Novel Statistical Methods for Restricted Mean Survival Time and Patient Preference Augmented Dynamic Treatment Regimes in Observational Studies

by

Yingchao Zhong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2020

Doctoral Committee:

Professor Douglas E. Schaubel, Co-Chair
Professor Lu Wang, Co-Chair
Professor Charles Francis Burant
Assistant Professor Zhenke Wu

Yingchao Zhong

zhongych@umich.edu

ORCID iD: 0000-0002-1878-3352

To my family.

ACKNOWLEDGEMENTS

This dissertation is the culmination of five years of scholarship in the Biostatistics department. As such, thanks are due to select individuals who made the past five years an extremely formative period of my life. The content and writing of this thesis took place during a challenging year filled with fears and uncertainty surrounding Covid-19, Black Lives Matters movements and protests, shifting national and international relationships, and numerous climate crises. It would not be exaggerating to say that this thesis could not have stayed its course without incredible encouragement and direction from my mentors, family, and friends.

First and foremost, I would like to thank my Co-chairs and advisors, Lu Wang and Douglas Schaubel, both of whom have taught me much in both scholarship and life. I'm very lucky to have been mentored by two amazing statisticians and to have learned from two different approaches to research and scholarship.

I began working with Doug soon after joining the Biostats department. Over the past few years it has been such a joy to have worked on both methodology and applied projects with him. Doug is an extremely patient mentor who has deep insights in his fields of expertise. I am in awe of his ability to think outside the box and come up with creative statistical solutions. I have learned so much through from him through the publication process. I am inspired by Doug and hope to have this kind of domain influence in my future career.

I worked with Lu for the past few years on both dynamic treatment regime projects of this dissertation. I found Lu to be an extremely affectionate professor who has been incredibly generous with her time and advice. I'm grateful for her unconditional support towards my goals, and her guidance and encouragement have been paramount in shaping my knowledge and practices as a contributing scholar. In Lu I am so grateful to have found an excellent role model for both my career and personal life.

I am also grateful to Zhenke Wu for his support over the past few years. Zhenke is a very approachable and generous faculty member. Through Zhenke's introduction, I was able to meet current Biostatisticians at Genentech and potential collaborators here at Michigan. His broad knowledge base and expertise in mobile health has also helped shaped this thesis with his thoughtful suggestions and insights.

I would also like to thank Dr. Charles Burant for his practical inputs and considerations into this dissertation. Some of the examples of patient trade-offs in projects 2 and 3 were directly generated from our conversations. I have also always appreciated Dr. Burant's jolly presence at many Taubman events and his candid perspectives regarding the physician scientist experience.

Other members of the Biostatistics community were also crucial in ensuring my smooth experience in the department. Special thanks goes to Nicole Fenech, who so often had to find workarounds because my MSTP status made many logistical things like course registration more challenging. Also many thanks to Mike Boehnke, who gifted me a summer experience in Biostatistics and encouraged me to pursue Biostatistics for graduate study. Thanks are also due to Kevin He and Sehee Kim, both of whom taught me much during their times as my GSRA advisors.

I am also indebted to the Medical Scientist Training Program here at Michigan, which has always been like a second home. Thank you to all the staff Justine, Liz,

Gretchen, Ellen, Hilkka, and Laurie for making the program such a welcoming and smooth running place. Dr. Ron Koenig has been a wonderful listener and so generous with his time and advice that I have never felt alone with the challenges that have come up over the years.

Last but not least, I would like to thank my family, Chonghua Zhong, Yufeng Shen, Mingjie Zhong, and Huai-Hsun Lien for their never ending love and support.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

In this dissertation, we develop three new statistical methods and estimating procedures in survival analysis with restricted mean survival time and in evaluating the optimal treatment decision rules by involving patient preference.

Restricted mean survival time (RMST) is a clinically interpretable and meaningful survival metric defined as the patient's mean survival time up to a pre-specified time horizon of interest, denoted as L. No existing RMST regression method allows for the covariate effects to be expressed as functions over time, which is a considerable limitation in light of the many hazard regression models that do accommodate such effects. To address this void in the literature, in the first project of my dissertation, we propose an inference framework for directly modeling RMST as a continuous function of L. We apply our method to kidney transplant data obtained from the Scientific Registry of Transplant Recipients (SRTR).

The second and third projects of my dissertation consider personalized treatment decision strategies in the management of chronic diseases, such as end stage renal disease, which typically consists of sequential and adaptive treatment decision making. This can be formulated through a dynamic treatment regime (DTR) framework, where the goal is to tailor treatment to each individual given their medical history in order to maximize a desirable health outcome. We develop a new method, Augmented Patient Preference incorporated Reinforcement Learning (APP-RL), to incorporate a patient's latent preferences through data augmentation into a tree-based reinforcement learning method to estimate optimal dynamic treatment regimes for multi-stage, multi-treatment settings. For each patient at each stage, we derive their posterior distribution of preferences given responses to a questionnaire, and then subsequently weight multiple outcomes with the estimated preferences to identify the optimal stage-wise personalized decision. APP-RL is robust, efficient, and leads to interpretable DTR estimation.

We further extend the APP-RL ideas into the survival setting with censored data in the last project. We investigate a two-stage treatment setting where patients have to decide between quality of life and survival restricted at maximal follow-up. We successfully develop a method that incorporates the latent patient preference into a weighted utility function that balances between quality of life and survival time, in a Q-learning model framework. We further propose a corresponding m-out-of-n Bootstrap procedure to accurately make statistical inferences and construct confidence intervals on the effects of tailoring variables, whose values can guide the personalized treatment strategies.

CHAPTER I

Introduction

Statistical formulation of decision making requires two critical components: a precisely defined and estimable outcome, and the determination of decision making rules that result in said desired outcome. We illustrate this decision making trajectory using chronic kidney disease (CKD) as an example. Management of chronic progressive diseases such as CKD is often a titration exercise requiring multiple sequential visits and management decisions that tailor to an individual's history, current status, and personal values and preferences. This requires an complex interplay between the physician, the standard of care knowledge, and the patient and his/her willingness to engage in the proposed treatment plan in order to obtain the most optimal care. If the patient's disease progresses, decisions must be made with respect to a new selection of treatment options. Finally, when the patient reaches end stage renal disease (ESRD), dialysis and transplantation become the main treatment options, and patients commonly want to know expected outcomes based on their current situation. The goal of this dissertation is to follow the patient through such a chronic disease trajectory course, providing both methodological advancements in determining the decision rules for managing the chronic disease, as well as methodological improvements for estimating a final outcome.

A commonly used outcome of interest in the transplantation is the restricted mean survival time (RMST). A clinically meaningful and interpretable metric defined as the mean survival time up to a fixed time, L, the RMST can be thought of as a $L-$year life expectancy. Originally proposed in 1949 by Irwin (*Irwin*, 1949), the RMST was originally meant as a substitute for the overall mean, which can be difficult to observe due to a long-tail of the survival distribution and a limited follow-up time. In time, however, the RMST has gained intellectual and practical interest for being a meaningful measure of its own right, especially as certain chronic diseases have important milestones (e.g. 5-year cancer free survival). Despite the dominance of the cox proportional hazards model (*Cox*, 1972) for doing analysis with covariate adjustment and using the hazard ratio (HR) as the summary measure, the RMST provides an attractive alternative, especially when the proportional hazards (PH) assumption is violated (*Struthers and Kalbfleisch* (1986), *Wei and Schaubel* (2008), *Schaubel and Wei* (2011), *Royston and Parmar* (2011)). In addition to its practical interpretation and its avoidance of the PH assumption, *Royston and Parmar* (2013), *Tian et al.* (2017), and *Huang and Kuan* (2018) found that under PH scenarios, RMST-based tests and log-rank tests perform similarly, while RMST-based tests perform better under non-PH scenarios, making it a highly efficient alternative to the HR.

In recent years, many methods have been proposed to directly estimate RMST (*Andersen et al.* (2004), *Tian et al.* (2014),*Wang and Schaubel* (2018)). However, no current method in the literature allows for covariate effects to be expressed as functions over time, a considerable limitation in light of the many hazard regression methods that do accommodate such effects. In Chapter II, we propose an inference framework for directly modeling of the RMST as a continuous function of L. We apply our method to kidney transplant data obtained from the Scientific Registry of Transplant Recipients (SRTR).

In Chapters III and IV, we turn our attention from deriving an efficient and robust estimator to deriving the decision making regimen required for managing a chronic disease, for both continuous and survival outcome contexts. The mathematical formulation of this problem is the main goal of the dynamic treatment regime (DTR) research community (*Chakraborty and Moodie* (2013), *Chakraborty and Murphy* (2014), *Murphy* (2003)). In recent years, a large number of methods have been proposed on how to find the optimal dynamic regime, including Q- and A-learning (*Schulte et al.*, 2014), G-estimation of structural nested mean models (*Robins*, 2004), Bayesian likelihood based methods (*Thall et al.*, 2007), and machine learning based methods (*Laber and Zhao* (2015), *Tao et al.* (2018), *Tao and Wang* (2017), *Zhang et al.* (2018)).

Despite the abundance of flavor and approaches, most of these approaches optimize a single desired outcome. In reality, clinical decisions, especially ones that address chronic illnesses that stretch out across time horizons, are multidimensional and often affect a whole cohort of outcomes, often in conflicting directions of desirability. One example can be found in hand surgery – patients often have to weigh the benefits and risks of taking a certain medication. Taking stronger pain medication might provide quicker and stronger pain relief, but might have higher risks of addiction and side effects. Finding the right balance often requires patient input, which reflects on patient values and needs. Despite recognition by the medical community of the importance of patient input(*Barry and Edgman-Levitan* (2012), *Basu and Meltzer* (2007)), the shared-decision making movement lacks a systematic optimization approach and is difficult to implement. In Chapter III, we develop a method where we model patient preference as a latent variable, which we then estimate it through an item response approach (*Embretson and Reise*, 2013). Our method incorporates a patient's latent preferences through data augmentation into a tree-based reinforcement learning method to estimate optimal dynamic treatment regimes for multi-stage, multi-treatment settings. For each patient at each stage, we derive their posterior distribution of preferences given responses to a questionnaire, and then subsequently weight multiple outcomes with the estimated preferences to identify the optimal stagewise personalized decision. Our proposed method, named Augmented Patient Preference incorporated Reinforcement Learning (APP-RL) is robust, efficient, and leads to interpretable DTR estimation.

In Chapter IV, we extend the APP-RL ideas in Chapter III into the survival setting of Chapter I. We look at the two-stage setting where patients have to decide between quality of life and survival length (*Torrance and Feeny*, 1989). Patients often receive a first treatment and is followed up after a short time to determine if the treatment needs to be adjusted. The patient is then followed until death or a certain maximal follow-up time. The presence of censored data complicates our scenario as compared to Chapter III. In this case, as in Chapter III, our primary goal is to estimate the optimal treatment regime that would maximize a patient preference weighted combination of quality of life and survival time. Our secondary objective is to provide an inference framework for more confident decision making, focusing on tailoring variables.

Inference in DTR methods is a known challenge with current active research. Due to the nonsmoothness caused by maximization when going through backward induction, the asymptotic distribution of the true coefficient oscillates between two asymptotic distributions, thus resulting in asymptotic bias and poor Wald-type confidence intervals. Even bootstrap-type estimators rely on smoothness and are also affected (*Chakraborty et al.*, 2010). Multiple approaches have been proposed to deal with the nonsmoothness, with varying degrees of adjustment. *Chakraborty et al.* (2010) proposed hard threshold and soft threshold estimators, while (*Laber et al.*, 2014) proposed an adaptive confidence interval for the first stage parameters by utilizing regular, uniformly convergent lower and upper bounds for the asymptotic distribution of interest, and later bootstrapping for the confidence set. *Shao* (1994) proposed an m-out-of-n bootstrap approach to adjust for general nonsmoothness, which was adapted to the DTR setting by (*Chakraborty et al.*, 2013). Our proposed method in Chapter IV is an amalgamation of the ideas of the Chapter III with specific improvements towards censoring and inference. We show that our method is promising in simulation studies through evaluating metrics like the accuracy rate of predicting the optimal treatment and inference coverage probabilities and measures of confidence.

CHAPTER II

Restricted Mean Survival Time as a Function of Restriction Time

2.1 Introduction

For time to event data with right censoring, the proportional hazards model (*Cox*, 1972) has long been the default for doing analysis with covariate adjustment. The principal summary measure that results from Cox regression is the hazard ratio (HR), which is routinely used to quantify between-group differences. This line of analysis relies on proportional hazards (PH), which is the assumption that the ratio of the two hazards are constant over time. Although the approach is convenient to implement, the PH assumption is frequently violated, leading to difficulties with interpretation (*Struthers and Kalbfleisch*, 1986; *Wei and Schaubel*, 2008).

A number of authors have advocated for using summary statistics beyond the hazard ratio in both clinical trial and observational data analyses, especially when the proportional hazards assumption has been called into doubt (*Royston and Parmar*, 2011; *Schaubel and Wei*, 2011; *Royston and Parmar*, 2013; *Uno et al.*, 2014; *Uno et al.*, 2015). In particular, the restricted mean survival time (RMST) has been suggested. Defined as the mean survival time up to a prespecified time horizon of interest, L, in a given population, the RMST can simply be thought of as a L-year life expectancy. Mathematically, it is written as the area under the survival curve up to time L (see Figure 2.1 for a schematic). First proposed in *Irwin* (1949), RMST was initially meant as a substitute for the overall mean, for settings where the presence of censoring prevented the estimation of the latter. More recently, it has come to be known as an interesting measure in its own right. Simulation studies have compared RMST treatment effect estimation and statistical power with HR-based tests both under proportional hazards and non-proportional hazards scenarios. *Royston and Parmar* (2013), *Tian et al.* (2017), and *Huang and Kuan* (2018) found that under PH scenarios, RMST-based and log-rank tests perform similarly (with a slight advantage for the log-rank test), while RMST performs better in non-PH scenarios. Hence, RMST is a clinically relevant and interpretable measure that does not depend on the PH assumption and requires little sacrifice in statistical power even when the PH assumption holds.

Most existing methods estimate RMST indirectly by integrating under an estimate of the survival curve. *Irwin* (1949) used the actuarial estimator for the survival probability and approximated the area under the curve using numerical quadrature methods. More recent methods that extend those of *Irwin* (1949) by incorporating covariates tend to proceed initially through hazard regression. *Karrison* (1987) introduced covariate adjustment for the RMST using a piece-wise exponential hazard model, assuming covariates affect the hazard in a multiplicative manner just as in the Cox model, subsequently obtaining the piecewise cumulative hazard, survival probability curve, and the restricted mean. *Zucker* (1998) followed a similar protocol, using a stratified Cox model instead. Even more recent approaches still require 4-5 sequential steps to obtain the restricted mean: estimate the regression parameter (e.g., through a Cox model); estimate the cumulative baseline hazard; transforming the subject-specific cumulative hazard, then integrate it to obtain the restricted mean

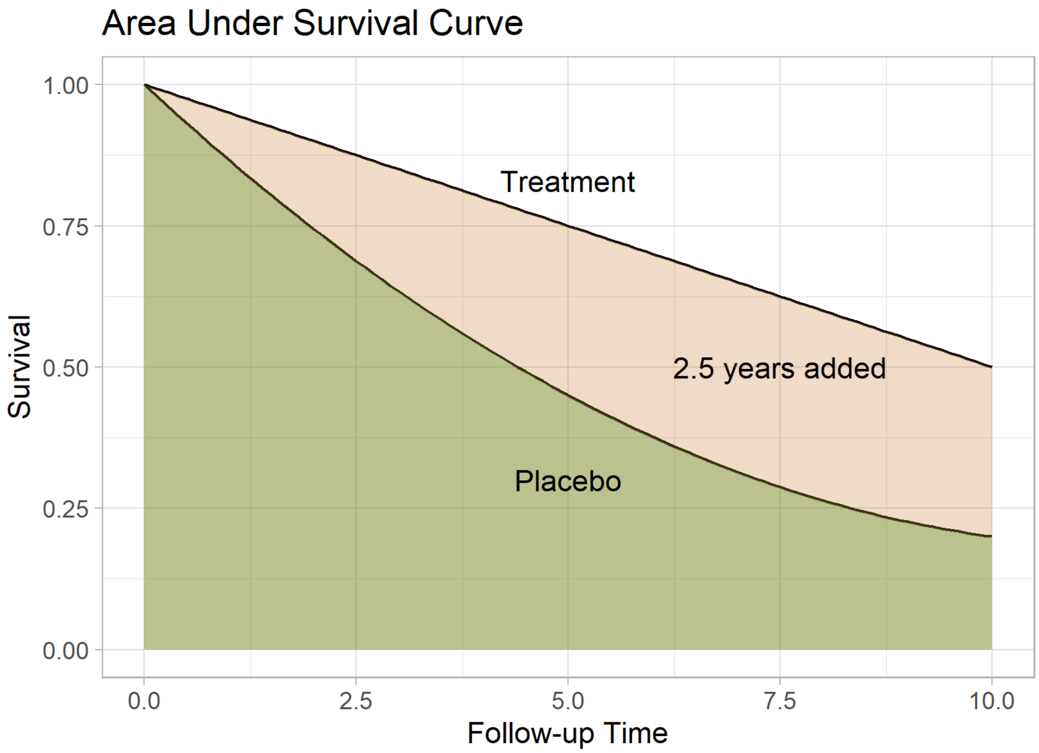

Figure 2.1: Schematic of two RMST curves

(*Chen and Tsiatis*, 2001; *Zhang and Schaubel*, 2011). This process is cumbersome and computationally expensive in large data sets, especially to obtain asymptotic standard errors. Furthermore, through the use of Cox model, this process also relies on the proportional hazards assumption, which, if untrue, can also lead to bias, inefficient estimation, and a challenging interpretation.

Hence, several authors have suggested to directly model the RMST itself. *Andersen et al.* (2004) and *Andersen and Pohar Perme* (2010) used imputation based on pseudo-observations to model the RMST directly using generalized linear models. *Tian et al.* (2014) employed a different but similarly direct approach by constructing estimating equations for RMST based on Inverse Probability of Censoring Weighting (*Robins and Rotnitzky*, 1992; *Robins*, 1993; *Robins and Finkelstein*, 2000), similar to the approach of Zhao et al. for quality adjusted life (*Zhao and Tsiatis*, 1997; *Zhao and Tsiatis*, 1999). *Wang and Schaubel* (2018) employed a similar modeling strategy, but further extended the method to accommodate dependent censoring.

To the best of our knowledge, no existing regression methods have been proposed for modeling RMST as a continuous function of the restriction time, *L. Zhao et al.* 2016 make a strong case for by looking at the entire RMST curve, in order to obtain a complete temporal picture, much like the survival function. We extend this concept to the regression setting, which has two important analytic implications. First, through our proposed approach, one can obtain RMST predictions for various restriction times through a single model. Second and much more importantly, models fitted though our proposed methods yield time-varying covariate effects. The second property is essential for RMST regression to be on more equal footing with hazard regression, since the latter is currently the strong default analysis when time-varying covariate effects are an objective.

The remainder of this report is organized as follows. In Section 2.2, we describe the proposed methods, formulating the notation, data structure, and list out the assumptions made. In Section 2.3, we present the derived asymptotic properties. In Section 2.4, we present results from simulation studies to evaluate the accuracy of the proposed methods. In Section 2.5, we apply the method to the Scientific Registry of Transplant Recipients (SRTR) kidney transplant data, illustrating the use of our method. We conclude this report in Section 2.6 with a discussion. Asymptotic derivations are provided in the Supplementary Materials.

2.2 Method for Estimation of RMST as a Function of L

Let D_i be the survival time for subject i, where $i = 1, \ldots, n$. Let C_i be the censoring time, assumed to be independent of D_i conditional on the baseline covariates. The observation time for subject i is $X_i = D_i \wedge C_i$, where $a \wedge b = \min\{a, b\}$. The at-risk indicator is denoted by $R_i(t) = I(X_i \geq t)$, and the event and censoring indicators are $\Delta_i^D = I(D_i \leq C_i)$ and $\Delta_i^C = I(C_i < D_i)$, respectively. We denote covariates predicting D_i and C_i by $\boldsymbol{Z}_i^D$ and $\boldsymbol{Z}_i^C$, respectively. Stacking these covariates and removing redundancy, we obtain $\boldsymbol{Z}_i$. Our observed data are then given by $\{X_i, \Delta_i^D, \Delta_i^C, \boldsymbol{Z}_i : i = 1, \ldots, n\}$.

Let $\tau = \max\{X_i : i = 1, \ldots, n\}$ be the end of follow-up time, and L_{max} be a prespecified maximal value of L after which estimation becomes potentially unstable and of little interest. Naturally, it is required that $\tau \geq L_{max}$. Let $\boldsymbol{L}$ be a vector of length K where $\boldsymbol{L} = (L_1, L_2, \ldots, L_K)'$ values sorted in ascending order. For a particular element of $\boldsymbol{L}$, say L_k, the restricted observation time is $Y_{ik} = X_i \wedge L_k$, and the corresponding observed-event indicator is $\Delta_{ik} = I(D_i \wedge L_k \leq C_i)$. Note that Δ_{ik} is analogous to a complete-case indicator, taking the value 1 if subject i either dies before $(C_i \wedge L_k)$ or lives (and remains uncensored) past L_k.

In general, for any arbitrary value of L, we are interested in the average survival time up to L, modeled through an individual's covariates:

$$\mu_i(L) := E\left\{D_i \wedge L | \boldsymbol{Z}_i^D\right\}.$$

As in *Wang and Schaubel* (2018), we assume the same direct relationship between the RMST and the baseline covariates. However, in addition, we assume that the covariate effects vary as a function of L in the following equation:

$$g\left[\mu_i(L)\right] \equiv g\left[E\left\{D_i \wedge L | \boldsymbol{Z}_i^D\right\}\right] = \boldsymbol{\beta}_D'(L)\boldsymbol{Z}_i^D, \tag{2.1}$$

where g is a strictly monotone link function with a continuous derivative within an open neighborhood $\boldsymbol{\mathcal{B}}_D(L)$ of $\boldsymbol{\beta}_D(L)$. Some conventional examples of $g(x)$ could be the identity link, log link, or logistic link. Without any adjustments, Equation (2.1) is an infinite dimensional problem and would generally be inconvenient to estimate. Instead, we address the problem by assuming that $\boldsymbol{\beta}_D(L)$, a vector of continuous and monotonic functions, is able to be parametrically modeled as a function of L. For example, denote this parametric model of L as $\boldsymbol{\beta}_D(L) = \boldsymbol{\alpha}_0 L_0(L) + \ldots + \boldsymbol{\alpha}_m L_m(L)$, where $L_0(L), L_1(L), \ldots, L_m(L)$ are functions of L, i.e. parametric or spline functions. Let $\boldsymbol{Z}_i = (1, Z_{i1}, \ldots, Z_{ip})'$ and $\boldsymbol{L}(L) = (L_0(L), L_1(L), \ldots, L_m(L))'$. Then we can re-express the covariate vector as follows:

$$\widetilde{\boldsymbol{Z}}_i^D(L) = \boldsymbol{Z}_i \otimes \boldsymbol{L}(L)$$

where $\otimes$ denotes the Kronecker product. Correspondingly, let $\boldsymbol{\alpha}_0 = (\alpha_{00}, \ldots, \alpha_{0m})', \ldots, \boldsymbol{\alpha}_p =$

$(\alpha_{p0}, \ldots, \alpha_{pm})'$, such that the new parameter vector can be written as

$$\widetilde{\boldsymbol{\beta}}_D = \begin{bmatrix} \boldsymbol{\alpha}_0 \\ \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_p \end{bmatrix}.$$

Hence, we can rewrite Equation (2.1) as:

$$g\left[\mu_i(L)\right] \equiv g\left[E\left\{D_i \wedge L | \boldsymbol{Z}_i^D\right\}\right] = \boldsymbol{\beta}'_D(L)\boldsymbol{Z}_i^D = \widetilde{\boldsymbol{\beta}}'_D \widetilde{\boldsymbol{Z}}_i^D(L). \tag{2.2}$$

This parametrization in effect reduces an infinite dimensional problem to a finite dimensional one, thereby making it more convenient to estimate the regression parameter. The specific parametrization of $\beta_k(L)$ requires careful consideration and should be supported by graphical evidence. For relationships that do not seem to be simply linear, the authors recommend fitting a spline as an initial choice. The knots of the spline should be pre-selected and evenly span across the represented data to ensure a comprehensive fit. Further exploration of this issue is given in Section 6.

Based on Equation (2.2), in the absence of censoring, we can derive the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \widetilde{\boldsymbol{Z}}_i^D(L_k)[Y_{ik} - g^{-1}\{\widetilde{\boldsymbol{\beta}}'\widetilde{\boldsymbol{Z}}_i^D(L_k)\}] = \mathbf{0}. \tag{2.3}$$

In effect, this is a stacked version of the estimating equation presented in *Wang and Schaubel* (2018), where each new iteration of the data (for each value of L_k) is stacked to make a complete vector of responses. Each individual is now represented in the data set K times through its relationship with individual $L_k s$. The complete response vector is then used to fit a model that incorporates each L_k as part of the covariate information. The fitted model is a generalized estimating equation (GEE). To retain flexibility and robustness, we utilize a working independence correlation structure for each individual.

As in most survival data, we are unlikely to observe D_i for all patients due to censoring. In this report, we will focus on independent censoring and make the standard assumption that $C_i \perp D_i | \boldsymbol{Z}_i$. We further assume that the hazard for censoring time C_i follows a proportional hazards model (*Cox*, 1972),

$$\lambda_i^C(t) = \lambda_0^C(t) \exp(\boldsymbol{\beta}_C' \boldsymbol{Z}_i^C). \tag{2.4}$$

Then, each subject-specific cumulative hazards is given by $\Lambda_i^C(t) = \int_0^t \lambda_i^C(u) du$ for $i = 1, \ldots, n$. In the presence of censoring, $E(\widetilde{\boldsymbol{Z}}_i^D(L_k)[Y_{ik} - g^{-1}\{\widetilde{\boldsymbol{\beta}}' \widetilde{\boldsymbol{Z}}_i^D(L_k)\}]) \neq \boldsymbol{0}$, but we can show that the IPCW weighted expectation $E(\widetilde{\boldsymbol{Z}}_i^D(L_k)\Delta_{ik} W_i^C(Y_{ik})[Y_{ik} - g^{-1}\{\widetilde{\boldsymbol{\beta}}' \widetilde{\boldsymbol{Z}}_i^D(L_k)\}]) = \boldsymbol{0}$, where $W_i^C(t) = \exp\{\Lambda_i^C(t)\}$.

We then present the following estimating equation, proven in the Appendix to be unbiased for $\widetilde{\boldsymbol{\beta}}_D'$:

$$\boldsymbol{\Phi}^*(\widetilde{\boldsymbol{\beta}}) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \widetilde{\boldsymbol{Z}}_i^D(L_k) \Delta_{ik} W_i^C(Y_{ik}) [Y_{ik} - g^{-1}\{\widetilde{\boldsymbol{\beta}}' \widetilde{\boldsymbol{Z}}_i^D(L_k)\}] = \boldsymbol{0}. \tag{2.5}$$

Because the cumulative censoring hazard is usually not known in real data settings, the following empirical estimating equation substitutes for $\Lambda_i^C(t)$ using the standard partial likelihood (*Cox*, 1975) and Breslow-Aalen (*Breslow*, 1972) estimator,

$$\boldsymbol{\Phi}(\widetilde{\boldsymbol{\beta}}) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \widetilde{\boldsymbol{Z}}_i^D(L_k) \Delta_{ik} \widehat{W}_i^C(Y_{ik}) [Y_{ik} - g^{-1}\{\widetilde{\boldsymbol{\beta}}' \widetilde{\boldsymbol{Z}}_i^D(L_k)\}] = \boldsymbol{0}. \tag{2.6}$$

The solution to Equation (2.6) is shown to provide for consistent estimation of $\widetilde{\boldsymbol{\beta}}_D$, and asymptotic properties are discussed in Section 2.3.

2.3 Asymptotic Properties

We specify the following regularity conditions (1)-(7):

1. $\{X_i, \Delta_i^D, \Delta_i^C, \boldsymbol{Z}_i\}, i = 1, 2, \dots, n$ are independently and identically distributed.

2. $P\{R_i(t) = 1\} > 0 \; for \; t \in (0, \tau), i = 1, \dots, n$

3. $|Z_{ik}| < M_Z < \infty$ for $i = 1, \dots, n$, where Z_{ik} is the k^{th} component of $\boldsymbol{Z}_i$

4. $\Lambda_i^C(\tau) < \infty$ and $\Lambda_i^C(t)$ is absolutely continuous for $t \in (0, \tau]$.

5. There exist neighborhoods $\boldsymbol{\mathcal{B}}_C$ of $\boldsymbol{\beta}_C$ such that for $k = 0, 1, 2$,

$$\sup_{t \in (0,\tau],\ \boldsymbol{\beta} \in \boldsymbol{\beta}_C} \left\| \frac{1}{n} \sum_{i=1}^{n} \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i^C) R_i(t) \boldsymbol{Z}_i^{C \otimes k} - \boldsymbol{r}_C^{(k)}(t; \boldsymbol{\beta}) \right\| \xrightarrow{p} 0,$$

   where $\boldsymbol{v}^{\otimes 0} = 1, \boldsymbol{v}^{\otimes 1} = \boldsymbol{v}, \boldsymbol{v}^{\otimes 2} = \boldsymbol{v}'\boldsymbol{v}$, and

$$\boldsymbol{r}_C^{(k)}(t; \boldsymbol{\beta}) = E[\exp(\boldsymbol{\beta}' \boldsymbol{Z}_i^C) R_i(t) \boldsymbol{Z}_i^{C \otimes k}].$$

6. Define $h(x) = \partial g^{-1}(x)/\partial x$, where h exists and is continuous in an open neighborhood $\boldsymbol{\mathcal{B}}_D(L)$ of $\widetilde{\boldsymbol{\beta}}_D(L)$.

