbf{H})$) if misclassification is inevitable. Ideally, for each patient, we would utilize all $K-1$ contrasts as weights to conduct treatment classification,

which, however, is challenging in practice. Meanwhile, given the inequality

$$0 \leq \mu_{(K)}(\mathbf{H}) - \mu_{(K-1)}(\mathbf{H}) \leq \mu_{(K)}(\mathbf{H}) - \mu_{(a)}(\mathbf{H}) \leq \mu_{(K)}(\mathbf{H}) - \mu_{(1)}(\mathbf{H}),$$

it is easy to show

$$\begin{aligned} E_{\mathbf{H}}\left[\sum_{a=1}^{K-1}\{\mu_{(K)}(\mathbf{H}) - \mu_{(a)}(\mathbf{H})\}I\{g(\mathbf{H}) = l_a(\mathbf{H})\}\right] &\geq E_{\mathbf{H}}\left[\sum_{a=1}^{K-1}\{C_1(\mathbf{H})I\{g(\mathbf{H}) = l_a(\mathbf{H})\}\right] \\ &= E_{\mathbf{H}}\left[C_1(\mathbf{H})I\{g(\mathbf{H}) \neq l_K(\mathbf{H})\}\right] \end{aligned}$$

and

$$\begin{aligned} E_{\mathbf{H}}\left[\sum_{a=1}^{K-1}\{\mu_{(K)}(\mathbf{H}) - \mu_{(a)}(\mathbf{H})\}I\{g(\mathbf{H}) = l_a(\mathbf{H})\}\right] &\leq E_{\mathbf{H}}\left[\sum_{a=1}^{K-1}\{C_2(\mathbf{H})I\{g(\mathbf{H}) = l_a(\mathbf{H})\}\right] \\ &= E_{\mathbf{H}}\left[C_2(\mathbf{H})I\{g(\mathbf{H}) \neq l_K(\mathbf{H})\}\right], \end{aligned}$$

where $C_1(\mathbf{H}) = \mu_{(K)}(\mathbf{H}) - \mu_{(K-1)}(\mathbf{H})$ and $C_2(\mathbf{H}) = \mu_{(K)}(\mathbf{H}) - \mu_{(1)}(\mathbf{H})$. These two contrasts indicate the minimum and maximum expected losses in the outcome, respectively, if a subject does not receive the optimal treatment, and thus are adaptive to each patient's own treatment effect ordering.

In the best (least conservative) case where sub-optimal treatments only lead to minimal expected losses in the outcome, g^{opt} is equal to

$$arg\min_{g\in\mathcal{G}} E_{\mathbf{H}}\left[C_1(\mathbf{H})I\{g(\mathbf{H}) \neq l_K(\mathbf{H})\}\right], \tag{3.2}$$

while in the worst (most conservative) case where sub-optimal treatments all lead to maximal expected losses in the outcome, g^{opt} is equal to

$$arg\min_{g\in\mathcal{G}} E_{\mathbf{H}}\left[C_2(\mathbf{H})I\{g(\mathbf{H}) \neq l_K(\mathbf{H})\}\right]. \tag{3.3}$$

We propose to estimate g^{opt} via (3.2) and (3.3) for the following reasons. By using the adaptive contrasts $C_1(\mathbf{H})$ and $C_2(\mathbf{H})$, (3.2) and (3.3) minimize, respectively, the lower and the upper bounds of the expected loss in the outcome due to sub-optimal treatments in the entire population of interest. Note that both the lower and the upper bounds of the expected loss have a limiting value of zero that can be reached with perfect classification, implying that (3.2) and (3.3) tend to g^{opt} as the expected loss goes to zero. Even when the classification is far from perfect, by minimizing the expected weighted misclassification error, (3.2) and (3.3) tend to classify as many patients as possible to their optimal treatment l_K with more focus on subjects with larger contrasts, which is consistent with g^{opt}. Therefore, we expect (3.2) and (3.3) to yield an optimal treatment regime similar, if not identical, to g^{opt}. Moreover, using the adaptive contrasts $C_1(\mathbf{H})$ and $C_2(\mathbf{H})$ simplifies the problem of optimization with multiple treatment comparisons to a weighted classification problem that many existing statistical learning methods can handle, for example, classification and regression tree (CART) (*Breiman et al.*, 1984) and SVM. These classification methods aim to reduce the difference between the true and the estimated classes by minimizing an objective function, which is the expected weighted misclassification error in our case.

The key to identifying the optimal treatment regime lies in the estimation of $\mu_A(\mathbf{H})$ and $l_A(\mathbf{H})$. *Wang et al.* (2016) show that given root-n consistent estimators $\hat{\mu}_k(\mathbf{H}), k = 1, \ldots, K$, the corresponding orders $\hat{l}_k(\mathbf{H})$ are also consistent. An intuitive approach is to posit a parametric regression model for $\mu_A(\mathbf{H}) = E(Y|A, \mathbf{H})$ to get the regression estimator $\hat{\mu}_A^{RG}(\mathbf{H})$, and then we can obtain $\hat{g}^{opt}(\mathbf{H}) = \hat{l}_K^{RG}(\mathbf{H})$ directly from $\hat{\mu}_A^{RG}(\mathbf{H})$. Alternatively, instead of using solely the regression model to infer g^{opt}, we could use it as the working model to estimate treatment effect ordering and adaptive contrasts, and then solve the weighted classification problems (3.2) and (3.3). However, both methods are susceptible to the misspecification of $\mu_A(\mathbf{H})$ by using $\hat{\mu}_A^{RG}(\mathbf{H})$. If sample size is sufficiently large, one may estimate $\mu_A(\mathbf{H})$ using nonparametric methods, for

example, random forest (*Breiman*, 2001). To balance robustness and efficiency, we propose to apply the AIPW estimator (*Robins et al.*, 1994; *Scharfstein et al.*, 1999). The K treatment options can be regarded as K arbitrary missing data patterns as in *Rotnitzky et al.* (1998). Given the estimated propensity score $\hat{\pi}_a(\mathbf{H})$, the AIPW estimator $\hat{\mu}_a^{AIPW}$ for $E\{Y^*(a)\}$ is calculated by solving

$$\mathbb{P}_n \left\{ \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})}[Y - E\{Y^*(a)\}] + U(\mathbf{H}) \right\} = 0$$

with the augmentation term

$$U(\mathbf{H}) = \sum_{k \neq a} \left\{ I(A=k) - \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})} \hat{\pi}_k(\mathbf{H}) \right\} \phi_k(\mathbf{H}).$$

Here $\phi_k(\mathbf{H})$ is an arbitrary function for treatment k, which could potentially improve the efficiency of the AIPW estimator and meanwhile does not affect the consistency of the AIPW estimator as long as the model for $\pi_a(\mathbf{H})$ is correctly specified. To incorporate the doubly robust property, we propose to set $\phi_k(\mathbf{H}) = \hat{\mu}_a(\mathbf{H}) - E\{Y^*(a)\}$ for all $k \neq a$, and then it is straightforward to show that

$$\hat{\mu}_a^{AIPW} = \mathbb{P}_n \left[\frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})} Y + \left\{ 1 - \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})} \right\} \hat{\mu}_a(\mathbf{H}) \right].$$

Notice $E\{Y^*(a)\} = E_{\mathbf{H}}\{\mu_a(\mathbf{H})\}$ under the foregoing casual assumptions and thus we define

$$\hat{\mu}_a^{AIPW}(\mathbf{H}) = \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})} Y + \left\{ 1 - \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})} \right\} \hat{\mu}_a(\mathbf{H}). \tag{3.4}$$

$\mathbb{P}_n\{\hat{\mu}_a^{AIPW}(\mathbf{H})\}$ converges to μ_a if either the model for $\pi_a(\mathbf{H})$ or the model for $\mu_a(\mathbf{H})$ is correctly specified, and thus the method is doubly robust. To apply the weighted classification problems (3.2) and (3.3), we obtain the working orders $\hat{l}_a^{AIPW}(\mathbf{H})$ by sorting $\hat{\mu}_1^{AIPW}(\mathbf{H}), \ldots, \hat{\mu}_K^{AIPW}(\mathbf{H})$ and calculate the AIPW adaptive contrasts $\hat{C}_1^{AIPW}(\mathbf{H}) = \hat{\mu}_{(K)}^{AIPW}(\mathbf{H}) - \hat{\mu}_{(K-1)}^{AIPW}(\mathbf{H})$ and $\hat{C}_2^{AIPW}(\mathbf{H}) = \hat{\mu}_{(K)}^{AIPW}(\mathbf{H}) - \hat{\mu}_{(1)}^{AIPW}(\mathbf{H})$.

For continuous outcomes, a simple and oftentimes reasonable $\hat{\mu}_a(\mathbf{H})$ can be obtained as the regression estimator $\hat{\mu}_a^{RG}(\mathbf{H})$ from a parametric linear model with coefficients dependent on treatment:

$$E(Y|A,\mathbf{H}) = \sum_{a=1}^{K} (\beta_a^\top \mathbf{H}^a) I(A=a), \tag{3.5}$$

where $\mathbf{H}^a, a = 1,\ldots,K$, are (potentially treatment dependent) summaries of the history $\mathbf{H}$ with the addition of a constant, or intercept, term, and β_a is a parameter vector for $\mathbf{H}^a$ under treatment a. For binary and count outcomes, it is straightforward to extend the method by using generalized linear models. For survival outcomes with non-informative censoring, one may use an accelerated failure time model to predict survival time for all patients. Survival outcomes with more complex censoring issues are beyond the scope of this study. The propensity score can be estimated via multinomial logistic regression (*Menard*, 2002). A working model could include all variables in $\mathbf{H}$ as linear main effect terms. Summary variables or interaction terms may also be included based on scientific knowledge.

3.2.3 ACWL with $T > 1$

The method proposed in Section 3.2.2 can be generalized to a multi-stage situation by estimating the treatment effect ordering and adaptive contrasts and applying weighted classification at each stage. Based on the idea of backward induction, we develop the following dynamic statistical learning procedure of ACWL.

For stage T, the assumptions and the way to derive the method are the same as in Section 3.2.2, except that we redefine the counterfactual outcome for a patient

following regime g_T as

$$Y^*(A_1, \ldots, A_{T-1}, g_T) = \sum_{a_T=1}^{K_T} Y^*(A_1, \ldots, A_{T-1}, a_T) I\{g_T(H_T) = a_T\},$$

where $Y^*(A_1, \ldots, A_{T-1}, a_T)$ is the counterfactual outcome for a patient treated with a_T conditional on previous treatments $(A_1, \ldots, A_{T-1})$. Let $\mu_{T,a_T}(\mathbf{H}_T)$ denote $E(Y|A_T = a_T, \mathbf{H}_T)$, we have

$$g_T^{opt} = arg \max_{g_T \in \mathcal{G}_T} E_{\mathbf{H_T}} \left[\sum_{a_T=1}^{K_T} \mu_{T,a_T}(\mathbf{H}_T) I\{g_T(\mathbf{H}_T) = a_T\} \right].$$

For stage j, $T-1 \geq j \geq 1$, we combined the method in Section 3.2.2 with machine learning methods to conduct backward induction. Following *Moodie et al.* (2012), the stage-specific pseudo-outcome PO_j for estimating treatment effect ordering and adaptive contrasts is a predicted counterfactual outcome under optimal treatments at all future stages, also known as the "optimal benefit-to-go" in *Murphy* (2005). Specifically, we have

$$PO_j = E\left\{Y^*(A_1, \ldots, A_j, g_{j+1}^{opt}, \ldots, g_T^{opt})\right\},$$

or in a recursive form,

$$PO_j = E\{PO_{j+1}|A_{j+1} = g_{j+1}^{opt}(\mathbf{H}_{j+1}), \mathbf{H}_{j+1}\}$$

and we set $PO_T = Y$. For $a_j = 1, \ldots, K_j$, let $\mu_{j,a_j}(\mathbf{H}_j)$ denote the conditional mean $E[PO_j|A_j = a_j, \mathbf{H}_j]$, and we have $PO_j = \mu_{j+1,g_{j+1}^{opt}}(\mathbf{H}_{j+1})$. We replace Y with PO_j to apply the method in Section 3.2.2 at stage j. Specifically, let $PO_j^*(a_j)$ denote the counterfactual pseudo-outcome for a patient with treatment a_j at stage j. We have the consistency assumption as $PO_j = \sum_{a_j=1}^{K_j} PO_j^*(a_j) I(A_j = a_j)$, NUCA as

$\{PO_j^*(1), \ldots, PO_j^*(K_j)\} \perp\!\!\!\perp \mathbf{H}_j$ and the positivity assumption as $\pi_{a_j}(\mathbf{H}_j) = Pr(A_j = a_j|\mathbf{H}_j)$ being bounded away from zero. With these three assumptions, we identify the optimal regime directly following Section 3.2.2 and get g_j^{opt} among all potential regimes $\mathcal{G}_j$ as

$$g_j^{opt} = arg \max_{g_j \in \mathcal{G}_j} E_{\mathbf{H}_j} \left[\sum_{a_j=1}^{K_j} \mu_{j,a_j}(\mathbf{H}_j) I\{g_j(\mathbf{H}_j) = a_j\} \right],$$

or equivalently,

$$g_j^{opt} = arg \min_{g_j \in \mathcal{G}_j} E_{\mathbf{H}_j} \left[\sum_{a_j=1}^{K_j-1} \{\mu_{j,(K)}(\mathbf{H}_j) - \mu_{j,(a)}(\mathbf{H}_j)\} I\{g_j(\mathbf{H}_j) = l_{a_j}(\mathbf{H}_j)\} \right], \quad (3.6)$$

where $\mu_{j,(1)}(\mathbf{H}_j) \leq \ldots \leq \mu_{j,(K)}(\mathbf{H}_j)$ denote the treatment effect ordering and the order $l_{a_j}(\mathbf{H}_j)$ means $\mu_{j,(a_j)}(\mathbf{H}_j) = \mu_{j,l_{a_j}}(\mathbf{H}_j)$.

Again, the optimization problem (3.6) is complicated by the multiple treatment comparisons. Therefore, we incorporate the adaptive contrasts as in Section 3.2.2 for each stage. Specifically, the adaptive contrasts are $C_{j,1}(\mathbf{H}_j) = \mu_{j,(K)}(\mathbf{H}_j) - \mu_{j,(K-1)}(\mathbf{H}_j)$ and $C_{j,2}(\mathbf{H}_j) = \mu_{j,(K)}(\mathbf{H}_j) - \mu_{j,(1)}(\mathbf{H}_j)$, which indicate respectively, the minimum and the maximum expected losses in the pseudo-outcome, if a patient does not receive the optimal treatment at stage j. Via the adaptive contrasts, we transform the problem of optimization with multiple treatment comparisons to a simpler weighted classification problem.

We start the estimation with stage T and conduct backward induction. Our ACWL algorithm starting with stage $j = T$ is carried out as follows:

Step 1: Fit regression model (3.5) with pseudo-outcome PO_j to obtain regression-based conditional mean estimator $\hat{\mu}_{j,a_j}^{RG}(\mathbf{H}_j)$.

Step 2: Fit the propensity score model to obtain $\hat{\pi}_{j,a_j}(\mathbf{H}_j)$.

Step 3: Calculate AIPW-based conditional mean estimator $\hat{\mu}^{AIPW}_{j,a_j}(\mathbf{H}_j)$ using $\hat{\mu}^{RG}_{j,a_j}(\mathbf{H}_j)$ and $\hat{\pi}_{j,a_j}(\mathbf{H}_j)$ as in (3.4).

Step 4: Calculate the AIPW-based working orders $\hat{l}^{AIPW}_{j,a_j}(\mathbf{H}_j)$ and adaptive contrasts $\hat{C}^{AIPW}_{j,1}(\mathbf{H}_j)$ and $\hat{C}^{AIPW}_{j,2}(\mathbf{H}_j)$ using $\hat{\mu}^{AIPW}_{j,a_j}(\mathbf{H}_j)$.

Step 5: Take $\hat{l}^{AIPW}_{j,K}(\mathbf{H}_j)$ as the class label, and $\hat{C}^{AIPW}_{j,1}(\mathbf{H}_j)$ and $\hat{C}^{AIPW}_{j,2}(\mathbf{H}_j)$ as the weights to solve problems (3.2) and (3.3) using existing classification techniques.

Step 6: If $j > 1$, set $j = j - 1$ and repeat steps 1 to 6. If $j = 1$, stop.

When the outcome is cumulative (e.g., the sum of longitudinally observed values or a single continuous final outcome), we modify the pseudo-outcomes to reduce accumulated bias from the conditional mean models, following *Huang et al.* (2015). For stage j, $T - 1 \geq j \geq 1$, instead of using only the model-based values under optimal future treatments, i.e., $\mu_{j+1,g^{opt}_{j+1}}(\mathbf{H}_{j+1})$, we use the actual observed outcomes plus the expected future loss due to sub-optimal treatments. Specifically, the modified pseudo-outcome is

$$PO^{'}_{j} = PO^{'}_{j+1} + \mu_{j+1,g^{opt}_{j+1}}(\mathbf{H}_{j+1}) - \mu_{j+1,a_{j+1}}(\mathbf{H}_{j+1}),$$

where a_{j+1} is the treatment that a patient actually received at stage $j + 1$, and $\mu_{j+1,g^{opt}_{j+1}}(\mathbf{H}_{j+1}) - \mu_{j+1,a_{j+1}}(\mathbf{H}_{j+1})$ is the expected loss due to sub-optimal treatments at stage $j + 1$ for a given patient, which is zero if $g^{opt}_{j+1}(\mathbf{H}_{j+1}) = a_{j+1}$ and positive otherwise. Again we set $PO'_T = Y$. This modification leads to more robustness against model misspecification and is less likely to accumulate bias from stage to stage during backward induction (*Huang et al.*, 2015).

