ABSTRACT

ROSE, ERIC JAMES. Methods for Dynamic Treatment Regimes. (Under the direction of Dr. Eric Laber and Dr. Marie Davidian.)

Precision medicine is a data-driven approach for leveraging patient heterogeneity to improve heath care by tailoring treatments for patients based on individual patient characteristics. In practice, clinicians have to make several decisions during the course of treating a disease to adapt to the evolving health status of the patient. A dynamic treatment regime operationalizes clinical decision making as a sequence of decision rules that map up-to-date patient information to a recommended intervention.

Sequential Multiple Assignment Randomized Trials (SMARTs) are considered the gold standard for estimation and evaluation of treatment regimes. In Chapter 2, we derive sample size procedures for a SMART that ensure: (i) sufficient power for comparing the optimal treatment regime with standard of care; and (ii) the estimated optimal regime is within a given tolerance of the true optimal regime with high-probability. We establish asymptotic validity of the proposed procedures and demonstrate their finite sample performance in a series of simulation experiments.

In the context of time-to-event data where the number of decisions depends on each patient's health trajectory, naive application of existing methods for estimating an optimal treatment regime can lead to bias. In Chapter 3, we propose a variant of Q-learning, a regression-based approximate dynamic programming method, to estimate an optimal treatment regime with time-to-event data subject to censoring.

In Chapter 4, we discuss a SMART that was used to study parent messaging strategies for reducing chronic absenteeism in elementary school children. We will compare the adaptive regimes embedded in the trial in addition to estimating an optimal treatment regime in which we tailor interventions based on family and school characteristics.

Methods for Dynamic Treatment Regimes

by
Eric James Rose

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

Statistics

Raleigh, North Carolina

2019

APPROVED BY:

Dr. Dennis Boos	Dr. Shu Yang
Dr. Eric Laber Co-chair of Advisory Committee	Dr. Marie Davidian Co-chair of Advisory Committee

DEDICATION

To my parents, Tom and Alice, and brother, Jason.

BIOGRAPHY

Eric James Rose grew up in Ridgefield, Connecticut where he graduated from Ridgefield High School in 2010. He attended the University of Connecticut where he received a Bachelor of Arts in mathematics and statistics in 2014. He then continued his education at North Carolina State University where he earned a Master of Statistics in 2017.

ACKNOWLEDGEMENTS

I would like to thank my advisors Dr. Eric Laber and Dr. Marie Davidian for all of the help they gave me throughout my graduate studies. Beginning with the courses I took of theirs and continuing with the guidance and encouragement they provided with my research, I am extremely grateful to have had the opportunity to work with them. The faculty and staff in the Department of Statistics at North Carolina State University do an incredible job at creating a great environment for their students. I would also like to thank James Gilman, Nick Kapur, and the rest of the members of Laber Labs who have provided not only helpful feedback on my work, but many friendships.

During my time in graduate school, I have been lucky to make many friendships inside and outside of the department that I credit much of my success to. In particular, the support Susheela and Nikhil Singh and Emily Gower have provided has meant so much to me over the past several years.

Finally, I would like to thank my parents and brother for the love and support they have provided throughout my life. I would have never been able to accomplish this without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

Precision medicine focuses on how to tailor treatments for patients based on individual patient characteristics. There is a significant amount of heterogeneity in patients due to differences in attributes such as genetics, lifestyle, and medical history. These differences affect how a patient will respond to the treatments they receive and should therefore be considered in the decision making process for assigning treatments. In precision medicine, we seek to formalize this process with data-driven methods for informing how to best treat patients based on these characteristics.

Conventional methods for studying treatment effects involve conducting a randomized clinical trial in which a set of candidate treatments are randomly assigned to patients. These patients are then followed to determine the expected outcome for each individual treatment option if it was assigned to the entire population of interest. This does not match the clinical

decision making process in practice. Typically clinicians will make several decisions over the course of treating a disease for a single patient. This could be because there are multiple key decision points corresponding to the development of the disease or during regular follow-ups the clinician may determine the patient is not responding to the initial treatment or experiencing an adverse reaction and decide to adjust the treatment plan. This process of synthesizing patient information and assigning treatments sequentially over time is formalized in what is called a dynamic treatment regime.