7. Matrices $\boldsymbol{A}(\widetilde{\boldsymbol{\beta}}_D), \boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)$ are both positive definite, and are defined below:

$$\boldsymbol{A}(\widetilde{\boldsymbol{\beta}}_D) = E\left[\sum_{k=1}^{K} \widetilde{\boldsymbol{Z}}_i^D(L_k)^{\otimes 2} h\{\widetilde{\boldsymbol{\beta}}_D' \widetilde{\boldsymbol{Z}}_i^D(L_k)\}\right]$$

$$\boldsymbol{\Omega}_C(\boldsymbol{\beta}_C) = E\left[\int_0^\tau \left\{ \frac{\boldsymbol{r}_C^{(2)}(t; \boldsymbol{\beta}_C)}{r_C^{(0)}(t; \boldsymbol{\beta}_C)} - \bar{\boldsymbol{z}}_C(t; \boldsymbol{\beta}_C)^{\otimes 2} \right\} dN_i^C(t) \right],$$

where

$$\bar{\boldsymbol{z}}_C(t;\boldsymbol{\beta}) = \frac{\boldsymbol{r}_C^{(1)}(t;\boldsymbol{\beta})}{r_C^{(0)}(t;\boldsymbol{\beta})}.$$

Condition (1) could be relaxed, but additional technical developments would be needed to compensate. Condition (2) is required for identifiability. Conditions (3) - (6) are required for the convergence of stochastic integrals in several proofs. In (7), matrices $\boldsymbol{A}(\widetilde{\boldsymbol{\beta}}_D), \boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)$ are at least non-negative definite and will be positive-definite provided the covariate vectors are specified sensibly.

The main asymptotic results are presented below, in Theorems (2.1) and (2.2). The proofs are presented in the supplementary appendix.

Theorem 2.1. *Under regularity conditions (1)-(7), as* $n \to \infty$, $\sqrt{n}\boldsymbol{\Phi}(\widetilde{\boldsymbol{\beta}}_D)$ *converges in distribution to* $Normal(\boldsymbol{0}, \boldsymbol{B}(\widetilde{\boldsymbol{\beta}}_D))$, *where*
$\boldsymbol{B}_i(\widetilde{\boldsymbol{\beta}}) = \sum_{k=1}^{K}\{\boldsymbol{\epsilon}_{ik}(\widetilde{\boldsymbol{\beta}}) + \boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)^{-1}\boldsymbol{U}_i^C(\boldsymbol{\beta}_C)\boldsymbol{K}_C(\widetilde{\boldsymbol{\beta}})\}$ *and* $\boldsymbol{B}(\widetilde{\boldsymbol{\beta}}) \equiv E\{\boldsymbol{B}_i(\widetilde{\boldsymbol{\beta}})^{\otimes 2}\}$, *where we define:*

$$\begin{aligned}
\boldsymbol{\epsilon}_{ik}(\widetilde{\boldsymbol{\beta}}_D) &= \widetilde{\boldsymbol{Z}}_i^D(L_k)\Delta_{ik}W_i^C(Y_{ik})[Y_{ik} - g^{-1}\{\widetilde{\boldsymbol{\beta}}'_D\widetilde{\boldsymbol{Z}}_i^D(L_k)\}] \\
\boldsymbol{U}_i^C(\boldsymbol{\beta}_C) &= \int_0^\tau \{\boldsymbol{Z}_i^C - \bar{\boldsymbol{z}}_C(u;\boldsymbol{\beta}_C)\}dM_i^C(u) \\
\boldsymbol{D}_i^C(t) &= \int_0^t \{\boldsymbol{Z}_i^C - \bar{\boldsymbol{z}}_C(u;\boldsymbol{\beta}_C)\}d\Lambda_i^C(u) \\
\boldsymbol{K}_C(\widetilde{\boldsymbol{\beta}}) &\equiv E[\boldsymbol{\epsilon}_{ik}(\widetilde{\boldsymbol{\beta}})\boldsymbol{D}_i^C(Y_{ik})'].
\end{aligned}$$

Proof of Theorem 2.1 uses results presented in *Zhang and Schaubel* (2011). Mainly, we borrow techniques for expressing the asymptotic empirical weight in terms of the true weight for independent censoring times. Theorem 2.1 sets the stage for the next theorem.

Theorem 2.2. *Under regularity conditions (1)-(7), as $n \to \infty$, $\widehat{\boldsymbol{\beta}}_D$ converges in probability to $\widetilde{\boldsymbol{\beta}}_D$, and $\sqrt{n}(\widehat{\boldsymbol{\beta}}_D - \widetilde{\boldsymbol{\beta}}_D)$ converges in distribution to $Normal(\mathbf{0}, \boldsymbol{A}(\widetilde{\boldsymbol{\beta}}_D)^{-1}\boldsymbol{B}(\widetilde{\boldsymbol{\beta}}_D)\boldsymbol{A}(\widetilde{\boldsymbol{\beta}}_D)^{-1})$.*

The proof of consistency follows from the use of the Inverse Function Theorem (*Foutz*, 1977). The asymptotic normality and variance follows from combination of Theorem 2.1 and a sequence of Taylor expansions.

We propose a variance estimator that is computationally more convenient than that derived in Theorem 2. Specifically, the weight function is treated as known, such the middle matrix involves only $\boldsymbol{\epsilon}_{ik}(\boldsymbol{\beta})$, which implies the following variance estimator,

$$\widehat{V}(\widehat{\boldsymbol{\beta}}_D) \quad = \quad \widehat{\boldsymbol{A}}(\widetilde{\boldsymbol{\beta}}_D)^{-1}\widehat{\boldsymbol{B}}^{\boldsymbol{*}}(\widetilde{\boldsymbol{\beta}}_D)\widehat{\boldsymbol{A}}(\widetilde{\boldsymbol{\beta}}_D)^{-1}, \tag{2.7}$$

where $\widehat{\boldsymbol{B}}^{*}(\boldsymbol{\beta}) = \widehat{E}\{(\sum_{k=1}^{K} \boldsymbol{\epsilon}_{ik}(\boldsymbol{\beta}))^{\otimes 2}\}$. Treating the IPCW weights as fixed has a long history, dating back at least to the works of Robins et al. (2000). Moreover, Wang and Schaubel (2018) demonstrated through simulation that there was no practical difference between standard errors that treated the weights as fixed versus random. The asymptotic standard error (ASE) estimator given in Equation (2.7) will be used in Sections 2.4 and 2.5. Computationally, Equation (2.7) can be quickly computed with built-in commands in standard software (e.g., R, SAS), using any function that can handle weighted GEE data structures.

2.4 Simulation Study

For each subject, $i = 1, \ldots, n$, we first generated a baseline covariate with two elements, $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2})'$, with each element generated from a Unif(-1,1) distribution.

The death time, D_i, was then generated from an exponential distribution with

$$E[D_i|\boldsymbol{Z}_i] \;=\; g^{-1}(\alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2}). \tag{2.8}$$

Parameter settings were chosen to cover a wide variety of realistic scenarios. For $g(x) = x$, we set $\boldsymbol{\alpha} = [4, 2.5, -2.5]'$ for the 'strong' covariate effect scenario, and set $\boldsymbol{\alpha} = [4, 0.75, -0.75]'$ to represent weaker covariate effects. Note that Cox regression under the 'strong' scenario yields hazard ratios of $HR_1 \approx 0.45$ and $HR_1 \approx 2.15$ for Z_{i1} and Z_{i2}, respectively; the 'weak' covariate setting lines up with $HR_1 \approx 0.80$ and $HR_2 \approx 1.20$. For $g(x) = \log(x)$, we set $\boldsymbol{\alpha} = [1.25, \log(2), -\log(2)]'$ and $\boldsymbol{\alpha} = [1.25, \log(1.25), -\log(1.25)]'$ for the strong and weak covariate effect scenarios, respectively. For the log link, the strong setting yields hazard ratios $HR_1 \approx 0.5$ and $HR_2 \approx 2.0$, while the weak setting corresponds to $HR_1 \approx 0.80$ and $HR_2 \approx 1.25$. Although we did not directly generate the restricted mean survival time ($D_i \wedge L$), we can induce its relationship with the two covariates through Monte Carlo methods (with population size 10 million for each configuration).

With respect to censoring, we examined scenarios with low (15% censored), moderate (30%) and high (45%) proportion censored. Independent censoring time, C_i, was generated from the following hazard,

$$\lambda_i^C(t) = \lambda_0^C \exp(\beta_{C1} Z_{i1} + \beta_{C2} Z_{i2}).$$

For all settings, $\beta_{C1} = \log(1.5)$ and $\beta_{C2} = -\log(1.5)$. We varied λ_0^C in order to generate the desired percent censored; censoring parameters are given in the table captions.

We present the results for sample size $n = 1000$, under low, moderate and high censoring scenarios. For each setting, we generated 1000 iterations. In Tables 2.1 and

2.2, we present results for the strong covariate setting for the linear and log links, respectively. For illustrative purposes, we will select $L = \{5, 7.5, 10\}$. Tables 2.1 and 2.2 contain the true values, bias, empirical standard deviation (ESD), the asymptotic standard error (ASE), and empirical coverage probabilities (CP) corresponding to the asymptotic 95% confidence intervals.

The general conclusion from Tables 2.1 and 2.2 is that, in moderate samples, the proposed estimator is approximately unbiased. Furthermore, the ESDs matched the ASEs very closely, supporting the accuracy of the derivations, and that treating the inverse probability censoring weights as known is adequate for maintaining estimation accuracy of the standard errors. The empirical coverage probabilities are similarly very close to the nominal level.

Figure 2.2 displays plots comparing $\widehat{\boldsymbol{\beta}}(L)$ with $\boldsymbol{\beta}(L)$ from the 30% censoring scenario shown in Table 2.1. The proposed estimator is quite accurate across all L values plotted, as evidenced by the fact that 'estimated' and 'true' lines are practically indistinguishable.

Additional simulation results are provided in the Appendix section of this chapter. In particular, we show results for weak covariate effects in Tables 2.3 and 2.4. Results are very similar those afore-described for Tables 2.1 and 2.2. We show results for smaller sample sizes ($n = 500$ and $n = 250$) in Tables 2.5 and 2.6. Results are acceptable, although residual bias is greater than that shown in Tables 2.1 and 2.2, and CP is a bit lower, as one would expect. In Tables 2.7 and 2.8, we compare the efficiency of the proposed methods with that of *Wang and Schaubel* (2018); efficiency is shown to be approximately equal for the two approaches. Finally, we evaluated the impact of increasing the number of k values (i.e., the number of stacked data sets) in Tables 2.9 and 2.10. It appears that slight gains in efficiency can be achieved by increasing K.

Table 2.1: Simulation results: linear link, strong covariate effect. Data were generated using $\boldsymbol{\beta}_D = [4, 2.5, -2.5]$. True $\boldsymbol{\beta}_D$ are given by $[2.621, 1.006, -1.006]$ for $L = 5$, $[3.140, 1.440, -1.440]$ for $L = 7.5$, and $[3.453, 1.753, -1.756]$ for $L = 10$. For low censoring (15%), λ_0^C=0.025, $\boldsymbol{\beta}_C = [-\log(1.5), \log(1.5)]$. For moderate censoring (30%), λ_0^C=0.1, $\boldsymbol{\beta}_C = [-\log(1.5), \log(1.5)]$. For high censoring (45%), $\lambda_0^C = 0.225$, $\boldsymbol{\beta}_C = [-\log(1.5), \log(1.5)]$.

L	Censor %	Parameter	BIAS	ESD	ASE	CP
5	15	β_0	-0.001	0.053	0.055	0.960
		β_1	-0.004	0.098	0.095	0.944
		β_2	-0.009	0.100	0.095	0.940
	30	β_0	-0.003	0.056	0.062	0.976
		β_1	0.004	0.101	0.106	0.958
		β_2	0.003	0.107	0.106	0.942
	45	β_0	-0.002	0.064	0.079	0.982
		β_1	-0.004	0.121	0.131	0.964
		β_2	-0.001	0.123	0.132	0.956
7.5	15	β_0	-0.005	0.073	0.078	0.964
		β_1	-0.011	0.138	0.136	0.944
		β_2	-0.008	0.141	0.136	0.958
	30	β_0	-0.009	0.081	0.094	0.974
		β_1	-0.003	0.145	0.160	0.964
		β_2	0.009	0.160	0.160	0.944
	45	β_0	-0.009	0.099	0.133	0.988
		β_1	-0.017	0.194	0.220	0.974
		β_2	0.003	0.195	0.220	0.968
10	15	β_0	-0.005	0.089	0.097	0.968
		β_1	-0.009	0.170	0.171	0.950
		β_2	-0.009	0.173	0.170	0.964
	30	β_0	-0.011	0.104	0.124	0.970
		β_1	-0.005	0.197	0.213	0.978
		β_2	0.008	0.210	0.211	0.954
	45	β_0	-0.016	0.132	0.197	0.992
		β_1	-0.037	0.290	0.323	0.982
		β_2	-0.011	0.287	0.323	0.964

Table 2.2: Simulation results: log link, strong covariate effect. Data were generated using $\boldsymbol{\beta}_D = [1.25, \log(2), -\log(2)]$. True $\boldsymbol{\beta}_D$ are given by $[0.923, 0.359, -0.359]$ for $L = 5$, $[1.074, 0.451, -0.451]$ for $L = 7.5$, and $[1.148, 0.515, -0.515]$ for $L = 10$. For low censoring (15%), $\lambda_0^C = 0.025$, $\boldsymbol{\beta}_C = [-\log(1.5), \log(1.5)]$. For moderate censoring (30%), $\lambda_0^C = 0.1$, $\boldsymbol{\beta}_C = [-\log(1.5), \log(1.5)]$. For high censoring (45%), $\lambda_0^C = 0.225$, $\boldsymbol{\beta}_C = [-\log(1.5), \log(1.5)]$.

L	Censor %	Parameter	BIAS	ESD	ASE	CP
5	15	β_0	-0.001	0.022	0.023	0.946
		β_1	0.001	0.035	0.036	0.956
		β_2	-0.004	0.034	0.036	0.964
	30	β_0	-0.001	0.023	0.026	0.974
		β_1	0.001	0.041	0.041	0.956
		β_2	0.001	0.037	0.041	0.974
	45	β_0	-0.001	0.028	0.034	0.986
		β_1	0.001	0.047	0.053	0.978
		β_2	-0.001	0.050	0.054	0.958
7.5	15	β_0	-0.003	0.026	0.027	0.952
		β_1	0.000	0.041	0.043	0.948
		β_2	-0.004	0.041	0.043	0.954
	30	β_0	-0.002	0.028	0.033	0.970
		β_1	0.002	0.049	0.051	0.974
		β_2	0.005	0.046	0.051	0.968
	45	β_0	-0.005	0.039	0.048	0.980
		β_1	0.004	0.066	0.073	0.960
		β_2	0.002	0.072	0.074	0.960
10	15	β_0	-0.003	0.029	0.029	0.956
		β_1	-0.000	0.045	0.048	0.962
		β_2	-0.005	0.047	0.048	0.948
	30	β_0	-0.001	0.031	0.039	0.976
		β_1	0.003	0.058	0.061	0.966
		β_2	0.007	0.056	0.061	0.956
	45	β_0	-0.008	0.052	0.062	0.966
		β_1	0.006	0.089	0.093	0.956
		β_2	0.001	0.096	0.095	0.952

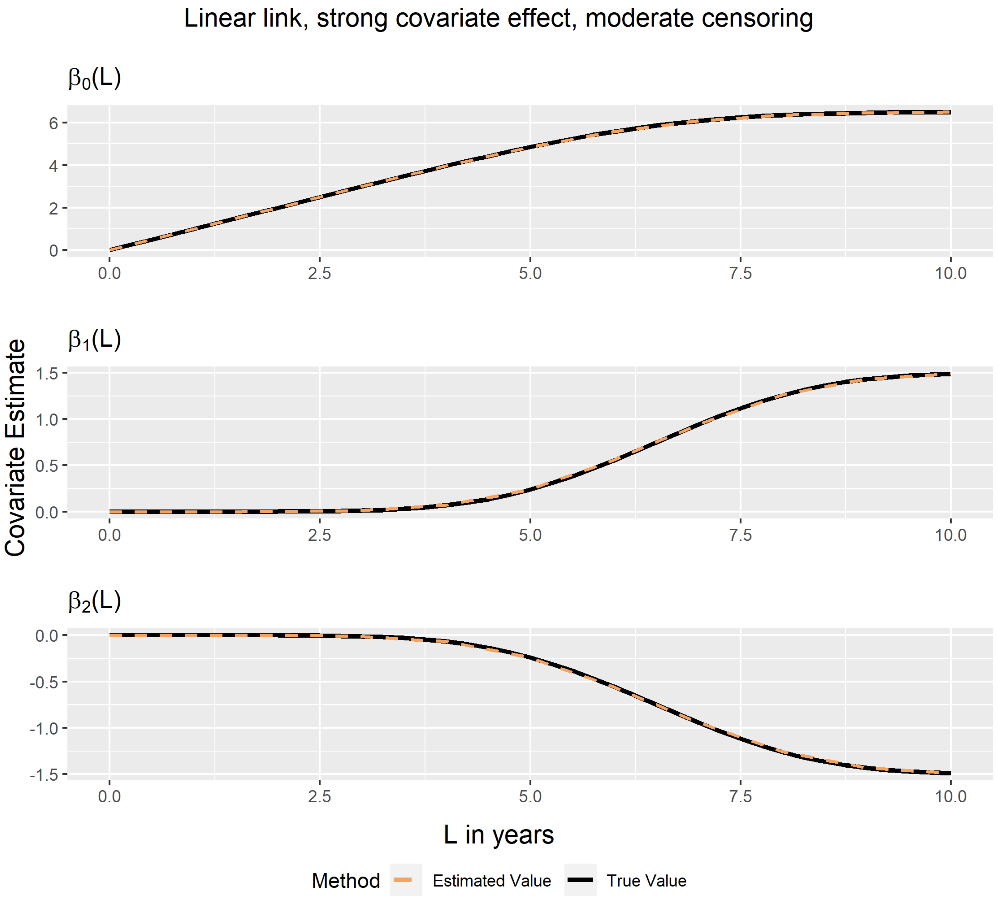

Figure 2.2: Comparison between true covariate values and estimated covariate values as a function of L for linear link. Data were generated using $\boldsymbol{\beta}_D = [4, 2.5, -2.5]$. Censoring was generated at (30%), with $\lambda_0^C = 0.1$, $\boldsymbol{\beta}_C = [-\log(1.5), \log(1.5)]$.

2.5 Application to USRDS Renal Transplantation Data

We applied our proposed method to estimating time to graft failure in kidney-transplantation recipients. The data was obtained from the Scientific Registry of Transplant Recipients (SRTR). The SRTR data system includes data on all donors, wait-listed candidates and transplant recipients in the U.S., as submitted by members of the Organ Procurement and Transplantation Network (OPTN), and has been described elsewhere. The Health Resources and Services Administration (HRSA), U.S. Department of Health and Human Services provides oversight to the activities of the OPTN and SRTR contractors.

The study population includes end stage renal disease patients who received a kidney transplant between January 01, 2000 and December 31, 2014. For this analysis, we included only those who received deceased donor kidneys, excluded all recipients younger than 18 years of age and those who have received a previous kidney transplants. Graft failure, our main event of interest, is defined as the minimum of death, transplant failure (return to dialysis), and re-transplantation. This is consistent with the majority of previous kidney transplant literature (*Zhong et al.*, 2019). Each patient was followed from the date of transplant to the earliest of graft failure or censoring date, or the end of observation period of December 31, 2014. Independent censoring occurred through a loss of follow-up or administrative censoring. For this analysis, a total of $n = 127,082$ patients were included in the study population. A total of $45,516(35.8\%)$ patients experienced graft failure. Of these, 48.6% of them died, 50.6% experienced transplant failure, and 0.8% had a re-transplant.

To illustrate our method, we chose a set of five baseline recipient covariates: age, gender, height, weight, and log of the kidney donor recipient index (log-KDRI) (*Rao et al.*, 2009). Age, height, weight, and log-KDRI are continuous, while gender is binary. We selected our $\boldsymbol{L} = [1, 2, 3, \ldots, 10]'$ years. In this case, the L_{max} is set to be 10 years. The data were replicated and assorted into an expanded dataset, with $Y_{ik} = Y_i \wedge L_k,\ k = 1, 2, \ldots, 10$. The same $\boldsymbol{L}$ was also used to fit the model, using individual $L_k's$ as knots in the parametric spline. Although we opted to use the same vector both to create the expanded data set as well as to fit the parametric spline model, the two could be chosen separately if desired.

In Figure 2.3, we plot time-varying effects for some of the more prominent covariates. Due to our having shifted continuous covariates, the intercept (top left panel) pertains to a 40-year old male who is 170cm tall, weighs 80kg, receives a kidney transplant from a deceased donor with KDRI=1 (approximately the 60th percentile of donor quality, per *Zhong et al.* (2019).

We chose here to use the linear link for its easy interpretability, but other link functions could also be used if desired. To find $\widehat{\beta}_{Z_k}(L)$ for a particular covariate Z_k using this spline parametrization, we follow the following formula:

$$\begin{aligned}\widehat{\beta}_{Z_k}(L) &= \widehat{\alpha}_{k0}L + \widehat{\alpha}_{k1}(L-1)_+ + \widehat{\alpha}_{k2}(L-2)_+ + \widehat{\alpha}_{k3}(L-3)_+ + \widehat{\alpha}_{k4}(L-4)_+ \\ &\quad +\widehat{\alpha}_{k5}(L-5)_+ + \widehat{\alpha}_{k6}(L-6)_+ + \widehat{\alpha}_{k7}(L-7)_+ + \widehat{\alpha}_{k8}(L-8)_+ \\ &\quad +\widehat{\alpha}_{k9}(L-9)_+,\end{aligned}$$

where $a_+ = \max\{a, 0\}$.

Similarly, we can obtain $\widehat{V}\{\widehat{\beta}_{Z_k}(L)\}$ by requesting the standard robust variance-covariance matrix from GEE software/functions. Since the spline terms are constant (zero variance), one can simply obtain the variance using relevant elements in the variance-covariance matrix and summing through the sum of variance formula.

In Figure 2.4 we present RMST predictions for various covariate patterns. Covariate sets 1 and 3 represent lower-risk patients, while 2 and 4 correspond to patients that are

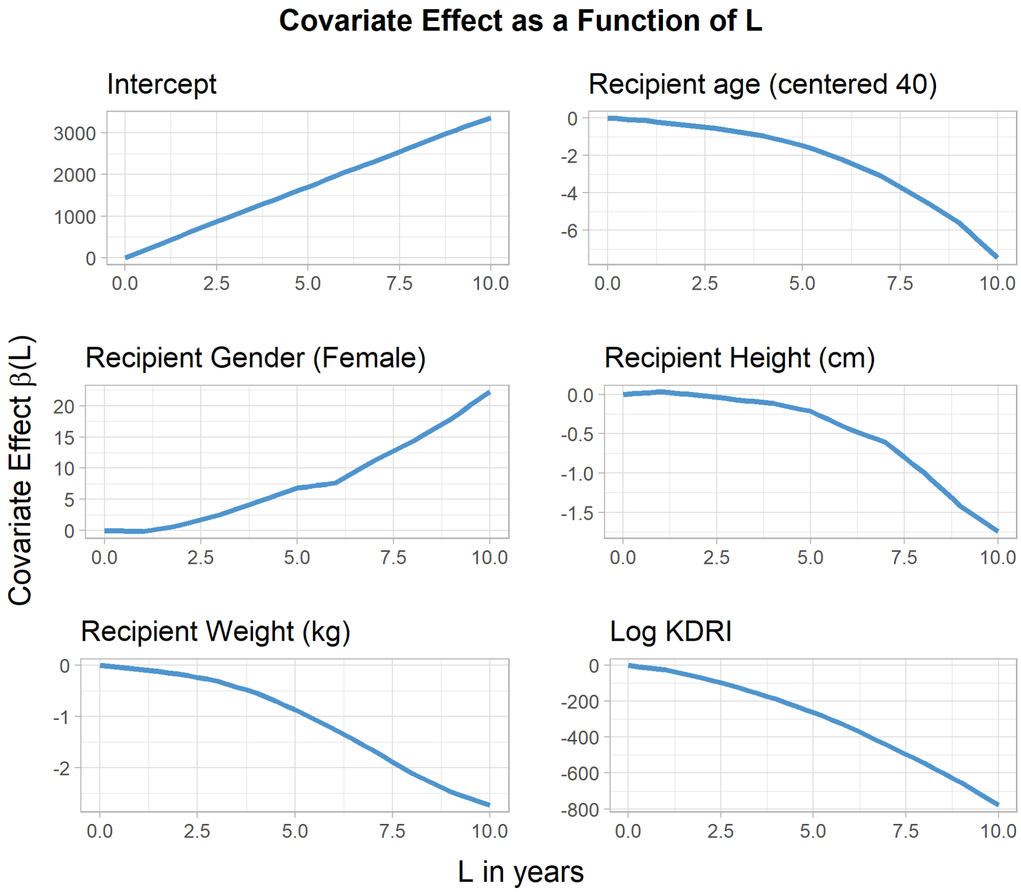

Figure 2.3: Analysis of SRTR data. Estimated covariate effects as a function of time since transplant

higher-risk. Contrasts between the predicted RMST values across patients generally become more pronounced as L increases. Note that, due to the use of spline terms, this pattern is not forced by the model.

2.6 Discussion and Conclusions

In this report, we developed a method for modeling restricted mean survival time as a function of the restriction time. Unlike existing methods, the proposed methods allow covariate effects to depend on restriction time. The methods also permit the analyst to obtain RMST predictions for several time horizons through a single model. Our method requires specifying a maximum 'reasonable' restriction time, $L_{\max}$, after which RMST is then modeled as a parametric function of L on $(0, L_{\max}]$. Our method amounts to developing a "super-model", through stacking data sets defined by L_k values which map out a grid over $(0, L_{\max}]$. Through our methods, one can create a flexible and temporal picture of covariate effects as a function of L. The proposed variance estimator is convenient to implement and was shown to work well in moderate samples. Furthermore, computational feasibility in larger data sets is implied by our method having easily been able to handle national organ transplant registry data.

The proposed methods allow the covariate effects to depend on time, which is a major advantage over Wang and Schaubel (2018). The flexibility to use time-varying effects is well-accepted and frequently utilized in the context of hazard regression. Moreover, the work of Zhao et al. (2016) underscores the importance of viewing RMST as a function of restriction time in comparing groups nonparametrically. The major advantage of our work over Zhao et al. is that our proposed methods utilize regression, while Zhao et al. (2016) uses nonparametric comparisons. Zhao et al. (2018) would generally not be applicable to observational studies requiring simultaneous estimation of many predictors and/or when some predictors are continuous; e.g., the transplant

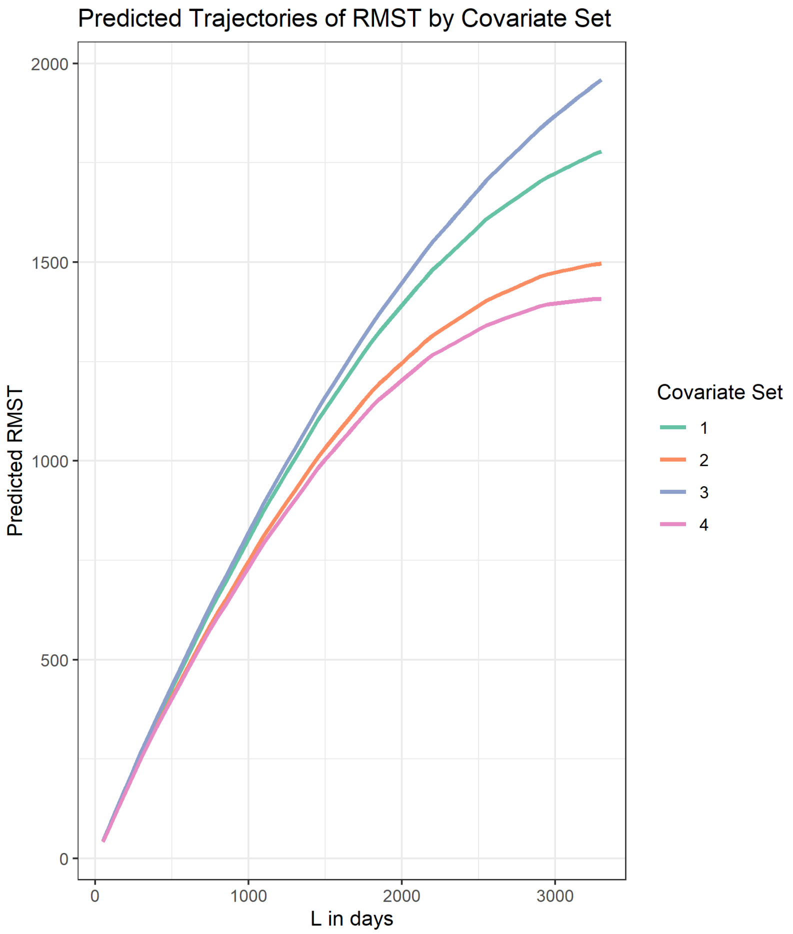

Figure 2.4: Predicted RMST projections by covariate pattern and time. Covariate Set 1 refers to a 65 year old female who received an organ with KDRI of 0.75. Covariate Set 2 is a 45 year old female who received an organ with KDRI of 1.35. Covariate Set 3 refers to a 30 year old male who received an organ with KDRI of 0.75. Covariate Set 4 refers to a 40 year old male who received an organ with KDRI of 1.5. All recipients were assumed to be 170cm in height and 80kg in weight.

registry study we analyzed in Section 2.5.

The proposed methods require IPCW, which is generally known to be subject to instability. It should be noted that small remaining-uncensored probabilities are less of an issue in RMST modeling, provided that a sensible value of L (or, in our case, L_{max}) is chosen. It is not necessary to compute the weight function too far into the tail of the observation time distribution, hence avoiding scenarios where there are very few subjects remaining at-risk (which leads to large and unstable weights). We illustrate this phenomenon in Figure 2.5, which shows box plots of the IPCW weight function versus L for the SRTR data analyzed in Section 2.5. The plot reveals that variability in the weight function increases as L increases, as does the maximum weight. However, unrealistically large weights are not observed, as the maximum weight observed is 27 at $L = 10$, and the vast majority of weights at $L = 10$ were less than three, which is very reasonable for a dataset with sample size of 127,082. In the event that unduly large weights did occur, one could cap the weight function.

In addition to choosing L_{max}, the two other main decisions involved in our method are the vector components of $\boldsymbol{L}$ used to create the expanded data-set, and the precise parametric model (including specification of knots, if appropriate) used to fit the expanded data-set. In our experience, for the first question, it is generally important to create a well spread out grid that includes copies of the data both smaller and larger than $L's$ of interest. For the second question, we propose that investigators fit separate models at a grid of L values to preliminarily determine the functional form of covariate effects and use that as a guide to determine the specific parametrization. For example, in the SRTR data set, it was clear after this step that a simple linear model would be deficient. On both of these topics, further research would help elucidate the pros and cons of particular approaches.