3.3 Simulation Studies

We conduct simulation studies to evaluate the performance of our proposed method in two aspects. First, we need to evaluate whether $\hat{g}^{opt}$ estimated through weighted classification with adaptive contrasts is close enough to the truth in numerical studies. Second, we aim to show the robustness of our methods with different levels of model misspecification. To achieve this, we purposely set all regression models μ to be misspecified, as is the case for most real data applications, and let the propensity model π be either correctly (e.g., randomized trials) or incorrectly (e.g., most observational studies) specified. We consider a single-stage scenario as in Section 3.2.2 and a multi-stage scenario as in Section 3.2.3, each with 500 replications. For both scenarios, we generate five baseline covariates $X_1, \ldots, X_5$ according to $N(0, 1)$, and set the expected counterfactual outcome under the optimal treatment regime, i.e., $E\{Y^*(\mathbf{g}^{opt})\}$, to be 8. We use CART to minimize the weighted misclassification error, which is implemented by the R package *rpart*.

3.3.1 Scenario 1: $T = 1$ and $K = 5$

In Scenario 1, we consider a single stage with five treatment options and sample size of 1000. We generate treatment A from $Multinomial(\pi_0/\pi_s, \pi_1/\pi_s, \pi_2/\pi_s, \pi_3/\pi_s, \pi_4/\pi_s)$, with $\pi_0 = 1$, $\pi_1 = \exp(0.5 - 0.5X_1)$, $\pi_2 = \exp(0.5X_1 + 0.2)$, $\pi_3 = \exp(0.5X_5 + 0.1)$, $\pi_4 = \exp(0.5X_5 - 0.1)$, and $\pi_s = \sum_{m=0}^{4} \pi_m$. We set A to take values in $\{0, \ldots, 4\}$ and generate outcomes as

$$Y = \exp[2.06 + 0.2X_3 - |X_1 + X_2|\varphi\{A, g^{opt}(\mathbf{H})\}] + \epsilon,$$

with $\varphi\{A, g^{opt}(\mathbf{H})\}$ taking the form of $\varphi^{(1)} = 3I\{A \neq g^{opt}(\mathbf{H})\}$ or $\varphi^{(2)} = \{A - g^{opt}(\mathbf{H})\}^2$, $g^{opt}(\mathbf{H}) = I(X_1 > -1)\{1 + I(X_2 > -0.4) + I(X_2 > 0.4) + I(X_2 > 1)\}$ and $\epsilon \sim N(0, 1)$.

The function $\varphi\{A, g^{opt}(\mathbf{H})\}$ indicates the penalty if a patient does not receive the optimal treatment. Given $\varphi^{(1)}$, misclassification to any of the four sub-optimal treatments leads to the same expected loss in the outcome for a given patient, which means that all $K - 1$ contrasts in (3.1) are actually the same for that patient. In this case, (3.2) and (3.3) are both identical to g^{opt} and we expect them to have good performances. With $\varphi^{(2)}$, we consider a more common situation where the differences among treatments vary, and misclassification to a treatment closer to the optimal one leads to a smaller expected loss in the outcome. In this case, the $K - 1$ contrasts are not all the same and therefore, (3.2) and (3.3) are not identical to g^{opt}. Simulation studies under $\varphi^{(1)}$ and $\varphi^{(2)}$ investigate the performance of ACWL and see how close (3.2) and (3.3) are tending to g^{opt}. Under each form of $\varphi\{A, g^{opt}(\mathbf{H})\}$, we further assess the robustness of our method. By using linear regression, we have a misspecified conditional mean model. For the propensity score, we consider both a correctly specified model $log(\pi_d/\pi_0) = \beta_{0d} + \beta_{1d}X_1 + \beta_{2d}X_5, d = 1, \ldots, 4$, and an incorrectly specified one $log(\pi_d/\pi_0) = \beta_{0d}$.

We apply the proposed ACWL algorithm to each simulated dataset and denote the methods using the two adaptive contrasts as ACWL-C_1 and ACWL-C_2, respectively. For comparison, we use the regression-based conditional mean models directly to infer the optimal DTRs and we denote this method as RG. We also use the contrasts and orders estimated from the conditional mean models to apply weighted classification (3.2) and (3.3), and denote these two methods as RG-C_1 and RG-C_2. Furthermore, we apply the OWL method by *Zhao et al.* (2012) with CART.

Table 3.1 summarizes the performances of all methods considered in Scenario 1, in

Table 3.1: Simulation results for Scenario 1 in Chapter III with a single stage and five treatment options. π is the propensity score model. $\varphi^{(1)}$ and $\varphi^{(2)}$ indicate equal and varying penalties for misclassification. *opt*% shows the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their optimal treatments. $\hat{E}\{Y^*(\hat{g}^{opt})\}$ shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime. $E\{Y^*(g^{opt})\} = 8$, 500 replications, and $n = 1000$.

π	Method	$\varphi^{(1)}$		$\varphi^{(2)}$	
		opt%	$\hat{E}\{Y^*(\hat{g}^{opt})\}$	*opt*%	$\hat{E}\{Y^*(\hat{g}^{opt})\}$
-	RG	58.1 (2.6)	5.39 (0.17)	59.5 (3.8)	5.99 (0.25)
	RG-C_1	55.1 (4.1)	5.20 (0.29)	58.2 (6.0)	6.00 (0.37)
	RG-C_2	55.7 (3.8)	5.24 (0.29)	58.4 (5.7)	6.00 (0.34)
Correct	OWL	83.2 (9.2)	6.92 (0.60)	74.6 (11.6)	6.80 (0.56)
	ACWL-C_1	94.2 (3.5)	7.69 (0.21)	88.7 (5.5)	7.60 (0.22)
	ACWL-C_2	90.4 (6.1)	7.38 (0.40)	86.4 (8.4)	7.36 (0.38)
Incorrect	OWL	60.0 (13.8)	5.57 (0.89)	52.0 (11.0)	5.89 (0.65)
	ACWL-C_1	92.5 (4.1)	7.60 (0.23)	84.2 (6.7)	7.47 (0.24)
	ACWL-C_2	90.2 (6.0)	7.37 (0.38)	85.6 (8.2)	7.35 (0.36)

terms of the percentage of subjects correctly classified to their optimal treatments, denoted as *opt*%, and the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime, denoted as $\hat{E}\{Y^*(\hat{g}^{opt})\}$. *opt*% shows how likely the estimated optimal regime is to assign a new patient to his or her real optimal treatment and $\hat{E}\{Y^*(\hat{g}^{opt})\}$ shows how much the entire population of interest will benefit from following $\hat{g}^{opt}$. The regression-based methods RG, RG-C_1 and RG-C_2 have relatively poor performances since the conditional mean model is misspecified. They classify $55 \sim 59\%$ patients to their optimal treatments, resulting in a $\hat{E}\{Y^*(\hat{g}^{opt})\}$ much smaller than the true value of 8. OWL has relatively good performance only when the propensity score model is correctly specified, as expected, and it is the least efficient among all methods considered with large empirical standard deviations (SDs) for both *opt*% and $\hat{E}\{Y^*(\hat{g}^{opt})\}$. Our proposed method classifies over 84% patients to their optimal treatments in all cases and achieves $\hat{E}\{Y^*(\hat{g}^{opt})\}$ close

to 8. ACWL is highly robust against model misspecification with only slight decrease in performance from using a correctly specified propensity score model to using an incorrectly specified one. Under $\varphi^{(1)}$ when all $K-1$ contrasts are the same, both (3.2) and (3.3) are equal to g^{opt} and thus yield satisfactory $opt\%$ and $\hat{E}\{Y^*(\hat{g}^{opt})\}$. From $\varphi^{(1)}$ to $\varphi^{(2)}$, the regression-based methods show improved $\hat{E}\{Y^*(\hat{g}^{opt})\}$ despite similar $opt\%$, indicating higher sensitivity to subjects with larger contrasts given varying expected losses due to sub-optimal treatments. Although $K-1$ contrasts are not all the same under $\varphi^{(2)}$, ACWL-C_1 and ACWL-C_2 show very slight deterioration in $opt\%$ and $\hat{E}\{Y^*(\hat{g}^{opt})\}$, compared to the results under $\varphi^{(1)}$, and are still much better than the other methods. These results confirm the feasibility of estimating g^{opt} via ACWL with adaptive AIPW contrasts.

3.3.2 Scenario 2: $T = 2$ and $K_1 = K_2 = 3$

In this section, we generate data under a two-stage DTR with three treatment options at each stage. We consider the outcome of interest as the sum of the rewards from each stage, i.e., $Y = R_1 + R_2$, and set φ to be the form as $\varphi^{(2)}$ in Scenario 1. We evaluate the performance of our proposed method given a misspecified conditional mean model through linear regression, while allowing the propensity score models to be either correctly or incorrectly specified. Furthermore, since we apply CART for classification, we consider both a tree-type underlying optimal DTR and a non-tree-type one. We consider sample sizes of 500 and 1000.

Treatment variables are set to take values in $\{0, 1, 2\}$ at each stage. For stage 1, we generate A_1 from $Multinomial(\pi_{10}, \pi_{11}, \pi_{12})$, with $\pi_{10} = 1/\{1 + \exp(0.5 - 0.5X_3) + \exp(0.5X_4)\}$, $\pi_{11} = \exp(0.5 - 0.5X_3)/\{1 + \exp(0.5 - 0.5X_3) + \exp(0.5X_4)\}$ and $\pi_{12} =$

$1-\pi_{10}-\pi_{11}$. We generate stage 1 reward as

$$R_1 = \exp[1.5 - |1.5X_1 + 2|\{A_1 - g_1^{opt}(\mathbf{H}_1)\}^2] + \epsilon_1,$$

with tree-type $g_1^{opt}(\mathbf{H}_1) = I(X_1 > -1)\{I(X_2 > -0.5) + I(X_2 > 0.5)\}$ or non-tree-type $g_1^{opt}(\mathbf{H}_1) = I(X_1 > -0.5)\{1 + I(X_1 - X_2 > 0)\}$, and $\epsilon_1 \sim N(0,1)$.

For stage 2, we have treatment $A_2 \sim Multinomial(\pi_{20}, \pi_{21}, \pi_{22})$, with $\pi_{20} = 1/\{1 + \exp(0.2R_1 - 1) + \exp(0.5X_4)\}$, $\pi_{21} = \exp(0.2R_1 - 1)/\{1 + \exp(0.2R_1 - 1) + \exp(0.5X_4)\}$ and $\pi_{22} = 1 - \pi_{20} - \pi_{21}$. We generate stage 2 reward as

$$R_2 = \exp[1.26 - |1.5X_3 - 2|\{A_2 - g_2^{opt}(\mathbf{H}_2)\}^2] + \epsilon_2,$$

with tree-type $g_2^{opt}(\mathbf{H}_2) = I(X_3 > -1)\{I(R_1 > 0.5) + I(R_1 > 3)\}$ or non-tree-type $g_2^{opt}(\mathbf{H}_2) = I(X_3 > 0) + I(X_3 + R_1 > 2.5)$, and $\epsilon_2 \sim N(0,1)$.

We apply the proposed ACWL algorithm with the modified pseudo-outcome to each simulated dataset. For comparison, we use the regression-based conditional mean models directly to infer the optimal DTR, which is Q-learning. We also apply the backward OWL (BOWL) method by *Zhao et al.* (2015) with CART. As BOWL does not involve outcome regression models, only subjects whose observed treatments are optimal at stage 2 can be used for identifying the optimal regime at stage 1, resulting in a significantly reduced sample size. Therefore, we also consider BOWL combined with Q-learning, denoted as BOWL-Q. Basically, at stage 1, we use the conditional mean model from Q-learning to predict the pseudo-outcome for patients whose observed treatments are not optimal at stage 2 and then apply OWL using all subjects to identify the optimal regime.

Results for Scenario 2 are shown in Table 3.2. The regression-based conditional mean

Table 3.2: Simulation results for Scenario 2 in Chapter III with two stages and three treatment options at each stage. π is the propensity score model. *opt*% shows the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their optimal treatments. $\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$ shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal DTR. $E\{Y^*(\mathbf{g}^{opt})\} = 8$, and 500 replications.

π	Method	Tree-type DTR				Non-tree-type DTR			
		$n = 500$		$n = 1000$		$n = 500$		$n = 1000$	
		opt%	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$	*opt*%	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$	*opt*%	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$	*opt*%	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$
-	Q-learning	51.2 (3.3)	5.83 (0.23)	53.3 (2.6)	5.97 (0.21)	55.5 (4.2)	6.07 (0.23)	58.0 (3.2)	6.24 (0.20)
Correct	BOWL	28.9 (6.1)	4.30 (0.44)	38.6 (7.9)	4.66 (0.52)	26.1 (5.8)	4.25 (0.40)	34.8 (7.6)	4.56 (0.50)
	BOWL-Q	38.1 (9.3)	5.04 (0.66)	63.2 (10.5)	6.41 (0.59)	31.7 (7.4)	4.69 (0.50)	49.1 (10.3)	5.49 (0.55)
	ACWL-C_1	85.1 (4.7)	7.29 (0.21)	93.3 (3.3)	7.57 (0.13)	74.1 (5.7)	6.68 (0.29)	83.3 (3.8)	7.10 (0.16)
	ACWL-C_2	85.4 (5.3)	7.31 (0.24)	93.7 (3.3)	7.60 (0.13)	77.8 (5.4)	6.83 (0.24)	86.6 (3.6)	7.25 (0.15)
Incorrect	BOWL	23.7 (5.9)	4.05 (0.42)	26.3 (6.5)	4.10 (0.42)	22.1 (4.9)	4.02 (0.34)	23.6 (5.6)	4.11 (0.38)
	BOWL-Q	26.4 (7.3)	4.31 (0.51)	30.3 (8.7)	4.43 (0.61)	24.7 (5.6)	4.30 (0.40)	26.4 (6.7)	4.41 (0.46)
	ACWL-C_1	84.1 (4.9)	7.27 (0.25)	91.8 (3.7)	7.42 (0.17)	72.3 (6.1)	6.60 (0.33)	81.1 (4.1)	7.03 (0.18)
	ACWL-C_2	83.8 (5.9)	7.25 (0.28)	91.8 (3.8)	7.43 (0.18)	76.9 (5.9)	6.65 (0.31)	82.9 (4.0)	7.09 (0.17)

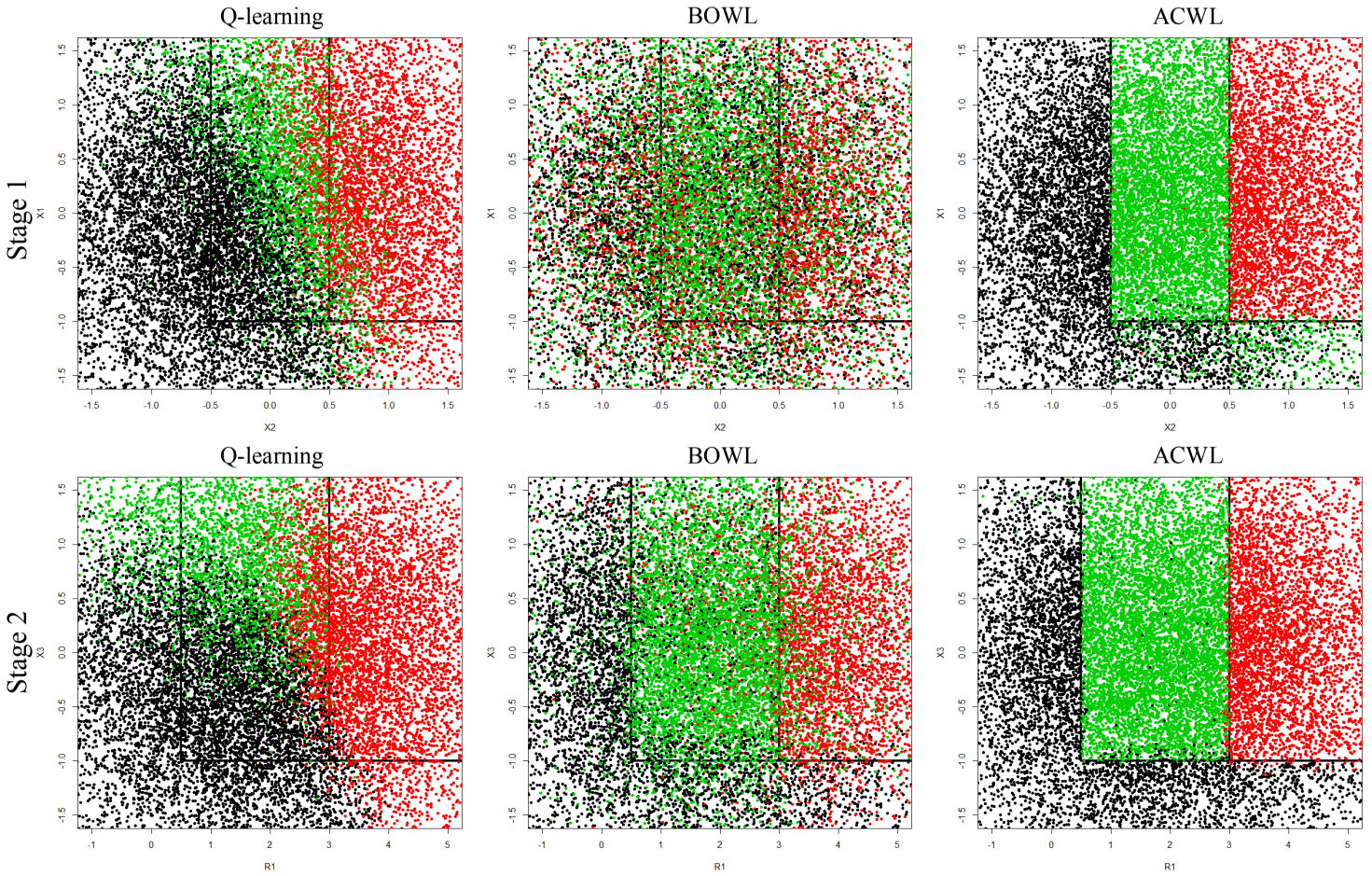

Figure 3.1: Predicted optimal treatments in simulation Scenario 2 of Chapter III with a tree-type underlying optimal DTR, correctly specified propensity score model and sample size of 1000. The true regions at stage 1 are red for $X_1 > -1$ and $X_2 > 0.5$, green for $X_1 > -1$ and $-0.5 < X_2 \leq 0.5$ and black elsewhere. The true regions at stage 2 are red for $R_1 > 3$ and $X_3 > -1$, green for $0.5 < R_1 \leq 3$ and $X_3 > -1$ and black elsewhere.