A dynamic treatment regime operationalizes clinical decision making as a sequence of decision rules that map up-to-date patient information to a recommend intervention. The goal in precision medicine is to estimate an optimal treatment regime where optimal is defined as the regime that maximizes some mean outcome of interest if all patients in the population of interest were to receive treatments via these decision rules.

There have been many different methods proposed for estimating optimal dynamic treatment regimes which can generally be classified into two different categories, direct or indirect methods. Direct estimation methods work by specifying a class of regimes and then maximizing the estimated expected outcome over all regimes within that class. Examples of direct methods include marginal structural mean models (Robins et al., 2008) and outcome weighted learning (Zhao et al., 2012). Indirect methods function by positing models for the outcome of interest and the optimal regime is then determined by assigning the treatment that would maximize the predicted expected outcome. Examples of this include A-learning (Robins, 2004a; Murphy, 2003a) and Q-learning (Murphy, 2005b; Qian et al., 2013). Here we will focus on the use of Q-learning for estimating an optimal dynamic treatment regime.

1.1 Q-learning

Q-learning is an approximate dynamic programming algorithm used for the estimation of an optimal dynamic treatment regime. It was developed in the context of reinforcement

learning for learning an optimal policy for a Markov decision process (Watkins, 1989) and has been used for sequential decision making problems in many different areas of application. It consists of performing a series of regressions each one with a subsequent maximization.

Suppose we are interested in estimating an optimal regime for a situation where we have K different key decision points. We will assume we observe data of the form $\{(\boldsymbol{X}_{1,i}, A_{1,i}, \boldsymbol{X}_{2,i}, A_{2,i}, \ldots, \boldsymbol{X}_{K,i}, A_{K,i}, Y_i)\}_{i=1}^n$ which is comprised of n $i.i.d$ trajectories of the form $(\boldsymbol{X}_1, A_1, \boldsymbol{X}_2, A_2, \ldots, \boldsymbol{X}_K, A_k, Y)$ such that $\boldsymbol{X}_1 \in \mathbb{R}^{p_1}$ is baseline covariate information, $A_k \in \{-1, 1\}$ is the treatment assigned at stage k, $\boldsymbol{X}_k \in \mathbb{R}^{p_k}$ is additional information recorded during the course of treatment $k-1$, and $Y \in \mathbb{R}$ is the outcome of interest coded such that higher values are better. Define $\boldsymbol{H}_k = (\boldsymbol{X}_1, A_1, \ldots, \boldsymbol{X}_k)$ to be the history of a patient at decision point k. Q-learning works by defining the Q-functions

$$
\begin{aligned}
Q_K(\boldsymbol{h}_K, a_K) &= \mathbb{E}(Y | \boldsymbol{H}_K = \boldsymbol{h}_K, A_K = a_k) \\
Q_k(\boldsymbol{h}_k, a_k) &= \mathbb{E}\Big\{ \max_{a_{k+1}} Q_{k+1}(\boldsymbol{H}_{k+1}, a_{k+1}) | \boldsymbol{H}_k = \boldsymbol{h}_k, A_k = a_k \Big\} \text{ for } k = 1, \ldots, K-1.
\end{aligned}
$$

It can then be shown via dynamic programming that $\pi_k^{\text{opt}}(\boldsymbol{h}_k) = \arg\max_{a_k} Q_k(\boldsymbol{h}_k, a_k)$ (Bellman, 1957). Models $Q_k(\boldsymbol{h}_k, a_k; \beta_k)$ are then posited for each Q-function and fit in a backwards iterative manner. First estimate $\hat{\beta}_K$ by regressing Y on $\boldsymbol{H}_K$ and A_K. For $k = K-1, \ldots, 1$ the Q-function is estimated by regressing $\max_{a_{k+1}} Q_{k+1}(\boldsymbol{H}_{k+1}, a_{k+1}; \hat{\beta}_{k+1})$ on $\boldsymbol{H}_k$ and A_k. This then gives a plug-in estimator for the optimal regime as $\hat{\pi}_k^{\text{opt}}(\boldsymbol{h}_k) = \arg\max_{a_k} Q_k(\boldsymbol{h}_k, a_k; \hat{\beta}_k)$.