Finally, to illustrate our method, we applied it to kidney transplantation data to study post-transplant outcomes. To our knowledge, this is the first paper to provide a temporal model of RMST in the kidney transplantation setting.

2.7 Appendix

2.7.1 Asymptotic Properties of the Proposed Estimator

2.7.1.1 Notations

Here are the notations needed for further discussion:
i: subject index, $i \in \{1, 2, ..., n\}$

D_i: death time

C_i: Independent censoring time

τ: end of follow up time

L_k: one pre-specified time point of interest, $L_k \leq \tau$

$\boldsymbol{L}$: vector of selected values of L sorted in ascending order, i.e. $L_1, L_2, ..., L_K$

K: length of vector $\boldsymbol{L}$

L_{max}: a maximal value of L beyond which estimation becomes difficult

$X_i = D_i \wedge C_i$: observation time

$Y_{ik} = X_i \wedge L_k$: restricted observation time by L_k

$\Delta_{ik} = I(D_i \wedge L_k \leq C_i)$: indicator for restricted survival time $D_i \wedge L_k$

$\Delta_i^D = I(D_i \leq C_i)$: death indicator

$\Delta_i^C = I(C_i < D_i)$: independent censoring indicator

$\mathbf{Z_i^D}$: covariates that predict death D_i

$\mathbf{Z_i^C}$: covariates that predict independent censoring C_i

$\mathbf{Z_i}$: a covariate set that stacks $\mathbf{Z_i^D}$ and $\mathbf{Z_i^C}$ together and removes redundacy

$\lambda_i^C(t)$: hazard rate for independent censoring C_i

$\Lambda_i^C(t) = \int_0^t \lambda_i^C(u)du$: cumulative hazard rate for independent censoring C_i

$N_i^D(t) = I(X_i \leq t, \Delta_i^D = 1)$: counting process for death

$N_i^C(t) = I(X_i \leq t, \Delta_i^C = 1)$: counting process for independent censoring

$R_i(t) = I(X_i \geq t)$: at risk process

$dM_i^C(t) = dN_i^C(t) - R_i(t)d\Lambda_i^C(t)$: zero mean process for independent censoring

$Z_{i1}, Z_{i2}, ..., Z_{ip}$: covariates for i^{th} individual

$L_0(L), L_1(L), ...L_m(L)$: functions of L that estimate $\boldsymbol{\beta}_D(L)$

2.7.1.2 Model Assumptions

We have made these assumptions in our paper:

1. Assume restricted mean lifetime conditional on covariates $\mu_i(L) := E\{D_i \wedge L|\mathbf{Z}_i^D\}$ follows the model structure as below,

$$g[\mu_i(L)] \equiv g\left[E\{D_i \wedge L|\mathbf{Z}_i^D\}\right] = \boldsymbol{\beta}_D'(L)\mathbf{Z}_i^D$$

   where $g(*)$ is a given smooth and strictly monotone link function and $\boldsymbol{\beta}_D(L)$ is our primary interest.

2. We assume that $\boldsymbol{\beta}_D(L)$, a vector of continuous and monotonic functions, can be parametrically modeled as a function of L. For example, denote this parametric model of L as $\boldsymbol{\beta}_D(L) = \boldsymbol{\alpha}_0 L_0(L) + ... + \boldsymbol{\alpha}_m L_m(L)$, where $L_0(L), L_1(L), ...L_m(L)$ are functions of L, i.e. parametric or spline functions.

3. Assume Cox proportional hazards model for independent censoring time C_i:

$$\lambda_i^C(t) = \lambda_0^C(t)\exp(\boldsymbol{\beta}_C' \boldsymbol{Z}_i^C)$$

4. Assume independent censoring time is independent of death time given covariates; i.e., $C_i \perp D_i | \boldsymbol{Z}_i$.

2.7.1.3 Infinite to Finite Dimensional

As before,

$$g[\mu_i(L)] \equiv g[E\{D_i \wedge L | \mathbf{Z}_i^D\}] = \boldsymbol{\beta}_D'(L)\mathbf{Z}_i^D$$

is an infinite dimensional problem. We can in general say that $\beta_k(L)$ is a linear combination of functions of m different functions of L denoted $L_0(L), ..., L_m(L)$. We would have:

$$\boldsymbol{\beta}(L) = \begin{bmatrix} \beta_0(L) \\ \beta_1(L) \\ \vdots \\ \beta_p(L) \end{bmatrix} = \begin{bmatrix} \alpha_{00}L_0(L) + \alpha_{01}L_1(L) + ... + \alpha_{0m}L_m(L) \\ \alpha_{10}L_0(L) + \alpha_{11}L_1(L) + ... + \alpha_{1m}L_m(L) \\ \vdots \\ \alpha_{p0}L_0(L) + \alpha_{p1}L_1(L) + ... + \alpha_{pm}L_m(L) \end{bmatrix}$$

Then we can rewrite the model in the following way:

$$
\begin{aligned}
g[\mu_i(L)] \equiv g[E\{D_i \wedge L|\mathbf{Z}_i^D\}] &= \boldsymbol{\beta}'_D(L)\boldsymbol{Z}_i^D \\
&= \begin{bmatrix} \beta_0(L) & \beta_1(L) & \dots & \beta_p(L) \end{bmatrix} \begin{bmatrix} 1 \\ Z_{i1} \\ \vdots \\ Z_{ip} \end{bmatrix} \\
&= \begin{bmatrix} \alpha_{00}L_0(L) + ... + \alpha_{0m}L_m(L) \\ \alpha_{10}L_0(L) + ... + \alpha_{1m}L_m(L) \\ \vdots \\ \alpha_{p0}L_0(L) + ... + \alpha_{pm}L_m(L) \end{bmatrix}^T \begin{bmatrix} 1 \\ Z_{i1} \\ \vdots \\ Z_{ip} \end{bmatrix} \\
&= \begin{bmatrix} \alpha_{00} & \dots & \alpha_{0m} & \alpha_{10} & \dots & \alpha_{1m} & \dots & \alpha_{p0} & \dots & \alpha_{pm} \end{bmatrix} \begin{bmatrix} L_0(L) \\ L_1(L) \\ \vdots \\ L_m(L) \\ L_0(L)Z_{i1} \\ L_1(L)Z_{i1} \\ \vdots \\ L_m(L)Z_{i1} \\ \vdots \\ L_0(L)Z_{ip} \\ L_1(L)Z_{ip} \\ \vdots \\ L_m(L)Z_{ip} \end{bmatrix}
\end{aligned}
$$

Denote our new covariate vector as

$$\tilde{\boldsymbol{Z}}_i^D(L) = \boldsymbol{Z}_i \otimes \boldsymbol{L}(L)$$

where $\otimes$ denotes the Kronecker product. Similarly, let $\boldsymbol{\alpha}_0 = (\alpha_{00}, \ldots, \alpha_{0m})', \ldots, \boldsymbol{\alpha}_p = (\alpha_{p0}, \ldots, \alpha_{pm})'$. Then the new coefficient vector can be written as

$$\tilde{\boldsymbol{\beta}}_D = \begin{bmatrix} \boldsymbol{\alpha}_0 \\ \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_p \end{bmatrix}$$

Hence, we can rewrite assumption 1 as:

$$g[\mu_i(L)] \equiv g[E\{D_i \wedge L|\mathbf{Z}_i^D\}] = \boldsymbol{\beta}_D'(L)\mathbf{Z_i^D} = \tilde{\boldsymbol{\beta}}_D'\tilde{\boldsymbol{Z}}_i^D(L)$$

2.7.1.4 Regularity Conditions

We specify the necessary regularity conditions (1)-(7) as below.

1. $\{X_i, \Delta_i^D, \Delta_i^C, \boldsymbol{Z}_i\}, i = 1, 2, \ldots, n$ are independently and identically distributed.

2. $P(R_i(t) = 1) > 0 \ for \ t \in (0, \tau), i = 1, ..., n$

3. $|Z_{ik}| < M_Z < \infty$ for $i = 1, ..., n$, where Z_{ik} are the k^{th} components of $\boldsymbol{Z}_i$

4. $\Lambda_i^C(\tau) < \infty$ and is absolutely continuous for $t \in (0, \tau]$.

5. There exist neighborhoods $\boldsymbol{\mathcal{B}}_C$ of $\boldsymbol{\beta}_C$ such that for $k = 0, 1, 2$,

$$\sup_{t\in(0,\tau],\ \boldsymbol{\beta}\in\boldsymbol{\beta}_C} \left\| \frac{1}{n}\sum_{i=1}^{n} \exp(\boldsymbol{\beta}'\boldsymbol{Z}_i^C)R_i(t)\boldsymbol{Z}_i^{C\otimes k} - \mathbf{r}_C^{(k)}(t;\boldsymbol{\beta}) \right\| \xrightarrow{p} 0,$$

where $\boldsymbol{v}^{\otimes 0} = 1, \boldsymbol{v}^{\otimes 1} = \boldsymbol{v}, \boldsymbol{v}^{\otimes 2} = \boldsymbol{v}'\boldsymbol{v}$, and

$$\mathbf{r}_C^{(k)}(t;\boldsymbol{\beta}) = E[\exp(\boldsymbol{\beta}'\boldsymbol{Z}_i^C)R_i(t)\boldsymbol{Z}_i^{C\otimes k}]$$

6. Define $h(x) = \partial g^{-1}(x)/\partial x$, where h exists and is continuous in an open neighborhood $\boldsymbol{\mathcal{B}}_D$ of $\tilde{\boldsymbol{\beta}}_D$.

7. Matrices $\mathbf{A}(\tilde{\boldsymbol{\beta}}_D), \boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)$ are both positive definite, and are defined below:

$$\mathbf{A}(\tilde{\boldsymbol{\beta}}_D) = E[\sum_{k=1}^{K} \tilde{\boldsymbol{Z}}_i^D(L_k)^{\otimes 2} h\{\tilde{\boldsymbol{\beta}}_D'\tilde{\boldsymbol{Z}}_i^D(L_k)\}]$$
$$\boldsymbol{\Omega}_C(\boldsymbol{\beta}_C) = E[\int_0^{\tau} \{\frac{\mathbf{r}_C^{(2)}(t;\boldsymbol{\beta}_C)}{r_C^{(0)}(t;\boldsymbol{\beta}_C)} - \bar{\mathbf{z}}_C(t;\boldsymbol{\beta}_C)^{\otimes 2}\} dN_i^C(t)]$$

where $\bar{\mathbf{z}}_C(t;\boldsymbol{\beta}) = \frac{\mathbf{r}_C^{(1)}(t;\boldsymbol{\beta})}{r_C^{(0)}(t;\boldsymbol{\beta})}$.

2.7.1.5 Estimating Equations

We can estimate with the following estimating equations.

$$\boldsymbol{\Phi}^*(\tilde{\boldsymbol{\beta}}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \boldsymbol{\Phi}_i^*(\tilde{\boldsymbol{\beta}}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \tilde{\boldsymbol{Z}}_i^D(L_k)\Delta_{ik}W_i^C(Y_{ik})[Y_{ik} - g^{-1}\{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{Z}}_i^D(L_k)\}] = \mathbf{0}$$

The empirical version:

$$\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\boldsymbol{\Phi}_i(\tilde{\boldsymbol{\beta}}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\tilde{\boldsymbol{Z}}_i^D(L_k)\Delta_{ik}\hat{W}_i^C(Y_{ik})[Y_{ik}-g^{-1}\{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{Z}}_i^D(L_k)\}] = \mathbf{0}$$

We are looking for the $\hat{\boldsymbol{\beta}}$ that would zero out the empirical estimating equation.

2.7.1.6 Unbiased Estimating Equation

Lemma 2.3. *Under regularity conditions (1)-(7) and for a given value of L, the estimating equation (2.7.1.5) is unbiased at the true value of $\boldsymbol{\beta}_D(L)$.*

From the reasoning above, we know that the true value, $\boldsymbol{\beta}_D(L)$ corresponds to a specific extended vector, denote as $\tilde{\boldsymbol{\beta}}_D$. We want to show that the estimating equation is zero at the true value of $\tilde{\boldsymbol{\beta}}_D$.

Proof. Let $\boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}}_D) = \tilde{\boldsymbol{Z}}_i^D(L_k)\Delta_{ik}W_i^C(Y_{ik})[Y_{ik} - g^{-1}\{\tilde{\boldsymbol{\beta}}_D'\tilde{\boldsymbol{Z}}_i^D(L_k)\}]$. Then we would like to show that $E[\boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}}_D)] = \mathbf{0}$.

$$\begin{aligned}
E[\boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}}_D)|\tilde{\boldsymbol{Z}}_i^D(L_k)] &= \tilde{\boldsymbol{Z}}_i^D(L_k)E[\Delta_{ik}W_i^C(Y_{ik})Y_{ik}|\tilde{\boldsymbol{Z}}_i^D(L_k)] \\
&- \tilde{\boldsymbol{Z}}_i^D(L_k)g^{-1}\{\tilde{\boldsymbol{\beta}}_D\tilde{\boldsymbol{Z}}_i^D(L_k)\}E[\Delta_{ik}W_i^C(Y_{ik})|\tilde{\boldsymbol{Z}}_i^D(L_k)] \\
&= \tilde{\boldsymbol{Z}}_i^D(L_k)E[E\{\Delta_{ik}W_i^C(Y_{ik})Y_{ik}|D_i\}|\tilde{\boldsymbol{Z}}_i^D(L_k)] \\
&- \tilde{\boldsymbol{Z}}_i^D(L_k)g^{-1}\{\tilde{\boldsymbol{\beta}}_D\tilde{\boldsymbol{Z}}_i^D(L_k)\}E[E\{\Delta_{ik}W_i^C(Y_{ik})|D_i\}|\tilde{\boldsymbol{Z}}_i^D(L_k)] \\
&= \tilde{\boldsymbol{Z}}_i^D(L_k)E[E\{\frac{I(D_i \wedge L_k \leq C_i)}{P(D_i \wedge L_k \leq C_i)}(C_i \wedge D_i \wedge L_k)|D_i\}|\tilde{\boldsymbol{Z}}_i^D(L_k)] \\
&- \tilde{\boldsymbol{Z}}_i^D(L_k)g^{-1}\{\tilde{\boldsymbol{\beta}}_D\tilde{\boldsymbol{Z}}_i^D(L_k)\}E[E\{\frac{I(D_i \wedge L_k \leq C_i)}{P(D_i \wedge L_k \leq C_i)}|D_i\}|\tilde{\boldsymbol{Z}}_i^D(L_k)] \\
&= \tilde{\boldsymbol{Z}}_i^D(L_k)E[E\{\frac{I(D_i \wedge L_k \leq C_i)}{P(D_i \wedge L_k \leq C_i)}(D_i \wedge L_k)|D_i\}|\tilde{\boldsymbol{Z}}_i^D(L_k)] \\
&- \tilde{\boldsymbol{Z}}_i^D(L_k)g^{-1}\{\tilde{\boldsymbol{\beta}}_D\tilde{\boldsymbol{Z}}_i^D(L_k)\}E[E\{\frac{I(D_i \wedge L_k \leq C_i)}{P(D_i \wedge L_k \leq C_i)}|D_i\}|\tilde{\boldsymbol{Z}}_i^D(L_k)] \\
&= \tilde{\boldsymbol{Z}}_i^D(L_k)E[D_i \wedge L_k|\tilde{\boldsymbol{Z}}_i^D(L_k)] - \tilde{\boldsymbol{Z}}_i^D(L_k)g^{-1}\{\tilde{\boldsymbol{\beta}}_D\tilde{\boldsymbol{Z}}_i^D(L_k)\} \\
&= \tilde{\boldsymbol{Z}}_i^D(L_k)[E[D_i \wedge L_k|\tilde{\boldsymbol{Z}}_i^D(L_k)] - g^{-1}\{\tilde{\boldsymbol{\beta}}_D\tilde{\boldsymbol{Z}}_i^D(L_k)\}] \\
&= \tilde{\boldsymbol{Z}}_i^D(L_k)[E[D_i \wedge L_k|\tilde{\boldsymbol{Z}}_i^D] - g^{-1}\{\tilde{\boldsymbol{\beta}}_D\tilde{\boldsymbol{Z}}_i^D(L_k)\}] \\
&= \mathbf{0} \text{ by assumption 1}
\end{aligned}$$

Averaging over individual terms whose expectations are $\mathbf{0}$, we obtain that the estimating equation is $\mathbf{0}$ at $\tilde{\boldsymbol{\beta}}_D$. $\square$

Theorem 2.1. *Under regularity conditions (1)-(7), as* $n \to \infty$, $\sqrt{n}\boldsymbol{\Phi}(\widetilde{\boldsymbol{\beta}}_D)$ *converges in distribution to* $Normal(\mathbf{0}, \boldsymbol{B}(\widetilde{\boldsymbol{\beta}}_D))$.

Proof. First we begin by defining a few terms:

$$\begin{aligned}
\mathbf{U}_i^C(\boldsymbol{\beta}_C) &= \int_0^\tau \{\boldsymbol{Z}_i^C - \bar{\mathbf{z}}_C(u;\boldsymbol{\beta}_C)\}dM_i^C(u) \\
\mathbf{D}_i^C(t) &= \int_0^t \{\boldsymbol{Z}_i^C - \bar{\mathbf{z}}_C(u;\boldsymbol{\beta}_C)\}d\Lambda_i^C(u)
\end{aligned}$$

For independent censoring time C_i, we can derive the following for the weights (Zhang and Schaubel, 2011):

$\sqrt{n}\{\hat{W}_i^C(t) - W_i^C(t)\} = \frac{1}{\sqrt{n}} W_i^C(t)\{\mathbf{D}_i^C(t)'\boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)^{-1}\sum_{j=1}^n \mathbf{U}_j^C(\boldsymbol{\beta}_C)\} + \mathbf{o_p(1)}$

Then let us rearrange the estimating equation in the following way:

$$\begin{aligned}
&\sqrt{n}\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}})\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{k=1}^K \Delta_{ik}[Y_{ik} - g^{-1}\{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{Z}}_i^D(L_k)\}]\tilde{\boldsymbol{Z}}_i^D(L_k)[W_i^C(Y_{ik}) + \{\hat{W}_i^C(Y_{ik}) - W_i^C(Y_{ik})\}]\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{k=1}^K \Delta_{ik}[Y_{ik} - g^{-1}\{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{Z}}_i^D(L_k)\}]\tilde{\boldsymbol{Z}}_i^D(L_k)W_i^C(Y_{ik})\\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{k=1}^K \Delta_{ik}[Y_{ik} - g^{-1}\{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{Z}}_i^D(L_k)\}]\tilde{\boldsymbol{Z}}_i^D(L_k)\{\hat{W}_i^C(Y_{ik}) - W_i^C(Y_{ik})\}
\end{aligned}$$

We know that the first part is just $\frac{1}{\sqrt{n}}\sum_{k=1}^K\sum_{i=1}^n \boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}})$.

The second equation can be written as

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{k=1}^K \Delta_{ik}[Y_{ik} - g^{-1}\{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{Z}}_i^D(L_k)\}]\tilde{\boldsymbol{Z}}_i^D(L_k)\{\hat{W}_i^C(Y_{ik}) - W_i^C(Y_{ik})\}\\
&= \frac{1}{n^{1.5}}\sum_{i=1}^n\sum_{k=1}^K \boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}})\{\mathbf{D}_i^C(Y_{ik})'\boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)^{-1}\sum_{j=1}^n \mathbf{U}_j^C(\boldsymbol{\beta}_C)\} + \mathbf{o_p(1)}\\
&= \frac{1}{n^{1.5}}\sum_{i=1}^n\sum_{k=1}^K\sum_{j=1}^n \boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}})\{\mathbf{D}_i^C(Y_{ik})'\boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)^{-1}\mathbf{U}_j^C(\boldsymbol{\beta}_C)\} + \mathbf{o_p(1)}\\
&= \frac{1}{\sqrt{n}}\sum_{j=1}^n\{\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K \boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}})\mathbf{D}_i^C(Y_{ik})'\}\boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)^{-1}\mathbf{U}_j^C(\boldsymbol{\beta}_C) + \mathbf{o_p(1)}
\end{aligned}$$

If we define $\mathbf{K}_C(\tilde{\boldsymbol{\beta}}) \equiv E[\boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}})\mathbf{D}_i^C(Y_{ik})']$, then we have
$\frac{1}{\sqrt{n}}\sum_{j=1}^n\sum_{k=1}^K \mathbf{K}_C(\tilde{\boldsymbol{\beta}})\boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)^{-1}\mathbf{U}_j^C(\boldsymbol{\beta}_C) + \mathbf{o_p(1)}$.

Putting the first and second parts together, we obtain

$$\begin{aligned}
&\sqrt{n}\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}}) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{K}\boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}}) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{K}\boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)^{-1}\mathbf{U}_i^C(\boldsymbol{\beta}_C)\mathbf{K}_C(\tilde{\boldsymbol{\beta}}) + \mathbf{o_p}(\mathbf{1}) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{K}\{\boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}}) + \boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)^{-1}\mathbf{U}_i^C(\boldsymbol{\beta}_C)\mathbf{K}_C(\tilde{\boldsymbol{\beta}})\} + \mathbf{o_p}(\mathbf{1})
\end{aligned}$$

Let $\mathbf{B}_i(\tilde{\boldsymbol{\beta}}) = \sum_{k=1}^{K}\{\boldsymbol{\epsilon}_{ik}(\tilde{\boldsymbol{\beta}}) + \boldsymbol{\Omega}_C(\boldsymbol{\beta}_C)^{-1}\mathbf{U}_i^C(\boldsymbol{\beta}_C)\mathbf{K}_C(\tilde{\boldsymbol{\beta}})\}$ and define $\mathbf{B}(\tilde{\boldsymbol{\beta}}) \equiv E\{\mathbf{B}_i(\tilde{\boldsymbol{\beta}})^{\otimes 2}\}$.

Then, $\sqrt{n}\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{B}_i(\tilde{\boldsymbol{\beta}}) + \mathbf{o_p}(\mathbf{1})$.

Then, by the central limit theorem, we have $\sqrt{n}\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}}_D) \xrightarrow{D} Normal(\mathbf{0}, \mathbf{B}(\tilde{\boldsymbol{\beta}}_D))$. □

2.7.1.7 Consistency

Theorem 2.2(a). *Under regularity conditions (1)-(7), $n \to \infty$, $\hat{\boldsymbol{\beta}}_D \xrightarrow{p} \tilde{\boldsymbol{\beta}}_D$.*

Proof. We will make use of the Inverse Function Theorem (*Foutz*, 1977) by first verifying the following conditions:

1. $\partial\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}})/\partial\tilde{\boldsymbol{\beta}}'$ exists and is continuous in an open neighborhood $\tilde{\boldsymbol{\mathcal{B}}}_D$ of $\tilde{\boldsymbol{\beta}}_D$. We can show that $\partial\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}})/\partial\tilde{\boldsymbol{\beta}}' = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\tilde{\boldsymbol{Z}}_i^D(L_k)^{\otimes 2}\Delta_{ik}W_i^C(Y_{ik})h\{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{Z}}_i^D(L_k)\}$, where $h(x) = \partial g^{-1}(x)/\partial x$

2. $-\partial\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}})/\partial\tilde{\boldsymbol{\beta}}'|_{\tilde{\boldsymbol{\beta}}=\tilde{\boldsymbol{\beta}}_D}$ is positive definite with probability 1 as $n \to \infty$.

3. $-\partial\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}})/\partial\tilde{\boldsymbol{\beta}}'$ converges in probability to a fixed function uniformly in an open neighborhood $\tilde{\boldsymbol{\mathcal{B}}}_D$ of $\tilde{\boldsymbol{\beta}}_D$.

4. The estimating function is asymptotically unbiased, i.e. $\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}}_D) \xrightarrow{p} \mathbf{0}$.

Verification of conditions:

1. The first condition is satisfied automatically because $h(x)$ is assumed to exist and be continuous in an open neighborhood $\tilde{\boldsymbol{\mathcal{B}}}_D$ of $\tilde{\boldsymbol{\beta}}_D$.

2. For the second condition,

$$
\begin{aligned}
&-\frac{\partial \boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}'}|_{\tilde{\boldsymbol{\beta}}=\tilde{\boldsymbol{\beta}}_D} \\
&= E[\sum_{k=1}^{K} \Delta_{ik} W_i^C(Y_{ik}) h\{\tilde{\boldsymbol{\beta}}_D' \tilde{\boldsymbol{Z}}_i^D(L_k)\} \tilde{\boldsymbol{Z}}_i^D(L_k)^{\otimes 2}] + \mathbf{o_p(1)} \\
&= E[E\{\sum_{k=1}^{K} \Delta_{ik} W_i^C(Y_{ik}) h\{\tilde{\boldsymbol{\beta}}_D \tilde{\boldsymbol{Z}}_i^D(L_k)\} \tilde{\boldsymbol{Z}}_i^D(L_k)^{\otimes 2} | D_i, \tilde{\boldsymbol{Z}}_i^D(L_k)\}] + \mathbf{o_p(1)} \\
&= E[\sum_{k=1}^{K} E\{\Delta_{ik} W_i^C(Y_{ik}) | D_i, \tilde{\boldsymbol{Z}}_i^D(L_k)\} h\{\tilde{\boldsymbol{\beta}}_D' \tilde{\boldsymbol{Z}}_i^D(L_k)\} \tilde{\boldsymbol{Z}}_i^D(L_k)^{\otimes 2}] + \mathbf{o_p(1)} \\
&= E[\sum_{k=1}^{K} E\{\frac{I(C_i \geq D_i \wedge L_k)}{P(C_i \geq D_i \wedge L_k)} | D_i, \tilde{\boldsymbol{Z}}_i^D(L_k)\} h\{\tilde{\boldsymbol{\beta}}_D' \tilde{\boldsymbol{Z}}_i^D(L_k)\} \tilde{\boldsymbol{Z}}_i^D(L_k)^{\otimes 2}] + \mathbf{o_p(1)} \\
&= E[\sum_{k=1}^{K} h\{\tilde{\boldsymbol{\beta}}_D' \tilde{\boldsymbol{Z}}_i^D(L_k)\} \tilde{\boldsymbol{Z}}_i^D(L_k)^{\otimes 2}] + \mathbf{o_p(1)} \\
&\equiv \mathbf{A}(\tilde{\boldsymbol{\beta}}_D) + \mathbf{o_p(1)}
\end{aligned}
$$

   We have previously assumed that $\mathbf{A}(\tilde{\boldsymbol{\beta}}_D)$ is positive definite, so the second condition holds as $n \to \infty$.

3. The third condition holds by the law of large numbers.

4. Since we have proven $\sqrt{n}\boldsymbol{\Phi}(\tilde{\boldsymbol{\beta}}_D) \xrightarrow{D} Normal(\mathbf{0}, \mathcal{B}(\tilde{\boldsymbol{\beta}}_D))$, this statement follows by Chebyshev's inequality.

Hence, having verified these conditions, we conclude that $\hat{\boldsymbol{\beta}}_D \xrightarrow{p} \tilde{\boldsymbol{\beta}}_D$ from the Inverse Function Theorem. □

2.7.1.8 Asymptotic Distribution

Theorem 2.2(b). *Under regularity conditions, as* $n \to \infty$*,*
$\sqrt{n}(\hat{\boldsymbol{\beta}}_D - \tilde{\boldsymbol{\beta}}_D) \xrightarrow{D} Normal(\mathbf{0}, \mathbf{A}(\tilde{\boldsymbol{\beta}}_D)^{-1}\mathbf{B}(\tilde{\boldsymbol{\beta}}_D)\mathbf{A}(\tilde{\boldsymbol{\beta}}_D)^{-1})$*.*

Proof. We do Taylor expansion of the estimating equation $\mathbf{\Phi}(\hat{\boldsymbol{\beta}}_D)$ around $\tilde{\boldsymbol{\beta}}_D$ and get

$\mathbf{0} = \mathbf{\Phi}(\hat{\boldsymbol{\beta}}_D) = \mathbf{\Phi}(\tilde{\boldsymbol{\beta}}_D) + \frac{\partial\mathbf{\Phi}(\boldsymbol{\beta})}{\partial\tilde{\boldsymbol{\beta}}'}|_{\tilde{\boldsymbol{\beta}}=\breve{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}_D - \tilde{\boldsymbol{\beta}}_D)$, where $\breve{\boldsymbol{\beta}}$ lies between $\hat{\boldsymbol{\beta}}_D$ and $\tilde{\boldsymbol{\beta}}_D$.

Then we have:

$$\begin{aligned}
&-\mathbf{\Phi}(\tilde{\boldsymbol{\beta}}_D) = \left\{\frac{\partial\mathbf{\Phi}(\tilde{\boldsymbol{\beta}})}{\partial\tilde{\boldsymbol{\beta}}}|_{\tilde{\boldsymbol{\beta}}=\breve{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}_D - \tilde{\boldsymbol{\beta}}_D)\right\}\\
&\Rightarrow -\mathbf{\Phi}(\tilde{\boldsymbol{\beta}}_D)\left\{\frac{\partial\mathbf{\Phi}(\tilde{\boldsymbol{\beta}})}{\partial\tilde{\boldsymbol{\beta}}}|_{\tilde{\boldsymbol{\beta}}=\breve{\boldsymbol{\beta}}}\right\}^{-1} = \hat{\boldsymbol{\beta}}_D - \tilde{\boldsymbol{\beta}}_D\\
&\Rightarrow \sqrt{n}\mathbf{\Phi}(\tilde{\boldsymbol{\beta}}_D)\left\{-\frac{\partial\mathbf{\Phi}(\tilde{\boldsymbol{\beta}})}{\partial\tilde{\boldsymbol{\beta}}}|_{\tilde{\boldsymbol{\beta}}=\breve{\boldsymbol{\beta}}}\right\}^{-1} = \sqrt{n}(\hat{\boldsymbol{\beta}}_D - \tilde{\boldsymbol{\beta}}_D)\\
&\Rightarrow \sqrt{n}\mathbf{\Phi}(\tilde{\boldsymbol{\beta}}_D)\mathbf{A}(\breve{\boldsymbol{\beta}})^{-1} = \sqrt{n}(\hat{\boldsymbol{\beta}}_D - \tilde{\boldsymbol{\beta}}_D)\\
&\Rightarrow \sqrt{n}(\hat{\boldsymbol{\beta}}_D - \tilde{\boldsymbol{\beta}}_D) = \mathbf{A}(\tilde{\boldsymbol{\beta}}_D)^{-1}\sqrt{n}\mathbf{\Phi}(\boldsymbol{\beta}_D) + \mathbf{o_p(1)}
\end{aligned}$$

Bringing in theorem 2, we have $\sqrt{n}(\hat{\boldsymbol{\beta}}_D - \tilde{\boldsymbol{\beta}}_D) \xrightarrow{D} Normal(\mathbf{0}, \mathbf{A}(\tilde{\boldsymbol{\beta}}_D)^{-1}\mathbf{B}(\tilde{\boldsymbol{\beta}}_D)\mathbf{A}(\tilde{\boldsymbol{\beta}}_D)^{-1})$

□

2.7.2 Computation of Standard Errors

2.7.2.1 Variance Calculations

The above formula gives the variance-covariance matrix of $\tilde{\boldsymbol{\beta}}$. We want to obtain the variance of $\hat{\boldsymbol{\beta}}(L)$.