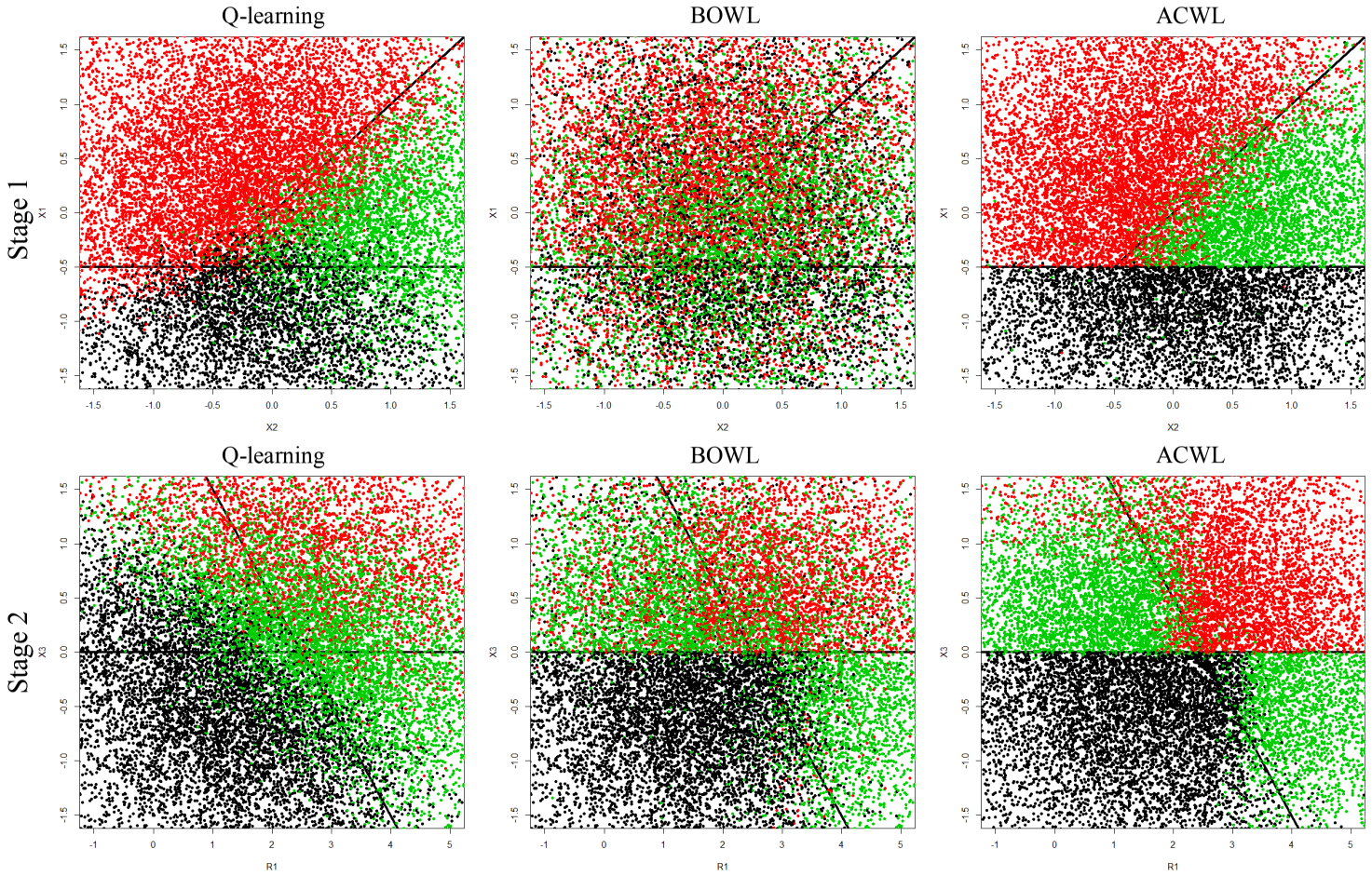

Figure 3.2: Predicted optimal treatments in simulation Scenario 2 of Chapter III with a non-tree-type underlying optimal DTR, correctly specified propensity score model and sample size of 1000. The true regions at stage 1 are red for $X_1 > 0$ and $X_1 > X_2$, black for $X_1 \leq 0$ and green elsewhere. The true regions at stage 2 are red for $X_3 > 0$ and $R_1 + X_3 > 2.5$, black for $X_3 \leq 0$ and $R_1 + X_3 \leq 2.5$, and green elsewhere.

models explain about 34% of the total variance at stage 2 and 20% of the total variance at stage 1. Q-learning is relatively stable with different sample sizes while all classification-based methods show clear improvement with an increased sample size. The two OWL methods are the least efficient with large empirical SDs. BOWL-Q has higher *opt*% and $\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$ but also larger variability than BOWL, implying a bias and variance trade-off by incorporating misspecified but informative regression models. Similarly as in Scenario 1, ACWL has the best performance among all methods considered with average *opt*% over 80% and $\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$ closest to 8 in all cases. ACWL is also very robust against misspecification of the propensity score model while BOWL and BOWL-Q have significant deterioration in performance with a misspecified propensity score model. From a tree-type underlying optimal DTR to a non-tree-type one, all CART-based methods show worse performance. For our proposed method, *opt*% decreases by approximately 10% and $\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$ drops $0.3 \sim 0.6$, yet still much better than all the other methods. Figures 3.1 and 3.2 further shows how the methods perform in predicting the optimal treatments for new subjects with correctly specified propensity score models, sample size of 1000 and the underlying optimal DTR being tree-type and non-tree-type, respectively. We only present the results from ACWL-C_2 given the similarity between ACWL-C_1 and ACWL-C_2. In Figure 3.1, ACWL leads to clear differentiation of the three regions, which matches the true underlying DTR, while in Figure 3.2, there are more misclassified cases near the borders, likely due to the use of CART for the non-tree-type underlying DTR. In both figures, ACWL shows superior performances compared to Q-learning and BOWL.

Notably, in both single-stage and multi-stage scenarios, ACWL is robust and efficient compared to the other methods, even with misspecified models for both outcome and propensity score. This may be due to the following reasons. First, the treatment effect ordering and adaptive contrasts are constructed using the doubly robust AIPW

estimator. Second, we utilize the flexible weighted classification, instead of using the orders and contrasts directly, to estimate the optimal DTR, which further improves robustness. Comparing ACWL-C_1 and ACWL-C_2, we do not have a clear conclusion on which one is better. We suggest implementing both and choosing the optimal DTR by taking the common part or by incorporating background knowledge. Additional simulation studies can be found in the Appendix A, which lead to a similar conclusion. ACWL becomes less efficient with more treatment options or more stages but still performs much better than the other competing methods.

3.4 Application to the Esophageal Cancer Example

As a further illustration, we apply ACWL to the esophageal cancer data collected by MD Anderson Cancer Center from 1998 to 2012. At baseline, we have $n = 1170$ patients with about 90% at overall cancer stage II or III (*Byrd et al.*, 2010). The general disease management strategy is chemotherapy or chemoradiation therapy (CRT) followed by surgery (*Lloyd and Chang*, 2014).

Figure 3.3 shows the two-stage disease management before surgery in our observational data. At baseline, all patients had records of basic characteristics and disease status, including a total 11 covariates, denoted by $\mathbf{X}_1 = (X_{1,1}, \ldots, X_{1,11})^\top$. At treatment stage 1, about 40% of the patients received induction chemotherapy (ICT), denoted by A_1 with YES for treated and NO for untreated. Tumor response was measured right before treatment stage 2, denoted as X_2, which is an intermediate variable. X_2 takes values from 0 to 5 with 0 being progression and 5 being complete response, compared to baseline tumor measures. At stage 2, all patients received CRT with one of three radiation modalities: 3D conformal radiotherapy (3DCRT, 39% of the total patients), intensity-modulated radiation therapy (IMRT, 45%) and proton

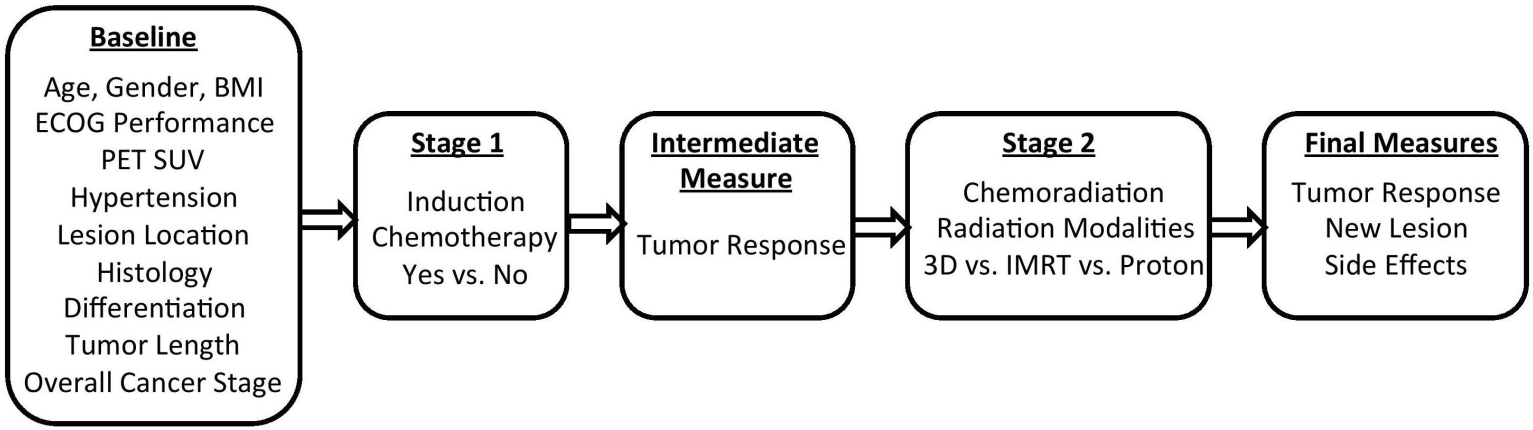

Figure 3.3: Two-stage disease management for esophageal cancer patients.

therapy (PT, 16%). We use A_2 to denote the stage 2 treatment variable. After CRT, tumor response and development of new lesions were measured within three months (before surgery), denoted as $R_{3,1}$ (same scale as X_2) and $R_{3,2}$ (0 for development of new lesions and 1 for none), respectively. We focus on these two stages to estimate the optimal DTR to decide whether a patient should receive ICT at stage 1 and what radiation modality should be used for CRT at stage 2. We define a single outcome $Y = R_{3,1} + 2R_{3,2}$ to measure the effectiveness of the two-stage treatments, and side effects (e.g., nausea, anorexia and fatigue) are not included in the evaluation because most of them would go away shortly after CRT. Missing data is imputed using IVEware (*Raghunathan et al.*, 2002).

We apply the ACWL algorithm to the data described as above. Specifically, the covariate and treatment history just prior to stage 2 treatment is $\mathbf{H}_2 = (\mathbf{X}_1, A_1, X_2)$ and the number of treatment options at stage 2 is $K_2 = 3$. We fit a linear regression model for $\mu_{2,A_2}(\mathbf{H}_2)$ as in (3.5) using Y as the outcome and all variables in $\mathbf{H}_2$ as predictors that interact with A_2. For the propensity score, we fit a multinomial logistic regression model including main effects of all variables in $\mathbf{H}_2$. We use CART with pruning for weighted classification. We repeat the same procedure for stage 1 except that we have $\mathbf{H}_1 = (X_{1,1}, \ldots, X_{1,11})$, $K_1 = 2$ and $\hat{PO}_1^{'} = Y + \hat{\mu}_{2,\hat{g}_2^{opt}}(\mathbf{H}_2) - \hat{\mu}_{2,a_2}(\mathbf{H}_2)$.

We find very similar results using ACWL-C_1 and ACWL-C_2, and thus combine the results by using variables that both methods identify as important (CART variable importance ≥ 15). For stage 1, the most important variables are tumor length (mm, continuous) and overall clinical stage (I/II vs. III/IV). For stage 2, the most important variables are stage 1 treatment A_1, intermediate tumor response X_2 and baseline tumor differentiation (well/moderate vs. poor). The estimated optimal DTR $\hat{\mathbf{g}}^{opt} =$

$c(\hat{g}_1^{opt}, \hat{g}_2^{opt})$ is

$$\hat{g}_1^{opt}(\mathbf{H}_1) = \begin{cases} \text{YES} & \text{if tumor length } \geq 36mm \text{ or stage} = \text{III/IV} \\ \text{NO} & \text{otherwise} \end{cases}$$

and

$$\hat{g}_2^{opt}(\mathbf{H}_2) = \begin{cases} \text{PT} & \text{if } A_1 = \text{NO and tumor differentiation} = \text{poor} \\ \text{IMRT} & \text{if } A_1 = \text{YES and intermediate tumor response } < 4 \\ \text{3DCRT} & \text{otherwise} \end{cases}$$

As suggested by the estimated optimal DTR, ICT is recommended at stage 1 for patients with larger tumor or worse clinical stage, which is consistent with clinical findings that the addition of ICT is appropriate for advanced disease with high risk for local or distant failure (*Haddad et al.*, 2013). Some studies have shown that ICT is beneficial overall for both tumor control and prolonging survival (*Jin et al.*, 2004) but there have not been randomized trials or studies focusing on subgroups of patients. At stage 2, our result suggests that patients who do not use ICT and have poor tumor differentiation should use PT in CRT, patients with ICT and minor or worse tumor response after ICT should use IMRT and all other patients should use 3DCRT. Currently, there has not been any large trial comparing the three radiation modalities. Some studies have shown that PT and IMRT are more efficient at targeting the tumors and less toxic than 3DCRT (*Lloyd and Chang*, 2014), which may explain why our result suggests PT or IMRT for patients with worse conditions.

3.5 Discussion

We have proposed a robust and efficient method ACWL to estimate the optimal DTR, which can effectively handle multiple treatment options at each stage. The adaptive contrasts we develop at each stage simplify the problem of optimization with multiple treatment comparisons to a dynamic weighted learning procedure, and our simulations studies show that this simplification leads to excellent numerical performances. Our method combines robust semiparametric regression estimators with flexible machine learning methods. With regression models at each stage, one can predict the future outcomes under optimal treatments for patients whose assigned treatments are not all optimal at future stages, thus improving efficiency if the regression models are well approximated. The doubly robust AIPW estimator and nonparametric classification method that we utilize help improve the robustness of ACWL against model misspecification. Therefore, our proposed method is capable of dealing with observational data. Moreover, the dynamic ACWL algorithm can be easily implemented with existing regression and classification methods.

Several improvements and extensions can be explored in future studies. Generalizing the ACWL method to handle informatively censored data is clinically meaningful as many studies focus on prolonging patients' survival. *Goldberg and Kosorok* (2012) has developed a method within the Q-learning framework by using inverse-probability-of-censoring weighting (IPCW). With ACWL, one may combine the probability of treatment with the probability of censoring in the AIPW estimator. Due to the flexibility in the ACWL algorithm, many other machine learning methods can be considered, for both the classification part (e.g., SVM or other tree-based learning methods) and the backward induction part (e.g., A-learning). Moreover, with high dimensional data, one can incorporate variable selection at each stage for the conditional mean models. In addition, it may be of great practical interest to explore

generalization of ACWL with continuous treatment options, such as radiation dose.

CHAPTER IV

Tree-based Reinforcement Learning for Estimating Optimal Dynamic Treatment Regimes

4.1 Introduction

Nowadays personalized medicine has become a popular concept in health care. Many clinical and intervention scientists are working on treatment optimization to account for patients' heterogeneity in response to treatments as well as the progressive nature of many chronic diseases and conditions. Dynamic treatment regimes (DTRs) (*Robins*, 1986, 1997, 2004; *Murphy*, 2003; *Chakraborty and Murhpy*, 2014), as sequential treatment decision rules, mathematically generalize personalized medicine to focus simultaneously on treatment individualization and adaptation over time. The identification of optimal DTRs offers an effective vehicle for personalized management of diseases, and helps physicians tailor the treatment strategies dynamically and individually based on clinical evidence, providing a key foundation for better health care (*Wagner et al.*, 2001). However, it is challenging to identify optimal DTRs in a multi-stage treatment setting due to the complex relationships between the alternating sequences of time-varying treatments and clinical outcomes. Diverse statistical methods have been developed including marginal structural models with

inverse probability weighting (IPW) (*Robins*, 2000; *Hernán et al.*, 2001; *Wang et al.*, 2012), G-estimation of structural nested mean models (*Robins*, 1986, 1989, 1997) and its generalizations (*Murphy*, 2003; *Robins*, 2004), targeted maximum likelihood estimators (*van der Laan and Rubin*, 2006), and likelihood-based approaches (*Thall et al.*, 2007).

Machine learning methods have become popular alternative approaches for estimating optimal DTRs. In particular, the problem of multi-stage decision-making has strong resemblance to reinforcement learning (RL), which is a branch of machine learning (*Chakraborty and Moodie*, 2013). Unlike supervised learning (SL) (e.g., regression and classification), the desired output value or the optimal decision, known as *label*, is not observed in RL, and the learning agent has to keep interacting with the environment to learn the best policy for decision-making. In the observational data with dynamic decisions, each patient only receives one treatment at a stage, which may not be the optimal decision, and such data are usually expensive and perhaps time-consuming to obtain. Therefore, one oftentimes has to apply parametric or semiparametric methods and pool over subject-level data to estimate the optimal DTR. This type of RL that works with one sample of data is called batch-mode RL (*Ernst et al.*, 2005), which is typical in a medical setting. Commonly employed batch-mode RL methods include Q-learning (*Watkins and Dayan*, 1992; *Sutton and Barto*, 1998) and A-learning (*Murphy*, 2003; *Schulte et al.*, 2014), both of which use backward induction (*Bather*, 2000). Q- and A-learning rely on maximizing or minimizing an objective function to indirectly infer the optimal DTRs and thus emphasize prediction accuracy of the clinical response model instead of directly optimizing the decision rule (*Zhao et al.*, 2012). In addition to RL methods, *Zhao et al.* (2015) propose outcome weighted learning (OWL) to convert the optimal DTR problem into an either sequential or simultaneous classification problem, and then apply SL methods such as support vector machines (SVM) (*Cortes and Vapnik*, 1995). However, OWL is susceptible

to trying to retain the actually observed treatments and its estimated individualized treatment rule is affected by a simple shift of the outcome (*Zhou et al.*, 2015). OWL is also susceptible to the misspecification of propensity score models since it is based on IPW. Moreover, most existing methods, including A-learning and OWL, are limited to binary treatment options. To deal with multi-stage multi-treatment decisions in a robust way, *Tao and Wang* (2016) propose adaptive contrast weighted learning (ACWL) which combines doubly robust augmented IPW (AIPW) estimators with classification algorithms. ACWL is another example of converting batch-mode RL into SL. Unlike OWL which uses the observed treatment as *label*, ACWL identifies *label* according to the treatment effect ordering estimated by AIPW. Nonetheless, the conversion from batch-mode RL to SL may induce additional uncertainty through the identification of *label*. ACWL also avoids the challenging multiple treatment comparisons by utilizing adaptive contrasts, which may not the most efficient way despite superior performances to all other existing methods.