For example, if we were interested in the k-th covariate, $\hat{\beta}_k(L)$,

$$\begin{aligned} Var(\hat{\beta}_k(L)) &= Var(\alpha_{k0}L_0(L) + ... + \alpha_{km}L_m(L)) \\ &= Var(\alpha_{k0})L_0(L) + ... + Var(\alpha_{km})L_m(L) + \sum_{i \neq j} 2Cov(\alpha_{ki}, \alpha_{kj})L_i(L)L_j(L) \end{aligned}$$

We can obtain both corresponding variances and covariances from matrix $\frac{1}{n}\widehat{\mathbf{A}}(\hat{\boldsymbol{\beta}}_D)^{-1}\widehat{\mathbf{B}}(\hat{\boldsymbol{\beta}}_D)\widehat{\mathbf{A}}(\hat{\boldsymbol{\beta}}_D)^{-1}$.

Computationally, we can obtain both $Var(\alpha_{ki})$ and $Cov(\alpha_{ki}, \alpha_{kj})$, where i, j are arbitrary indexes for enumerating coefficients, directly from GEE model outputs.

2.7.3 Supplementary Tables and Figures

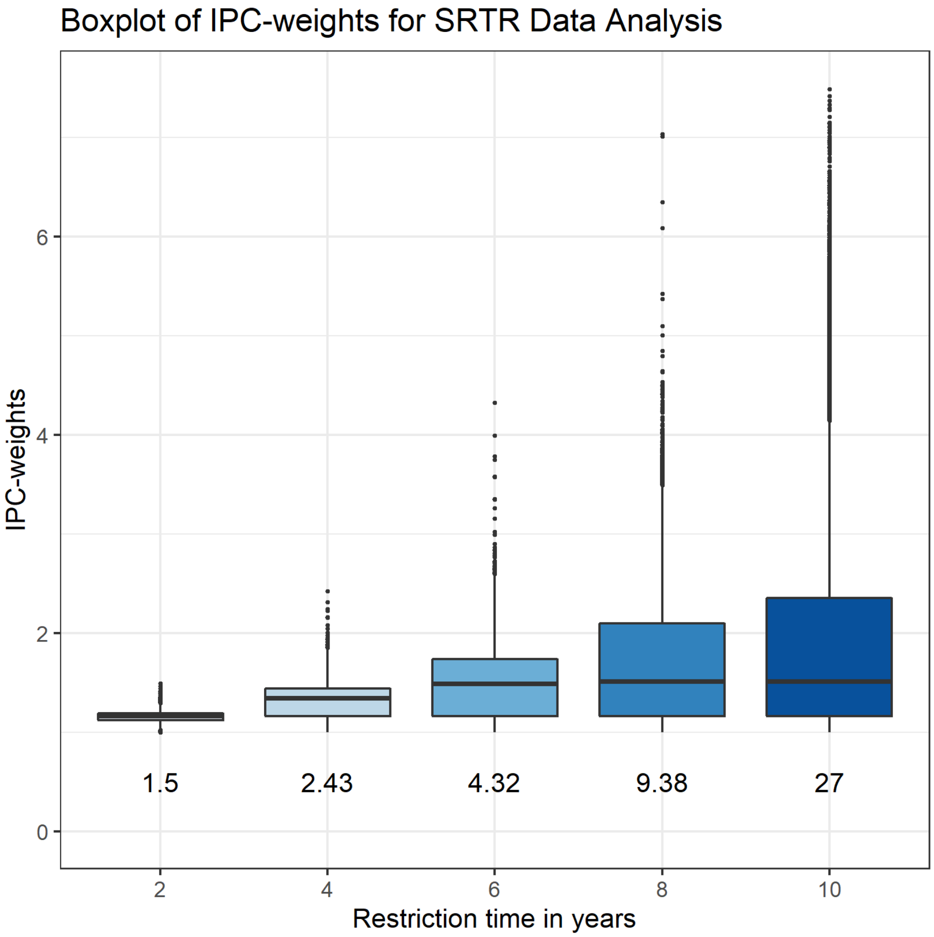

Figure 2.5: Boxplots of Inverse Probability Censoring Weights by restriction time, L for the SRTR kidney transplantation data analysis. The numbers below the boxplots indicate maximum weight.

Table 2.3: Simulation results: linear link, weak covariate effect. Data were generated using $\boldsymbol{\beta}_D = [4, 0.75, -0.75]$.

L	Censor %	Parameter	BIAS	ESD	ASE	CP
5	15	β_0	-0.001	0.058	0.059	0.956
		β_1	0.003	0.099	0.101	0.960
		β_2	0.001	0.098	0.101	0.960
	30	β_0	-0.001	0.060	0.067	0.976
		β_1	0.001	0.116	0.115	0.940
		β_2	-0.001	0.111	0.115	0.958
	45	β_0	-0.004	0.069	0.081	0.978
		β_1	0.004	0.131	0.143	0.952
		β_2	-0.002	0.133	0.143	0.968
7.5	15	β_0	-0.005	0.083	0.084	0.952
		β_1	0.004	0.139	0.145	0.966
		β_2	0.005	0.136	0.145	0.958
	30	β_0	-0.006	0.089	0.103	0.978
		β_1	0.002	0.170	0.178	0.954
		β_2	-0.000	0.167	0.179	0.966
	45	β_0	-0.017	0.113	0.142	0.984
		β_1	0.013	0.223	0.246	0.960
		β_2	-0.002	0.237	0.248	0.962
10	15	β_0	-0.004	0.101	0.104	0.952
		β_1	0.003	0.172	0.179	0.966
		β_2	0.003	0.167	0.179	0.960
	30	β_0	-0.006	0.112	0.138	0.988
		β_1	-0.003	0.221	0.237	0.978
		β_2	-0.004	0.224	0.239	0.956
	45	β_0	-0.040	0.153	0.210	0.970
		β_1	0.036	0.347	0.356	0.950
		β_2	-0.034	0.379	0.356	0.908

Table 2.4: Simulation results: log link, weak covariate effect. Data were generated using $\boldsymbol{\beta}_D = [1.25, \log(1.25), -\log(1.25)]$.

L	Censor %	Parameter	BIAS	ESD	ASE	CP
5	15	β_0	-0.001	0.022	0.023	0.952
		β_1	0.001	0.037	0.039	0.964
		β_2	0.000	0.036	0.039	0.968
	30	β_0	-0.001	0.023	0.026	0.970
		β_1	0.001	0.044	0.045	0.954
		β_2	-0.001	0.043	0.045	0.958
	45	β_0	-0.003	0.028	0.034	0.982
		β_1	0.002	0.052	0.058	0.958
		β_2	-0.001	0.053	0.058	0.968
7.5	15	β_0	-0.003	0.026	0.027	0.950
		β_1	0.002	0.045	0.047	0.962
		β_2	0.000	0.044	0.046	0.964
	30	β_0	-0.003	0.029	0.035	0.984
		β_1	-0.000	0.056	0.059	0.966
		β_2	0.000	0.055	0.059	0.956
	45	β_0	-0.011	0.040	0.050	0.974
		β_1	0.009	0.079	0.084	0.960
		β_2	-0.005	0.084	0.085	0.942
10	15	β_0	-0.002	0.029	0.031	0.960
		β_1	0.001	0.050	0.053	0.968
		β_2	0.001	0.049	0.053	0.968
	30	β_0	-0.004	0.034	0.043	0.988
		β_1	0.000	0.066	0.072	0.970
		β_2	-0.000	0.068	0.072	0.952
	45	β_0	-0.021	0.054	0.067	0.950
		β_1	0.018	0.112	0.109	0.920
		β_2	-0.012	0.117	0.109	0.918

Table 2.5: Comparison between n=250 vs. n=500: linear link, strong covariate effect, moderate censoring. Data were generated using $\boldsymbol{\beta}_D = [4, 2.5, -2.5]$. Censoring was generated at (30%).

L	n	Parameter	BIAS	ESD	ASE	CP
5	250	β_0	0.006	0.118	0.125	0.958
		β_1	0.006	0.210	0.212	0.948
		β_2	-0.010	0.189	0.213	0.960
	500	β_0	0.002	0.084	0.088	0.958
		β_1	0.006	0.146	0.150	0.962
		β_2	-0.009	0.148	0.150	0.938
7.5	250	β_0	0.001	0.171	0.188	0.962
		β_1	0.001	0.305	0.319	0.950
		β_2	-0.005	0.285	0.320	0.960
	500	β_0	-0.003	0.119	0.133	0.968
		β_1	0.002	0.209	0.226	0.964
		β_2	-0.010	0.212	0.226	0.962
10	250	β_0	-0.006	0.219	0.247	0.970
		β_1	0.003	0.401	0.422	0.954
		β_2	-0.011	0.390	0.422	0.974
	500	β_0	-0.003	0.150	0.175	0.978
		β_1	-0.001	0.273	0.301	0.966
		β_2	-0.008	0.273	0.300	0.972

Table 2.6: Comparison between n=250 vs. n=500: log link, strong covariate effect, moderate censoring. Data were generated using $\boldsymbol{\beta}_D = [1.25, \log(2), -\log(2)]$. Censoring was generated at (30%).

L	n	Parameter	BIAS	ESD	ASE	CP
5	250	β_0	-0.006	0.049	0.052	0.976
		β_1	-0.001	0.079	0.083	0.946
		β_2	-0.003	0.081	0.083	0.954
	500	β_0	-0.002	0.032	0.037	0.976
		β_1	0.007	0.054	0.059	0.974
		β_2	0.000	0.057	0.059	0.950
7.5	250	β_0	-0.011	0.058	0.065	0.972
		β_1	0.003	0.097	0.102	0.948
		β_2	-0.001	0.097	0.102	0.950
	500	β_0	-0.006	0.039	0.046	0.982
		β_1	0.007	0.069	0.072	0.968
		β_2	0.000	0.070	0.072	0.958
10	250	β_0	-0.013	0.068	0.076	0.960
		β_1	0.003	0.118	0.118	0.940
		β_2	0.000	0.116	0.119	0.944
	500	β_0	-0.008	0.044	0.054	0.978
		β_1	0.006	0.082	0.085	0.954
		β_2	-0.005	0.081	0.085	0.966

Table 2.7: Simulation results comparing proposed method to Wang and Schaubel 2018 for linear link. Data were generated using $\boldsymbol{\beta}_D = [4, 2.5, -2.5]$. Censoring was generated at a moderate level (30%).

		$L = 5$			$L = 7.5$			$L = 10$		
Method	Parameter	BIAS	ESD	$\sqrt{MSE}$	BIAS	ESD	$\sqrt{MSE}$	BIAS	ESD	$\sqrt{MSE}$
Proposed	β_0	-0.003	0.062	0.062	-0.004	0.089	0.089	-0.009	0.113	0.113
	β_1	-0.007	0.105	0.105	-0.014	0.154	0.154	-0.018	0.198	0.199
	β_2	0.007	0.111	0.111	0.007	0.164	0.164	0.005	0.214	0.214
W and S	β_0	-0.003	0.062	0.062	-0.004	0.089	0.089	-0.009	0.113	0.113
	β_1	-0.007	0.105	0.105	-0.014	0.155	0.155	-0.018	0.198	0.199
	β_2	0.007	0.111	0.111	0.008	0.165	0.165	0.005	0.214	0.214

Table 2.8: Simulation results comparing proposed method to Wang and Schaubel 2018 for log link. Data were generated using $\boldsymbol{\beta}_D = [1.25, \log(2), -\log(2)]$. Censoring was generated at a moderate level (30%).

		$L = 5$			$L = 7.5$			$L = 10$		
Method	Parameter	BIAS	ESD	$\sqrt{MSE}$	BIAS	ESD	$\sqrt{MSE}$	BIAS	ESD	$\sqrt{MSE}$
Proposed	β_0	-0.001	0.025	0.025	-0.002	0.029	0.029	-0.004	0.032	0.032
	β_1	0.000	0.039	0.039	0.001	0.046	0.046	0.002	0.054	0.054
	β_2	0.001	0.040	0.040	0.000	0.049	0.049	0.002	0.058	0.058
W and S	β_0	-0.001	0.025	0.025	-0.002	0.029	0.029	-0.004	0.032	0.032
	β_1	0.000	0.039	0.039	0.000	0.047	0.047	0.002	0.054	0.054
	β_2	0.001	0.040	0.040	0.000	0.049	0.049	0.002	0.058	0.058

Table 2.9: Simulation results comparing efficiency between stacking $k = 10$ vs $k = 20$ datasets for linear link. Data were generated using $\boldsymbol{\beta}_D = [4, 2.5, -2.5]$. Censoring was generated at a moderate level (30%).

Number of stackings	Parameter	L=5 $\sqrt{MSE}$	L=7.5 $\sqrt{MSE}$	L=10 $\sqrt{MSE}$
k=10	β_0	0.062	0.089	0.113
	β_1	0.105	0.154	0.199
	β_2	0.111	0.163	0.214
k=20	β_0	0.058	0.084	0.106
	β_1	0.100	0.151	0.200
	β_2	0.105	0.147	0.190

Table 2.10: Simulation results comparing efficiency between stacking $k = 10$ vs $k = 20$ datasets for log link. Data were generated using $\boldsymbol{\beta}_D = [1.25, \log(2), -\log(2)]$. Censoring was generated at a moderate level (30%).

Number of stackings	Parameter	L=5 $\sqrt{MSE}$	L=7.5 $\sqrt{MSE}$	L=10 $\sqrt{MSE}$
k=10	β_0	0.025	0.029	0.032
	β_1	0.039	0.046	0.054
	β_2	0.040	0.049	0.058
k=20	β_0	0.023	0.028	0.033
	β_1	0.039	0.046	0.054
	β_2	0.040	0.048	0.057

CHAPTER III

Augmented Patient Preference Incorporated Reinforcement Learning to Estimate the Optimal Dynamic Treatment Regime

3.1 Introduction

Personalized health care aims to predict patients' responses to targeted therapy using patient characteristics through a multifaceted approach (*Hamburg and Collins*, 2010). Instead of one-size-fits-all, personalized medicine hopes to concentrate therapeutic interventions on "those who will benefit, sparing expense and side effects for those who will not" (*Council et al.*, 2011). Patient responses to treatments may vary due to different levels of heterogenities, such as genetics, environmental factors, and the interplay between the two. Because of this, an appropriately personalized treatment plan needs to be sensitive and adaptive to a patient's evolving condition, especially in the case of chronic diseases. Dynamic treatment regimes (DTRs) are sequences of treatment decision rules, in which treatment decisions are adapted over time in response to an individual's treatment response and trajectory (*Chakraborty and Moodie* (2013), *Chakraborty and Murphy* (2014), *Murphy* (2003)). A data-driven adaptation of the reinforcement learning problem (*Sutton and Barto*, 2018), DTRs play an important role in evidence based medicine by mathematically formulating the complicated problem of making multiple decisions at multiple stages in order to maximize a clinically meaningful outcome (*Moodie et al.*, 2012).

A large number of methods have been proposed to evaluate the optimal DTR, including Q- and A-learning (*Murphy* (2003), *Schulte et al.* (2014)), marginal structural model with inverse probability weighting (IPW) (*Robins*, 2000), G-estimation of structural nested mean models (*Robins*, 2004), and other likelihood based methods (*Thall et al.*, 2007). In more recent years, machine learning flavored methods have joined the arsenal of available methods for optimizing DTRs, including tree-based methods (*Laber and Zhao* (2015), *Tao and Wang* (2017), *Tao et al.* (2018)), and list-based method (*Zhang et al.*, 2018). Despite the abundant selection of methods to find optimal DTRs, all stated methods rely on pre-specifying a single metric of interest (e.g. survival time, adjusted quality of life, tumor response, etc), thereby forcibly simplifying complex medical scenarios in the formulation of the problem.

Evidence based optimization of one outcome further ignores an important and personalizable part of a patient's experience. In reality, clinical decisions often result in a plethora of outcomes, often in competing directions of desirability to the patient. The hand surgery field provides an illustrative example. In light of the current opioid epidemic, surgical specialties are re-examining current post-operative opioid prescription habits. Patients have to weigh between the benefits (strong pain relief) and risks (addiction, side effects, etc) of using opioid medications vs non-opioid pain relief (weaker pain relief, but potentially fewer risks). In this case, patient preference could help guide whether to prescribe opioid medications. Similarly, we can illustrate the challenges with another example from neurology. Anti-epileptic drugs (AEDs) often come with side effects such as sedation, somnolence, distractibility, insomnia, and dizziness (*Ortinski and Meador*, 2004). Even with all of these side effects, patients might willingly tolerate their AEDs if these drugs control their seizures sufficiently such that they could retain more autonomy, (i.e. by maintaining driving privileges) (*Krumholz*, 2009). In this case, patients have to balance between value for autonomy (i.e. lower risk of seizures but higher burden of side effects) with their need for mental clarity (higher risk of seizures, but more mental clarity). This example illustrates a complex system of pros and cons that patients need to navigate using their value system, previous experience, and changing needs. A college student in an urban college environment might be willing to forgo his driving privileges in return for mental acuity to meet the demands of school, only to reverse the decision after moving to the suburbs, while a patient with a history of opioid addiction might opt for non-opioid pain medications for fear of relapse. In both of these cases, optimizing one outcome (pain relief or risk of seizure) fits neatly within the evidence based paradigm, but the inclusion of patient preference to coordinate between multiple outcomes does not. Appropriate coordination using patient preference then necessitates the search for another framework.

In recent decades, the medical community has also recognized this gap and has shifted from a paternalistic approach to advocating for patient input through a shared decision-making (SDM) framework (*Barry and Edgman-Levitan* (2012), *Basu and Meltzer* (2007)). Augmentation with patient preferences and values into decision making contributes to a more holistic approach to patient care. A survey of SDM literature shows that patient input has a positive correlation with satisfaction scores and quality of life outcomes (*Kashaf and McGill* (2015), *Shay and Lafata* (2015)). SDM has also been shown to reduce costs with unnecessary procedures (*Oshima Lee and Emanuel*, 2013), thus making it an actionable policy for reducing cost. Despite these documented benefits from the literature, physicians have struggled to practice shared decision making, partly because preference is challenging to quantify and incorporate. Current popular shared decision making approaches include using decision aids and increased communication between the clinician and patient (*Barry and Edgman-Levitan*, 2012). Diverse approaches cited here are difficult to model, making it challenging to draw conclusions from data methodically.

We propose a modeling approach to incorporate patient preference. To properly optimize an outcome representative of patient preference, accurate estimation of the preference itself is paramount. We endorse modeling preference as a latent variable and estimating it through an item response approach (*Embretson and Reise*, 2013). Patients could communicate their preference through responses on a questionnaire. These questions may ask patients to rate their agreement with certain statements, or ask patients to rate the importance of certain activities in their lives. *Butler et al.* (2018) used this approach to estimate patient preference from surveys with binary $\{0, 1\}$ responses. Their method, designed to select the optimal treatment between two potential choices at one decision time point, combined estimated preferences along with Q-learning to find the optimal individualized treatment regime (ITR).

For our scenario, chronic diseases require treatment plans to adapt to patient trajectories over a changing disease course. Furthermore, as the number of treatment stages increases, the number of treatment options often also increases, making two treatment options limiting for realistic settings. As of writing, the authors are unaware of methods in the literature that are able to accommodate multiple stages with multiple treatment options. In this paper, we propose a method that augments the combination of two potentially competing outcomes with estimated patient preference. As an illustrative example, we will consider competing outcomes of efficacy and toxicity. Our proposed method accommodates the selection between more than two treatments per stage, and allows patient preference to evolve through the stages. Finally, we propose modeling preference through a polytomous latent variable model (*Moustaki and Knott*, 2000; *Bartholomew et al.*, 2011), which allows us to estimate preferences more precisely through questions with categorical responses, an extension from binary responses in *Butler et al.* (2018). We combine competing outcomes with a linear utility function weighted by estimated patient preferences, and seek for the treatment decision that would provide the largest patient satisfaction using a tree-based reinforcement learning (T-RL) method. The remainder of this report is organized as follows. In Section 3.2, we introduce the notation, problem and goal. In Section 3.3, we describe our estimation procedure and required assumptions. In Section 3.4, we describe the implementation details. We evaluate our method in Section 3.5 through simulations of multiple scenarios. Finally, we conclude with discussions and ideas for future directions in Section 3.6.

3.2 Patient Preference Incorporated Dynamic Treatment Regimes

Let T denote the number of treatment stages, and let K_j be the number of treatment options at the j^{th} stage. Let A_{ij} be the treatment indicator for the i^{th} patient at the j^{th} stage, with the observed treatment denoted by a_{ij}. Note that when referring to a specific treatment, we will drop the subscript i, i.e. treatment a_j. Let $\mathbf{X}_{ij}$ denote patient characteristics prior to treatment assignment at stage j. In addition, we assume that each patient will have an evolving preference h_{ij}, which can be derived from answers $\mathbf{W}_{ij}$ to a questionnaire at stage j. Finally, we assume that each patient will have two observed outcomes at each stage, efficacy F_{ij} and toxicity S_{ij}. In general, we denote all history, or history up to stage K for a given variable with an overhead bar (i.e., $\overline{\mathbf{W}}_i$ and $\overline{\mathbf{W}}_{iK}$, respectively).

We will assume a utility of the form $U(F, S; h) = \Phi(h)F + \{1 - \Phi(h)\}S$ to designate the utility function at each stage, where $\Phi(\cdot)$ denotes the cumulative distribution function of a normal random variable. The choice of utility function is flexible, but assuming a linear weighted sum utility function is a common approach in multiobjective optimization (*Marler and Arora* (2004), *Lizotte et al.* (2012)). This utility is also intuitive in that a patient with preference h cares $\Phi(h)/\{1-\Phi(h)\}$ more about F than about S. Let the overall outcome of interest that we would like to optimize be $Y = f(U_1, \ldots, U_T)$, where $f(\cdot)$ is a pre-specified function. We assume that Y is bounded, and that higher values are more desirable. Going forward, we will proceed with $f(\cdot)$ as the sum, i.e. the sum: $Y = U_1 + \ldots + U_T$, although other functions of f could also be optimized in a similar manner.

Let $g_j(\overline{\mathbf{X}}_{ij}, \overline{\mathbf{W}}_{ij})$ be a function that maps from covariate and survey history to the domain of treatment assignment A_{ij}. The expected potential reward of stage specific decision rule $g_j(\overline{\mathbf{X}}_{ij}, \overline{\mathbf{W}}_{ij})$ for patient i is therefore defined as
$\mathbb{E}\left[\sum_{a_j=1}^{K_j}\{\Phi(h_{ij})F^*_{ij}(a_j) + [1-\Phi(h_{ij})]S^*_{ij}(a_j)\}I\{g_j(\overline{\mathbf{X}}_{ij}, \overline{\mathbf{W}}_{ij}) = a_j\}\right]$, where $F^*_{ij}(a_j) = F^*_{ij}(A_{i1}, \ldots, A_{i,j-1}, a_j)$ denotes the counterfactual outcome where the patient is assumed to have taken treatment a_j at stage j, conditional on previous treatment decisions $A_{i1}, \ldots, A_{i,j-1}$, and equivalently for $S^*_{ik}(a_j)$. Our goal is to find a sequence of individualized decision rules, $\mathbf{g}(\overline{\mathbf{X}}_{\mathbf{i}}, \overline{\mathbf{W}}_{\mathbf{i}}) = (g_1(\overline{\mathbf{X}}_{i1}, \overline{\mathbf{W}}_{i1}), \ldots, g_T(\overline{\mathbf{X}}_{iT}, \overline{\mathbf{W}}_{iT}))$, that optimize the potential outcome of Y_i. For the sake of brevity going forward, let us abbreviate $g_j(\overline{\mathbf{X}}_{ij}, \overline{\mathbf{W}}_{ij})$ with g_j and drop the patient index i when there is no room for confusion.

3.3 Optimization of g using proposed iterative Augmented Patient Preference incorporated Reinforcement Learning (APP-RL)

3.3.1 Bridging the counterfactual framework to the observational data

Stage T

At stage T, the counterfactual potential reward under decision rule g_T and conditional on previous treatments $(A_1, \ldots, A_{T-1})$ is $R^*_T(g_T) = \Phi(H_T)F^*_T(g_T) + \{1 - \Phi(H_T)\}S^*_T(g_T)$, where $F^*_T(g_T) = \sum_{a_T=1}^{K_T} F^*_T(a_T)I(g_T = a_T)$ and $S^*_T(g_T) = \sum_{a_T=1}^{K_T} S^*_T(a_T)I(g_T = a_T)$. The performance of g_T is measured by the expected counterfactual outcome $E\{R^*_T(g_T)\}$. The optimal rule, g_T^{opt}, then satisfies $E\{R^*_T(g_T^{opt})\} \geq E\{R^*_T(g_T)\} \forall\ g_T \in \mathscr{G}_T$, where $\mathscr{G}_T$ is the class of all potential regimes. In order to relate the observed data to counterfactual outcomes (*Murphy et al.* (2001), *Orellana et al.* (2010), *Robins and Hernán* (2009)), we make the following assumptions.

1. Consistency: the observed outcome is the same as the counterfactual outcome under a patient's received treatment, i.e., $F_T = \sum_{a_T=1}^{K_T} F^*_T(a_T)I\{A_T = a_T\}$ and $S_T = \sum_{a_T=1}^{K_T} S^*_T(a_T)I\{A_T = a_T\}$

2. No unmeasured confounding: treatment A_T is randomly assigned with probability possibly dependent on $\overline{\mathbf{X}}_T$ and $\overline{\mathbf{W}}_T$, i.e., $\{F^*_T(1), \ldots, F^*_T(K_T)\} \perp A_T | \overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T$, and $\{S^*_T(1), \ldots, S^*_T(K_T)\} \perp A_T | \overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T$

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

4. Latent variable independence: $H_T \perp (A_T, F^*_T(a_T), S^*_T(a_T)) | \overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T$

The first three assumptions are standard assumptions in causal inference used to connect observed data with counterfactual framework. The last assumption is proposed to facilitate separate modeling of preference and outcomes, but can be weakened at the expense of more complicated models and estimation procedure (*Butler et al.*, 2018).

Notice that

$$E\left[R_T^*(g_T)\right] = \mathbb{E}_T\left[\sum_{a_T=1}^{K_T}\Big[E\left\{\Phi(H_T)|\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T\right\}\mu_{T,a_T}^F(\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T) + E\left\{1-\Phi(H_T)|\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T\right\}\mu_{T,a_T}^S(\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T)\Big] I\{g_T = a_T\}\right], \quad (3.1)$$

where $\mu_{T,a_T}^F(\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T) = E(F_T|A_T = a_T, \overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T)$, and likewise $\mu_{T,a_T}^S(\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T) = E(S_T|A_T = a_T, \overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T)$. Note that the left hand side (LHS) of Equation (3.1) is expectation of potential outcome, which can be estimated using only observed data as shown by the right hand side (RHS). Furthermore, note that we can separately model preference and outcomes using assumption (4). To find the optimal regime, we want to find $g_T^{opt} = arg\ max_{g_T \in \mathscr{G}_\mathscr{T}}$ RHS of Equation(3.1).

Stage j

At stage j, $T-1 \geq j \geq 1$, g_j^{opt} can be derived from the observed data via backward induction. Following presumption of maximizing the summation of stage-specific rewards from stage j onward, we define the following stage j reward, which is a cumulative sum of stage j to T utility functions:

$R_j^*(a_j) = \mathbf{\Phi}(\mathbf{H}_{j+}) \otimes \mathbf{F}_{j+}^*(a_j) + \{1 - \mathbf{\Phi}(\mathbf{H}_{j+})\} \otimes \mathbf{S}_{j+}^*(a_j)$, where $\mathbf{\Phi}(\mathbf{H}_{j+}) = [\mathbf{\Phi}(H_j), \mathbf{\Phi}(H_{j+1}), \ldots, \mathbf{\Phi}(H_T)]$, $\mathbf{F}_{j+}^*(a_j) = [F_j^*(a_j)\ F_{j+1}^*(a_j)\ \ldots F_T^*(a_j)]$, $\mathbf{S}_{j+}^*(a_j) = [S_j^*(a_j)\ S_{j+1}^*(a_j)\ \ldots S_T^*(a_j)]$, and $\otimes$ denotes the dot product. Note that for $k = j$, $F_k^*(a_j)$ and $S_k^*(a_j)$ are as defined previously and for $k > j$, $F_k^*(a_j) = F_k^*(A_1, \ldots, A_{j-1}, a_j, g_{j+1}^{opt}, \ldots, g_k^{opt})$ denotes a counterfactual outcome given future optimized treatments and conditional on $A_1, \ldots, A_{j-1}$ and taking treatment a_j at stage j (similarly for $S_k^*(a_j)$). Then the optimal regime at stage j satisfies $E\{R_j^*(g_j^{opt})\} \geq E\{R_j^*(g_j)\}$ for all $g_j \in \mathscr{G}_j$, where $\mathscr{G}_j$ is the class of all potential regimes at stage j.

For stage $j < T$, we define a stage-specific APP-pseudo-outcome as $F_j(A_1, \ldots, A_{j-1}, A_j, g_{j+1}^{opt}, \ldots, g_T^{opt})$, and similarly $S_j(A_1, \ldots, A_{j-1}, A_j, g_{j+1}^{opt}, \ldots, g_T^{opt})$. Note that the difference between the APP-pseudo-outcome and the future optimized counterfactual outcome defined previously is that the future optimized counterfactual outcome assumes the patient takes specific treatment a_j during stage j, while the APP-pseudo-outcome is the observed data equivalent of F_T but future optimized for stages after stage j. However, note that the future optimized component makes the expectation of APP-pseudo-outcome an estimable but unobserved entity.

Again we make the following assumptions to link observed data to their counterfactual versions:

1. Consistency:

$$\sum_{a_j=1}^{K_j} F_k^*(a_j) I\{A_j = a_j\} = \begin{cases} F_k & \text{when } k = j \\ F_k(A_1, \ldots, A_j, g_{j+1}^{opt}, \ldots, g_k^{opt}) & \text{when } k > j \end{cases}$$

$$\sum_{a_j=1}^{K_j} S_k^*(a_j) I\{A_j = a_j\} = \begin{cases} S_k & \text{when } k = j \\ S_k(A_1, \ldots, A_j, g_{j+1}^{opt}, \ldots, g_k^{opt}) & \text{when } k > j \end{cases}$$

2. No unmeasured confounding: $\{F_k^*(1), \ldots, F_k^*(K_j)\} \perp A_j | \overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j$ and $\{S_k^*(1), \ldots, S_k^*(K_j)\} \perp A_j | \overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j$, where $k \geq j$. Furthermore, $\{\overline{\mathbf{X}}_{j+1}, \ldots, \overline{\mathbf{X}}_T, \overline{\mathbf{W}}_{j+1}, \ldots, \overline{\mathbf{W}}_T\} \perp A_j | \overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j$

3. Positivity: $\pi_{a_j}(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j) = Pr(A_j = a_j | \overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)$ is bounded away from zero and one, where $k \geq j$

4. Latent variable independence: $H_k \perp (A_j, F_k^*(a_j), S_k^*(a_j)) | \overline{\mathbf{X}}_k, \overline{\mathbf{W}}_k$, where $k \geq j$

Combining all these above,

$$
\begin{aligned}
&E[R_j^*(g_j)] = \\
&\quad \mathbb{E}_j\left[\sum_{a_j=1}^{K_j}\Big[E\left\{\Phi(H_j)|\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j\right\}\mu_{j,a_j}^F(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j) + E\left\{1-\Phi(H_j)|\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j\right\}\mu_{j,a_j}^S(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)\Big]I\{g_j = a_j\}\right] \\
&+\sum_{t=j+1}^{T}\mathbb{E}_t\left[\sum_{a_j=1}^{K_j}\Big[E\{\Phi(H_t)|\overline{\mathbf{X}}_t, \overline{\mathbf{W}}_t\}\mu_{t,a_j}^F(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j) + E\{1-\Phi(H_t)|\overline{\mathbf{X}}_t, \overline{\mathbf{W}}_t\}\mu_{t,a_j}^S(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)\Big]I\{g_j = a_j\}\right],
\end{aligned}
\tag{3.2}
$$

where $\mu_{j,a_j}^F(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j) = E(F_j|A_j = a_j, \overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)$ and $\mu_{t,a_j}^F(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)$ denotes $E\left[F_t(A_1, \ldots, A_j, g_{j+1}^{opt}, \ldots, g_t^{opt})|A_j = a_j, \overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j\right]$ (equivalently for $\mu_{j,a_j}^S(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)$ and $\mu_{t,a_j}^S(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)$). Notice again that the RHS can be estimated from observed data only and is a combination of separately estimated preference and outcomes. Under these assumptions, the optimization problem at stage j, among all potential regimes $\mathscr{G}_j$, can be written as $g_j^{opt} = arg\ max_{g_j \in \mathscr{G}_j}$ RHS of Eqn (3.2).

3.3.2 APP-RL to solve observational data based optimization

Our proposed method, named Augmented Patient Preference incorporated Reinforcement Learning (APP-RL), is summarized as follows. Through iteration and sequential estimation, APP-RL is able to combine elements of patient preference and observed outcomes to get the optimal decision rules for each stage. APP-RL first uses survey information from each stage to estimate patient preferences through an expectation maximization algorithm. APP-RL then combines estimated preferences with observed outcomes of toxicity and side effects into previously mentioned utility function and uses a tree-based reinforcement learning method to find the optimal decision rule for that stage. Finally, APP-RL moves backwards through the stages, obtaining optimal decision rule first for stage T, next for stage $T-1$, etc, and lastly for stage 1. Estimated decision rules for j-th stage are used in the estimation of earlier $(j-1)$ stage in a backward induction manner.

Although the tree-based reinforcement learning component is similar to traditional CART methods, there are important differences that separate the two. Traditional CART methods are supervised learning methods that repeatedly split a parent node into child nodes, generally resulting in purer (fewer misclassification) nodes. Commonly used purity measures include the Gini index, information gain, and least squares deviation (*Breiman*, 2017). In this framework, each observation carries a label, and the goal of the CART method is to use covariates to correctly classify each subject with its observed label. In contrast, the estimation target of dynamic treatment regime problem, the optimal treatment, is not observed (a patient often does not get the most optimal treatment). Rather, the optimal treatment needs to be inferred indirectly from other patients' treatments and response trajectories. Instead of trying to classify correctly each patient to their assigned treatment, our goal is to optimize the counterfactual mean outcome for the every patient. Aligned with this goal, we propose to use an augmented inverse probability weighted (AIPW) estimator for the counterfactual outcome to be used within our purity measure.

3.3.2.1 APP-AIPW estimators

We assume that the K treatment options are of arbitrary missing data patterns. Our APP-AIPW estimator based off of the AIPW estimator (*Rotnitzky et al.*, 1998), a doubly robust and consistent estimator which in our case takes into account both observed outcome and estimated patient preference.

For stage T, the APP-AIPW estimator for $E\{R_T^*(g_T)\} =$ $\mathbb{P}_n\left[\frac{I(A_T=g_T)}{\hat{\pi}_{T,A_T}(\overline{\mathbf{X}}_T,\overline{\mathbf{W}}_T)}R_T + \{1 - \frac{I(A_T=g_T)}{\hat{\pi}_{T,A_T}(\overline{\mathbf{X}}_T,\overline{\mathbf{W}}_T)}\}\hat{\mu}_{T,g_T}(\overline{\mathbf{X}}_T,\overline{\mathbf{W}}_T)\right]$, where $R_T = \Phi(\hat{h}_T)F_T + \{1 - \Phi(\hat{h}_T)\}S_T$, where $\hat{h}_T$ is the estimated preference. $\hat{\mu}_{T,g_T}(\overline{\mathbf{X}}_T,\overline{\mathbf{W}}_T)$ can be any model of R_T as a function of observed covariates and survey data accumulated by stage T. Under previously listed causal assumptions, if either the propensity model $\pi_{T,A_T}(\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T)$ or the conditional model $\mu_{T,g_T}(\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T)$ are correctly specified, then this estimator is a consistent estimator for $E\{R_T^*(g_T)\}$.

For stage $j < T$, the stage-specific APP-pseudo-outcome analog of the stage j reward is $\widetilde{PO}_j = R_j(A_1, \ldots, A_{j-1}, A_j, g_{j+1}^{opt}, \ldots, g_T^{opt})$, which is a weighted combination of stage-specific pseudo-outcomes summed across stages j to T. In this formulation, we add the observed value of stage j (including an estimated value of h_j), and follow the convention in *Huang et al.* (2015), where instead of only using the model-based values under optimal future treatments, we use actual observed outcomes adjusted by expected future loss due to non-optimal treatment. This approach prevents bias accumulation from stage to stage as compared to only using model based estimates.

We can estimate the APP-pseudo-outcome recursively as

$$\widetilde{PO}_j = U(F_j, S_j; h_j) + \sum_{t=j+1}^{T} \{U(F_t, S_t; h_t) + \hat{\mu}_{t,g_t^{opt}}(\overline{\mathbf{X}}_t, \overline{\mathbf{W}}_t) - \hat{\mu}_{t,A_t}(\overline{\mathbf{X}}_t, \overline{\mathbf{W}}_t)\} \quad (3.3)$$

where $\hat{\mu}_{t,g_t^{opt}}(\overline{\mathbf{X}}_t, \overline{\mathbf{W}}_t) - \hat{\mu}_{t,A_t}(\overline{\mathbf{X}}_t, \overline{\mathbf{W}}_t)$ is the expected cumulative loss from stage j onwards of not following the optimal regime during stage t. Both $\hat{\mu}_{t,g_t^{opt}}(\overline{\mathbf{X}}_t, \overline{\mathbf{W}}_t)$ and $\hat{\mu}_{t,A_t}(\overline{\mathbf{X}}_t, \overline{\mathbf{W}}_t)$ can come from the same prediction model, which can take many forms, commonly parametric regression or random forests. Then, the proposed APP-AIPW estimator for $E\{R_j^*(g_j)\}$ is $\mathbb{P}_n\left[\frac{I(A_j=g_j)}{\hat{\pi}_{j,A_j}(\overline{\mathbf{X}}_j,\overline{\mathbf{W}}_j)}\widetilde{PO}_j + \{1 - \frac{I(A_j=g_j)}{\hat{\pi}_{j,A_j}(\overline{\mathbf{X}}_j,\overline{\mathbf{W}}_j)}\}\hat{\mu}_{j,g_j}(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)\right]$, where $\widetilde{PO}_j$ takes the place of R_T from stage T and $\hat{\mu}_{j,g_j}(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)$ can be any model of $\widetilde{PO}_j$ that uses accumulated information up to stage j.

The propensity score $\pi_{j,a_j}(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)$ for all stages j can be estimated via multinomial logistic regression. In the above estimating equations, each R_j requires an estimated h_j. Although for the sake of generality we suggest that π and μ functions can be functions of both $\overline{\mathbf{X}}_j$ and $\overline{\mathbf{W}}_j$, we envision that $\overline{\mathbf{W}}_j$ will mostly contribute through its effect h_j. We will discuss our proposed method to estimate h_j in the next subsection.

3.3.2.2 Implementation to obtain APP-weights

We assume going forward that information given by $\mathbf{W}_j$ will dominate information obtained from previous surveys (i.e. $\overline{\mathbf{W}}_{j-1}$), and other covariate information ($\overline{\mathbf{X}}_j$). Hence, following *Butler et al.* (2018), we assume here that $H_j \perp \overline{\mathbf{X}}_j|\mathbf{W}_j$. This assumption can be weakened at the sake of more complicated models, which would be more burdensome to implement but follow an identical approach. Furthermore, we assume a latent traits model (*Moustaki and Knott*, 2000) and that latent patient preferences are connected to items on the questionnaire through modified Rasch model (*Rasch* (1961), *Rasch* (1960)).

We assume the underlying generating form for a binary response is $logit\{P(W_{jl} = 1|H_j = h_j)\} = \alpha_{l0} + \alpha_{l1}h_j$ where j is for stage and l for the question number. If we wanted to relax the assumption that H_j does not depend on $\overline{\mathbf{X}}_j$ given $\overline{\mathbf{W}}_j$, we can for example use the model $logit\{P(W_{jl} = 1|H_j = h_j, \mathbf{X}_j = \mathbf{x}_j)\} = \alpha_{l0} + \alpha_{l1}h_j + \boldsymbol{\gamma}_l^T\mathbf{x}_j$, giving explicit dependence on covariate information. For cases where questions had more than two possible responses (e.g. three per question), assuming the first category is the reference (i.e. coefficients are 0), we use the generating model $\log(\pi_{lb}(h_j)/\pi_{la}(h_j)) = \alpha_{lb0} + \alpha_{lb1}h_j$, and $\log(\pi_{lc}(h_j)/\pi_{la}(h_j)) = \alpha_{lc0} + \alpha_{lc1}h_j$.

Algorithm 1: EM algorithm for estimating patient preference $\hat{h}_j$

```
Result: Obtain $p(h_j|\mathbf{w}_j)$ for patient $i$
Guess initial value of $h_j$ for all subjects to estimate an initial guess of $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1$;
while not reached convergence do
    Using MH, get an updated estimate of $p(h_j|\mathbf{w}_j)$;
    Approximate likelihood integral using Gauss-Hermite quadrature with $k$ abscissae
      $h_t$ and weights $p(h_t)$;
    Solve likelihood equations using Newton-Raphson to get updated estimates of
      $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1$;
end
```

Algorithm 1 outlines the algorithm for estimating patient preference $\hat{h}_j$. The APP-weights we propose are $\Phi(\hat{h}_j)$. Essentially, the Expectation-Maximization algorithm (*Moon*, 1996) iterates between estimates of $\boldsymbol{\alpha}$, the questionnaire coefficients, and h_j, individual patient preferences at stage j. In the process of derivation, we will use Gauss-Hermite quadrature to approximate the integral numerically, and estimate $P(h_j|\mathbf{W}_j) \propto P(\mathbf{W}_j|h_j)P(h_j)$ through the Metropolis Hastings algorithm.

3.3.2.3 APP-purity measure

The APP-purity measure that we propose is the following: $\mathcal{P}_j(\Omega, \omega) = max_{a_1,a_2 \in \mathcal{A}_j} \mathbb{P}_n \left[\sum_{a_j=1}^{K_j} \hat{\mu}_{j,a_j}^{AIPW}(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j) I\{g_{j,\omega,a_1,a_2}(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j) = a_j\} I\{\overline{\mathbf{X}}_j \in \Omega\} \right]$. The node Ω here is the space that divides each individual in our dataset, and can be a factor of all observed data (in our case, Ω will depend on patient covariates $\overline{\mathbf{X}}$). For a given partition ω and ω^c of node Ω, g_{j,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 at stage $j(T \leq j \leq 1)$. Then, the two treatments a_1 and a_2 that yield the largest purity measure are selected to constitute the purity measure. Finally, the APP-RL algorithm uses this purity measure at each node to decide whether to split the tree.

3.4 Algorithmic Implementation

As previously mentioned, $\mathcal{P}_j(\Omega, \omega)$ is the APP-purity measure of a potential split assigning treatment a_1 to patients in ω, and a_2 to patients in ω^c. Equivalently, $\mathcal{P}_j(\Omega)$ is the APP-purity measure when everyone in the node is assigned the single best treatment. The difference between $\mathcal{P}_j(\Omega, \omega)$ and $\mathcal{P}_j(\Omega)$ will provide primary guidance on if and how the node of the tree should split.

To prevent overfitting, λ is given to represent threshold for practical significance, and n_0 is given as minimal node size. The choice of λ can be obtained through cross-validation or using domain knowledge, and n_0 could be selected a priori.

Under this set-up, we propose the following stopping rules:

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

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

3. If maximum purity improvement $\mathcal{P}_j(\Omega, \hat{w}^{opt}) - \mathcal{P}_j(\Omega)$ is less than λ, where $\hat{w}^{opt} = arg\ max_\omega \left[\mathcal{P}_j(\Omega, \omega) : min\{n\mathbb{P}_n I(\overline{\mathbf{X}}_j \in \omega), n\mathbb{P}_n I(\overline{\mathbf{X}}_j \in \omega^c)\} \geq n_0\right]$, the node will not be split.

4. If the current tree depth reaches the user-specified maximum depth, the tree growing process will stop

5. Finally, if none of the stopping rules were triggered, split Ω into ω and ω^c.

This process is repeated at each node Ω, until all of the potential nodes are terminated by a stopping rule. Note that $\mathcal{P}_j(\Omega)$ at each terminal node is also the expected counterfactual utility outcome (or expected APP-pseudo-outcome for stage $j < T$), which takes into account our chosen utility, observed toxicity and side effects, weighted by patient preference. The final tree will therefore use patient characteristics to assign each patient to a terminal node, which will determine their optimal stage specific decision that maximizes the preference weighted counterfactual utility outcome.

Algorithm 2: $APP - RL$ Algorithm Implementation

```
Result: $\mathbf{g^{opt}} = (g_1^{opt}, \ldots, g_T^{opt})$
Initialize stage $j = T$;
while Stage $j \geq 1$ do
    Estimate $\hat{h}_j$ for each patient from $\mathbf{W}_j$ (all patients) using EM (See Alg. 1);
    if $j = T$ then
        Combine $\hat{h}_T, F_T, S_T$ into $R_T$ ;
        Obtain $\hat{\pi}_{T,a_T}(\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T), \hat{\mu}_{T,g_T}(\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T)$ and combine with $R_T$ to obtain
         $\hat{\mu}_{T,a_T}^{AIPW}(\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T)$ ;
        Set $m = 1$ at root node $\Omega_{T,m}$ ;
        At node $\Omega_{T,m}$, evaluate the Stopping Rules. If stop, assign the best treatment
         $arg\max_{a_T \in \mathcal{A}_T} \mathbb{P}_n \left[\hat{\mu}_{T,a_T}^{AIPW}(\overline{\mathbf{X}}_T, \overline{\mathbf{W}}_T) I\{\overline{\mathbf{X}}_T \in \Omega_{T,m}\}\right]$. Otherwise, split $\Omega_{T,m}$
         into child nodes $\Omega_{T,2m}$ and $\Omega_{T,2m+1}$ by $\hat{\omega}^{opt}$ ;
        Set $m = m + 1$ and repeat until all nodes are terminal → Obtain $g_T^{opt}$ ;
    else
        Combine $\hat{h}_j, F_j, S_j$ into $R_j$ ;
        Estimate $\hat{\pi}_{j,a_j}(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)$. Using $g_{j+1}^{opt}, \ldots, g_T^{opt}$ estimated previously, estimate
         $\hat{\mu}_{j,g_j}(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)$ and $\widetilde{PO}_j$. Combine together to obtain $\hat{\mu}_{j,a_j}^{AIPW}(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j)$ ;
        Set $m = 1$ at root node $\Omega_{j,m}$ ;
        At node $\Omega_{j,m}$, evaluate the Stopping Rules. If stop, assign the best treatment
         $arg\max_{a_j \in \mathcal{A}_j} \mathbb{P}_n \left[\hat{\mu}_{j,a_j}^{AIPW}(\overline{\mathbf{X}}_j, \overline{\mathbf{W}}_j) I\{\overline{\mathbf{X}}_j \in \Omega_{j,m}\}\right]$. Otherwise, split $\Omega_{j,m}$ into
         child nodes $\Omega_{j,2m}$ and $\Omega_{j,2m+1}$ by $\hat{\omega}^{opt}$ ;
        Set $m = m + 1$ and repeat until all nodes are terminal → Obtain $g_j^{opt}$ ;
    end
    Set $j = j - 1$;
end
```

Algorithm 2 provides a pseudo-code schematic for the layout of the entire algorithm. As we can see, the algorithm follows a backward induction strategy. At each stage, the algorithm updates patient preference estimates to those most recent and combines with observed outcomes of that stage into an APP-AIPW estimator of the expected counterfactual reward for each stage. The APP-AIPW estimator is then fed into the tree-based reinforcement learning algorithm to get the stage specific optimal rule.

3.5 Numerical Demonstration

We conduct simulation studies to investigate the performance of our proposed method. We first consider a single-stage scenario with two treatments to facilitate comparison with *Butler et al.* (2018) in Section 3.5.1. Then, we also simulate a one-stage with three treatments in Section 3.5.2 to evaluate how sensitive the performance of our proposed method depends on the number of treatment options, and a multi-stage with three treatments per stage in Section 3.5.3 to assess the performance of the proposed method in multiple stage case. For all scenarios, we generate five independent baseline covariates $X_1, \ldots, X_5 \sim N(0, 1)$.

We simulate questionnaire responses as well as efficacy and side-effect outcomes for each individual. We then estimate both the patient preferences and optimal dynamic treatment from these responses and outcomes. For one stage, we obtain the percentage of subjects correctly classified to their true optimal treatment as $\%opt$. For multi-stage, $\%opt$ represents the percentage of subjects correctly classified to their true optimal regime (correct treatment for all stages).

3.5.1 Scenario 1: $T = 1$ and $K = 2$

In Scenario 1, we consider one-stage, two treatment options, and sample sizes of 300, 500, and 1000. The observed treatment A was generated from a $Bernoulli(\pi)$, where $\pi = \exp(0.5X_1 + 0.5X_4)/(1 + \exp(0.5X_1 + 0.5X_4))$. The true underlying optimal rule is defined by:

$$g^{opt}(\mathbf{H}) \begin{cases} 1, & X_1 > -0.25, X_2 \leq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

We generate the observed outcomes for efficacy and side effect (F and S, respectively) as $F = 1+X_4+qf*(g^{opt} == A)+\epsilon_1$, and $S = 1+X_5+qs*(g^{opt} == A)+\epsilon_2$. To ensure that the distributions of F and S are similar, a $Bernoulli(0.5)$ random variable was used to determine whether qs or qf would be simulated first and set to $3X_3$. The latter was set to $(-\Phi(h)*qf/(1-\Phi(h)))+(1/(1-\Phi(h)))$ or $((\Phi(h)-1)*qs/\Phi(h)+(1/\Phi(h))$ for qs and qf, respectively, where $\Phi(\cdot)$ is the cumulative density distribution of a standard normal variable. This complicated form for qf, qs simply ensures that the overall reward $R = \Phi(h)F + \{1 - \Phi(h)\}S$ is greater by 1 when the correct treatment is chosen. The concordance between optimal regime for F and S was approximately $25-30\%$, indicating that for over 75% of cases, there is a trade-off between side-effect and efficacy.

We looked at the performance of our method under two different patient preference distribution settings. Patient preferences were generated from either the $Uniform(-1, 1)$ or $0.5Normal(X_2, SD = 0.2)$ distribution, where preference depended on an underlying covariate. Each patient answered a questionnaire of ten questions, and each question allowed a binary $\{0, 1\}$ response. For each question l we assume a latent traits model, where $W_{jl} \sim Bernoulli\{expit(\alpha_{l0} + \alpha_{l1}H_j)\}$. The coefficients for each of the questions for Scenario 1 is below:

In calculating the APP-purity measure, we specified the $\pi-$model with covariates X_1 and X_4 used to assign the treatment. Similarly, we specified the $\mu-$model in the augmentation term using a linear regression framework with $R = \Phi(\hat{h})F+\{1-\Phi(\hat{h})\}S$ as the outcome, and treatment indicator A, patient characteristics $X_1, \ldots, X_5$ and

Table 3.1: Scenario 1 questionnaire coefficients for the latent traits model

l	α_0	α_1
1	-0.25	0.92
2	-0.84	0.64
3	-1.64	2.35
4	-0.78	0.82
5	-0.89	1.10
6	0.42	0.16
7	1.27	2.96
8	-0.61	0.56
9	0.09	0.30
10	-0.40	1.35

their interaction terms with A as covariates. Table 3.2 shows results of simulation for the methods in Scenario 1. We compared our method against that proposed in *Butler et al.* (2018). As suggested in their paper, we fit linear working models with all covariates, questionnaire responses, and their interaction terms with treatment. For $n = 300$ and when h followed an uniform distribution, APP-RL method was able to select 87.1% of optimal treatment, which increased to 96.6% when $n = 1000$. In contrast, the Q-learning method lags behind in accuracy by over 15%, at 70.6% and 77.3% respectively. However, both methods saw reduction in standard errors as sample size increased. The patterns are similar when patient preference was generated with a normal distribution.

3.5.2 Scenario 2: $T = 1$ and $K = 3$

In this scenario, we consider one-stage with three treatment options. We again investigate sample sizes of 300, 500, and 1000. In this case, treatment A could take values in $\{0, 1, 2\}$ generated from $Multinomial(\boldsymbol{\pi})$, where $\boldsymbol{\pi} = [\pi_0, \pi_1, \pi_2] = [1/(1+\exp(0.5X_4+0.5X_1)+\exp(0.5X_5-0.5X_1)), \exp(0.5X_4+0.5X_1)/(1+\exp(0.5X_4+0.5X_1) + \exp(0.5X_5 - 0.5X_1)), 1 - \pi_0 - \pi_1]$.

Table 3.2: % optimal chosen for 1-stage, 2 treatment, binary responses, 200 iterations. Setting 1 refers to $h \sim U(-1, 1)$, and Setting 2 refers to $h \sim 0.5N(X_2, SD = 0.2)$

		n=300	n=500	n=1000
		% opt (sd)	% opt (sd)	% opt (sd)
Setting 1	APP-RL	87.06 (9.70)	92.65 (6.88)	96.56 (4.76)
	Q-learning	70.64 (4.12)	73.75 (3.63)	77.26 (2.20)
Setting 2	APP-RL	87.86 (10.45)	90.65 (8.57)	96.44 (5.35)
	Q-learning	71.12 (5.62)	73.71 (4.49)	75.82 (3.74)

APP-RL: Augmented Patient Preference incorporated Reinforcement Learning. Q-learning refers to the method by *Butler et al.* (2018)

The true underlying optimal rule is defined by:

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

F and S were generated in the same way as in Scenario 1, but the first of qf or qs was set to $2X_3 - (1.25X_2)^2$. The second variable (qs/qf) was generated similarly to that of stage 1 to ensure an advantage of 1.5 on the reward if the optimal treatment was chosen.

Patient preferences were likewise generated from either an uniform $Uniform(-1, 1)$ or a $0.5Normal(X_2, SD = 0.2)$ distribution, and each patient answered ten questions with three categorical responses in a survey. For each question l, assuming the first category is the reference (i.e. coefficients are 0), we used the generating model $\log(\pi_{lb}(h)/\pi_{la}(h)) = \alpha_{lb0} + \alpha_{lb1}h$, and $\log(\pi_{lc}(h)/\pi_{la}(h)) = \alpha_{lc0} + \alpha_{lc1}h$. Hence, the vector of probabilities we feed into a multinomial random generator to generate $\mathbf{W}_l$

is:

$\boldsymbol{\pi}_l(h) = [\pi_{la}(h), \pi_{lb}(h), \pi_{lc}(h)]$, where

$\pi_{la}(h) = 1/\{1 + \exp(\alpha_{lb0} + \alpha_{lb1}h) + \exp(\alpha_{lc0} + \alpha_{lc1}h)\}$,

$\pi_{lb}(h) = \{\exp(\alpha_{lb0} + \alpha_{lb1}h)\}/\{1 + \exp(\alpha_{lb0} + \alpha_{lb1}h) + \exp(\alpha_{lc0} + \alpha_{lc1}h)\}$, and

$\pi_{lc}(h) = \{\exp(\alpha_{lc0} + \alpha_{lc1}h)\}/\{1 + \exp(\alpha_{lb0} + \alpha_{lb1}h) + \exp(\alpha_{lc0} + \alpha_{lc1}h)\}$. For each person and each question, we generate $\mathbf{W}_l \sim Multinomial(\pi_{la}(h), \pi_{lb}(h), \pi_{lc}(h))$.

The exact coefficients for each of the questions for Scenario 2 is shown below:

Table 3.3: Scenario 2 questionnaire coefficients for the latent traits model

l	α_{b0}	α_{b1}	α_{c0}	α_{c1}
1	0.19	1.61	0.15	1.76
2	0.01	-0.51	-0.47	3.12
3	-0.77	2.87	0.38	2.49
4	0.68	1.08	0.83	1.74
5	-0.77	-0.30	0.74	0.84
6	-0.48	0.24	0.49	0.72
7	0.23	2.26	0.83	1.48
8	0.91	0.33	0.89	2.01
9	0.72	-0.12	-0.56	0.70
10	0.88	0.84	-0.30	0.66

We estimated the working models in an equivalent way to Scenario 1. Table 3.4 showcases results from our APP-RL method. Because the Q-learning based method of *Butler et al.* (2018) is unable to handle more than two treatments and more than two responses per question, it was not possible to directly compare with our method. In general, we see that the algorithm does better at selecting the optimal regime given a smaller range of patient preference (i.e. the uniform distribution), but the difference is small. We can also see that both accuracy and efficiency increases as sample size increases, with over 83% selected as optimal for $n = 300$, and over 95% for $n = 1000$.

Table 3.4: % optimal chosen for 1-stage, 3 treatment, 3-responses per question, 200 iterations.

	n=300	n=500	n=1000
	% opt (sd)	% opt (sd)	% opt (sd)
Setting 1	85.26 (13.72)	92.74 (7.81)	97.48 (4.47)
Setting 2	82.72 (15.79)	90.04 (11.39)	95.18 (7.94)

Setting 1 refers to $h \sim U(-1, 1)$, and Setting 2 refers to $h \sim 0.5N(X_2, SD = 0.2)$

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

In this scenario, we consider a two-stage set-up, with three treatment options at each stage and 3-category response per question on the survey. We simulated sample sizes of 1000 and 2000. The outcome to be maximized is the sum of expected rewards of each stage. As in scenario 2, treatment A_1 could take values in $\{0, 1, 2\}$ generated from $Multinomial(\boldsymbol{\pi})$, where $\boldsymbol{\pi} = [\pi_0, \pi_1, \pi_2] = [1/(1 + \exp(0.5X_4 + 0.5X_1) + \exp(0.5X_5 - 0.5X_1)), (\exp(0.5X_4+0.5X_1)/(1+\exp(0.5X_4+0.5X_1)+\exp(0.5X_5-0.5X_1)), 1-\pi_0-\pi_1]$.

For this scenario, we looked at two forms of true underlying rules, tree-type and non-tree type. The stage 1 tree-type optimal regime is defined by $g_1^{opt} = I(X_1 > -0.65)\{I(X_2 > -0.75) + I(X_2 > 0.1)\}$, while the non-tree type is defined by $g_1^{opt} = I(X_1 > -0.3)\{1 + I(X_1 + X_2 > 0.3)\}$. We also investigated performance under both equal (where selection of the optimal treatment guarantees a uniform reward advantage over the other two treatment options) and varying penalty setting (where selection of the optimal treatment has differing reward advantages as compared to the other treatments selected). For the equal penalty setting, $F_1 = 1 + X_4 + 0.7X_1 + qf \cdot I(A_1 \neq g_1^{opt}) + \epsilon_1$ and $S_1 = 1 + X_5 + 0.7X_1 + qs \cdot I(A_1 \neq g_1^{opt}) + \epsilon_2$. For the varying penalty scenario, $F_1 = 1 + X_4 + 1.3X_1 + qf \cdot \left|A_1 - g_1^{opt}\right| + \epsilon_1$ and $S_1 = 1 + X_5 + 1.3X_1 + qs \cdot \left|A_1 - g_1^{opt}\right| + \epsilon_2$. As before, the first of qs and qf to be simulated was set to $2X_3+(1+2X_5)^2$, and the latter was set to $-\phi(h_1)*qf/(1-\phi(h_1)-(2.25/(1-\phi(h_1))$

and $(\phi(h_1) - 1) * qs/\phi(h_1) - 2.25/\phi(h_1)$, for qs and qf respectively, where $\phi(\cdot)$ is the cumulative density distribution for a $N(0, SD = 3)$ random variable.