In this paper, we aim to develop a batch-mode RL method without the hassle of conversion to SL and to directly handle the problem of optimization with multiple treatment comparisons. As an example for a single-stage decision problem, *Laber and Zhao* (2015) propose a novel tree-based method, which we denote as LZ, to directly estimate optimal treatment regimes in a multi-treatment setting. Typically, a decision tree is a SL method that uses tree-like graphs or models to map observations about an item to conclusions about the item's target value, for example, the classification and regression tree (CART) algorithm by *Breiman et al.* (1984). LZ fits the batch-mode RL task into a decision tree with a purity measure that is unsupervised, and meanwhile maintains the advantages of a decision tree, such as simplicity for understanding and interpretation, and capability of handling various types of data (e.g., continuous or categorical) without distributional assumptions. However, similar to OWL, LZ is also susceptible to propensity model misspecification.

ACWL and LZ have inspired us to develop a tree-based RL (T-RL) algorithm that enjoys the advantages of both existing methods. In summary, T-RL has the following strengths. 1) Through the use of decision trees, T-RL is capable of handling multinomial or ordinal treatments. T-RL incorporates multiple treatment comparisons directly, while ACWL relies on the simplification by adaptive contrasts that indicate the minimum or maximum expected loss in the outcome given any sub-optimal treatment. Moreover, T-RL maintains the nature of RL without the extra step of conversion to SL as in ACWL. Therefore, we expect T-RL to perform better than ACWL given a tree-type underlying DTR. 2) We replace the IPW estimators in LZ with AIPW estimators. Similar to ACWL, we expect T-RL to be robust and efficient by combining robust semiparametric regression estimators with nonparametric machine learning methods. 3) T-RL works for multiple fixed treatment stages by using backward induction, as opposed to a single stage in LZ.

The remainder of this paper is organized as follows. In Section 4.2, we formalize the problem of estimating the optimal DTR in a multi-stage multi-treatment setting using the counterfactual framework, derive purity measures for decision trees at multiple stages and describes the recursive tree growing process. The performance of our proposed method in various scenarios is evaluated by simulation studies in Section 4.3. We further illustrate our method in Section 4.4 using a case study to identify dynamic substance abuse treatment regimes for adolescents. Finally, we conclude with some discussions and suggestions for future research in Section 4.5.

4.2 Tree-based Reinforcement Learning (T-RL)

4.2.1 Dynamic Treatment Regimes (DTRs)

Consider a multi-stage decision problem with T decision stages and K_j ($K_j \geq 2$) treatment options at the j^{th} ($j = 1, \ldots, T$) stage. Data could come from either a randomized trial or an observational study. Let A_j denote the multi-categorical or ordinal treatment indicator with observed value $a_j \in \mathcal{A}_j = \{1, \ldots, K_j\}$. Let $\mathbf{X}_j$ denote the vector of patient characteristics history just prior to treatment assignment A_j, and $\mathbf{X}_{T+1}$ denote the entire characteristics history up to the end of stage T. Let R_j be the reward following A_j, which depends on the precedent covariate history $\mathbf{X}_j$ and treatment history $A_1, \ldots, A_j$, and is also a part of the covariate history $\mathbf{X}_{j+1}$. We consider the overall outcome of interest as $Y = f(R_1, \ldots, R_T)$, where $f(\cdot)$ is a prespecified function (e.g., sum), and we assume that Y is bounded and preferable with larger values. The observed data are $\{(A_{1i}, \ldots, A_{Ti}, \mathbf{X}_{T+1,i}^\top, Y_i)\}_{i=1}^n$, assumed to be independent and identically distributed for n subjects from a population of interest. For brevity, we suppress the subject index i in the following text when no confusion exists.

A DTR is a sequence of individualized treatment rules, $\mathbf{g} = (g_1, \ldots, g_T)$, where g_j is a map from the domain of covariate and treatment history $\mathbf{H}_j = (A_1, \ldots, A_{j-1}, \mathbf{X}_j^\top)^\top$ to the domain of treatment assignment A_j, and we set $A_0 = \emptyset$. To define and identify the optimal DTR, we consider the counterfactual framework for causal inference (*Robins*, 1986). At stage T, let $Y^*(A_1, \ldots, A_{T-1}, a_T)$, or $Y^*(a_T)$ for brevity, denote the counterfactual outcome for a patient treated with $a_T \in \mathcal{A}_T$ conditional on previous treatments $(A_1, \ldots, A_{T-1})$, and define $Y^*(g_T)$ as the counterfactual outcome

under regime g_T, i.e.,

$$Y^*(g_T) = \sum_{a_T=1}^{K_T} Y^*(a_T) I\{g_T(\mathbf{H}_T) = a_T\}.$$

The performance of g_T is measured by the counterfactual mean outcome $E\{Y^*(g_T)\}$, and the optimal regime, g_T^{opt}, satisfies $E\{Y^*(g_T^{opt})\} \geq E\{Y^*(g_T)\}$ for all $g_T \in \mathcal{G}_T$, where $\mathcal{G}_T$ is the class of all potential regimes. To connect the counterfactual outcomes with the observed data, we make the following three standard assumptions (*Murphy et al.*, 2001; *Robins and Hernán*, 2009; *Orellana et al.*, 2010a).

Assumption 1 *Consistency.* The observed outcome is the same as the counterfactual outcome under the treatment a patient is actually given, i.e., $Y = \sum_{a_T=1}^{K_T} Y^*(a_T) I(A_T = a_T)$, where $I(\cdot)$ is the indicator function that takes the value 1 if $\cdot$ is true and 0 otherwise. It also implies that there is no interference between subjects.

Assumption 2 *No unmeasured confounding.* Treatment A_T is randomly assigned with probability possibly dependent on $\mathbf{H}_T$, i.e.,

$$\{Y^*(1), \ldots, Y^*(K_T)\} \perp\!\!\!\perp A_T | \mathbf{H}_T,$$

where $\perp\!\!\!\perp$ denotes statistical independence.

Assumption 3 *Positivity.* There exists constants $0 < c_0 < c_1 < 1$ such that, with probability 1, the propensity score $\pi_{a_T}(\mathbf{H}_T) = Pr(A_T = a_T | \mathbf{H}_T) \in (c_0, c_1)$.

Following the derivation in *Tao and Wang* (2016) under the three assumptions, we

have

$$E\{Y_T^*(g_T)\} = E_{\mathbf{H}_T}\left[\sum_{a_T=1}^{K_T} E(Y|A_T = a_T, \mathbf{H}_T) I\{g_T(\mathbf{H}_T) = a_T\}\right],$$

where $E_{\mathbf{H}_T}(\cdot)$ denotes expectation with respect to the marginal joint distribution of $\mathbf{H}_T$. If we denote the conditional mean $E(Y|A_T = a_T, \mathbf{H}_T)$ as $\mu_{T,a_T}(\mathbf{H}_T)$, we have

$$g_T^{opt} = arg \max_{g_T \in \mathcal{G}_T} E_{\mathbf{H}_T}\left[\sum_{a_T=1}^{K_T} \mu_{T,a_T}(\mathbf{H}_T) I\{g_T(\mathbf{H}_T) = a_T\}\right]. \tag{4.1}$$

At stage j, $T-1 \geq j \geq 1$, g_j^{opt} can be expressed in terms of the observed data via backward induction (*Bather*, 2000). Following *Murphy* (2005) and *Moodie et al.* (2012), we define a stage-specific pseudo-outcome PO_j for estimating g_j^{opt}, which is a predicted counterfactual outcome under optimal treatments at all future stages, also known as the value function. Specifically, we have

$$PO_j = E\left\{Y^*(A_1, \ldots, A_j, g_{j+1}^{opt}, \ldots, g_T^{opt})\right\},$$

or in a recursive form,

$$PO_j = E\{PO_{j+1}|A_{j+1} = g_{j+1}^{opt}(\mathbf{H}_{j+1}), \mathbf{H}_{j+1}\}$$

and we set $PO_T = Y$.

For $a_j = 1, \ldots, K_j$, let $\mu_{j,a_j}(\mathbf{H}_j)$ denote the conditional mean $E[PO_j|A_j = a_j, \mathbf{H}_j]$, and we have $PO_j = \mu_{j+1,g_{j+1}^{opt}}(\mathbf{H}_{j+1})$. Let $PO_j^*(a_j)$ denote the counterfactual pseudo-outcome for a patient with treatment a_j at stage j. For the three assumptions, we have *positivity* as $PO_j = \sum_{a_j=1}^{K_j} PO_j^*(a_j) I(A_j = a_j)$, *no unmeasured confounding* as $\{PO_j^*(1), \ldots, PO_j^*(K_j)\} \perp\!\!\!\perp \mathbf{H}_j$ and *positivity* as $\pi_{a_j}(\mathbf{H}_j) = Pr(A_j = a_j|\mathbf{H}_j)$ being bounded away from zero. Under these three assumptions, the optimization problem

at stage j, among all potential regimes $\mathcal{G}_j$, can be written as

$$g_j^{opt} = arg \max_{g_j \in \mathcal{G}_j} E_{\mathbf{H}_j} \left[\sum_{a_j=1}^{K_j} \mu_{j,a_j}(\mathbf{H}_j) I\{g_j(\mathbf{H}_j) = a_j\} \right]. \tag{4.2}$$

4.2.2 Purity Measures for Decision Trees at Multiple Stages

We propose to use a tree-based method to solve (4.1) and (4.2). Typically, a CART tree is a binary decision tree constructed by splitting a parent node into two child nodes repeatedly, starting with the root node which contains the entire learning samples. The basic idea of tree growing is to choose a split among all possible splits at each node so that the resulting child nodes are the purest. Thus the purity or impurity measure is crucial to the tree growing. Traditional classification and regression trees are supervised learning methods, with the goal of inferring a function that describes the relationship between the outcome and covariates. The outcome, also known as *label*, is the observed truth and can be used directly to measure purity. Commonly used impurity measures include Gini index and information index for categorical outcomes, and least squares deviation for continuous outcomes (*Breiman et al.*, 1984).

However, the estimation target of a DTR problem, which is the optimal treatment for a patient with characteristics $\mathbf{H}_j$ at stage j, i.e., $g_j^{opt}(\mathbf{H}_j), j = 1, \ldots, T$, is not directly observed. Information about $g_j^{opt}(\mathbf{H}_j)$ can only be inferred indirectly through the observed treatments and outcomes. Using the causal framework and the foregoing three assumptions, we can pool over all subject-level data to estimate the expected counterfactual outcomes given all possible treatments. With the overall goal of maximizing the counterfactual mean outcome in the entire population of interest, the selected split at each node should also improve the counterfactual mean outcome,

which can serve as a measure of purity in DTR trees. Figure 4.1 shows a decision tree for a single-stage optimal treatment rule with $\mathcal{A} = \{0, 1, 2\}$. Let $\Omega_m, m = 1, 2, \ldots,$ denote the nodes which are regions defined by the covariate space following all precedent binary splits, with the root node $\Omega_1 = \mathbb{R}^p$. We number the rectangular region $\Omega_m, m \geq 2$, so that its parent node is $\Omega_{\lceil m/2 \rceil}$, where $\lceil \cdot \rceil$ means taking the smallest integer not less than $\cdot$. Figure 4.1 shows the chosen covariate and best split at each node, as well as the expected counterfactual outcome after assigning a single optimal treatment to that node. The splits are selected to increase the counterfactual mean outcome. At the root node, if we select a single treatment for all subjects, treatment 1 is the most beneficial overall, yielding a counterfactual mean outcome of 0.7. Splitting via X_1 and X_2, the optimal regime g^{opt} is to assign treatment 2 to region $\Omega_3 = \{X_1 > 0\}$, treatment 0 to region $\Omega_4 = \{X_1 \leq 0, X_2 \leq 0.5\}$, and treatment 1 to region $\Omega_5 = \{X_1 \leq 0, X_2 > 0.5\}$. We can see that this tree is fundamentally different from a CART tree as it does not attempt to describe the relationship between the outcome and covariates or the rule for the assignment of the observed treatments, and instead it describes the rule by which treatments should be assigned to future subjects in order to maximize purity.

Laber and Zhao (2015) propose a measure of node purity based on the IPW estimator of the counterfactual mean outcome (*Zhang et al.*, 2012a; *Zhao et al.*, 2012),

$$E\{Y^*(g)\} = E_{\mathbf{H}}\left[\frac{I(A = g(\mathbf{H}))}{\pi_A(\mathbf{H})} Y\right],$$

for a single-stage ($T = 1$) decision problem. They assume that the propensity score $\pi_A(\mathbf{H})$ is known, and propose a purity measure $\mathcal{P}^{LZ}(\Omega, \omega)$ as

$$\max_{a_1, a_2 \in \mathcal{A}} \mathbb{P}_n\left[\frac{\{Y - \hat{m}(\mathbf{H})\} I\{A = g_{\omega, a_1, a_2}(\mathbf{H})\}}{\pi_A(\mathbf{H})} I(\mathbf{H} \in \Omega)\right] \left(\mathbb{P}_n\left[\frac{I\{A = g_{\omega, a_1, a_2}(\mathbf{H})\}}{\pi_A(\mathbf{H})} I(\mathbf{H} \in \Omega)\right]\right)^{-1},$$

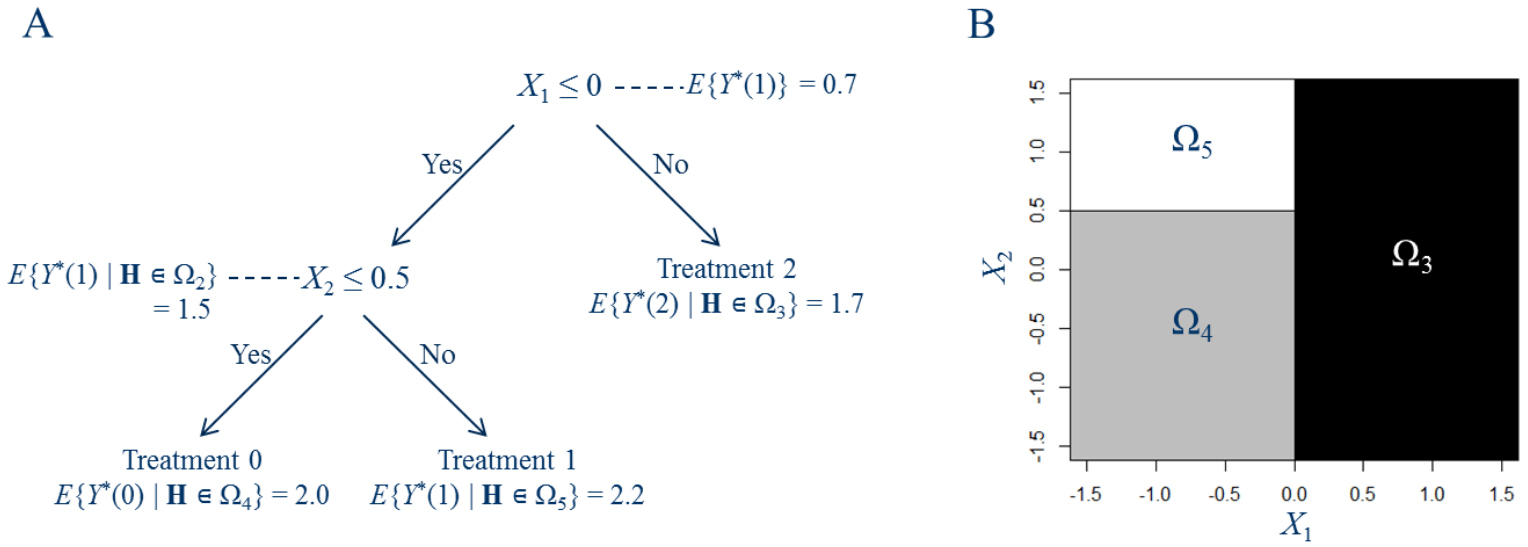

Figure 4.1: (A) A decision tree for optimal treatment rules and the expected counterfactual outcome by assigning a single best treatment to each node that represents a subset covariate space. (B) Regions divided by the terminal nodes in the decision tree indicating different optimal treatments.

where $\mathbb{P}_n$ is the empirical expectation operator, $\hat{m}(\mathbf{H}) = \max_{a\in\mathcal{A}} \hat{\mu}_a(\mathbf{H})$, Ω denotes the node to be split, ω and ω^c is a partition of Ω, and for a given partition ω and ω^c, g_{ω,a_1,a_2} denotes the decision rule that assigns treatment a_1 to subjects in ω and treatment a_2 to subjects in ω^c. However, in an observational study where $\pi_A(\mathbf{H})$ is unknown, $\mathcal{P}^{LZ}(\Omega,\omega)$ is subject to misspecification of the propensity model.