Stage 2 parameters followed the same pattern as those from stage 1. Treatment A_2 could take values in $\{0, 1, 2\}$ generated from $Multinomial(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = [\gamma_0, \gamma_1, \gamma_2] = [1/(1 + \exp(0.2U_1 - 0.5) + \exp(0.5X_2)), (\exp(0.2U_1 - 0.5)/(1 + \exp(0.2U_1 - 0.5) + \exp(0.5X_2)), 1 - \gamma_0 - \gamma_1]$, where U_1 is the utility of stage 1. The stage 2 tree-type optimal rule is defined by $g_2^{opt} = I(X_2 > -0.5)\{I(U_1 > -7) + I(U_1 > 0)\}$, while the non-tree type is defined by $g_1^{opt} = I(X_2 > 0.15)\{1 + I(X_2 + U_1 > 1.25)\}$. For the equal penalty setting, $F_2 = 1 + 1.5X_3 + qf \cdot 0.7I(A_2 \neq g_2^{opt}) + \epsilon_1$ and $S_2 = 1 + 1.5X_3 + qs \cdot 0.7I(A_2 \neq g_2^{opt}) + \epsilon_2$. For the varying penalty setting, $F_2 = 1 + 1.5X_3 + qf \cdot \left|A_2 - g_2^{opt}\right| + \epsilon_1$ and $S_2 = 1 + 1.5X_3 + qs \cdot \left|A_2 - g_2^{opt}\right| + \epsilon_2$. As before, the first of qs and qf to be simulated was set to $1.25X_4 + (1.7X_1)^2$, and the latter was set to $-\phi(h_2) * qf/(1 - \phi(h_2)) - (1.5/(1 - \phi(h_2))$ and $(\phi(h_2) - 1) * qs/(\phi(h_2)) - (1.5/(1 - \phi(h_2))$, for qs and qf respectively.

For both stages, the concordance between F and S was approximately $17 - 20\%$, indicating for vast majority of cases, the best treatment for maximizing efficiency was not the same as the best treatment for minimizing toxicity.

Patient preferences at stage 1 was generated from a $Normal(0, SD = 0.5)$ distribution for all subjects. Stage 2 preferences were generated from $0.3Normal(U_1, SD = 0.4)$ distribution, thereby assuming that stage 2 preferences are influenced by stage 1 outcomes and satisfaction levels. We assume the same latent model for questionnaire parameters as in scenario 2. The first stage coefficients are identical to those in scenario 2, and the stage 2 questionnaire coefficients are below:

Model specification for stage 2 involves specifying the π_2-model using $\hat{U}_1$ and X_2. The μ_2- model in the augmentation term also uses $\hat{U}_2$ as the outcome, and A_1, A_2 and

Table 3.5: Stage 2 questionnaire coefficients for the latent traits model

l	α_{b0}	α_{b1}	α_{c0}	α_{c1}
1	-0.09	1.10	-0.00	-0.28
2	-0.52	0.85	-1.64	-0.62
3	-1.69	1.76	-0.43	-0.52
4	1.86	1.74	0.45	1.20
5	-2.08	1.36	0.07	-2.95
6	2.18	1.32	0.52	1.01
7	0.73	1.38	-0.84	-3.07
8	-0.60	2.14	-1.93	-1.30
9	0.69	0.86	0.67	-1.62
10	-0.29	1.08	-1.81	-2.20

patient characteristics $X_1, \ldots, X_5, \hat{U}_1$, and the interaction terms between A_2 and all the rest of the terms as covariates. This set up allows us to obtain the optimal decision tree for stage 2. The π_1-model uses X_1, X_4, and X_5 as covariates. In calculating the pseudo-outcome $\widetilde{PO}_1 = \hat{U}_1 + \hat{U}_2 + \hat{\mu}_{2,g_2^{opt}}(\overline{\mathbf{X}}_2) - \hat{\mu}_{2,A_2}(\overline{\mathbf{X}}_2)$, we obtain both $\hat{\mu}_{2,g_2^{opt}}(\overline{\mathbf{X}}_2)$ and $\hat{\mu}_{2,A_2}(\overline{\mathbf{X}}_2)$ through predictions obtained through a random forest model with $A_1, A_2, X_1, \ldots, X_5, \hat{U}_1$ as covariates, and $\hat{U}_2$ as the outcome. Finally, the μ_1-model in the first stage augmentation term used $A_1, X_1, \ldots, X_5$ and the interaction terms between A_1 and the patient characteristics as covariates.

Figure 3.1 shows simulation results for each of the settings in this scenario. In all settings, our method was able to select the correct regimen (the correct treatment for both Stage 1 and for Stage 2) over 80% of the time, much higher than that could have been gotten by random chance (approximately $(1/3)^2 = 1/9$). Furthermore, we can see that when the underlying distribution is tree-based, the method does $\sim 10\%$ better than when the underlying optimal regime is not tree-based. Finally, we can see that generally increased sample size improves accuracy and reduces variation for tree-based distribution, but the improvement is less obvious with non-tree based distribution.

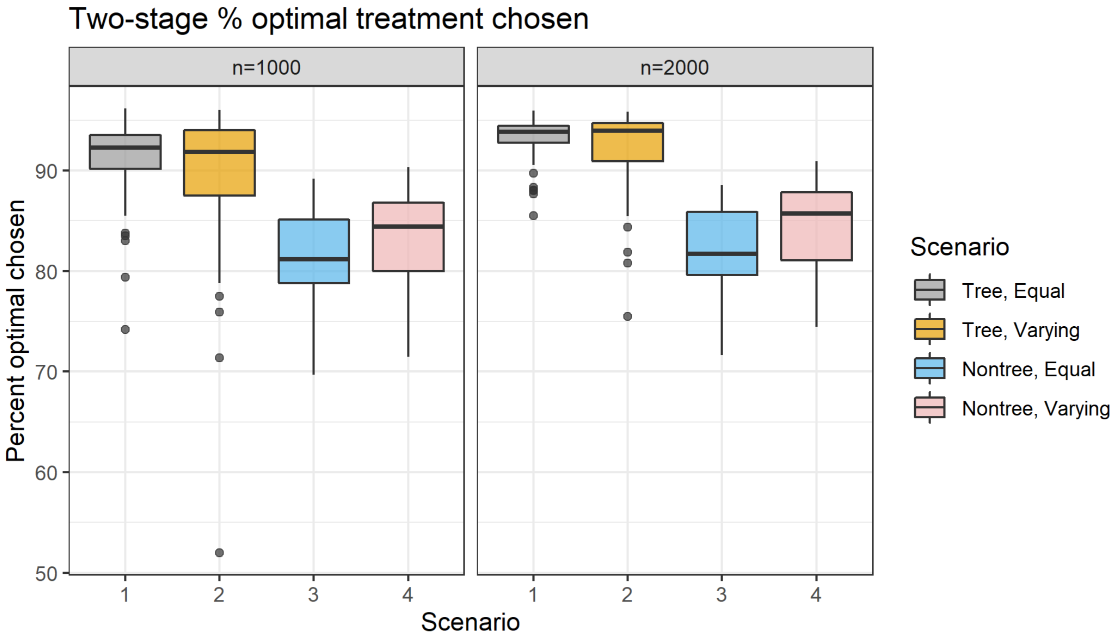

Figure 3.1: Results for 2-stage, 200 iterations. We looked at the case where the true underlying reward structure follows a tree and non-tree structure, as well as the equal and varying penalty settings for the reward.

Figure 3.2 shows the observed rewards as compared with predicted rewards, where a patient hypothetically follows the predicted optimal treatments obtained through our algorithm. Both rewards are calculated using true patient preferences but predicted optimal treatment assignments were obtained using estimated patient preferences. In aggregate across the four scenarios, the median difference between predicted optimal reward and observed true reward is 2.722, and over 85% of patients on average derived benefit from following the predicted optimal treatment as compared to their observed treatment.

3.6 Discussion

In this report, we propose a method that estimates an optimal dynamic treatment regime that maximizes a patient preferred utility function using a tree-based reinforcement learning approach. The vast majority of dynamic treatment regime methods

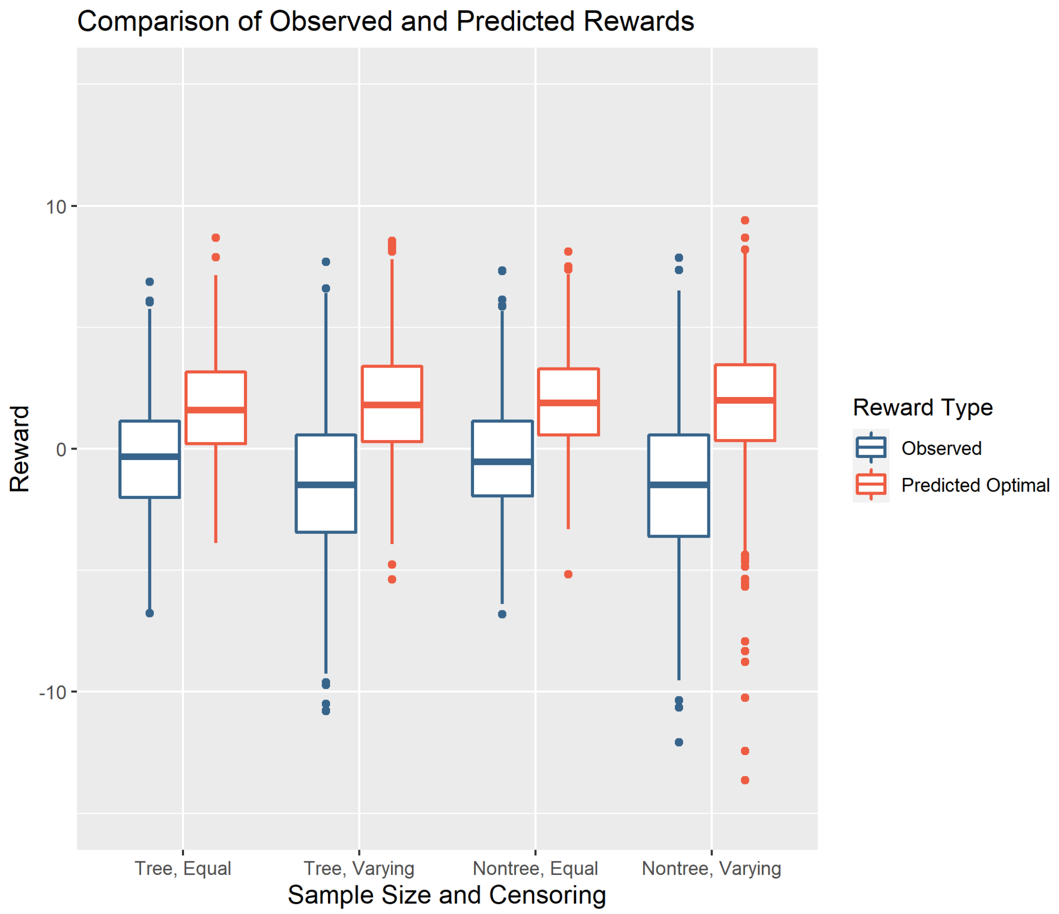

Figure 3.2: Comparison of observed cumulative reward vs predicted cumulative reward if following APP-RL's treatment predictions for the case of $n = 1000$. Both rewards are calculated using simulated true preferences.

in the literature optimize a single outcome. The main published method that the authors are aware of with an endpoint that incorporates patient preference is work by *Butler et al.* (2018). However, the authors wish to highlight a few key differences between the our proposed method and (*Butler et al.*, 2018). First, the method in *Butler et al.* (2018) is for single stage scenarios and is not designed to handle multiple stage decision making. Secondly, their method is designed to select between two potential treatments, which is clinically limiting. Thirdly, their method estimates patient preference from a survey of questions with binary choices, which we have extended to surveys with categorical choices. In summary, our method brings in patient preference incorporation into the chronic disease, multi-stage, multiple treatment option setting. Furthermore, incorporation of APP-AIPW into the purity measure endows our model with the doubly robust feature, providing a safety net against model misspecification.

As shown in Table 3.2, our algorithm has strong performance across the board for both one-stage, binary treatment settings and outperforms traditional Q-learning approaches when the underlying functional form for benefit follows a tree-based structure. Similarly, we show promising results in Table 3.4 for the one stage, three treatment options, where we obtain over 80% for all settings. In both of these tables, we can see that the prediction accuracy of APP-RL increases with sample size and its associated variance decreases with increasing sample size. For the two-stage scenario, Figure 3.1 shows that our method does best when the underlying distribution is tree-type, which is unsurprising given the algorithmic architecture. However, it still does respectively well even when the underlying distribution is not tree-type, indicating that our method is generally robust and applicable to more than one setting. In general, equal penalty seems to be associated with smaller variability, but the type of penalty seems to slightly influence the prediction accuracy of our method. In general, no matter the type of underlying distribution or type of penalty, the expected increase in rewards show similar patterns across the board, with the vast majority of cases gaining improved reward by following APP-RL's recommended treatment sequence. In summary, the numerical demonstration results indicate that APP-RL is a robust, efficient algorithm that is able to predict the optimal treatment in a myriad number of settings.

There are a number of potential improvements and extensions that we could explore in future studies. Generalizing our method and researching potential utility functions to accommodate more than two competing outcomes would be one improvement. Even more ambitious would be to move this augmentation patient preference framework into multi-objective optimization, where we could directly optimize in n-dimensional space. Instead of obtaining an unique solution, the goal would be to produce a set of non-dominated (where no one solution is better than the others in all ways) solutions. More precise and efficient ways of estimating patient preference would be of value, as the reliance on questionnaire and subsequent sampling and numerical methods are both labor and time intensive. Finally, the incorporation of continuous stages (i.e. mobile health interventions) which are heavily influenced by personal decisions would be of interest in this work, as it would bring together instantaneous preference estimation and decision making for more general and timely clinical scenarios.

CHAPTER IV

Augmented Patient Preference Incorporated Reinforcement Learning in Survival Settings

4.1 Introduction

For chronic illnesses, patients often have to navigate a series of treatment decisions. Increasingly, there is recognition that due to patient heterogeneities due to genetics, environmental factors, and various other factors and interplay between the factors, a good treatment plan needs to be both personalized and adaptive to a patient's changing clinical course. Dynamic treatment regimes (DTRs) are algorithmic solutions to this clinical problem. The treatment rules obtained through DTR algorithms adapt over time in response to an individual's response and trajectory.

A large number of methods have been proposed for the evaluation of the optimal DTR. Some of the earlier and foundational work include marginal structural model with inverse probability weighting (IPW) (*Robins*, 2000), G-estimation of structural nested mean models (*Robins*, 2004), Q-learning (*Murphy* (2003), *Murphy* (2005), *Moodie et al.* (2012)) and A-learning (*Schulte et al.*, 2014). More recently, along with the development of data science, machine learning flavored methods were also developed for DTR estimation, including tree-based and list-based methods (*Laber* *and Zhao* (2015), *Tao and Wang* (2017), *Tao et al.* (2018), *Zhang et al.* (2018)), and classification type methods (*Zhao et al.* (2012), *Zhang and Zhang* (2018)).

Despite the large number of methods that can be used to calculate the optimal DTR, the majority of methods rely on pre-specifying a single endpoint of interest. Often times, a single clinical decision can affect multiple outcomes, often in opposing directions of desirability. The classic example in this case is toxicity vs. efficacy. A newly introduced drug that is highly efficacious might come with a larger burden of undesirable side effects. In recent years, a few proposed methods have tackled the delicate balance between multiple outcomes of a proposed treatment. *Butler et al.* (2018) balances treatment efficacy and toxicity using patient derived preference using a Q-learning approach, while *Zhao et al.* (2009) assigned differing rewards based on survival status, wellness (a measure of toxicity) and tumor size (a measure of drug efficacy) at each stage.

The particular scenario that we would like to address in this work is a delicate balance of quality vs quantity, a dilemma commonly encountered for patients at the end of life (*Torrance and Feeny*, 1989). Patients often receive a first treatment and is followed up after a short time to determine if the treatment needs to be adjusted. The patient is then followed until death or a certain maximal follow-up time. Our primary goal is to estimate the optimal treatment regime that would maximize a patient preference weighted combination of quality of life and survival time. Secondly, we would like to provide an inference framework for more confident decision making. Although similar in flavor to the above-mentioned works, our scenario brings with it a unique set of distinct technical challenges.

The main challenge in this scenario is the presence of censored data. Because of the long tail of survival distributions, and because of other logistical reasons (patients move, dropout due to deteriorating health, etc), it is common to not observe the outcomes of a significant fraction of the population. However, partially observed information from censored subjects can still contribute important information and give power to the analysis if analyzed correctly.

An enormous body of work has been developed to estimate the optimal rule or regime in the presence of censoring. A non-exhaustive and overlapping list includes methods for single stage (*Cui et al.* (2017), *Zhu and Kosorok* (2012), *Zhao et al.* (2015)), methods for multiple stages (*Goldberg and Kosorok* (2012), *Hager et al.* (2018), *Zhao et al.* (2018), *Jiang et al.* (2017a)), inverse probability weighted censoring (IPCW) adjustment methods (*Goldberg and Kosorok* (2012), *Zhao et al.* (2018), *Zhao et al.* (2015)), Q-learning based methods (*Goldberg and Kosorok* (2012), *Zhao et al.* (2018)), tree-based methods (*Zhu and Kosorok* (2012), *Cui et al.* (2017)), survival probability based methods (*Jiang et al.* (2017a)), accelerated failure time (AFT) based models (*Huang and Ning* (2012), *Huang et al.* (2014)), and doubly robust methods (*Zhang and Schaubel* (2012), *Hager et al.* (2018), *Jiang et al.* (2017b)), just to name some of the numerous ways these methods differ in scope and direction.

A secondary but equally important goal of our method is to provide inference on stage specific parameters, particularly for tailoring variables. Inference in DTR methods is challenging due to the known issue of nonregularity caused by non-smooth functions that get carried forward through backward induction (*Chakraborty et al.*, 2010). When the degree of nonregularity is large (in other words, a larger fraction of covariate space doesn't have a treatment effect), the asymptotic distribution of the true coefficient oscillates between two asymptotic distributions, resulting in asymptotic bias and poor Wald-type confidence intervals. In the same paper, *Chakraborty et al.* (2010) proposed hard-threshold estimator and soft-threshold estimator to adjust for this poor coverage. *Laber et al.* (2014) proposed an adaptive confidence interval for first stage parameters by utilizing regular, uniformly convergent lower and upper bounds for the asymptotic distribution of interest, and later bootstrapping for the confidence set. *Chakraborty et al.* (2013) proposed an adaptive bootstrap based method that adjusts for the bias and coverage by adjusting the bootstrap sample size.

In this work, we explored, synthesized, and adapted existing methods in the literature to create a method for estimating optimal treatment regimen and stage-specific confidence intervals that fits our scenario. We utilize IPCW to enable a complete data analysis and used the Q-learning framework to estimate the optimal treatment regimen. We further adapted the m-out-of-n bootstrap to accommodate censoring in order to obtain the covariate specific confidence intervals for inference. We illustrate the performance of our method through simulation studies.

4.2 Set up and notation

We look at a two-stage setting where a patient, upon diagnosis, receives a stage 1 treatment. Shortly after at a scheduled follow-up time, the patient will be assessed for stage 2 treatment. Following stage 2 treatment, the patient is at risk of death, denoted by time D_i. Furthermore, patient information might be lost to follow-up, either due to administrative censoring (surpassed maximal follow-up time) or due to patient factors.

Let T_{1i}, T_{2i} denote the times of treatment for stage $1, 2$, respectively. Let τ be the maximal administrative follow-up time. Let S_{1i} and S_{2i} denote the amount of time survived in each stage, i.e. $S_{1i} = T_{2i} - T_{1i}$, and $S_{2i} = D_i - T_{2i}$. We assume that per protocol, everyone's S_{1i} should be the same (i.e. routine assessment following stage 1 treatment at a prespecified time interval). Let K_j be the number of treatment options in the j^{th} stage, $j = 1, 2$. Let A_{ij} denote the treatment indicator for i^{th} patient in the j^{th} stage, with a_{ij} denoting the observed treatment. Let $\mathbf{X}_{ij}$ denote patient characteristics prior to treatment assignment at stage j. In addition, we assume that

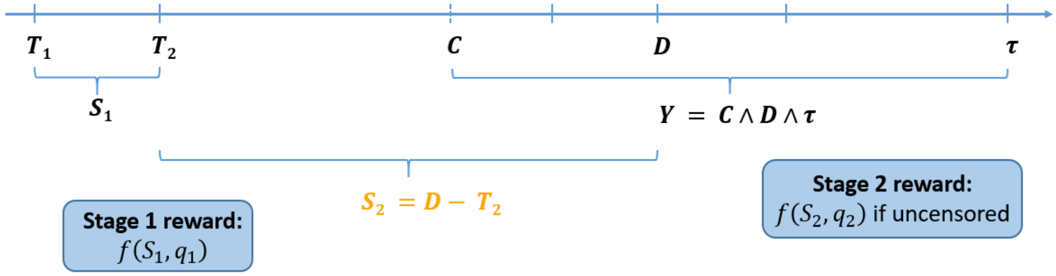
T_1
T_2
C
D
τ
S_1
$Y = C \wedge D \wedge \tau$
$S_2 = D - T_2$
Stage 1 reward: $f(S_1, q_1)$
Stage 2 reward: $f(S_2, q_2)$ if uncensored

Figure 4.1: Schematic of clinical progression timeline

each patient will have an evolving preference h_{ij}, which can be derived from answers $\mathbf{W}_{ij}$ to a questionnaire at stage j.

At each stage, we assume that each patient i will have two observed outcomes, q_{ij} for the average quality of life during stage j, and S_{ij}, the amount of time spent in stage j. q_{ij} allows us to calculate $Qu_{ij} = q_{ij} * S_{ij}$, the quality adjusted life years during stage j. However, since Qu_{ij} is dependent on S_{ij}, it is also subject to censoring at stage 2.

We assume a utility function of the form $Qu + \{1 - \Phi(h)\}|S - Qu|$, where $\Phi(\cdot)$ denotes the cumulative distribution function of a normal random variable. Intuitively, this utility function is a sliding scale between S and Qu, and a patient's preference would dictate where he/she would fall. Let the overall outcome of interest that we would like to optimize be $R_{1i} + R_{2i}$, where $R_{ij} = Qu_{ij} + \{1 - \Phi(h_{ij})\}|S_{ij} - Qu_{ij}|$. This is a cumulative preference adjusted quality of life years experienced on a given regime. If one's preference is such that $\Phi(h_i) = 0$ for both stages, then $R_{1i} + R_{2i}$ is the total survival time. Similarly, if a patient has preference $\Phi(h_i) = 1$ for both stages, then $R_{1i} + R_{2i}$ would be the total quality adjusted life years. In general, we denote all history, or history up to stage K for a given variable with an overhead bar (i.e., $\overline{\mathbf{W}}_i$ and $\overline{\mathbf{W}}_{iK}$, respectively).

Let $g_j(\overline{\mathbf{X}}_{ij}, \overline{\mathbf{W}}_{ij})$ be a function that maps from covariate and survey history to the domain of treatment assignment A_{ij}. At stage j, the expected potential reward of following decision rule $g_j(\overline{\mathbf{X}}_{ij}, \overline{\mathbf{W}}_{ij})$ for patient i is defined as
$\mathbb{E}\left[\sum_{a_j=1}^{K_j}\left[Qu^*_{ij}(a_j) + \{1 - \Phi(h_{ij})\}|S^*_{ij}(a_j) - Qu^*_{ij}(a_j)|\right] I\{g_j(\overline{\mathbf{X}}_{ij}, \overline{\mathbf{W}}_{ij}) = a_j\}\right]$, where $S^*_{ij}(a_j) = S^*_{ij}(A_{i1}, \ldots, A_{i,j-1}, a_j)$ and $Qu^*_{ij}(a_j) = Qu^*_{ij}(A_{i1}, \ldots, A_{i,j-1}, a_j)$ denotes the counterfactual survival outcome and quality adjusted survival, respectively, where the patient is assumed to have taken treatment a_j at stage j, conditional on previous treatment decisions $A_{i1}, \ldots, A_{i,j-1}$. In our case, our primary goal is to find a sequence of individualized decision rules, $\mathbf{g}(\overline{\mathbf{X}}_{\mathbf{i}}, \overline{\mathbf{W}}_{\mathbf{i}}) = (g_1(\overline{\mathbf{X}}_{i1}, \overline{\mathbf{W}}_{i1}), g_2(\overline{\mathbf{X}}_{i2}, \overline{\mathbf{W}}_{i2}))$, that optimize the potential outcome of $R_{1i} + R_{2i}$.

A second but equally important objective is to conduct inference on coefficients, with particular emphasis on tailoring variables (variables that interact with treatment selection). Inference on tailoring variables is important because it obviates the need to collect data for covariates that have no evidence of significant deviation from zero. Furthermore, inference allows us to know when there is insufficient evidence to support one treatment over another so that treatment decisions could be made using other factors important to the patient. In this work, along with stage-specific decision rules, we present a censoring adapted method of obtaining confidence intervals for the covariates in both stages of the model. For the sake of brevity going forward, let us abbreviate $g_j(\overline{\mathbf{X}}_{ij}, \overline{\mathbf{W}}_{ij})$ with g_j and drop the patient index i when there is no room for confusion.

4.3 Censoring adapted Q-learning

4.3.1 Traditional Q-learning

First, we introduce traditional Q-learning, a form of approximate dynamic programming originally proposed by *Murphy* (2003). Q-learning estimates the optimal DTR by postulating regression models for Q-functions and subsequently taking the solutions that would yield largest rewards. The Q-functions for the two stages are defined as:

$Q_2(\overline{\mathbf{X}}_2, A_2) = E[R_2|\overline{\mathbf{X}}_2, A_2]$

$Q_1(\overline{\mathbf{X}}_1, A_1) = E[R_1 + \max_{a_2} Q_2(\overline{\mathbf{X}}_2, a_2)|\overline{\mathbf{X}}_1, A_1]$

The decision rules can be written as $d_j(\overline{\mathbf{X}}_j) = \arg\max_{a_j} Q_j(\overline{\mathbf{X}}_j, a_j)$. Generally, we do not know the true Q-functions and so we consider linear working models for Q-functions of the form $Q_j(\overline{\mathbf{X}}_j, A_j; \boldsymbol{\beta}_j, \boldsymbol{\psi}_j) = \boldsymbol{\beta}_j^T \overline{\mathbf{Z}}_{j,0} + (\boldsymbol{\psi}_j^T \overline{\mathbf{Z}}_{j,1})A_j$, where $\overline{\mathbf{Z}}_{j,0}$ and $\overline{\mathbf{Z}}_{j,1}$ possibly contain different components of the history $\overline{\mathbf{X}}_j$.

The 2-stage Q-learning algorithm works as follows:

1. Stage 2 regression is obtained by
   $(\hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\psi}}_2) = \arg\min_{\beta_2, \psi_2} \sum_{i=1}^{n} (R_2 - Q_2(\overline{\mathbf{X}}_2, A_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2))^2$

2. Stage 1 pseudo-outcome is given by $\widetilde{PO}_1 = R_1 + \max_{a_2} Q_2(\overline{\mathbf{X}}_2, a_2; \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\psi}}_2), i = 1, \ldots, n$

3. Stage 1 regression: $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\psi}}_1) = \arg\min_{\boldsymbol{\beta}_1, \boldsymbol{\psi}_1} \sum_{i=1}^{n} (\widetilde{PO}_1 - Q_1(\overline{\mathbf{X}}_1, A_1; \boldsymbol{\beta}_1, \boldsymbol{\psi}_1))^2$

The decision rules can be simplified to be
$d_j(\overline{\mathbf{X}}_j) = \arg\max_{a_j} Q_j(\overline{\mathbf{X}}_j, a_j; \hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\psi}}_j) = \text{sign}(\hat{\boldsymbol{\psi}}_j^T \overline{\mathbf{Z}}_{j,1})$ when we have the particular case that $A_j \in \{-1, 1\}$. We will assume that we have binary treatment options for both stages for convenience, although this can be relaxed with further assumptions.

4.3.2 Censoring adapted Q-learning

Our stage 2 optimization objective is complicated by the fact that some S_2 may be unobserved due to censoring. Let C denote time of censoring, which started from time of stage 2 treatment. We assume that $S_2 \perp C|\overline{\mathbf{A}}_2, \overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2$ (conditional independence).

4.3.2.1 Stage 2

Let $S_2^*(a_2)$ be the counterfactual outcome of survival starting from stage 2 treatment conditional on previous treatment A_1. Correspondingly $S_2^*(g_2)$ is the counterfactual outcome under decision rule g_2, i.e. $S_2^*(g_2) = \sum_{a_2=1}^{K_2} S_2^*(a_2)I\{g_2(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2) = a_2\}$. Similarly, we can obtain $Qu_2^*(a_2)$ through $q_2^*(a_2)$ and $S_2^*(a_2)$.

Using the linear utility function defined in Section 4.2 and conditional on previous treatment A_1, $R_2^*(a_2) = Qu_2^*(a_2) + \{1 - \Phi(H_2)\}|S_2^*(a_2) - Qu_2^*(a_2)|$.

Correspondingly,

$$R_2^*(g_2) = Qu_2^*(g_2) + \{1 - \Phi(H_2)\}|S_2^*(g_2) - Qu_2^*(g_2)|$$

is the counterfactual utility, conditional on previous treatments A_1 under decision rule g_2.

The optimal regime, g_2^{opt}, satisfies $E\{R_2^*(g_2^{opt})\} \geq E\{R_2^*(g_2)\} \forall\ g_2 \in \mathscr{G}_2$, where $\mathscr{G}_2$ is the class of all potential decision rules for stage 2.

We make the following assumptions to connect the counterfactual outcomes with those observed in our data:

1. Consistency:
   $S_2 = \sum_{a_2=1}^{K_2} S_2^*(a_2)I\{A_2 = a_2\}$ and $q_2 = \sum_{a_2=1}^{K_2} q_2^*(a_2)I\{A_2 = a_2\}$

2. No unmeasured confounding:
   Treatment A_2 is randomly assigned with probability possibly dependent on $\overline{\mathbf{X}}_2$ and $\overline{\mathbf{W}}_2$, i.e., $\{S_2^*(1), \ldots, S_2^*(K_2)\} \perp A_2|\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2$ and $\{q_2^*(1), \ldots, q_2^*(K_2)\} \perp$

$A_2|\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2$

3. Positivity:

   There exists constants $0 < c_0 < c_1$ such that, with probability 1, the propensity score $\pi_{a_2}(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2) = Pr(A_2 = a_2|\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2) \in (c_0, c_1)$.