To improve robustness, we propose to use an AIPW estimator for the counterfactual mean outcome as in *Tao and Wang* (2016). By regarding the K treatment options as K arbitrary missing data patterns (*Rotnitzky et al.*, 1998), the AIPW estimator for $E\{Y^*(a)\}$ is $\mathbb{P}_n\{\hat{\mu}_a^{AIPW}(\mathbf{H})\}$, with

$$\hat{\mu}_a^{AIPW}(\mathbf{H}) = \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})}Y + \left\{1 - \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})}\right\}\hat{\mu}_a(\mathbf{H}). \tag{4.3}$$

Lemma 1 (Double Robustness)**.** $\mathbb{P}_n\{\hat{\mu}_a^{AIPW}(\mathbf{H})\}$ *is a consistent estimator of* $E\{Y^*(a)\}$ *if either the propensity model* $\pi_a(\mathbf{H})$ *or the conditional mean model* $\mu_a(\mathbf{H})$ *is correctly specified.*

In our multi-stage setting, for stage T, given estimated conditional mean $\hat{\mu}_{T,a_T}^{AIPW}(\mathbf{H}_T)$ and estimated propensity score $\hat{\pi}_{T,A_T}(\mathbf{H}_T)$, the proposed estimator for (4.1) is

$$\begin{aligned}\hat{g}_T^{opt} &= arg\max_{g_T\in\mathcal{G}_T} \mathbb{P}_n\left[\sum_{a_T=1}^{K_T}\hat{\mu}_{T,a_T}^{AIPW}(\mathbf{H}_T)I\{g_T(\mathbf{H}_T)=a_T\}\right]\\ &= arg\max_{g_T\in\mathcal{G}_T} \mathbb{P}_n\left[\frac{I(A_T=g_T(\mathbf{H}_T))}{\hat{\pi}_{T,A_T}(\mathbf{H}_T)}Y + \left\{1-\frac{I(A_T=g_T(\mathbf{H}_T))}{\hat{\pi}_{T,A_T}(\mathbf{H}_T)}\right\}\hat{\mu}_{T,g_T}(\mathbf{H}_T)\right].\end{aligned}$$

For stage j $(T-1\le j\le 1)$, the proposed estimator for (4.2) is

$$\hat{g}_j^{opt} = arg\max_{g_j\in\mathcal{G}_j} \mathbb{P}_n\left[\frac{I(A_j=g_j(\mathbf{H}_j))}{\hat{\pi}_{j,A_j}(\mathbf{H}_j)}\hat{PO}_j + \left\{1-\frac{I(A_j=g_j(\mathbf{H}_j))}{\hat{\pi}_{j,A_j}(\mathbf{H}_j)}\right\}\hat{\mu}_{j,g_j}(\mathbf{H}_j)\right],$$

where $\hat{\pi}_{j,A_j}(\mathbf{H}_j)$ is the estimated propensity score, $\hat{\mu}_{j,a_j}^{AIPW}(\mathbf{H}_j)$ is the estimated conditional mean, and $\hat{PO}_j = \hat{\mu}_{j+1,\hat{g}_{j+1}^{opt}}(\mathbf{H}_{j+1})$ is the estimated pseudo-outcome.

Similar to the derivation in *Laber and Zhao* (2015), for a given partition ω and ω^c of node Ω, let g_{j,ω,a_1,a_2} denote the decision rule that assigns treatment a_1 to subjects in ω and treatment a_2 to subjects in ω^c at stage j ($T \leq j \leq 1$), and we define the purity measure as

$$\mathcal{P}_j(\Omega, \omega) = \max_{a_1, a_2 \in \mathcal{A}_j} \mathbb{P}_n \left[\sum_{a_j=1}^{K_j} \hat{\mu}_{j,a_j}^{AIPW}(\mathbf{H}_j) I\{g_{j,\omega,a_1,a_2}(\mathbf{H}_j) = a_j\} I(\mathbf{H}_j \in \Omega) \right].$$

We can see that $\mathcal{P}_j(\Omega, \omega)$ evaluates the performance of the best decision rule which assigns a single treatment for each of the two arms under partition.

4.2.3 Recursive Partitioning

As we have mentioned, the purity measures for our T-RL are different from the ones in supervised decision trees. However, after defining $\mathcal{P}_j(\Omega, \omega), j = 1, \ldots, T$, the recursive partitioning to grow the tree is similar. Each split depends on the value of only one covariate. A nominal covariate with C categories has $2^{C-1} - 1$ possible splits and an ordinal or continuous covariate with L different values has $L - 1$ unique splits. Therefore, at a given node Ω, a possible split ω indicates either a subset of categories for a nominal covariate or values no larger than a threshold for an ordinal or continuous covariate. The best split ω^{opt} is chosen to maximize the improvement in the purity, $\mathcal{P}_j(\Omega, \omega) - \mathcal{P}_j(\Omega)$, where $\mathcal{P}_j(\Omega)$ means to assign a single best treatment to all subjects in Ω without splitting. It is straightforward to see that $\mathcal{P}_j(\Omega, \omega) \geq \mathcal{P}_j(\Omega)$. In order to control overfitting as well as to make meaningful splitting, a positive constant λ is given to represent a threshold for practical significance and another positive integer n_0 is given as the minimal node size which is dictated by problem-specific considerations. Under these conditions, we first evaluate the following three *Stopping Rules* for node Ω.

*R*ule 1. If the node size is less than $2n_0$, the node will not be split.

*R*ule 2. If all possible splits of a node result in a child node with size smaller than n_0, the node will not be split.

*R*ule 3. If the current tree depth reaches the user-specified maximum depth, the tree growing process will stop.

If none of the foregoing *Stopping Rules* are met, we compute the best split by

$$\hat{\omega}^{opt} = arg\max_{\omega} \left[\mathcal{P}_j(\Omega, \omega) : \min\{n\mathbb{P}_n I(\mathbf{H}_j \in \omega), n\mathbb{P}_n I(\mathbf{H}_j \in \omega^c)\} \geq n_0\right].$$

Before deciding whether or not to spliti Ω into ω and ω^c, we evaluate the following *Stopping Rule* 4.

*R*ule 4. If the maximum purity improvement $\mathcal{P}_j(\Omega, \hat{\omega}^{opt}) - \mathcal{P}_j(\Omega)$ is less than λ, the node will not be split.

We split Ω into ω and ω^c if none of the four stopping rules apply.

When there is no clear scientific guidance on λ to indicate practical significance, one approach is to choose a relatively small positive value to build a complete tree and then prune the tree back in order to minimize a measure of cost for the tree. Following the CART algorithm, the cost is a measure of the total impurity of the tree with a penalty term on the number of terminal nodes, and the complexity parameter for the penalty term can be tuned by cross-validation (CV) (*Breiman et al.*, 1984). Alternatively, we propose to select λ directly by CV, similar to the method by *Laber and Zhao* (2015). As a direct measure of purity is not available in RL, we again incorporate the idea of maximizing the counterfactual mean outcome and use a 10-fold CV estimator of the counterfactual mean outcome. Theoretically, CV can be conducted at each

stage separately and one can use a potentially different λ for each stage. To reduce modeling uncertainty in the pseudo-outcomes and also simplify the process, we carry out CV only at stage T using the overall outcome Y directly. Specifically, we use nine subsamples as training data to estimate the function of $\mu_{T,a_T}(\cdot)$ with $\hat{\mu}^{AIPW}_{T,a_T}(\mathbf{H}_T)$ and $g^{opt}_T(\cdot)$ using T-RL for a given λ, and then plug in $\mathbf{H}_T$ of the remaining subsample to get $\hat{\mu}^{AIPW,CV}_{T,a_T}(\mathbf{H}_T)$ and $\hat{g}^{opt,CV,\lambda}_T(\mathbf{H}_T)$. We repeat the process ten times with each subsample being the test data once. Then the CV-based counterfactual mean outcome under λ is

$$\hat{E}\{Y^*(\hat{g}^{opt,CV,\lambda}_T)\} = \mathbb{P}_n\left[\sum_{a_T=1}^{K_T} \hat{\mu}^{AIPW,CV}_{T,a_T}(\mathbf{H}_T) I\{\hat{g}^{opt,CV,\lambda}_T(\mathbf{H}_T) = a_T\}\right],$$

and the best value for λ is $\hat{\lambda} = arg\max_\lambda \hat{E}\{Y^*(\hat{g}^{opt,CV,\lambda}_T)\}$. As the scale of the outcome affects the scale of $\mathcal{P}_j(\Omega,\omega) - \mathcal{P}_j(\Omega)$, we search over a sequence of candidate λ's as a sequence of percentages of $\mathcal{P}_T(\Omega_1)$, i.e., the estimated counterfactual mean outcome under a single best treatment for all subjects (Ω_1 is the root node).

4.2.4 Implementation of T-RL

The AIPW estimator $\hat{\mu}^{AIPW}_{j,a_j}(\mathbf{H}_j), j = 1,\ldots,T, a_j = 1,\ldots,K_j$, consists of three parts to be estimated, the pesudo-outcome PO_j, the propensity score $\pi_{j,a_j}(\mathbf{H}_j)$ and the conditional mean model $\mu_{j,a_j}(\mathbf{H}_j)$.

We start the estimation with stage T and conduct backward induction. At stage T, we use the outcome Y directly, i.e., $PO_T = Y$. For stage $j, T-1 \geq j \geq 1$, given a cumulative outcome (e.g., the sum of longitudinally observed values or a single continuous final outcome), we use a modified version of pseudo-outcomes to reduce accumulated bias from the conditional mean models (*Huang et al.*, 2015). Instead of using only the model-based values under optimal future treatments, i.e.,

$\mu_{j+1,g_{j+1}^{opt}}(\mathbf{H}_{j+1})$, we use the actual observed outcomes plus the expected future loss due to sub-optimal treatments, which means

$$PO_j^{'} = PO_{j+1}^{'} + \mu_{j+1,g_{j+1}^{opt}}(\mathbf{H}_{j+1}) - \mu_{j+1,a_{j+1}}(\mathbf{H}_{j+1}),$$

where a_{j+1} is the treatment that a patient actually received at stage $j+1$, and $\mu_{j+1,g_{j+1}^{opt}}(\mathbf{H}_{j+1}) - \mu_{j+1,a_{j+1}}(\mathbf{H}_{j+1})$ is the expected loss due to sub-optimal treatments at stage $j+1$ for a given patient, which is zero if $g_{j+1}^{opt}(\mathbf{H}_{j+1}) = a_{j+1}$ and positive otherwise. Given $PO'_T = Y$, it is easy to see that

$$PO_j^{'} = Y + \sum_{t=j+1}^{T} \{\mu_{t,g_t^{opt}}(\mathbf{H}_t) - \mu_{t,a_t}(\mathbf{H}_t)\}.$$

This modification leads to more robustness against model misspecification and is less likely to accumulate bias from stage to stage during backward induction (*Huang et al.*, 2015). In our simulations, we estimate $PO_j^{'}$ by using random forests-based conditional mean estimates (*Breiman*, 2001).

The propensity score $\pi_{j,a_j}(\mathbf{H}_j)$ can be estimated via multinomial logistic regression (*Menard*, 2002). A working model could include linear main effect terms for all variables in $\mathbf{H}_j$. Summary variables or interaction terms may also be included based on scientific knowledge.

The conditional mean estimate $\hat{\mu}_{j,a_j}(\mathbf{H}_j)$ in the augmentation term of $\hat{\mu}_{j,a_j}^{AIPW}(\mathbf{H}_j)$ can be obtained from a parametric regression model. For continuous outcomes, a simple and oftentimes reasonable example is the parametric linear model with coefficients dependent on treatment:

$$E(\hat{PO}_j^{'}|A_j, \mathbf{H}_j) = \sum_{a_j=1}^{K_j} (\beta_{a_j}^{\top}\mathbf{H}_j) I(A_j = a_j), \tag{4.4}$$

where β_a is a parameter vector for $\mathbf{H}_j$ under treatment a_j. For binary and count outcomes, it is straightforward to extend the method by using generalized linear models. For survival outcomes with non-informative censoring, one may use an accelerated failure time model to predict survival time for all patients. Survival outcomes with more complex censoring issues are beyond the scope of the current study.

The T-RL algorithm starting with stage $j = T$ is carried out as follows:

*S*tep 1. Obtain AIPW estimates $\hat{\mu}^{AIPW}_{j,a_j}(\mathbf{H}_j), a_j = 1, \ldots, K_j$, using full data.

*S*tep 2. At root node $\Omega_{j,m}, m = 1$, set values for λ and n_0.

*S*tep 3. At node $\Omega_{j,m}$, evaluate the four *Stopping Rules*. If any of the *Stopping Rules* is satisfied, assign a single best treatment $arg\max_{a_j \in \mathcal{A}_j} \mathbb{P}_n[\hat{\mu}^{AIPW}_{j,a_j}(\mathbf{H}_j) I(\mathbf{H}_j \in \Omega_{j,m})]$ to all subject in $\Omega_{j,m}$. Otherwise, split $\Omega_{j,m}$ into child nodes $\Omega_{j,2m}$ and $\Omega_{j,2m+1}$ by $\hat{\omega}^{opt}$.

*S*tep 4. Set $m = m + 1$ and repeat Step 3 until all nodes are terminal.

*S*tep 5. If $j > 1$, set $j = j - 1$ and repeat steps 1 to 4. If $j = 1$, stop.

Similar to the CART algorithm, T-RL is greedy as it chooses splits only at the current node for purity improvement, which may not lead to a global maximum. One way to potentially enhance the performance is lookahead (*Murthy and Salzberg*, 1995). We test this in our simulation by fixed-depth lookahead: evaluating the purity improvement after splitting the parent node as well as its two child nodes, comparing the total purity improvement via up to four nodes to the improvement with no split at the parent node, and finally deciding whether or not to split the parent node. We denote this method as T-RL-LH.

4.3 Simulation Studies

We conduct simulation studies to investigate finite sample performance of our proposed method. We set all regression models μ to be misspecified, which is the case for most real data applications, while allowing the specification of the propensity model π be either correct (e.g., randomized trials) or incorrect (e.g., most observational studies). We consider first a single-stage scenario so as to facilitate the comparison with existing methods, particularly *Laber and Zhao* (2015), and then a multi-stage scenario. For each scenario, we consider sample sizes of either 500 or 1000 for the training datasets and 1000 for the test datasets, and repeat the simulation 500 times. We use the training datasets to estimate the optimal regime and then predict the optimal treatments in the test datasets, where the underlying truth is known. We denote the percentage of subjects correctly classified to their optimal treatments as $opt\%$. We also use the true outcome model and the estimated optimal regime in the test datasets to estimate the counterfactual mean outcome, denoted as $\hat{E}\{Y^*(\hat{g}^{opt})\}$. For both scenarios, we generate five baseline covariates $X_1, \ldots, X_5$ according to $N(0,1)$, and for Scenario 1, we further consider a setting with additional covariates $X_6, \ldots, X_{20}$ simulated independently from $N(0,1)$.

4.3.1 Scenario 1: $T = 1$ and $K = 3$

In Scenario 1, we consider a single stage with three treatment options and sample size of 500. Treatment variables are set to take values in $\{0, 1, 2\}$, and we generate A from $Multinomial(\pi_0, \pi_1, \pi_2)$, with $\pi_0 = 1/\{1+\exp(0.5X_1+0.5X_4)+\exp(-0.5X_1+0.5X_5)\}$, $\pi_1 = \exp(0.5X_1 + 0.5X_4)/\{1 + \exp(0.5X_1 + 0.5X_4) + \exp(-0.5X_1 + 0.5X_5)\}$ and

$\pi_2 = 1 - \pi_0 - \pi_1$. The underlying optimal regime is

$$g^{opt}(\mathbf{H}) = \begin{cases} 0 & X_1 \leq 0, X_2 \leq 0.5 \\ 2 & X_1 > 0, X_3 \leq 0.5 \\ 1 & \text{otherwise} \end{cases}$$

For the outcomes, we first consider equal penalties for sub-optimal treatments through outcome generating model (a), which is

$$Y = 2 + X_4 + X_5 + \sum_{a=0}^{2} \left[I(A = a)\{2I(g^{opt} = a) - 1\} \right] + \epsilon.$$

Then we consider varying penalties for sub-optimal treatments through outcome generating model (b), which is

$$Y = 0.79 + X_4 + X_5 + 2I(A = 0)\{2I(g^{opt} = 0) - 1\} + 1.5I(A = 2)\{2I(g^{opt} = 2) - 1\} + \epsilon.$$

In both outcome models, we have $\epsilon \sim N(0, 1)$ and $E\{Y^*(g^{opt})\} = 2$.

In the application of the proposed T-RL algorithm, we consider both a correctly specified model $log(\pi_d/\pi_0) = \beta_{0d} + \beta_{1d}X_1 + \beta_{2d}X_4 + \beta_{3d}X_5, d = 1, 2$, and an incorrectly specified one $log(\pi_d/\pi_0) = \beta_{0d}$. We also apply T-RL-LH to Scenario 1 as mentioned in Section 4.2.4. For comparison, we use both the regression-based and random forests-based conditional mean models to infer the optimal regimes, which we denote as RG and RF, respectively. We also apply the tree-based method LZ by *Laber and Zhao* (2015). Furthermore, we apply the OWL method by *Zhao et al.* (2012), and the ACWL algorithm by *Tao and Wang* (2016), denoted as ACWL-C_1 and ACWL-C_2, where C_1 and C_2 indicate respectively the minimum and maximum expected loss in the outcome given any sub-optimal treatment for each patient. Given outcome model (a), all sub-optimal treatments have the same expected loss in the outcome and we

Table 4.1: Simulation results for Scenario 1 in Chapter IV with a single stage, three treatment options and five baseline covariates. π is the propensity score model. (a) and (b) indicate equal and varying penalties for treatment misclassification in the generative outcome model. *opt*% shows the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their optimal treatments. $\hat{E}\{Y^*(\hat{g}^{opt})\}$ shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime. $E\{Y^*(g^{opt})\} = 2$, 500 replications, and $n = 500$.

π	Method	(a)		(b)	
		opt%	$\hat{E}\{Y^*(\hat{g}^{opt})\}$	*opt*%	$\hat{E}\{Y^*(\hat{g}^{opt})\}$
-	RG	74.2 (2.3)	1.49 (0.07)	68.8 (4.0)	1.42 (0.09)
	RF	75.3 (4.5)	1.51 (0.11)	81.1 (4.5)	1.69 (0.10)
Correct	OWL	44.3 (7.6)	0.89 (0.16)	47.1 (8.1)	0.89 (0.21)
	LZ	91.5 (7.5)	1.83 (0.16)	89.4 (9.5)	1.81 (0.18)
	ACWL-C_1	93.7 (4.1)	1.87 (0.10)	89.1 (5.3)	1.80 (0.11)
	ACWL-C_2	94.7 (3.3)	1.89 (0.09)	87.8 (5.5)	1.79 (0.11)
	T-RL	97.2 (3.3)	1.95 (0.08)	95.1 (5.6)	1.92 (0.11)
	T-RL-LH	97.5 (3.1)	1.96 (0.08)	96.1 (4.0)	1.94 (0.08)
Incorrect	OWL	33.5 (6.0)	0.67 (0.13)	36.7 (5.7)	0.64 (0.19)
	LZ	87.8 (12.0)	1.75 (0.25)	81.8 (14.7)	1.68 (0.27)
	ACWL-C_1	92.1 (4.7)	1.84 (0.10)	87.9 (5.6)	1.79 (0.11)
	ACWL-C_2	94.7 (3.4)	1.89 (0.09)	86.5 (6.1)	1.78 (0.12)
	T-RL	97.9 (1.8)	1.96 (0.06)	92.9 (7.2)	1.89 (0.13)
	T-RL-LH	98.3 (1.6)	1.97 (0.06)	93.7 (6.2)	1.91 (0.10)

expect ACWL to perform similarly well as T-RL. However, given outcome model (b) when the sub-optimal treatments have different expected losses in the outcome, we expect T-RL to perform better as it incorporates multiple treatment comparison. Both OWL and ACWL are implemented using the R package *rpart* for classification.

Table 4.1 summarizes the performances of all methods considered in Scenario 1 with five baseline covariates. T-RL-LH has the best performance among all the methods considered, classifying over 93% of subjects to their optimal treatments. However, lookahead has led to significant increase in computational time compared to T-RL, while the improvement is only moderate with $\leq 1\%$ more subjects being correctly

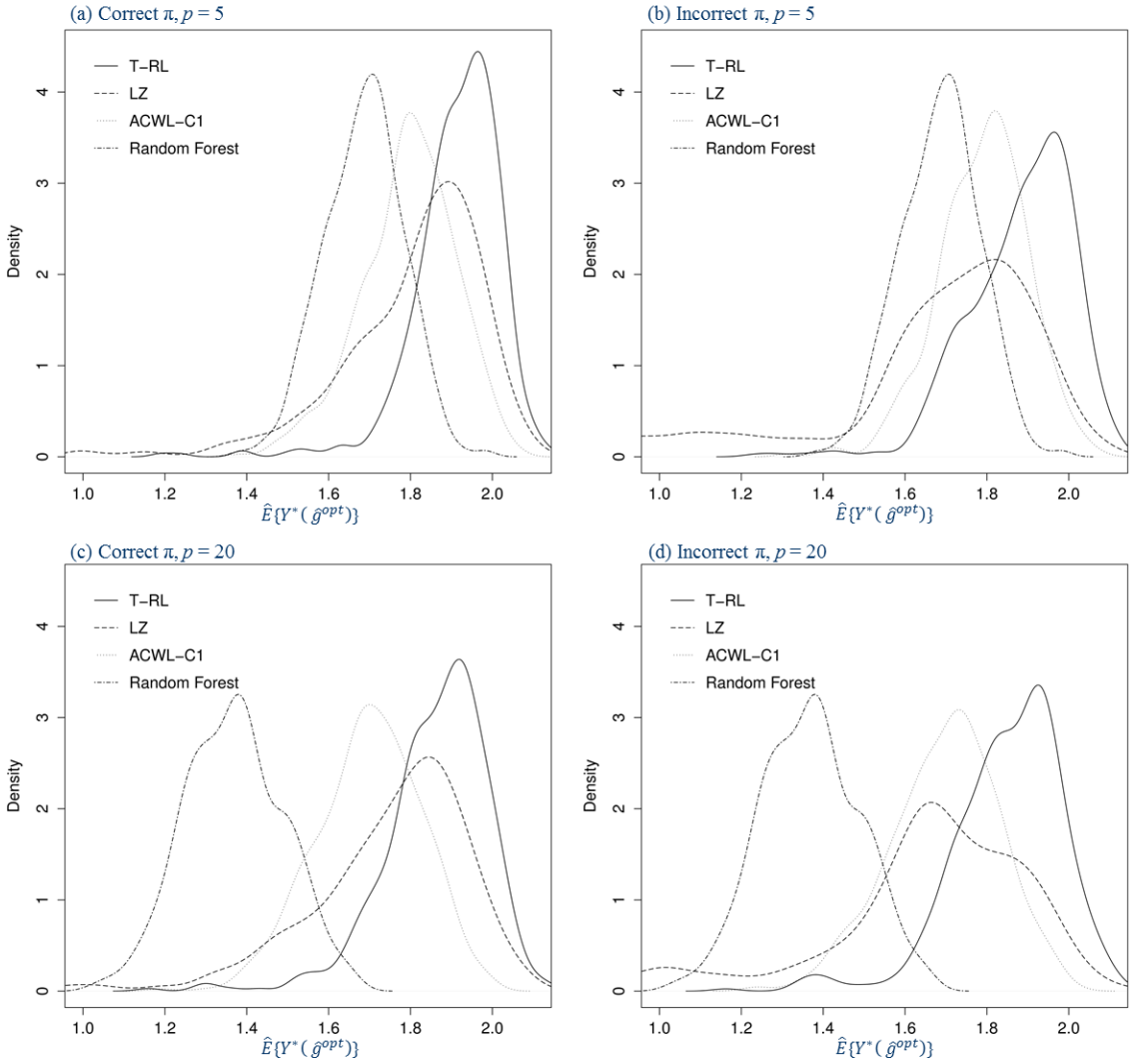

Figure 4.2: Density plots for the estimated counterfactual mean outcome in Scenario 1 of Chapter IV with varying penalties for misclassification in the generative outcome model (500 replications, $n = 500$). The four panels are under correctly or incorrectly specified propensity model and five or twenty baseline covariates.

Table 4.2: Simulation results for Scenario 1 in Chapter IV with a single stage, three treatment options and twenty baseline covariates. π is the propensity score model. (a) and (b) indicate equal and varying penalties for treatment misclassification in the generative outcome model. $opt\%$ shows the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their optimal treatments. $\hat{E}\{Y^*(\hat{g}^{opt})\}$ shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime. $E\{Y^*(g^{opt})\} = 2$, 500 replications, and $n = 500$.

π	Method	(a)		(b)	
		$opt\%$	$\hat{E}\{Y^*(\hat{g}^{opt})\}$	$opt\%$	$\hat{E}\{Y^*(\hat{g}^{opt})\}$
-	RG	66.7 (2.8)	1.34 (0.08)	63.5 (3.4)	1.30 (0.09)
	RF	51.6 (5.7)	1.03 (0.13)	62.7 (5.8)	1.37 (0.12)
Correct	OWL	36.3 (4.2)	0.73 (0.10)	38.4 (5.4)	0.63 (0.17)
	LZ	88.6 (9.4)	1.77 (0.20)	85.5 (0.11)	1.74 (0.21)
	ACWL-C_1	89.6 (5.0)	1.79 (0.11)	83.7 (6.0)	1.70 (0.13)
	ACWL-C_2	90.7 (4.6)	1.82 (0.11)	82.5 (6.2)	1.70 (0.13)
	T-RL	96.3 (4.1)	1.93 (0.10)	91.9 (6.7)	1.86 (0.13)
	T-RL-LH	96.8 (3.9)	1.94 (0.09)	92.8 (5.4)	1.89 (0.10)
Incorrect	OWL	32.6 (4.0)	0.65 (0.10)	34.5 (4.3)	0.56 (0.15)
	LZ	85.9 (12.6)	1.72 (0.26)	78.4 (15.4)	1.62 (0.30)
	ACWL-C_1	87.8 (5.5)	1.76 (0.12)	82.6 (6.3)	1.70 (0.13)
	ACWL-C_2	90.8 (4.3)	1.82 (0.10)	81.7 (6.3)	1.70 (0.13)
	T-RL	97.4 (2.4)	1.95 (0.07)	90.7 (7.7)	1.85 (0.14)
	T-RL-LH	97.9 (2.0)	1.96 (0.07)	92.0 (6.5)	1.87 (0.11)

classified. T-RL also has an estimated counterfactual mean outcome very close to the true value 2. As expected, ACWL-C_1 and ACWL-C_2 have performances comparable to T-RL under outcome model (a) with equal penalties for treatment misclassification, and the performance discrepancy gets larger under outcome model (b) with varying penalties, due to the approximation by adaptive contrasts C_1 and C_2. Similar results can be found in the Appendix B. LZ, using an IPW-based decision tree, works well only when the propensity score model is correctly specified and is less efficient than T-RL with larger empirical standard deviations (SDs). In contrast, T-RL-LH, T-RL, ACWL-C_1 and ACWL-C_2 are all highly robust to model misclassification, thanks to the combination of doubly robust AIPW estimators and flexible machine learning methods. OWL performs far worse than all other competing methods likely due to the low percentage of truly optimal treatments in the observed treatments, the shift in the outcome, which was intended to ensure positive weights, and its moderate efficiency.

After the inclusion of more noise covariates in Table 4.1, all methods have worse performances compared to Table 4.2, with RF suffering the most. T-RL and T-RL-LH have the slightest decreases in $opt\%$ and $\hat{E}\{Y^*(\hat{g}^{opt})\}$, showing satisfactory stability against noise interference. Thanks to the built-in variable selection feature of decision trees, LZ and ACWL with CART are also relatively stable. Figure 4.2 shows the density plots for $\hat{E}\{Y^*(\hat{g}^{opt})\}$ under outcome model (b), with each panel showing correctly or incorrectly specified propensity model and five or twenty baseline covariates. LZ is the least efficient method with the density plots more spread out. T-RL has the least density in lower values of $\hat{E}\{Y^*(\hat{g}^{opt})\}$ and the highest density in higher values.

4.3.2 Scenario 2: $T = 2$ and $K_1 = K_2 = 3$

In Scenario 2, we generate data under a two-stage DTR with three treatment options at each stage and consider sample sizes of 500 and 1000. The outcome of interest is the sum of the rewards from each stage, i.e., $Y = R_1 + R_2$. Furthermore, we consider both a tree-type underlying optimal DTR and a non-tree-type one.

Treatment variables are set to take values in $\{0, 1, 2\}$ at each stage. For stage 1, we generate A_1 from the same model as A in Scenario 1, and generate stage 1 reward as

$$R_1 = \exp[1.5 + 0.3X_4 - |1.5X_1 - 2|\{A_1 - g_1^{opt}(\mathbf{H}_1)\}^2] + \epsilon_1,$$

with tree-type $g_1^{opt}(\mathbf{H}_1) = I(X_1 > -1)\{I(X_2 > -0.5) + I(X_2 > 0.5)\}$ or non-tree-type $g_1^{opt}(\mathbf{H}_1) = I(X_1 > -0.5)\{1 + I(X_1 + X_2 > 0)\}$, and $\epsilon_1 \sim N(0, 1)$.

For stage 2, we have treatment $A_2 \sim Multinomial(\pi_{20}, \pi_{21}, \pi_{22})$, with $\pi_{20} = 1/\{1 + \exp(0.2R_1 - 0.5) + \exp(0.5X_2)\}$, $\pi_{21} = \exp(0.2R_1 - 0.5)/\{1 + \exp(0.2R_1 - 0.5) + \exp(0.5X_2)\}$ and $\pi_{22} = 1 - \pi_{20} - \pi_{21}$. We generate stage 2 reward as

$$R_2 = \exp[1.18 + 0.2X_2 - |1.5X_3 + 2|\{A_2 - g_2^{opt}(\mathbf{H}_2)\}^2] + \epsilon_2,$$

with tree-type $g_2^{opt}(\mathbf{H}_2) = I(X_3 > -1)\{I(R_1 > 0) + I(R_1 > 2)\}$ or non-tree-type $g_2^{opt}(\mathbf{H}_2) = I(X_3 > -0.5)\{1 + I(X_3 + R_1 > 2)\}$, and $\epsilon_2 \sim N(0, 1)$.

We apply the proposed T-RL algorithm with the modified pseudo-outcomes. For comparison, we use the regression-based conditional mean models directly to infer the optimal regimes, which is Q-learning. We also apply the backward OWL (BOWL) method by *Zhao et al.* (2015) and the ACWL algorithm, both of which are implemented using the R package *rpart* for classification. In this scenario, we attempt to see

how sample size and tree- or non-tree-type underlying DTRs affect the performances of various methods.

Results for Scenario 2 are shown in Table 4.3. ACWL and T-RL both work much better than Q-learning and BOWL in all settings. Given a tree-type underlying DTR, T-RL has the best performance among all methods considered with average *opt*% over 90% and $\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$ closest to the truth 8. The results are a bit more complex when the underlying DTR is non-tree-type. The tree-based methods of ACWL with CART and T-RL both have tree-type misspecification and thus show less satisfactory performances. However, ACWL seems more robust to tree-type misspecification with ACWL-C_2 showing larger *opt*% and $\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$ in all settings except when sample size is 500 and π is misspecified, in which case T-RL's stronger robustness to propensity score misspecification dominates. With non-rectangular boundaries in a non-tree-type DTR, a split may not improve the counterfactual mean estimates at the current node but may achieve such a goal in the future nodes. T-RL, with a purity measure based on $E\{Y^*(g)\}$, will terminate the splitting as soon as the best split of the current node fails to improve the counterfactual mean outcome. In contrast, the misclassification error-based impurity measure in CART may continue the recursive partitioning as the best split may still reduce misclassification error without improving the counterfactual mean outcome at the current node. In other words, T-RL may be more myopic when it comes to non-tree-type DTRs. Additional simulation results can be found in the Appendix B, which leads to similar conclusions.

4.4 Illustrative Data Example

As a further illustration, we apply T-RL to the data of 2870 adolescents entering community-based substance abuse treatment programs, which are pooled from sev-

Table 4.3: Simulation results for Scenario 2 in Chapter IV with two stages and three treatment options at each stage. π is the propensity score model. *opt*% shows the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their optimal treatments. $\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$ shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal DTR. $E\{Y^*(\mathbf{g}^{opt})\} = 8$, and 500 replications.

π	Method	Tree-type DTR				Non-tree-type DTR			
		$n = 500$		$n = 1000$		$n = 500$		$n = 1000$	
		opt%	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$	*opt*%	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$	*opt*%	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$	*opt*%	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$
-	Q-learning	53.2 (2.9)	5.80 (0.19)	55.4 (2.5)	5.90 (0.16)	67.5 (4.3)	6.60 (0.21)	70.6 (3.5)	6.72 (0.15)
Correct	BOWL	22.4 (5.1)	4.23 (0.38)	27.9 (5.7)	4.45 (0.43)	25.3 (6.0)	4.53 (0.42)	34.8 (7.0)	4.98 (0.47)
	ACWL-C_1	83.6 (5.9)	7.58 (0.18)	92.1 (4.4)	7.82 (0.11)	80.1 (6.1)	7.40 (0.18)	88.3 (3.4)	7.65 (0.11)
	ACWL-C_2	81.3 (6.6)	7.53 (0.20)	89.1 (5.7)	7.80 (0.12)	83.3 (5.5)	7.51 (0.16)	89.2 (3.0)	7.68 (0.11)
	T-RL	90.5 (7.0)	7.75 (0.22)	95.7 (3.6)	7.88 (0.11)	82.2 (4.6)	7.45 (0.14)	84.7 (2.9)	7.54 (0.11)
Incorrect	BOWL	16.4 (4.3)	4.20 (0.30)	16.5 (4.9)	4.29 (0.32)	16.3 (4.9)	4.29 (0.34)	17.9 (6.0)	4.56 (0.37)
	ACWL-C_1	80.9 (5.9)	7.55 (0.18)	89.1 (4.9)	7.80 (0.11)	73.4 (6.7)	7.30 (0.20)	81.0 (6.3)	7.56 (0.14)
	ACWL-C_2	80.4 (6.7)	7.54 (0.19)	86.4 (6.2)	7.76 (0.13)	79.8 (5.9)	7.46 (0.17)	86.3 (4.2)	7.66 (0.11)
	T-RL	90.2 (7.4)	7.74 (0.17)	93.9 (6.0)	7.87 (0.11)	82.2 (6.3)	7.47 (0.15)	84.8 (3.8)	7.55 (0.11)

eral adolescent treatment studies funded by the Center for Substance Abuse Treatment (CSAT) of the Substance Abuse and Mental Health Services Administration (SAMHSA). The measurements on individual characteristics and functioning are collected at baseline and at the end of 3 and 6 months. We use subscript values $t = 0, 1, 2$ to denote baseline, Month 3, and Month 6 respectively.

Substance abuse treatments were given twice, first during $0 \sim 3$ months, denoted as A_1 and second during $3 \sim 6$ months, denoted as A_2. At both stages, patients were either not treated, or given one of the two types of treatments: non-residential treatment (outpatient only, more freedom to return to patients' own living and work environments after intervention) and residential treatment (i.e., inpatient rehab) (*Marlatt and Donovan*, 2005). We denote the three treatment options as 0, 1 and 2, respectively. At stage 1, 93% of the subjects received treatment, either residential (56%), or non-residential (27%), while at stage 2, only 28% and 13% were treated residentially or non-residentially. The baseline covariate vector that determines the assignment of A_1 is denoted as $\mathbf{X}_1$ and the covariate history just before assigning A_2 is denoted as $\mathbf{X}_1$ (including $\mathbf{X}_0$). The detailed list of variables can be found in *Almirall et al.* (2012). The outcome of interest is the Substance Frequency Scale (SFS) collected during $6 \sim 9$ months (mean and SD: 0.09 and 0.13), with higher values indicating increased frequency of substance use in terms of days used, days staying high most of the day, and days causing problems. We take $Y = -1 \times \text{SFS}$ so that higher values are more desired, making it consistent with our foregoing notation and method derivation. Missing data is imputed using IVEware (*Raghunathan et al.*, 2002).