4. Latent variable independence:

   $H_2 \perp (\mathbf{A}_2, Qu_2^*(a_2), S_2^*(a_2))|\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2$

The first three assumptions are standard assumptions in causal inference. The last assumption facilitates separate modeling of outcomes and preferences and can be weakened at the expense of more complicated models (*Butler et al.*, 2018).

We denote the marginal expectation with respect to $\overline{\mathbf{X}}_t, \overline{\mathbf{W}}_t$ ($E_{\overline{\mathbf{X}}_t,\overline{\mathbf{W}}_t}$) as $\mathbb{E}_t$. Furthermore let us denote $\mu^S_{2,a_2}(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2) \equiv E\left\{S_2|A_2 = a_2, \overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2\right\}$ and $\mu^q_{2,a_2}(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2) \equiv E\left\{q_2|A_2 = a_2, \overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2\right\}$. As in traditional Q-learning, we assume linear working models for each of our outcomes of interest (i.e. q_2, S_2 can be generated through underlying models of predictive and tailoring variables $\boldsymbol{\beta}_2^T\mathbf{Z}_{20} + (\boldsymbol{\psi}_2^T\mathbf{Z}_{21})A_2$, where $\mathbf{Z}_{20}$ and $\mathbf{Z}_{21}$ are some possibly different components of $\overline{\mathbf{X}}_2$ and $\overline{\mathbf{W}}_2$).

Using causal assumptions above, we link observed data to their counterfactual outcomes:

$$\begin{aligned}E\left[R_2^*(g_2)\right] = \mathbb{E}_2\Bigg[\sum_{a_2\in\{-1,1\}} \Big[&\mu^q_{2,a_2}(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2)\mu^S_{2,a_2}(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2)\\ +E\left\{1-\Phi(H_2)|\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2\right\}|\mu^S_{2,a_2}(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2)-&\mu^q_{2,a_2}(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2)\mu^S_{2,a_2}(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2)|\Big]I\{g_2 = a_2\}\Bigg],\end{aligned} \tag{4.1}$$

where the separate modeling of preference and outcomes is allowed by the fourth assumption, $\mu^S_{2,a_2}(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2) = E(S_2 \,|A_2 = a_2, \overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2)$, and $\mu^q_{2,a_2}(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2) = E(q_2 \,|A_2 = a_2, \overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2)$.

With censoring, it is unlikely that all S_2 will be observed. We propose the following estimator that re-weights observed complete data using inverse probability of censoring weighting (IPCW):

$$\underset{\boldsymbol{\beta}_2, \boldsymbol{\psi}_2}{\operatorname{argmin}} \mathbb{P}_n \left(\left[R_2 - E(R_2(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2, A_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2))\right]^2 \frac{\Delta}{\widehat{Pr}\{\Delta = 1 | \overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2, A_2\}} \right),$$

where $E(R_2(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2, A_2; \boldsymbol{\beta}_2, \boldsymbol{\psi}_2))$ denotes the model estimate for R_2 using observed data and covariates. $\Delta = I(S_2 < C)$ is the event indicator and $\widehat{Pr}\{\Delta = 1 | \overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2, A_2\}$ is a working estimator of the probability that the individual has not been censored by their event time.

Denote our Q-function here to be $Q_2(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2, A_2) = E[R_2 | \overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2, A_2]$. The derived treatment rule is $\hat{g}^{opt}_2(\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2) = \operatorname{argmax}_{a_2} Q_2(\overline{\mathbf{W}}_2, \overline{\mathbf{X}}_2, A_2; \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\psi}}_2) = \operatorname{sign}(\hat{\boldsymbol{\psi}}_2^T \mathbf{Z}_{21})$, where $\mathbf{Z}_{21}$ represent tailoring variables in the stage 2 model.

4.3.2.2 Stage 1

$g^{opt}_1(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1)$ can be derived from the observed data using backward induction. Assuming that stage-specific rewards have been maximized after stage 1, we define the following stage 1 reward:

$R^*_1(a_1) = q^*_1(a_1)S^*_1(a_1) + \{1 - \boldsymbol{\Phi}(\mathbf{H}_1)\}|S^*_1(a_1) - q^*_1(a_1)S^*_1(a_1)| +$
$q^*_2(a_1)S^*_2(a_1) + \{1 - \boldsymbol{\Phi}(\mathbf{H}_2)\}|S^*_2(a_1) - q^*_2(a_1)S^*_2(a_1)|$. For stage 1, $S^*_1(a_1)$ is as defined previously and for stage 2, $S^*_2(a_1) = S^*_2(a_1, g^{opt}_2)$ denotes a counterfactual outcome given future optimized treatments and taking treatment a_1 at stage 1 (and similarly

for $q^*_{1/2}(a_1)$). The optimal regime at stage 1 satisfies $E\{R^*_1(g^{opt}_1)\} \geq E\{R^*_1(g_1)\}$ for all $g_1 \in \mathscr{G}_1$, where $\mathscr{G}_1$ is the class of all potential regimes at stage 1.

Again we make the following three standard assumptions to link observed data to their counterfactual versions (we use S_1 as an example, but we assume the same for q_1):

1. Consistency:

$$\sum_{a_1=1}^{K_1} S^*_k(a_1) I(A_1 = a_1) = \begin{cases} S_1 & \text{when } k = 1 \\ S_k(A_1, g^{opt}_2) & \text{when } k = 2 \end{cases}$$

2. No unmeasured confounding: $\{S^*_1(1), \ldots, S^*_1(K_1)\} \perp A_1 | \overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1$ and $\{S^*_2(1), \ldots, S^*_2(K_1)\} \perp A_1 | \overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2$. Furthermore, $\{\overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2\} \perp A_1 | \overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1$

3. Positivity: $\pi_{a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1) = Pr(A_1 = a_1 | \overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1)$ is bounded away from zero and one

4. Latent variable independence: $H_1 \perp (\mathbf{A}_1, S^*_k(a_1), Qu^*_k(a_1) | \overline{\mathbf{X}}_k, \overline{\mathbf{W}}_k$ where $k = 1, 2$

By linking counterfactual outcomes to those observed,

$$\begin{aligned} E[R^*_1(g_1)] = \mathbb{E}_1\Bigg[& \sum_{a_1 \in \{-1,1\}} \Big[\mu^q_{1,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1)\mu^S_{1,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1) + E\left\{1 - \Phi(H_1) | \overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1\right\} \times \\ & |\mu^S_{1,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1) - \mu^q_{1,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1)\mu^S_{1,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1)|\Big] I\{g_1 = a_1\}\Bigg] \\ + \mathbb{E}_2\Big[& \sum_{a_1 \in \{-1,1\}} \Big[\mu^q_{2,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1)\mu^S_{2,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1) + E\{1 - \Phi(H_2) | \overline{\mathbf{X}}_2, \overline{\mathbf{W}}_2\} \times \\ & |\mu^S_{2,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1) - \mu^q_{2,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1)\mu^S_{2,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1)|\Big] I\{g_1 = a_1\}\Big] \end{aligned} \tag{4.2}$$

where $\mu^S_{1,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1) = E(S_1|A_1 = a_1, \overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1)$, $\mu^S_{2,a_1}(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1)$ denotes $E\left[S_2(A_1, g_2^{opt})|A_1 = a_1, \overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1\right]$, and similarly for the equivalents of q. Notice again that the RHS can be completely estimated from observed data. Under these assumptions, the optimization problem at stage 1, among all potential regimes $\mathscr{G}_1$, can be written as $g_1^{opt} = arg\ max_{g_1 \in \mathscr{G}_1}$ RHS of Equation 4.3.2.2.

We maximize the stage 1 outcome through a pseudo-outcome defined as:

$$\widetilde{PO}_1 = q_1 S_1 + \{1 - \Phi(H_1)\}|S_1 - q_1 S_1| + \max_{a_2} Q_2(\overline{\mathbf{W}}_2, \overline{\mathbf{X}}_2, a_2; \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\psi}}_2)$$

Our proposed estimator for $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\psi}}_1)$ is $\text{argmin}_{\boldsymbol{\beta}_1, \psi_1} \mathbb{P}_n \left(\widetilde{PO}_1 - Q_1(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1, A_1; \boldsymbol{\beta}_1, \boldsymbol{\psi}_1)\right)^2$, where $Q_1(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1, A_1; \boldsymbol{\beta}_1, \boldsymbol{\psi}_1)$ can be modeled with $\boldsymbol{\beta}_1^T \mathbf{Z}_{10} + (\boldsymbol{\psi}_1^T \mathbf{Z}_{11}) A_1$. The first stage estimated optimal rule is given by $\hat{g}_1^{opt} = \text{argmax}_{a_1} Q_1(\overline{\mathbf{X}}_1, \overline{\mathbf{W}}_1, a_1; \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\psi}}_1) = \text{sign}(\hat{\boldsymbol{\psi}}_1^T \mathbf{Z}_{11})$, where $\mathbf{Z}_{11}$ represent tailoring variables in the stage 1 model.

4.3.3 Inference

In addition to obtaining the optimal stage-specific decisions, our second goal is to do inference on each stage's covariates. Particular emphasis was placed on tailoring variables. To that end, we propose using an censoring adjusted version of the m-out-of-n method presented by *Chakraborty et al.* (2013). Inference for dynamic treatment regime type problems is challenging due to the problem of non-regularity (*Chakraborty et al.*, 2010). In stage-1 optimization, the pseudo-outcome $\widetilde{PO}_1 = q_1 S_1 + \{1 - \Phi(H_1)\}|S_1 - q_1 S_1| + \hat{\boldsymbol{\beta}}_2^T \mathbf{Z}_{20} + |\hat{\boldsymbol{\psi}}_2^T \mathbf{Z}_{21}|$ is a nonsmooth function of $\hat{\boldsymbol{\psi}}_2$. In particular, if $P[\mathbf{Z}_{21} : \boldsymbol{\psi}_2^T \mathbf{Z}_{21} = 0] = 0$, then first stage covariates will converge to a normal distribution. However, if $P[\mathbf{Z}_{21} : \boldsymbol{\psi}_2^T \mathbf{Z}_{21} = 0] > 0$, the estimator $\hat{\boldsymbol{\psi}}_1$ oscillates between the two asymptotic distributions across samples (*Chakraborty et al.*, 2010). Hence, direct estimation results in asymptotically biased estimator and poor performance of usual Wald type confidence intervals. Even bootstrap-based approaches suffer from underlying nonsmoothness.

The m-out-of-n bootstrap was developed to address the bootstrap inconsistency due to nonsmoothness (*Shao*, 1994). Although conceptually very similar to the original bootstrap, the resample size m (which needs to depend on n, tends to infinity with n, and $m = o(n)$) is selected to be a smaller order than n. The general idea is that asymptotically, the empirical distribution can then tend to the true generative distribution at a faster rate than the bootstrap empirical distribution can tend to the empirical distribution. If the empirical distribution reaches its limit first, then bootstrapped empirical distributions would essentially be sampling from the true generative distribution. *Chakraborty et al.* (2013) showed through simulation studies that their m-out-of-n approach obtained desirable coverage probabilities for the two-stage DTR problem for first stage tailoring variables. Because censoring reduces the size of observed stage 2 data in our scenario, we further adapted the m-out-of-n algorithm to accommodate censoring. Our algorithm works as follows.

We adopted the functional form of m as presented in *Chakraborty et al.* (2013),

$$m = n^{\frac{1+\xi(1-\hat{p})}{1+\xi}} \tag{4.3}$$

Let n be the total number of subjects in the dataset (including those who were censored). For stage 2, we create a bootstrap sample of size n and fit a regression model using the complete data weighted by IPC-weights within the bootstrapped sample to obtain stage 2 coefficient estimates. Stage 2 95% confidence intervals are obtained after getting $\hat{l}_2$ and $\hat{u}_2$, the $\alpha/2$ ×100 and $(1-\alpha/2)$×100 percentiles of $\sqrt{n}(\hat{\theta}^{(b)}_{2,n} - \hat{\theta}_{2,n})$, where α is the desired significance level, $\hat{\theta}^{(b)}_{2,n}$ is the bootstrap estimate of stage 2 coefficients with bootstrap specific re-estimated censoring weights, and

$\hat{\theta}_{2,n}$ is the plug-in estimator obtained using weighted regression from the empirical dataset. The confidence interval is given by $(\hat{\theta}_{2,n} - \hat{u}_2/\sqrt{n}, \hat{\theta}_{2,n} - \hat{l}_2/\sqrt{n})$. For stage 1, we first generate bootstrap samples of size m, which is calculated using Equation 4.3 after calculating a sample specific $\hat{p}$. We further use each bootstrap sample to re-estimate IPC-weights, and fit a weighted lm model to obtain a bootstrap specific stage 2 estimate. The stage 2 coefficients from each bootstrap sample were then used to calculate pseudo-outcomes, which were then used to fit a stage 1 model to obtain $\hat{\theta}^{(b)}_{1,\hat{m}}$. As in stage 1, we obtain the $\hat{l}_1$ and $\hat{u}_1$, the $\alpha/2$ ×100 and $(1-\alpha/2)\times 100$ percentiles of $\sqrt{\hat{m}}(\hat{\theta}^{(b)}_{1,\hat{m}} - \hat{\theta}_{1,n})$, where $\hat{\theta}_{1,n}$ is the plug-in estimator obtained using the complete empirical dataset, while $\hat{\theta}^{(b)}_{1,\hat{m}}$ is the estimate obtained from each bootstrap sample of size m. The confidence set is given by $(\hat{\theta}_{1,n} - \hat{u}_1/\sqrt{\hat{m}}, \hat{\theta}_{1,n} - \hat{l}_1/\sqrt{\hat{m}})$.

We further selected the value of ξ to be 0.01, which provided stable coverage in simulations with complete data. The calculation of $\hat{p} = \mathbb{P}I\{n[\mathbf{Z}_{21}^T\hat{\psi}_{2,n}]^2 \leq \tau_n(\mathbf{Z}_{21})\}$ relies on a selection of $\tau_n(\mathbf{Z}_{21})$. We opted to use the plug-in estimator for $\tau_n(\mathbf{Z}_{21}) = (\mathbf{Z}_{21}^T\hat{\Sigma}_{21}\mathbf{Z}_{21}) \cdot \chi^2_{1,1-\nu}$ just as in *Chakraborty et al.* (2013), where $\hat{\Sigma}_{21}$ is the plug-in sandwich estimator of $nCov(\hat{\psi}_{2,n}, \hat{\psi}_{2,n})$, and $\nu = 0.01$.

4.3.4 SAPP-weights

Going forward, we assume that information in the lastest stage survey will override information from previous stages as well as other covariate information (i.e. $\mathbf{W}_j$ will override $\mathbf{W}_{j-1}$ and $\overline{\mathbf{X}}_j$). To model survey information as a function of latent preference, we assume a latent traits model (*Moustaki and Knott*, 2000). We further assume that the latent preferences are related to survey responses through a modified Rasch model (*Rasch*, 1961, *Rasch*, 1960).

For our scenario, we assume we have $numQ$ questions on a survey, each soliciting binary answer choices from the patient. For each binary response, we assume that the underlying generating form is of the form $logit\{P(W_{jk} = 1|H_j = h_j)\} = \alpha_{0,k} + \alpha_{1,k}h_j$ where j indicates the stage and k the question number.

```
Algorithm 3: EM algorithm for estimating patient preference $\hat{h}_j$
Result: Obtain $p(h_j|\mathbf{w}_j)$ for patient $i$
Guess initial value of $h_j$ for all subjects to estimate an initial guess of $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1$;
while not reached convergence do
    Using MH, get an updated estimate of $P(h_j|\mathbf{w}_j)$;
    Approximate likelihood integral using Gauss-Hermite quadrature with $k$
      abscissae $h_t$ and weight $p(h_t)$;
    Solve likelihood equations using Newton-Raphson to get updated estimates
      of $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1$;
end
```

Algorithm 3 outlines the algorithm for estimating patient preference $\hat{h}_j$. The APP-weights are the transformed estimated preferences $\Phi(\hat{h}_{ij})$. Essentially, we use the Expectation-Maximization algorithm (*Moon*, 1996) to iterate between estimates of $\boldsymbol{\alpha}$, the questionnaire coefficients, and h_j, individual patient preferences at stage j. We use Gauss-Hermite quadrature to numerically approximate the integral, and estimate $P(h_j|\mathbf{W}_j) \propto P(\mathbf{W}_j|h_j)P(h_j)$ using the Metropolis Hastings algorithm.

4.4 Numerical Demonstration

We conduct simulation studies to investigate the performance of our proposed method. We look at two scenarios, differing by degree of nonregularity (the estimated probability that stage 2 treatment does not provide a significant difference), p. The first scenario is an example of low nonregularity, where approximately for 25% of people will get similar results with both treatment. Scenario 2 is an example of higher nonregularity, where approximately 75% of patients could get similar results with

Table 4.1: Coefficients for the latent model used to solicit preferences. $\boldsymbol{\alpha}_{10}$ and $\boldsymbol{\alpha}_{11}$ are the coefficients for stage 1 questionnaires, while $\boldsymbol{\alpha}_{20}$ and $\boldsymbol{\alpha}_{21}$ are for stage 2 coefficients

	$\boldsymbol{\alpha}_{10}$	$\boldsymbol{\alpha}_{11}$	$\boldsymbol{\alpha}_{20}$	$\boldsymbol{\alpha}_{21}$
Q1	1.00	1.00	0.90	0.72
Q2	0.00	-1.00	-0.19	1.82
Q3	-1.64	2.35	1.47	-1.63
Q4	0.54	-2.35	-0.50	2.11
Q5	-0.88	-1.10	1.30	1.69
Q6	0.75	1.25	-0.40	-1.69
Q7	1.27	2.96	0.04	2.37
Q8	-1.50	2.00	-1.27	1.60
Q9	0.09	-1.50	0.62	-1.17
Q10	-0.55	1.35	-0.23	2.00

either stage 2 treatments. The true value of p was estimated through complete data (assuming no censoring) and true preferences.

For both scenarios, we generate baseline covariates $X_1, X_2 \sim N(0,1)$, censoring time C, quality of life q_1, q_2 and survival times S_1, S_2. Preferences of both stages were generated from $N(0, SD = 0.5)$ distribution, and ten binary preference derived questionnaire responses $\mathbf{W}_1, \mathbf{W}_2$ were generated according to our latent model with coefficients as in Table 4.1.

The two scenarios differ in terms of Stage 2 parameters but share common stage 1 settings. Stage 1 treatment assignment A_1 was randomly assigned with probability 0.5. The stage 1 quality of life outcome $q_1 \in [0,1]$ was generated from $N(\alpha_0 + \alpha_1 X_1 + \alpha_2 A_1 + \alpha_3 X_1 A_1, \sigma^2)$, where $\boldsymbol{\alpha} = [0.55, 0.03, 0.06, -0.09]$, and $\sigma = 0.03$. As mentioned previously, S_1 for everyone indicates a routine follow-up time of 30 days. The outcome of stage 1 is the weighted combination of q_1 and S_1 through the equation $R_1 = q_1 S_1 + (1 - \Phi(h_1))|S_1 - q_1 S_1|$, which can be interpreted as an quality of life weighted survival during the initial follow-up time. Although the true

R_1 is not observed, the true R_1 was used to assign treatment A_2, where if someone had a high R_1, they were more likely to remain on the same treatment as A_1, i.e. $Bernoulli(\{\exp(-3+R_1/3)\}/\{1+\exp(-3+R_1/3)\})$. In this simulation setting, $\Phi(h)$ represents the cumulative distributive function of a standard normal variable.

Using these simulated datasets, we estimate patient preferences, the optimal dynamic treatment regime, and the confidence intervals of estimated DTR coefficients from these responses and outcomes. We evaluate each scenario through measures of bias, coverage probability, optimal mean response, and the percent of subjects correctly classified to their true optimal treatment %opt.

4.4.1 Scenario 1: Low nonregularity (p=0.25)

We generated stage 2 outcome $q_2 \sim N(\beta_0 + \beta_1 X_2 + \beta_2 X_2 A_2, SD = 0.03)$, where $\boldsymbol{\beta} = [0.5, 0.07, 0.06]$. Similarly, we generated $S_2 \sim N(\gamma_0 + \gamma_1 R_1 + \gamma_2 (R_1 - c) A_2, SD = 5)$, where $\boldsymbol{\gamma} = [50, 10, -2.5]$ and $c = 20$. q_2 and S_2 disagreed on the optimal stage 2 treatment approximately half the time, indicating that half of all patients had to made a choice between quality and quantity of life. In a randomly generated dataset, treatment 1 gave 73.4% of patients a better stage 1 reward, while treatment 1 gave 31% of patients a better stage 2 reward. The range of S_2 varied from 169 to 361 days.

Unlike stage 1, stage 2 survival times could be subject to censoring time C. We generated censoring $C \sim \exp(\log \lambda_0^C + X_2 \beta_C)$, where $\beta_C = 0.01$ and $\lambda_0^C = 0.00058$ for 15% censoring and $\lambda_0^C = 0.0013$ for 30% censoring. τ was set to be a year after initiation of stage 2 treatment.

Because S_2 and q_2 are functions of $X_2, A_2,$ and R_1, the reward combination can be rearranged as

$$
\begin{aligned}
R_2 &= q_2 S_2 + \{1 - \Phi(h_2)\}(S_2 - q_2 S_2) \\
&= \gamma_0 + (\beta_0\gamma_0 - \gamma_0)\Phi(h_2) \\
&+ \{\gamma_1 + (\beta_0\gamma_1 - \gamma_1)\Phi(h_2)\}R_1 \\
&+ \{(\beta_1\gamma_0 - \beta_2\gamma_2 c)\Phi(h_2)\}X_2 \\
&+ \{(\beta_1\gamma_1 + \beta_2\gamma_2)\Phi(h_2)\}X_2 R_1 \\
&+ \{(\gamma_2 c - \beta_0\gamma_2 c)\Phi(h_2) - \gamma_2 c\}A_2 \\
&+ \{\gamma_2 + (\beta_0\gamma_2 - \gamma_2)\Phi(h_2)\}R_1 A_2 \\
&+ \{(\beta_1\gamma_2 + \beta_2\gamma_1)\Phi(h_2)\}R_1 X_2 A_2 \\
&+ \{(\beta_2\gamma_0 - \beta_1\gamma_2 c)\Phi(h_2)\}X_2 A_2
\end{aligned}
$$

Hence, stage 2 regression of the reward on covariates contains eight terms (intercept, $R_1, X_2, X_2R_1, A_2, R_1A_2, R_1X_2A_2, X_2A_2$). On the other hand, stage 1 coefficients are obtained by regressing a pseudo-outcome, $R_1+\hat{\boldsymbol{\beta}}_2^T\mathbf{Z}_{20}+|\hat{\boldsymbol{\psi}}_2^T\mathbf{Z}_{21}|$ on $X_1, A_1,$ and X_1A_1, so stage 1 regression has four covariate terms (including intercept). We obtain both true stage 1 and stage 2 coefficients by performing Monte Carlo sampling regressions on a sample size of 10 million.

Tables 4.2 and 4.3 lists the true covariate values, bias of our estimates, the empirical standard deviations, the mean bootstrap standard deviations, mean widths of confidence interval, and the coverage probabilities of our method. The coverage probabilities ranged from 0.86 to 0.96, with the majority between 0.92 to 0.96. Furthermore, we see general agreement between the empirical SD and the average bootstrap SD. In terms of trends, we see slightly larger ESD, mean bootstrap SD, and mean width when censoring is increased from 15% to 30%, while we see a reduction in all three

Table 4.2: Stage 2 simulation results for Scenario 1 (p=0.25). True parameters are $[7.49, 3.23, 37.58, 0.28, -1.88, 3.21, 0.21]$

n	% Censor	Parameter	Bias	ESD	AvgBootSD	MeanWidth	CP
500	15	R_1	-0.32	0.48	0.50	1.94	0.92
		X_2	0.16	11.94	12.18	47.75	0.96
		A_2	-4.93	11.72	11.57	45.17	0.92
		$R_1 : X_2$	-0.01	0.52	0.52	2.04	0.96
		$R_1 : A_2$	0.22	0.50	0.50	1.94	0.92
		$X_2 : A_2$	0.38	12.51	12.20	47.68	0.93
		$R_1 : X_2 : A_2$	-0.01	0.54	0.52	2.04	0.94
	30	R_1	-0.31	0.54	0.55	2.15	0.93
		X_2	0.38	13.39	13.57	53.19	0.95
		A_2	-4.83	13.03	12.80	50.05	0.92
		$R_1 : X_2$	-0.02	0.58	0.58	2.28	0.94
		$R_1 : A_2$	0.21	0.56	0.55	2.15	0.93
		$X_2 : A_2$	0.06	13.56	13.57	53.12	0.95
		$R_1 : X_2 : A_2$	0.00	0.58	0.58	2.28	0.95
1000	15	R_1	-0.29	0.36	0.34	1.34	0.86
		X_2	-0.14	8.23	8.34	32.64	0.95
		A_2	-4.73	8.38	8.01	31.26	0.88
		$R_1 : X_2$	0.01	0.35	0.36	1.40	0.95
		$R_1 : A_2$	0.21	0.36	0.34	1.34	0.88
		$X_2 : A_2$	0.10	8.55	8.34	32.59	0.94
		$R_1 : X_2 : A_2$	0.00	0.37	0.36	1.39	0.94
	30	R_1	-0.28	0.40	0.38	1.49	0.87
		X_2	0.04	9.38	9.25	36.17	0.96
		A_2	-4.94	8.95	8.85	34.58	0.90
		$R_1 : X_2$	0.00	0.40	0.40	1.55	0.96
		$R_1 : A_2$	0.22	0.39	0.38	1.48	0.89
		$X_2 : A_2$	0.44	9.67	9.24	36.11	0.94
		$R_1 : X_2 : A_2$	-0.02	0.41	0.40	1.55	0.94

Table 4.3: Stage 1 simulation results for Scenario 1 (p=0.25). True parameters are $[3.77, 8.17, -12.37]$

n	% Censor	Parameter	Bias	ESD	AvgBootSD	MeanWidth	CP
500	15	X_1	-0.10	0.93	0.98	3.84	0.95
		A_1	-0.30	1.05	1.07	4.18	0.93
		$X_1 : A_1$	0.41	1.21	1.23	4.78	0.94
	30	X_1	-0.10	0.93	1.00	3.93	0.96
		A_1	-0.29	1.07	1.11	4.34	0.94
		$X_1 : A_1$	0.40	1.28	1.29	5.05	0.94
1000	15	X_1	-0.04	0.67	0.68	2.66	0.96
		A_1	-0.20	0.73	0.74	2.90	0.94
		$X_1 : A_1$	0.31	0.85	0.85	3.32	0.92
	30	X_1	-0.04	0.68	0.69	2.71	0.95
		A_1	-0.21	0.75	0.77	2.99	0.94
		$X_1 : A_1$	0.31	0.88	0.90	3.49	0.94

when we increase n from 500 to 1000. Using covariate A_2 as an example, the ESD for $n = 500$ and 15% censoring is 11.72, which increased to 13.03 when censoring increased to 30% but decreased to 8.38 when n increased to 1000. We can see similar patterns with the average bootstrap SD and mean width of confidence interval.

We further investigated the distributions of the observed total reward, as well as the predicted optimal reward of one randomly selected simulation, which we illustrate in Figure 4.2. Aggregating means across the four sub-scenarios (based on sample size and censoring), the average observed reward is 233.77, while the average predicted optimal reward is 254.19, indicating an expected increase of 20.42 reward when everyone follows the regime assigned to them by our algorithm. It is also evident in the figure that the variability of the observed rewards is significantly larger than the variability of the predicted optimal reward. Aggregated SD of all observed rewards in this scenario is 39.2, while aggregated SD of all predicted rewards is 5.66.

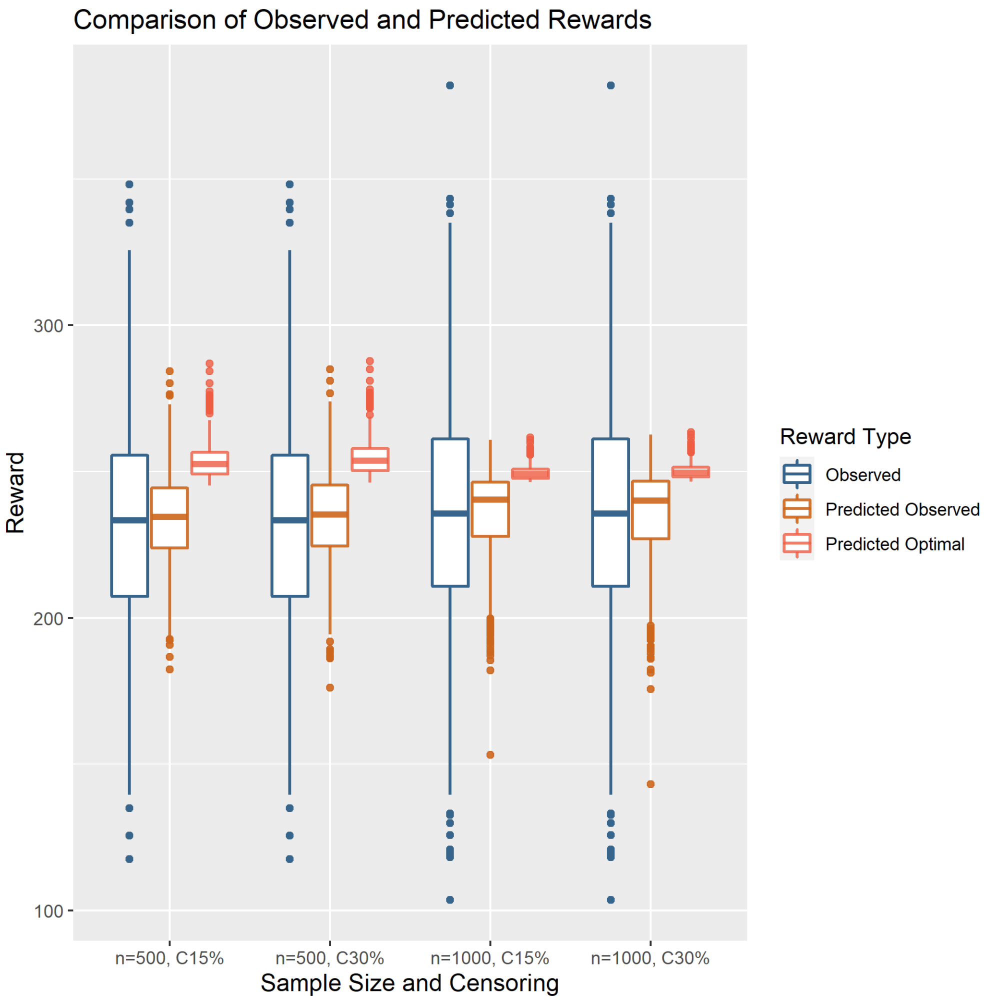

Figure 4.2: Comparison of Observed Reward and Predicted Optimal Reward for Scenario 1 (p=0.25).