We apply the T-RL algorithm to the data described above. Specifically, the covariate and treatment history just prior to stage 2 treatment is $\mathbf{H}_2 = (\mathbf{X}_1^\top, A_1)^\top$ and the number of treatment options at stage 2 is $K_2 = 3$. We fit a linear regression model for $\mu_{2,A_2}(\mathbf{H}_2)$ similar to (3.5) using Y as the outcome and all variables in $\mathbf{H}_2$ as predictors

that interact with A_2. For the propensity score, we fit a multinomial logistic regression model including main effects of all variables in $\mathbf{H}_2$. We set the minimal node size to be 50 and maximum tree depth to be 5, and use a 10-fold CV to select λ, the minimum purity improvement for splitting. We repeat the same procedure for stage 1 except that we have $\mathbf{H}_1 = \mathbf{X}_0$, $K_1 = 3$ and $\hat{PO}_1^{'} = Y + \hat{\mu}_{2,\hat{g}_2^{opt}}(\mathbf{H}_2) - \hat{\mu}_{2,a_2}(\mathbf{H}_2)$.

At stage 2, the variables that construct the tree are yearly substance dependence scale measured at the end of Month 3 (sdsy3, median (range): 3 $(0-7)$), age (median (range): 16 $(12-25)$ years), and yearly substance problem scale measured at baseline (spsy0, median (range): 8 $(0-16)$). For 1st stage treatment, the variables involved in the tree building are emotional problem scale measured at baseline (eps7p0, median (range): 0.22 $(0-1)$), drug crime scale measured at baseline (dcs0, median (range): 0 $(0-5)$), and environmental risk scale measured at baseline (ers0, median (range): 35 $(0-77)$). All these scale variables have higher values indicating more risk or problems. The estimated optimal DTR $\hat{\mathbf{g}}^{opt} = c(\hat{g}_1^{opt}, \hat{g}_2^{opt})$ is

$$\hat{g}_2^{opt}(\mathbf{H}_1) = \begin{cases} \text{residential} & \text{if sdsy3 } > 0, \text{ or sdsy3} = 0 \text{ \& age} < 16 \text{ \& spsy0 } > 5 \\ \text{non-residential} & \text{otherwise} \end{cases}$$

and

$$\hat{g}_1^{opt}(\mathbf{H}_2) = \begin{cases} \text{no treatment} & \text{if eps7p0} \leq 0.286 \text{ \& ers0} \leq 46 \\ \text{non-residential} & \text{if eps7p0} \geq 0.286 \text{ \& dcs0} \leq 2 \\ \text{residential} & \text{otherwise.} \end{cases}$$

According to the estimated optimal DTR, all patients should be treated at stage 2. Patients with higher yearly substance dependence as well as those with no yearly substance dependence but younger age and more yearly substance problems should go with residential treatment, i.e., receiving treatment in rehab facilities. In contrast, patients with older age or fewer yearly substance problems should receive the outpatient

treatment. At stage 1, patients with fewer emotional problems and lower environmental risk do not need to be treated, while those with more emotional problems but lower drug crime scale should go with outpatient treatment only. The majority of patients at both stages would benefit the most from residential treatment. In our data, about 70% of the patients at stage 1 have the estimated optimal treatment to be residential treatment and the number goes up to 85% at stage 2. Residential treatment is generally more intensive and patients are in a safe and structured environment, which may explain why patients with more substance, emotional or environmental problems would benefit more from this type of treatment. Existing studies have found a moderate level of evidence for the effectiveness of residential treatment for substance use disorders (*Reif et al.*, 2014). Outpatient programs allow patients to return to their own environments after treatment and require a greater amount of diligence. Patients are provided with a strong support network of non-using peers and sponsors and can automatically apply the lessons learned from outpatient treatment programs to their daily experiences (*Gifford*, 2015). Therefore, more self-disciplined patients with fewer existing problems and less environmental risk would likely benefit more from this type of treatment.

4.5 Discussion

We have developed T-RL, which utilizes a sequence of decision trees with backward induction, to handle multi-stage multi-treatment decision-making. The decision trees are unsupervised and thus maintain the nature of batch-model RL. T-RL enjoys the advantages of typical tree-based methods as being straightforward to understand and interpret, and capable of handling various types of data without distributional assumptions. T-RL can also handle multinomial or ordinal treatments by incorporating multiple treatment comparisons directly in the purity measure for node splitting,

and thus works better than ACWL when the underlying optimal DTR is tree-type. Moreover, T-RL maintains the robust and efficient property of ACWL by virtue of the combination of robust semiparametric regression estimators with flexible machine learning methods, which is superior to IPW-based methods such as LZ. However, when the true optimal DTR is non-tree-type, ACWL has slightly more robust performances.

Several improvements and extensions can be explored in future studies. As shown by the simulation, the fixed-depth lookahead is costly and only brings moderate improvement. Alternatively, one can use embedded models to select splitting variables which also enjoys the lookahead feature (*Zhu et al.*, 2015), or consider other variants of lookahead methods (*Elomaa and Malinen*, 2003; *Esmeir and Markovitch*, 2004). The method by *Zhu et al.* (2015) enables progressively muting noise variables as one goes further down a tree, which facilitates the modeling in high-dimensional sparse settings, and it also incorporates linear combination splitting rules, which may improve the identification of non-tree-type optimal DTRs. Furthermore, it is of great importance to explore how to handle continuous treatment options in the proposed T-RL framework. One way is to follow LZ to use a kernel smoother in the purity measure, which may suffer from the difficulty in selecting the optimal bandwidth. A simpler approach is to discretize the continuous treatments by certain quantiles and consider it as ordinal treatments, which may improve estimation stability and is also of practical interest as medical practitioners tend to prescribe treatments by several fixed levels instead of a continuous fashion.

CHAPTER V

Summary and Future Work

In this dissertation, we have explored statistical methods for the identification of optimal DTRs using observational data when complexities arise in either decision stages or treatment options. We employ classical semiparametric regression methods as well as a combination of semiparametric regression and machine learning methods to achieve flexibilities and robustness in estimation.

The framework for continuous or multiple random decision points proposed in Chapter II is an importation addition to the research on multi-stage and continuous decision-making. It is capable of handling the more practical cases where different subjects may have different treatment schedules. It may have greater potential with the rise of mobile health when more frequent biomarker measurements are available and more timely decisions need to be made. Chapters III and IV address another important problem of having more than two treatment options in multi-stage decision-making. The methodological novelty lies in the combination of semiparametric regression with machine learning, which yields robust and efficient estimates of optimal DTRs. The use of machine learning also relaxes the assumptions about the structure of candidate DTRs and allows the consideration of a large number of covariates. The method ACWL in Chapter III is able to incorporate existing regression and classification methods to handle both tree-type and non-tree-type optimal DTRs while the method

T-RL in Chapter IV focuses more on tree-type ones.

For future research on continuous or multiple random decision points, an important direction is to enhance the estimation of the weights, i.e., the inverse probability of adherence to a given DTR, for example, using the random survival forest (*Ishwaran et al.*, 2008; *Bou-Hamad et al.*, 2011). Furthermore, with either little background knowledge about the structure of candidate DTRs or a larger number of variables that may affect treatment decisions, one may apply a more exploratory approach first with only several fixed decision points discretized from the continuous decision trajectory, for example, using Q- or A-learning with variable selection (*Lu et al.*, 2013). In addition, instead of prespecifying a somewhat arbitrary univariate utility function, one may explore with a multivariate utility function, which requires searching over a multi-dimensional plane to find the optimal DTR that achieves the best joint payoff. Computational complexity and interpretability may be challenging in this case.

An important generalization of both ACWL and T-RL is to explore continuous treatment options such as radiation doses. An example is to use a kernel smoother to consider treatment options within a given bandwidth as in *Laber and Zhao* (2015). For the tree-based method in Chapter IV, another important improvement would be to reduce its greediness by lookahead while keeping the computational cost low, for example, the method by *Zhu et al.* (2015). We have mostly considered batch-model RL, where the sample of data has been fully collected. It is the most common case in medical studies. However, in the field of mobile health, subjects now have medical data that can keep updating. Therefore, it is of interest to extend our methods to be applied to this type of data.

APPENDICES

APPENDIX A

Supplementary Materials for Chapter III

Additional Simulation 1

This simulation follows Scenario 1 in Chapter III but with treatment assignment fully random. Specifically, we have

$$A \sim Multinomial(0.2, 0.2, 0.2, 0.2, 0.2),$$

and

$$Y = \exp[2.06 + 0.2X_3 - |X_1 + X_2|\varphi\{A, g^{opt}(\mathbf{H})\}] + \epsilon,$$

with $\varphi\{A, g^{opt}(\mathbf{H})\}$ taking the form of $\varphi^{(2)} = \{A - g^{opt}(\mathbf{H})\}^2$,

$$g^{opt}(\mathbf{H}) = I(X_1 > -1)\{1 + I(X_2 > -0.4) + I(X_2 > 0.4) + I(X_2 > 1)\}$$

and $\epsilon \sim N(0, 1)$.

The results are shown in Table A.1.

Table A.1: Additional simulation results for Scenario 1 in Chapter III with $\varphi^{(2)}$ and fully randomized treatment assignments. $E\{Y^*(g^{opt})\} = 8$, 500 replications, $n = 1000$.

π	Method	$\varphi^{(2)}$	
		$opt\%$	$\hat{E}\{Y^*(\hat{g}^{opt})\}$
Correct	OWL	75.8 (11.1)	6.91 (0.60)
	ACWL-C_1	89.2 (6.1)	7.63 (0.34)
	ACWL-C_2	87.9 (7.5)	7.39 (0.42)

Additional Simulation 2

This simulation follows Scenario 2 in in Chapter III but with the treatment models dependent on X_1 and X_2, so that the treatment models and the optimal treatment models are more related than Scenario 2. Specifically, we have $A_1 \sim \text{Multinomial}(\pi_{10}, \pi_{11}, \pi_{12})$, with $\pi_{10} = 1/\{1 + \exp(0.5 - 0.5X_1) + \exp(0.5X_2)\}$, $\pi_{11} = \exp(0.5 - 0.5X_1)/\{1 + \exp(0.5-0.5X_1)+\exp(0.5X_2)\}$, and $\pi_{12} = 1-\pi_{10}-\pi_{11}$, and $A_2 \sim \text{Multinomial}(\pi_{20}, \pi_{21}, \pi_{22})$, with $\pi_{20} = 1/\{1+\exp(0.2R_1-1)+\exp(0.5X_2)\}$, $\pi_{21} = \exp(0.2R_1-1)/\{1+\exp(0.2R_1-1) + \exp(0.5X_2)\}$, and $\pi_{22} = 1 - \pi_{20} - \pi_{21}$.

The outcome models are

$$R_1 = \exp[1.5 - |1.5X_1 + 2|\{A_1 - g_1^{opt}(\mathbf{H}_1)\}^2] + \epsilon_1,$$

with $g_1^{opt}(\mathbf{H}_1) = I(X_1 > -1)\{I(X_2 > -0.5) + I(X_2 > 0.5)\}$ and $\epsilon_1 \sim N(0, 1)$, and

$$R_2 = \exp[1.26 - |1.5X_3 - 2|\{A_2 - g_2^{opt}(\mathbf{H}_2)\}^2] + \epsilon_2,$$

with $g_2^{opt}(\mathbf{H}_2) = I(X_3 > -1)\{I(R_1 > 0.5) + I(R_1 > 3)\}$ and $\epsilon_2 \sim N(0, 1)$.

The results are shown in Table A.2.

Table A.2: Additional simulation results based on Scenario 2 in Chapter III, with treatment assignment models more related to optimal treatment models. $E\{Y^*(\mathbf{g}^{opt})\} = 8$, 500 replications, $n = 1000$.

π	Method	Tree-type DTR	
		$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$
-	Q-learning	54.6 (2.9)	6.10 (0.24)
Correct	BOWL	40.3 (8.2)	4.80 (0.53)
	BOWL-Q	66.0 (10.1)	6.57 (0.53)
	ACWL-C_1	92.5 (3.2)	7.50 (0.13)
	ACWL-C_2	92.7 (3.3)	7.54 (0.12)
Incorrect	BOWL	33.1 (7.9)	4.85 (0.48)
	BOWL-Q	41.4 (9.9)	5.48 (0.58)
	ACWL-C_1	91.6 (3.5)	7.48 (0.12)
	ACWL-C_2	90.9 (3.3)	7.47 (0.11)

Additional Simulation 3

This simulation is for a more complex scenario with 2 stages and 5 treatment options at each stage. Specifically, we have

$$A_1 \sim Multinomial(\pi_{10}/\pi_{1s}, \pi_{11}/\pi_{1s}, \pi_{12}/\pi_{1s}, \pi_{13}/\pi_{1s}, \pi_{14}/\pi_{1s}),$$

with $\pi_{10} = 1$, $\pi_{11} = \exp(0.4 - 0.5X_3)$, $\pi_{12} = \exp(0.5X_4)$, $\pi_{13} = \exp(0.5X_3 - 0.4)$, $\pi_{14} = \exp(-0.5X_4)$, and $\pi_{1s} = \sum_{m=0}^{4} \pi_{1m}$, and

$$A_2 \sim Multinomial(\pi_{20}/\pi_{2s}, \pi_{21}/\pi_{2s}, \pi_{22}/\pi_{2s}, \pi_{23}/\pi_{2s}, \pi_{24}/\pi_{2s}),$$

with $\pi_{20} = 1$, $\pi_{21} = \exp(-0.2R_1)$, $\pi_{22} = \exp(0.5X_3 - 0.4)$, $\pi_{23} = \exp(-0.5X_3)$, $\pi_{24} = \exp(0.2R_1 - 1)$, and $\pi_{2s} = \sum_{m=0}^{4} \pi_{2m}$.

Table A.3: Additional simulation results for two stages and five treatment options at each stage. $E\{Y^*(\mathbf{g}^{opt})\} = 8$, 500 replications, $n = 1000$.

π	Method	Tree-type DTR	
		$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$
-	Q-learning	31.7 (3.8)	4.83 (0.32)
Correct	BOWL	15.7 (4.5)	3.53 (0.47)
	BOWL-Q	34.0 (11.3)	4.90 (0.73)
	ACWL-C_1	68.7 (8.7)	6.64 (0.47)
	ACWL-C_2	67.9 (8.7)	6.66 (0.43)
Incorrect	BOWL	9.8 (3.9)	3.04 (0.43)
	BOWL-Q	12.8 (5.9)	3.35 (0.52)
	ACWL-C_1	59.8 (9.9)	6.11 (0.60)
	ACWL-C_2	63.6 (9.2)	6.40 (0.50)

The outcome models are

$$R_1 = \exp[1.5 - |X_1 + X_3|\{A_1 - g_1^{opt}(\mathbf{H}_1)\}^2] + \epsilon_1,$$

with $g_1^{opt}(\mathbf{H}_1) = I(X_1 > -1)\{1 + I(X_4 > -0.4) + I(X_4 > 0.4) + I(X_4 > 1)\}$ and $\epsilon_1 \sim N(0, 1)$, and

$$R_2 = \exp[1.26 - |1.5X_3 - 2|\{A_2 - g_2^{opt}(\mathbf{H}_2)\}^2] + \epsilon_2,$$

with $g_2^{opt}(\mathbf{H}_2) = I(R_1 > 0)\{1 + I(X_3 > -0.4) + I(X_3 > 0.4) + I(X_3 > 1)\}$ and $\epsilon_2 \sim N(0, 1)$.

The results are shown in Table A.3.

APPENDIX B

Supplementary Materials for Chapter IV

Proof of Lemma 1

By the *law of large numbers*, $\mathbb{P}_n\{\hat{\mu}_a^{AIPW}(\mathbf{H})\}$ estimates

$$E\left[\frac{I(A=a)}{\pi_a(\mathbf{H})}Y+\left\{1-\frac{I(A=a)}{\pi_a(\mathbf{H})}\right\}\mu_a(\mathbf{H})\right], \tag{B.1}$$

which is equal to

$$\begin{aligned}&E\left[\frac{I(A=a)}{\pi_a(\mathbf{H})}Y^*(a)+\left\{1-\frac{I(A=a)}{\pi_a(\mathbf{H})}\right\}\mu_a(\mathbf{H})\right]\\ =&E_{\mathbf{H}}\left[\frac{Pr(A=a|\mathbf{H})}{\pi_a(\mathbf{H})}E\{Y^*(a)|\mathbf{H}\}+\left\{1-\frac{Pr(A=a|\mathbf{H})}{\pi_a(\mathbf{H})}\right\}\mu_a(\mathbf{H})\right],\end{aligned}$$

under the foregoing causal assumptions.

If the postulated propensity score model $\pi_a(\mathbf{H})$ is correct, i.e., $\pi_a(\mathbf{H})=Pr(A=a|\mathbf{H})$, then (B.1) $=E_{\mathbf{H}}\left[E\{Y^*(a)|\mathbf{H}\}\right]=E\{Y^*(a)\}$.

If the conditional mean model $\mu_a(\mathbf{H})$ is correctly specified, then under the foregoing

causal assumptions, $E\{Y^*(a)|\mathbf{H}\} = E\{Y^*(a)|A=a,\mathbf{H}\} = E(Y|A=a,\mathbf{H}) = \mu_a(\mathbf{H})$. Therefore,

$$\begin{aligned}(1) &= E_{\mathbf{H}}\left\{\frac{Pr(A=a|\mathbf{H})}{\pi_a(\mathbf{H})}[E\{Y^*(a)|\mathbf{H}\} - \mu_a(\mathbf{H})] + \mu_a(\mathbf{H})\right\} \\ &= E_{\mathbf{H}}\{\mu_a(\mathbf{H})\} \\ &= E\{Y^*(a)\}.\end{aligned}$$

Additional Simulation 1

This simulation follows Scenario 1 in Chapter III. Specifically, we have treatment A from $Multinomial(\pi_0/\pi_s, \pi_1/\pi_s, \pi_2/\pi_s, \pi_3/\pi_s, \pi_4/\pi_s)$, with $\pi_0 = 1$, $\pi_1 = \exp(0.5 - 0.5X_1)$, $\pi_2 = \exp(0.5X_1 + 0.2)$, $\pi_3 = \exp(0.5X_5 + 0.1)$, $\pi_4 = \exp(0.5X_5 - 0.1)$, and $\pi_s = \sum_{m=0}^{4}\pi_m$. We set A to take values in $\{0,\ldots,4\}$ and generate outcomes as

$$Y = \exp[2.06 + 0.2X_3 - |X_1 + X_2|\varphi\{A, g^{opt}(\mathbf{H})\}] + \epsilon,$$

with $\varphi\{A, g^{opt}(\mathbf{H})\}$ taking the form of $\varphi^{(1)} = 3I\{A \neq g^{opt}(\mathbf{H})\}$ or $\varphi^{(2)} = \{A - g^{opt}(\mathbf{H})\}^2$, $g^{opt}(\mathbf{H}) = I(X_1 > -1)\{1 + I(X_2 > -0.4) + I(X_2 > 0.4) + I(X_2 > 1)\}$ and $\epsilon \sim N(0,1)$.