4.4.2 Scenario 2: High nonregularity (p=0.75)

Baseline covariates, preferences, treatment assignment mechanisms, stage 1 quality of life and S_1 were generated as in scenario 1. We generated stage 2 outcome $q_2 \sim N(\beta_0 + \beta_1 X_2 + \beta_2 X_2 A_2, SD = 0.03)$, where $\boldsymbol{\beta} = [0.5, 0.07, 0.021]$, and $S_2 \sim N(\gamma_0 + \gamma_1 R_1 + \gamma_2 (R_1 - c) A_2, SD = 5)$, where $\boldsymbol{\gamma} = [135, 8, -0.5]$ and $c = 20$. We adjusted the coefficients such that the magnitude of γ_2 and β_2 (which influence the effect of A_2), are smaller compared to Scenario 1. q_2 and S_2 also disagreed on the optimal stage 2 treatment approximately half the time. In a randomly generated dataset, treatment 1 gave 74.8% of patients a better stage 1 reward, while treatment 1 gave 41.6% of patients a better stage 2 reward. The range of S_2 varied from 235 to 368 days.

Stage 2 survival times are similarly subject to censoring time C. Censoring was generated with model $C \sim \exp(\log \lambda_0^C + X_2 \beta_C)$, where $\beta_C = 0.01$ and $\lambda_0^C = 0.00048$ for 15% censoring and $\lambda_0^C = 0.0011$ for 30% censoring. As before, τ was set to be a year after initiation of stage 2 treatment.

Tables 4.4 and 4.5 lists the equivalent information as for Scenario 1. As with before, the coverage probabilities range from 0.86 to 0.97, with most hovering around 95%. We further see decreases in mean width, SD, and mean bootstrap SD with decreasing censoring and increasing sample size. Again using A_2 as an example, the ESD across simulations of its coefficient was 12.38 when $n = 500$ and censoring was at 15%, which increased to 13.85 when censoring increased to 30%, and decreased to 8.84 when n increased to 1000.

As in Section 4.4.1, we investigated the distribution of the observed total reward and the predicted optimal reward of one random simulation in Figure 4.3. Aggregated means across the four sub-scenarios (based on sample size and censoring), the average observed reward is 263.62, while the average predicted reward is 275.35, indicating

Table 4.4: Stage 2 simulation results for Scenario 2 (p=0.75). True parameters are $[6.00, 4.74, 7.53, 0.28, -0.38, 1.66, 0.07]$

n	% Censor	Parameter	Bias	ESD	AvgBoot SD	MeanWidth	CP
500	15	R_1	-0.33	0.51	0.52	2.05	0.92
		X_2	0.31	12.42	12.83	50.29	0.97
		A_2	-3.47	12.38	12.27	47.86	0.93
		$R_1 : X_2$	-0.01	0.54	0.55	2.14	0.96
		$R_1 : A_2$	0.16	0.53	0.52	2.05	0.94
		$X_2 : A_2$	0.17	12.99	12.84	50.18	0.95
		$R_1 : X_2 : A_2$	-0.01	0.55	0.55	2.14	0.95
	30	R_1	-0.31	0.56	0.58	2.27	0.93
		X_2	0.41	13.92	14.32	56.06	0.95
		A_2	-3.35	13.85	13.60	53.13	0.92
		$R_1 : X_2$	-0.02	0.60	0.61	2.39	0.95
		$R_1 : A_2$	0.15	0.59	0.58	2.28	0.91
		$X_2 : A_2$	-0.08	14.13	14.31	55.97	0.94
		$R_1 : X_2 : A_2$	0.00	0.60	0.61	2.39	0.95
1000	15	R_1	-0.30	0.39	0.36	1.42	0.86
		X_2	-0.10	8.69	8.78	34.34	0.95
		A_2	-3.52	8.84	8.51	33.24	0.91
		$R_1 : X_2$	0.00	0.37	0.37	1.47	0.96
		$R_1 : A_2$	0.16	0.38	0.36	1.42	0.90
		$X_2 : A_2$	0.08	8.99	8.78	34.27	0.94
		$R_1 : X_2 : A_2$	0.00	0.38	0.37	1.46	0.94
	30	R_1	-0.29	0.42	0.40	1.57	0.88
		X_2	0.24	9.79	9.75	38.13	0.95
		A_2	-3.76	9.46	9.40	36.69	0.93
		$R_1 : X_2$	-0.01	0.42	0.42	1.63	0.94
		$R_1 : A_2$	0.17	0.41	0.40	1.57	0.93
		$X_2 : A_2$	0.37	10.25	9.73	38.04	0.93
		$R_1 : X_2 : A_2$	-0.02	0.44	0.42	1.62	0.94

Table 4.5: Stage 1 simulation results for Scenario 2 (p=0.75). True parameters are $[3.19, 6.39, -9.61]$

n	% Censor	Parameter	Bias	ESD	AvgBootSD	MeanWidth	CP
500	15	X_1	-0.33	0.83	0.91	3.54	0.94
		A_1	-0.42	0.95	1.01	3.94	0.94
		$X_1 : A_1$	0.61	1.14	1.19	4.63	0.92
	30	X_1	-0.35	0.83	0.93	3.62	0.94
		A_1	-0.42	0.99	1.05	4.11	0.94
		$X_1 : A_1$	0.60	1.20	1.26	4.91	0.93
1000	15	X_1	-0.22	0.61	0.63	2.46	0.95
		A_1	-0.33	0.68	0.70	2.75	0.93
		$X_1 : A_1$	0.48	0.80	0.83	3.25	0.92
	30	X_1	-0.23	0.62	0.64	2.50	0.94
		A_1	-0.34	0.69	0.73	2.86	0.93
		$X_1 : A_1$	0.49	0.84	0.88	3.43	0.92

an increase of 11.73 reward when everyone follows the regime assigned to them by our algorithm. Aggregated SD of observed rewards is 37.77, while aggregated SD of all predicted rewards is 4.39.

4.4.3 Optimality

Besides looking at performance of inference, we also looked at the number of times our algorithm chose the correct treatment for each patient at each stage, across the scenarios we have visited ($n = 500, 1000$ across two levels of censoring at $15\%, 30\%$).

Table 4.6 shows the simulation results. In this table we also included the average stage 1 bootstrap resample sizes for each sub-scenario. As expected from Equation 4.3, increasing p indicates increasing nonregularity, and decreases m. We can see that for Scenario 1 ($p = 0.25$), the algorithm chose the optimal treatment for stage 1 over 93% of the time, and over 83% of the time for stage 2. Our algorithm was able to assign the correct regime to a patient over 78% across all both sample sizes and

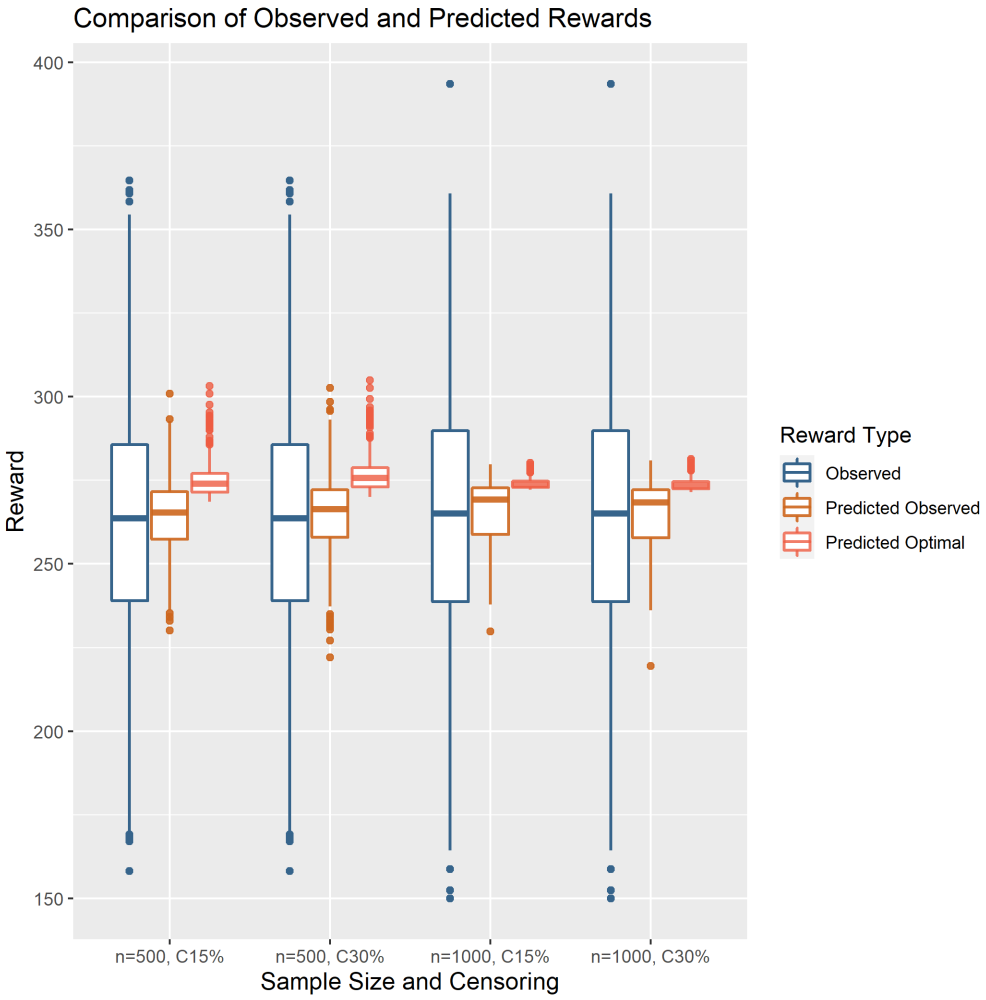

Figure 4.3: Comparison of Observed Reward and Predicted Optimal Reward for Scenario 2 (p=0.75).

Table 4.6: Simulation results for percent optimal treatment chosen

Scenario	n	% Censor	Stage1 %opt (sd)	Stage2 %opt (sd)	Overall %opt (sd)
p=0.25	500	15	0.936 (0.012)	0.839 (0.020)	0.785 (0.022)
	500	30	0.936 (0.012)	0.837 (0.022)	0.783 (0.023)
	1000	15	0.939 (0.008)	0.844 (0.012)	0.793 (0.013)
	1000	30	0.939 (0.008)	0.843 (0.013)	0.792 (0.014)
p=0.75	500	15	0.935 (0.014)	0.667 (0.051)	0.623 (0.050)
	500	30	0.934 (0.014)	0.662 (0.053)	0.618 (0.051)
	1000	15	0.938 (0.009)	0.687 (0.029)	0.644 (0.028)
	1000	30	0.938 (0.009)	0.682 (0.033)	0.640 (0.032)

censoring levels. This is significantly higher than the random guess approach, which would have landed us around 25%.

In contrast, Scenario 2's performance was weaker, coming in at over 93% for stage 1, around 66−69% for stage 2, and with an overall correct regime assignment percentage of between 61 − 65%. In this scenario, we could also see a general increase in SD as compared to Scenario 1, indicating higher uncertainty in our decision making process.

4.5 Discussion

In this report we proposed a method that estimates the optimal regimen for a two-stage treatment scenario subject to censoring. We propose to treat the balance between quality of life and quantity using a sliding scale function adjusted using patient preference. Through simulation studies, we have shown that our proposed method is capable of choosing the optimal treatment and regime a majority of the time, as well as provide convincing confidence intervals for each of the coefficients in question.

The simulation results of Scenarios 1 and 2 are notable in the following ways. Most importantly, we can see that the coverage probabilities mostly hover around 95%, showing that our confidence interval has the combination of adequate width and minimal bias required for a good coverage probability. The general congruence between the empirical SE and the mean bootstrap SE is further support that our method is sampling at the appropriate width. Generally, there is a slight increase in ESD and mean bootstrap SD when increasing censoring from 15% to 30%, indicating the decreased certainty, but the difference is slight (e.g. for parameter A_2 in Scenario 1, ESD increased from 11.72 to 13.03). The decrease in ESD is more significant between $n = 500$ and $n = 1000$, where for Scenario 1 and 15% censoring, the ESD of the estimated coefficient for parameter A_2 decreased from 11.72 to 8.38. In the boxplots in Figures 4.2 and 4.3, we can see much larger variability in the observed rewards as compared to the predicted rewards. This is expected for two reasons. First, observed rewards contain an error component that is not present in expected (predicted) rewards. Secondly, observed rewards includes individuals who have by chance obtained their optimal reward, as well as those who did not. Variability invariably reduces when more individuals are predicted to their optimal reward. As expected, the distribution of the predicted optimal rewards can be mapped to the upper part of the observed rewards. The same general patterns are observed for Scenario 2.

The difference in simulation results between Scenario 1 and Scenario 2 illustrates a few interesting points. Our adapted Q-learning algorithm chose over 78% of the correct regime in Scenario 1, but this decreased to around 65% for the various cases in Scenario 2, with the decrease particularly prominent in choosing the correct stage 2 decision. This is unsurprising, since a higher value of p indicates exactly that the exact choice of A_2 is less important for those in Scenario 2, making it more difficult for our algorithm to pick up the best decision. This hesitancy is further supported by the larger SDs seen for Scenario 2 as compared to Scenario 1. Again using A_2 as an example, for $n = 500$ and 15% censoring, the ESD in Scenario 1 is 11.72 while the ESD in Scenario 2 is 12.38. Finally, the increased difficulty in selecting the optimal regime when p increases is further illustrated in Figures 4.2 and 4.3, where the mean expected increases are much smaller in Scenario 2 (11.73) as compared to Scenario 1 (20.42). While the mean of the predicted optimal rewards are at the top whisker of observed rewards in Figure 4.2, they are slightly lower with respect to the distribution of observed rewards in Figure 4.3. This relative decrease in predicted increased rewards can be explained by the lower percentage of patients selected to their optimal regime.

One main challenge in this work, as is the case in *Chakraborty et al.* (2013), is the selection of m, which is a crucial factor in determining the coverage probability and confidence interval. In our current approach, we selected parameters ν and ξ using background knowledge and reference simulation results and used Equation 4.3 to select m. This approach is recommended by *Chakraborty et al.* (2013) and is straightforward and easy to implement, but risks inappropriate values of m if either ν or ξ are selected inadequately. Hence, one area of improvement would be to explore approaches for selection of m that are less reliant on tuning parameters. One approach that can be adapted towards our scenario is the double bootstrap procedure, where we take our empirical dataset estimator as the truth, and build confidence intervals using nested bootstrap samples of size m from empirical bootstrap samples of size n. We could then look across a range of m and select the one providing desired coverage. Similarly, another potential idea to further improve coverage across all covariates could be to select distinct values of m for each covariate. Preliminary simulation studies are underway for both of these directions, and adaptations of these approaches could potentially yield more intuitive and robust approaches for the selection of m.

The authors are aware of two works in the literature that balance between two outcomes and would like to highlight certain differences at this time. *Zhao et al.* (2009) looks at the dosage effect of a cancer drug on tumor size and toxicity. Each reward function is separated into three parts: survival status, wellness, and tumor size effects. In simulation studies, the reward was assigned to be -60 if the patient died, 15 if the patient's tumor shrunk to zero, and 5 or -5 if tumor size/wellness improved and deteriorated, respectively. The method of assigning rewards is particular to the example they proposed, and there is no clear indication of how to perform inference using this method.

The comparison between our method and *Butler et al.* (2018) is a bit more direct. Their work uses patient preference estimation to weigh between toxicity and efficacy, both continuous outcomes. There are definite similarities between this work and

Butler et al. (2018), especially in terms of preference estimation and the common Q-learning framework. However, most importantly, *Butler et al.* (2018) would not accommodate censored data, nor did it provide any framework for inference, both important contributions of this work.

This work can be improved in a couple of directions for increased generalizability and impact. First, the vast amount of methods for survival data is one indication of the complexity of generalizing the various time to event scenarios. One main direction for extension would be to accommodate multiple stages, with potential censoring and event times that could happen at any stage, similar to the set up of *Goldberg and Kosorok* (2012). Another realistic directions could be allowing the subset of treatments to change depending on patient outcomes, as in *Hager et al.* (2018). Generalizing binary treatment options to a continuous version (i.e. dosage), or increasing the number of outcomes we balance are both meaningful directions for future extensions.

CHAPTER V

Summary and Future Work

In this dissertation, we explored improvements to methods that follow a patient trajectory and estimate predicted outcomes using available covariates, as well as make preference driven patient decisions for different types of clinical scenarios.

In Chapter II, we developed a method for modeling the restricted mean survival time as a function of the restriction time. Our method comes with a simulation supported inference framework. Different from other methods in the literature, our method allows any researcher interested in the RMST to obtain a longitudinal time profile of the entire scenario, as well as how each covariate effect and significance changes as a function of time. This obviates the need to do multiple RMST analysis at each time point of interest, which requires a somewhat arbitrary selection of cutoff-times as well as a risk of multiple testing if there are many points of interest. In Chapter III, we proposed a method that augments two potentially competing outcomes with preference estimated through a latent variable model in order to find the optimal dynamic treatment regime. Our biggest contribution in this work is the incorporation of patient preference into the multiple stage setting, which is more fitting for a chronic clinical course. Patients with chronic disease often are faced with more potential choices than those in more acute medical scenarios. Hence, our method expands upon previous methods that allow only binary decision choices to accommodate more potential decisions. In Chapter IV, we bring patient preference into the realm of survival setting, where patients often have to choose between quality of life or length of time before an event. The major challenges addressed in this setting are the technical difficulties that come with the presence of censored data. Our method also provides a process for inference, which addresses both the complications caused by censoring as well as the nonregularity which affects all DTR type methods.

Each of these chapters naturally lends itself to future research ideas and challenges. In Chapter II, we currently approximate the time profile curve of the RMST as a function of the cut off time with parametric splines. One interesting advancement would be modeling using a nonparametric approach so that the RMST curve would be more sensitive to actual data perturbations. In addition, it would be useful to expand this method to accommodate dependent censoring in addition to independent censoring.

Chapter IV could be improved by moving to more than two stages and allowing for censoring and event times to happen before the last stage, greatly complicating the number of scenarios the method needs to accommodate. We showed through simulation studies that our method works well with specific values of hyperparameters. It would be worthwhile, however, to investigate an approach that would allow for systemic recommendations of values of m, similar to the double bootstrap approach suggested in *Chakraborty et al.* (2013), but which was computationally intractable in our situation. Further general extensions and improvements would apply to both Chapters III and IV. Both chapters could move into the realm of multi-objective optimization, where the optimization would take place in n-dimensional space, where instead of identifying a single best solution, the goal would be to produce a set of non-dominated solutions. Another audacious direction of interest would be the area of mobile health. While a phone application can easily interact with its owner, mobile health is highly reliant on preference on how much and when to engage with the application. Hence, a marriage between patient preference estimation and mobile health methods, which essentially reshapes the DTR treatment time horizon into a continuous one, would be an interesting and potentially very impactful and practical direction.

BIBLIOGRAPHY

BIBLIOGRAPHY

Andersen, P. K., and M. Pohar Perme (2010), Pseudo-observations in survival analysis, *Statistical methods in medical research*, *19*(1), 71–99.

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

Barry, M. J., and S. Edgman-Levitan (2012), Shared decision making—the pinnacle of patient-centered care, *New England Journal of Medicine*, *366*(9), 780–781.

Bartholomew, D. J., M. Knott, and I. Moustaki (2011), *Latent variable models and factor analysis: A unified approach*, vol. 904, John Wiley & Sons.

Basu, A., and D. Meltzer (2007), Value of information on preference heterogeneity and individualized care, *Medical Decision Making*, *27*(2), 112–127.

Breiman, L. (2017), *Classification and regression trees*, Routledge.

Breslow, N. E. (1972), Contribution to discussion of papeer by dr cox, *J. Roy. Statist. Assoc., B*, *34*, 216–217.

Butler, E. L., E. B. Laber, S. M. Davis, and M. R. Kosorok (2018), Incorporating patient preferences into estimation of optimal individualized treatment rules, *Biometrics*, *74*(1), 18–26.

Chakraborty, B., and E. Moodie (2013), *Statistical methods for dynamic treatment regimes*, Springer.

Chakraborty, B., and S. A. Murphy (2014), Dynamic treatment regimes, *Annual review of statistics and its application*, *1*, 447–464.

Chakraborty, B., S. Murphy, and V. Strecher (2010), Inference for non-regular parameters in optimal dynamic treatment regimes, *Statistical methods in medical research*, *19*(3), 317–343.

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

Chen, P.-Y., and A. A. Tsiatis (2001), Causal inference on the difference of the restricted mean lifetime between two groups, *Biometrics*, *57*(4), 1030–1038.

Council, N. R., et al. (2011), *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*, National Academies Press.

Cox, D. R. (1972), Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Cox, D. R. (1975), Partial likelihood, *Biometrika*, *62*(2), 269–276.

Cui, Y., R. Zhu, and M. Kosorok (2017), Tree based weighted learning for estimating individualized treatment rules with censored data., *Electronic journal of statistics*, *11*(2), 3927–3953.

Embretson, S. E., and S. P. Reise (2013), *Item response theory*, Psychology Press.

Foutz, R. V. (1977), On the unique consistent solution to the likelihood equations, *Journal of the American Statistical Association*, *72*(357), 147–148.

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

Hager, R., A. A. Tsiatis, and M. Davidian (2018), Optimal two-stage dynamic treatment regimes from a classification perspective with censored survival data, *Biometrics*, *74*(4), 1180–1192.

Hamburg, M. A., and F. S. Collins (2010), The path to personalized medicine, *New England Journal of Medicine*, *363*(4), 301–304.

Huang, B., and P.-F. Kuan (2018), Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event end point, *Pharmaceutical statistics*, *17*(3), 202–213.

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

Huang, X., J. Ning, and A. S. Wahed (2014), Optimization of individualized dynamic treatment regimes for recurrent diseases, *Statistics in medicine*, *33*(14), 2363–2378.

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), 3424–3443.

Irwin, J. (1949), The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice, *The Journal of hygiene*, *47*(2), 188.

Jiang, R., W. Lu, R. Song, and M. Davidian (2017a), On estimation of optimal treatment regimes for maximizing t-year survival probability, *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *79*(4), 1165.

Jiang, R., W. Lu, R. Song, M. G. Hudgens, and S. Naprvavnik (2017b), Doubly robust estimation of optimal treatment regimes for survival data—with application to an hiv/aids study, *The annals of applied statistics*, *11*(3), 1763.

Karrison, T. (1987), Restricted mean life with adjustment for covariates, *Journal of the American Statistical Association*, *82*(400), 1169–1176.

Kashaf, M. S., and E. McGill (2015), Does shared decision making in cancer treatment improve quality of life? a systematic literature review, *Medical decision making*, *35*(8), 1037–1048.

Krumholz, A. (2009), Driving issues in epilepsy: past, present, and future, *Epilepsy Currents*, *9*(2), 31–35.

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

Laber, E. B., D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy (2014), Dynamic treatment regimes: Technical challenges and applications, *Electronic journal of statistics*, *8*(1), 1225.

Lizotte, D. J., M. Bowling, and S. A. Murphy (2012), Linear fitted-q iteration with multiple reward functions, *Journal of Machine Learning Research*, *13*(Nov), 3253–3295.

Marler, R. T., and J. S. Arora (2004), Survey of multi-objective optimization methods for engineering, *Structural and multidisciplinary optimization*, *26*(6), 369–395.

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.

Moon, T. K. (1996), The expectation-maximization algorithm, *IEEE Signal processing magazine*, *13*(6), 47–60.

Moustaki, I., and M. Knott (2000), Generalized latent trait models, *Psychometrika*, *65*(3), 391–411.

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

Orellana, L., A. Rotnitzky, and J. M. Robins (2010), Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content, *The international journal of biostatistics*, *6*(2).

Ortinski, P., and K. J. Meador (2004), Cognitive side effects of antiepileptic drugs, *Epilepsy & Behavior*, *5*, 60–65.

Oshima Lee, E., and E. J. Emanuel (2013), Shared decision making to improve care and reduce costs, *New England Journal of Medicine*, *368*(1), 6–8.

Rao, P. S., D. E. Schaubel, M. K. Guidinger, K. A. Andreoni, R. A. Wolfe, R. M. Merion, F. K. Port, and R. S. Sung (2009), A comprehensive risk quantification score for deceased donor kidneys: the kidney donor risk index, *Transplantation*, *88*(2), 231–236.

Rasch, G. (1960), *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*, Nielsen & Lydiche.

Rasch, G. (1961), On general laws and the meaning of measurement in psychology, in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 4, pp. 321–333.

Robins, J., and A. Rotnitzky (1992), Recovery of information and adjustment for dependent censoring using surrogate markers, in *AIDS Epidemiology*, edited by N. Jewell, K. Dietz, and V. Farewell, pp. 297–331, Birkhäuser Boston, doi:10.1007/978-1-4757-1229-2_14.

Robins, J. M. (1993), Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers, in *Proceedings of the Biopharmaceutical Section, American Statistical Association*, pp. 24–33, San Francisco CA.

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, 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, Springer.

Robins, J. M., and D. M. Finkelstein (2000), Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests, *Biometrics*, pp. 779–788.

Robins, J. M., and M. A. Hernán (2009), Estimation of the causal effects of time-varying exposures, *Longitudinal data analysis*, *553*, 599.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998), Semiparametric regression for repeated outcomes with nonignorable nonresponse, *Journal of the american statistical association*, *93*(444), 1321–1339.

Royston, P., and M. K. Parmar (2011), The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt, *Statistics in medicine*, *30*(19), 2409–2421.

Royston, P., and M. K. Parmar (2013), Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome, *BMC medical research methodology*, *13*(1), 152.

Schaubel, D. E., and G. Wei (2011), Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring, *Biometrics*, *67*(1), 29–38.

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: a review journal of the Institute of Mathematical Statistics*, *29*(4), 640.

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

Shay, L. A., and J. E. Lafata (2015), Where is the evidence? a systematic review of shared decision making and patient outcomes, *Medical Decision Making*, *35*(1), 114–131.

Struthers, C. A., and J. D. Kalbfleisch (1986), Misspecified proportional hazard models, *Biometrika*, *73*(2), 363–369.

Sutton, R. S., and A. G. Barto (2018), *Reinforcement learning: An introduction*, MIT press.

Tao, Y., and L. Wang (2017), Adaptive contrast weighted learning for multi-stage multi-treatment decision-making, *Biometrics*, *73*(1), 145–155.

Tao, Y., L. Wang, and D. Almirall (2018), Tree-based reinforcement learning for estimating optimal dynamic treatment regimes, *The annals of applied statistics*, *12*(3), 1914.

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), 4687–4702.

Tian, L., L. Zhao, and L. Wei (2014), Predicting the restricted mean event time with the subject's baseline covariates in survival analysis, *Biostatistics*, *15*(2), 222–233.

Tian, L., H. Fu, S. J. Ruberg, H. Uno, and L.-J. Wei (2017), Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations, *Biometrics*.

Torrance, G. W., and D. Feeny (1989), Utilities and quality-adjusted life years, *International journal of technology assessment in health care*, *5*(4), 559–575.

Uno, H., et al. (2014), Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis, *Journal of clinical Oncology*, *32*(22), 2380.

Uno, H., et al. (2015), Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies, *Annals of internal medicine*, *163*(2), 127–134.

Wang, X., and D. E. Schaubel (2018), Modeling restricted mean survival time under general censoring mechanisms, *Lifetime data analysis*, *24*(1), 176–199.

Wei, G., and D. E. Schaubel (2008), Estimating cumulative treatment effects in the presence of nonproportional hazards, *Biometrics*, *64*(3), 724–732.

Zhang, B., and M. Zhang (2018), C-learning: A new classification framework to estimate optimal dynamic treatment regimes, *Biometrics*, *74*(3), 891–899.

Zhang, M., and D. E. Schaubel (2011), Estimating differences in restricted mean lifetime using observational data subject to dependent censoring, *Biometrics*, *67*(3), 740–749.

Zhang, M., and D. E. Schaubel (2012), Contrasting treatment-specific survival using double-robust estimators, *Statistics in medicine*, *31*(30), 4255–4268.

Zhang, Y., E. B. Laber, M. Davidian, and A. A. Tsiatis (2018), Interpretable dynamic treatment regimes, *Journal of the American Statistical Association*, *113*(524), 1541–1549.

Zhao, H., and A. A. Tsiatis (1997), A consistent estimator for the distribution of quality adjusted survival time, *Biometrika*, *84*(2), 339–348.

Zhao, H., and A. A. Tsiatis (1999), Efficient estimation of the distribution of quality-adjusted survival time, *Biometrics*, *55*(4), 1101–1107.

Zhao, L., B. Claggett, L. Tian, H. Uno, M. A. Pfeffer, S. D. Solomon, L. Trippa, and L. Wei (2016), On the restricted mean survival time curve in survival analysis, *Biometrics*, *72*(1), 215–221.

Zhao, Y., M. R. Kosorok, and D. Zeng (2009), Reinforcement learning design for cancer clinical trials, *Statistics in medicine*, *28*(26), 3294–3315.

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.-Q., D. Zeng, E. B. Laber, R. Song, M. Yuan, and M. R. Kosorok (2015), Doubly robust learning for estimating individualized treatment with censored data, *Biometrika*, *102*(1), 151–168.

Zhao, Y.-Q., R. Zhu, G. Chen, and Y. Zheng (2018), Constructing stabilized dynamic treatment regimes, *arXiv preprint arXiv:1808.01332*.

Zhong, Y., D. E. Schaubel, J. D. Kalbfleisch, V. B. Ashby, P. S. Rao, and R. S. Sung (2019), Reevaluation of the kidney donor risk index, *Transplantation*, *103*(8), 1714–1721.

Zhu, R., and M. R. Kosorok (2012), Recursively imputed survival trees, *Journal of the American Statistical Association*, *107*(497), 331–340.

Zucker, D. M. (1998), Restricted mean life with covariates: modification and extension of a useful survival analysis method, *Journal of the American Statistical Association*, *93*(442), 702–709.