The results are shown in Table B.1.

Table B.1: Simulation results for a single stage and five treatment options. π is the propensity score model. $\varphi^{(1)}$ and $\varphi^{(2)}$ indicate equal and varying penalties for misclassification. *opt*% shows the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their optimal treatments. $\hat{E}\{Y^*(\hat{g}^{opt})\}$ shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime. $E\{Y^*(g^{opt})\} = 8$, 500 replications, $n = 1000$.

π	Method	$\varphi^{(1)}$		$\varphi^{(2)}$	
		opt%	$\hat{E}\{Y^*(\hat{g}^{opt})\}$	*opt*%	$\hat{E}\{Y^*(\hat{g}^{opt})\}$
Correct	ACWL-C_1	94.2 (3.5)	7.69 (0.21)	88.7 (5.5)	7.60 (0.22)
	ACWL-C_2	90.4 (6.1)	7.38 (0.40)	86.4 (8.4)	7.36 (0.38)
	T-RL	95.2 (3.1)	7.74 (0.20)	92.9 (3.7)	7.72 (0.18)
Incorrect	ACWL-C_1	92.5 (4.1)	7.60 (0.23)	84.2 (6.7)	7.47 (0.24)
	ACWL-C_2	90.2 (6.0)	7.37 (0.38)	85.6 (8.2)	7.35 (0.36)
	T-RL	95.2 (2.8)	7.74 (0.17)	91.0 (4.3)	7.68 (0.16)

Additional Simulation 2

This simulation follows Scenario 1 in Chapter IV with five baseline covariates, the same treatment model and the same optimal treatment model but different outcome model. The outcome model indicates arbitrary penalties for misclassification, which is

$$Y = \exp[1.5 + 0.3X_4 - |1.5X_1 - 1|I(A \neq g^{opt})\{4I(A = 0) + I(A = 1) + 2I(A = 2)\}] + \epsilon,$$

with $\epsilon \sim N(0, 1)$.

The results are shown in Table B.2.

Table B.2: Additional simulation results based on Scenario 1 in Chapter IV with five baseline covariates and outcome model indicating arbitrary penalties for misclassification. $E\{Y^*(g^{opt})\} = 4.69$, 500 replications, $n = 500$.

π	Method	Tree-type	
		$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$
-	RG	69.7 (3.3)	3.71 (0.11)
Correct	OWL	63.3 (10.1)	3.54 (0.37)
	LZ	95.2 (6.5)	4.54 (0.19)
	ACWL-C_1	90.6 (4.7)	4.49 (0.12)
	ACWL-C_2	90.4 (5.3)	4.47 (0.13)
	T-RL	96.0 (5.1)	4.58 (0.14)
Incorrect	OWL	48.6 (8.0)	3.05 (0.34)
	LZ	84.4 (17.9)	4.24 (0.51)
	ACWL-C_1	88.2 (4.1)	4.46 (0.12)
	ACWL-C_2	88.5 (4.9)	4.46 (0.13)
	T-RL	96.0 (7.8)	4.58 (0.21)

Additional Simulation 3

This simulation follows Scenario 1 in Chapter IV with five baseline covariates, the same treatment model and the same outcome model (b) (i.e., varying penalties for treatment misclassification) but different optimal treatment model, which has a non-tree-type

$$g^{opt}(\mathbf{H}) = I(X_1 > 0) + I(X_1 + X_2 > 0).$$

The results are shown in Table B.3.

Table B.3: Additional simulation results based on Scenario 1 in Chapter IV with five baseline covariates, outcome model (b) and non-tree-type optimal treatment regime. $E\{Y^*(g^{opt})\} = 2$, 500 replications, $n = 500$.

π	Method	Non-tree-type	
		$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$
-	RG	75.5 (3.5)	1.67 (0.09)
Correct	OWL	46.4 (7.6)	0.98 (0.21)
	LZ	78.6 (6.9)	1.72 (0.13)
	ACWL-C_1	81.5 (4.7)	1.76 (0.11)
	ACWL-C_2	83.0 (4.8)	1.81 (0.10)
	T-RL	82.1 (4.3)	1.79 (0.10)
Incorrect	OWL	35.1 (5.7)	0.71 (0.19)
	LZ	75.2 (9.5)	1.67 (0.51)
	ACWL-C_1	81.4 (4.9)	1.77 (0.11)
	ACWL-C_2	82.0 (5.1)	1.80 (0.10)
	T-RL	81.1 (4.9)	1.78 (0.10)

BIBLIOGRAPHY

BIBLIOGRAPHY

Almirall, D., D. F. McCaffrey, B. A. Griffin, R. Ramchand, R. A. Yuen, and S. A. Murphy (2012), Examining moderated effects of additional adolescent substance use treatment: Structural nested mean model estimation using inverse-weighted regression-with-residuals, *Tech. Rep. 12-121*, Penn State University, University Park, PA.

American Diabetes Association (2014), Standards of medical care in diabetes, *Diabetes Care*, *37*(Suppl 1), S14–S80.

Bather, J. (2000), *Decision Theory: An Introduction to Dynamic Programming and Sequential Decisions*, Wiley, New York, NY.

Bou-Hamad, I., D. Larocque, and H. Ben-Ameur (2011), Discrete-time survival trees and forests with time-varying covariates application to bankruptcy data, *Statistical Modelling*, *11*(5), 429–446.

Breiman, L. (2001), Random forests, *Machine Learning*, *45*(1), 5–32.

Breiman, L., J. H. Freidman, R. A. Olshen, and C. J. Stone (1984), *Classification and Regression Trees*, Wadsworth, Belmont, CA.

Breslow, N. E. (1972), Comment on regression and life tables by d.r. cox, *Journal of the Royal Statistical Society, Series B*, *34*, 216–217.

Breslow, N. E. (1974), Covariance analysis of censored survival data, *Biometrics*, *30*(1), 89–99.

Buja, A., T. Hastie, and R. Tibshirani (1989), Linear smoothers and additive models, *The Annals of Statistics*, *17*(2), 453–510.

Byrd, D. R., C. C. Compton, A. G. Fritz, F. L. Greene, and A. Trotti (2010), *AJCC Cancer Staging Manual (Vol. 649)*, Springer, New York, NY.

Chakraborty, B., and E. E. Moodie (2013), *Statistical Methods for Dynamic Treatment Regimes*, Springer, New York, NY.

Chakraborty, B., and S. Murhpy (2014), Dynamic treatment regimes, *Annual Review of Statistics and Its Application*, *1*, 447–464.

Cortes, C., and V. Vapnik (1995), Support-vector netowrks, *Machine Learning*, *20*(3), 273–297.

De Boor, C. (1978), *A Practical Guide to Splines*, Springer, New York, NY.

Elomaa, T., and T. Malinen (2003), On lookahead heuristics in decision tree learning, in *International Symposium on Methodologies for Intelligent Systems, Lecture Notes in Artificial Intelligence 2871*, pp. 445–453, Springer, Heidelberg.

Ernst, D., P. Geurts, and L. Wehenkel (2005), Tree-based batch mode reinforcement learning, *Journal of Machine Learning Research*, *6*, 503–556.

Esmeir, S., and S. Markovitch (2004), Lookahead-based algorithms for anytime induction of decision trees, in *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 257–264, ACM, New York, NY.

Free, C., G. Phillips, L. Watson, L. Galli, L. Felix, P. Edwards, V. Patel, and A. Haines (2013), The effectiveness of mobile-health technologies to improve health care service delivery processes: a systematic review and meta-analysis, *PLoS medicine*, *10*(1), e1001,363.

Gatzoulis, L., and I. Iakovidis (2007), Wearable and portable ehealth systems, *Engineering in Medicine and Biology Magazine, IEEE*, *26*(5), 51–56.

Gerstein, H. C., et al. (2008), Effects of intensive glucose lowering in type 2 diabetes, *New England Journal of Medicine*, *358*(24), 2545–2559.

Gifford, S. (2015), Difference between outpatient and inpatient treatment programs, *Psych Central.*, retrieved on July 6, 2016, from http://psychcentral.com/lib/differences-between-outpatient-and-inpatient-treatment-programs.

Goldberg, Y., and M. R. Kosorok (2012), Q-learning with censored data, *Annals of Statistics*, *40*(1), 529–560.

Goodall, G., E. M. Sarpong, C. Hayes, and W. J. Valentine (2009), The consequences of delaying insulin initiation in uk type 2 diabetes patients failing oral hyperglycaemic agents: a modelling study, *BMC Endocrine Disorders*, *9*(1), 1–9.

Gray, R. J. (1992), Fexible methods for analyzing survival data using splines, with applications to breast cancer prognosis, *Journal of the American Statistical Association*, *87*(420), 942–951.

Haddad, R., et al. (2013), Induction chemotherapy followed by concurrent chemoradiotherapy (sequential chemoradiotherapy) versus concurrent chemoradiotherapy alone in locally advanced head and neck cancer (PARADIGM): a randomised phase 3 trial, *The Lancet Oncology*, *14*(3), 257–264.

Hayward, R. A., W. G. Manning, S. H. Kaplan, E. H. Wagner, and S. Greenfield (1997), Starting insulin therapy in patients with type 2 diabetes: effectiveness, complications, and resource utilization, *Journal of the American Medical Association*, *278*(20), 1663–1669.

Henderson, R., P. Ansell, and D. Alshibani (2010), Regret-regression for optimal dynamic treatment regimes, *Biometrics*, *66*(4), 1192–1201.

Hernán, M. A., B. Brumback, and J. M. Robins (2001), Marginal structural models to estimate the joint causal effect of nonrandomized treatments, *Journal of the American Statistical Association*, *96*(454), 440–448.

Huang, X., and J. Ning (2012), Analysis of multi-stage treatments for recurrent diseases, *Statistics in Medicine*, *31*(24), 2805–2821.

Huang, X., S. Choi, L. Wang, and P. F. Thall (2015), Optimization of multi-stage dynamic treatment regimes utilizing accumulated data, *Statistics in Medicine*, *34*(26), 3423–3443.

Hurvich, C. M., J. S. Simonoff, and C. Tsai (1998), Smoothing parameter selection in nonparametric regression using an improved akaike information criterion, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(2), 271–293.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer (2008), Random survival forests, *The Annals of Applied Statistics*, *2*(3), 841–860.

Jin, J., et al. (2004), Induction chemotherapy improved outcomes of patients with resectable esophageal cancer who received chemoradiotherapy followed by surgery, *International Journal of Radiation Oncology*Biology*Physics*, *60*(2), 427–436.

Johnson, B. A., and A. A. Tsiatis (2005), Semiparametric inference in observational duration-response studies, with duration possibly right-censored, *Biometrika*, *92*(3), 605–618.

Laber, E. B., and Y. Zhao (2015), Tree-based methods for individualized treatment regimes, *Biometrika*, *102*(3), 501–514.

Lloyd, S., and B. W. Chang (2014), Current strategies in chemoradiation for esophageal cancer, *Journal of Gastrointestinal Oncology*, *5*(3), 156–165.

Lok, J. J. (2008), Statistical modeling of causal effects in continuous time, *Annals of Statistics*, *36*(3), 1464–1507.

Lok, J. J., and V. DeGruttola (2012), Impact of time to start treatment following infection with application to initiating haart in hiv-positive patients, *Biometrics*, *68*(3), 745–754.

Lu, W., H. H. Zhang, and D. Zeng (2013), Variable selection for optimal treatment decision, *Statistical Methods in Medical Research*, *22*(5), 493–504.

Marlatt, G. A., and D. M. Donovan (2005), *Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors*, Guilford Press, New York, NY.

Menard, S. (2002), *Applied Logistic Regression Analysis*, 2nd ed., Sage, Thousand Oaks, CA.

Moodie, E. E., B. Chakraborty, and M. S. Kramer (2012), Q-learning for estimating optimal dynamic treatment rules from observational data, *Canadian Journal of Statistics*, *40*(4), 629–645.

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

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

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

Murthy, S., and S. Salzberg (1995), Lookahead and pathology in decision tree induction, in *Proceedings of Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1025–1031, San Francisco, CA: Morgan Kaufmann.

Orellana, L., A. Rotnitzky, and J. M. Robins (2010a), Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content, *The International Journal of Biostatistics*, *6*(2), 8.

Orellana, L., A. Rotnitzky, and J. M. Robins (2010b), Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part ii: Proofs of results, *The International Journal of Biostatistics*, *6*(2), 8.

Pantelopoulos, A., and N. G. Bourbakis (2010), A survey on wearable sensor-based systems for health monitoring and prognosis, *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE*, *40*(1), 1–12.

Patel, A., et al. (2008), Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes, *New England Journal of Medicine*, *358*(24), 2560–2572.

Raghunathan, T. E., P. Solenberger, and J. Van Hoewyk (2002), *IVEware: Imputation and Variance Estimation Software User Guide*, Survey Methodology Program, University of Michigan, Ann Arbor, Michigan.

Reif, S., P. George, L. Braude, R. H. Dougherty, A. S. Daniels, S. S. Ghose, and M. E. Delphin-Rittmon (2014), Residential treatment for individuals with substance use disorders: assessing the evidence, *Psychiatric Services*, *65*(3), 301–312.

Robins, J. (1986), A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect, *Mathematical Modelling*, *7*(9), 1393–1512.

Robins, J. M. (1989), The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies, *Health service research methodology: a focus on AIDS*, *113*, 159.

Robins, J. M. (1997), Causal inference from complex longitudinal data, in *Latent Variable Modeling and Applications to Causality*, pp. 69–117, New York: Springer.

Robins, J. M. (2000), Marginal structural models versus structural nested models as tools for causal inference, in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 95–133, New York: Springer.

Robins, J. M. (2004), Optimal structural nested models for optimal sequential decisions, in *Proceedings of the Second Seattle Symposium in Biostatistics*, pp. 189–326, New York: Springer.

Robins, J. M., and M. A. Hernán (2009), Estimation of the causal effects of time-varying exposures, in *Longitudinal Data Analysis*, pp. 553–599, Chapman and Hall/CRC Press: Boca Raton.

Robins, J. M., A. Rotnitzky, and L. Zhao (1994), Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, *89*(427), 846–866.

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

Rotnitzky, A., J. Robins, and D. Scharfstein (1998), Semiparametric regression for repeated outcomes with nonignorable nonresponse, *Journal of the American Statistical Association*, *93*(444), 1321–1339.

Scharfstein, D., A. Rotnitzky, and J. Robins (1999), Adjusting for nonignorable drop-out using semiparametric nonresponse models, *Journal of the American Statistical Association*, *94*(448), 1096–1120.

Schulte, P. J., A. A. Tsiatis, E. B. Laber, and M. Davidian (2014), Q-and a-learning methods for estimating optimal dynamic treatment regimes, *Statistical Science*, *29*(4), 640–661.

Sun, J., D. H. Park, L. Sun, and X. Zhao (2005), Semiparametric regression analysis of longitudinal data with informative observation times, *Journal of the American Statistical Association*, *100*(471), 882–889.

Sun, J., L. Sun, and D. Liu (2007), Regression analysis of longitudinal data in the presence of informative observation and censoring times, *Journal of the American Statistical Association*, *102*(480), 1397–1406.

Sutton, R., and A. Barto (1998), *Reinforcement Learning: An Introduction*, MIT Press, Cambridge.

Tao, Y., and L. Wang (2016), Adaptive contrast weighted learning for multi-stage multi-treatment decision-making, *Biometrics*, in press.

Thall, P. F., L. H. Wooten, C. J. Logothetis, R. E. Millikan, and N. M. Tannir (2007), Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring, *Statistics in Medicine*, *26*(26), 4687–4707.

Tsiatis, A. A. (1981), A large sample study of coxs regression model, *Annals of Statistics*, *9*, 93–108.

Turner, R. C., C. A. Cull, V. Frighi, R. R. Holman, and UK Prospective Diabetes Study (UKPDS) Group (1999), Glycemic control with diet, sulfonylurea, metformin, or insulin in patients with type 2 diabetes mellitus: progressive requirement for multiple therapies (UKPDS 49), *Journal of the American Medical Association*, *281*(21), 2005–2012.

van der Laan, M. J., and D. Rubin (2006), Targeted maximum likelihood learning, *The International Journal of Biostatistics*, *2*(1), 1–40.

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

Wang, F., L. Wang, and P. X. Song (2016), Fused lasso with the adaptation of parameter ordering (FLAPO) in merging multiple studies with repeated measurements, *Biometrics*, doi:10.1111/biom.12496.

Wang, L., A. Rotnitzky, X. Lin, R. E. Millikan, and P. F. Thall (2012), Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer, *Journal of the American Statistical Association*, *107*(498), 493–508.

Watkins, C. J., and P. Dayan (1992), Q-learning, *Machine Learning*, *8*(3-4), 279–292.

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

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

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

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012), Estimating individualized treatment rules using outcome weighted learning, *Journal of the American Statistical Association*, *107*(499), 1106–1118.

Zhao, Y., D. Zeng, E. B. Laber, and M. R. Kosorok (2015), New statistical learning methods for estimating optimal dynamic treatment regimes, *Journal of the American Statistical Association*, *110*(510), 583–598.

Zhou, X., N. Mayer-Hamblett, U. Khan, and M. R. Kosorok (2015), Residual weighted learning for estimating individualized treatment rules, *Journal of the American Statistical Association*, doi:10.1080/01621459.2015.1093947.

Zhu, R., D. Zeng, and M. R. Kosorok (2015), Reinforcement learning trees, *Journal of the American Statistical Association*, *110*(512), 1770–1784.