This can be implemented with a large variety of different regression methods for the Q-functions. The choice of regression model determines the form of the decision rule though so choosing simple models such as linear models will lead to simple, interpretable regimes. More flexible, nonparametric regression models could provide a better fit, but may lead to unintelligible regimes.

1.2 Sequential Multiple Assignment Randomized Trials

To be able to estimate an optimal dynamic treatment regime we need data of the correct form. We cannot take data from different clinical trials and observational studies that each focus on one individual decision point and piece them together. This is because there could be delayed effects of the early treatments that are not seen until later. Therefore we need data on the same set of patients through all the decision points.

Data from observational studies can lead to biased results because of confounding that occurs between the treatment assigned and the patient outcome. For example, sicker patients may be more likely to receive one of the treatment options and therefore worse outcomes for these patients are observed. In the case of multiple decision points we also have time-dependent confounding which further complicates inference. This has led to the preference of using experimental data for the use of studying dynamic treatment regimes.

The Sequential Multiple Assignment Randomized Trial (SMART) is the gold standard for gathering data for the study of sequential decision making for precision medicine. In a SMART, patients are randomized to the different feasible treatment options at each of the potential decision points. This provides high quality data for the estimation of an optimal regime while also mimicking the adaptive nature of assigning treatments that is observed in practice.

One example of a potential SMART with two decision points is shown in figure 1.1. In this example, in the first stage all the patients are randomized to receive either the new treatment which is denoted as treatment A or standard of care, denoted by treatment B. At the end of the first stage we check whether a patient is responding to the initial treatment. For example, a patient may be considered a responder if their symptom severity decreases to below some given threshold. All of the patients that are considered responders maintain their initial treatment during the second stage. If a patient were to receive treatment A during the first stage and is deemed a non-responder at the end of the first-stage, then

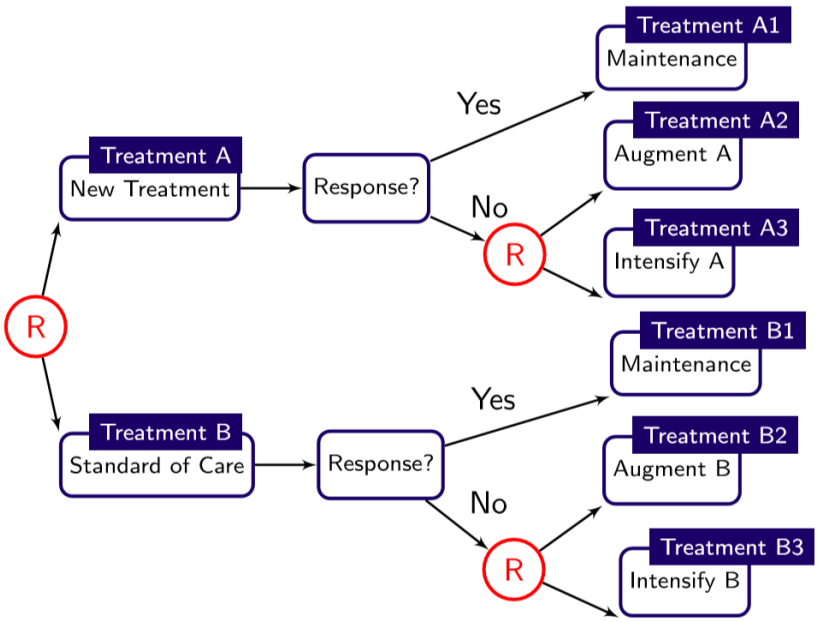

Figure 1.1 Example two-stage SMART design

they are re-randomized to receive either augment A or intensify A during the second stage. Similarly non-responsive patients receiving treatment B are re-randomized in the second stage to augment B or intensify B. This is just one possible design for a SMART among many different potential options. How to determine the best design for a SMART is dictated by the scientific questions of interest that researchers plan on examining using the data collected.

Chapter 2 covers how to conduct sample size calculations for a two-stage SMART where we are trying to size the trial for the estimation of an optimal dynamic treatment regime. Chapter 3 develops an adaptation of Q-learning for when the outcome of interest is a time-to-event. Chapter 4 discusses a SMART that was used to test a parent messaging system to reduce absenteeism in elementary school kids.

CHAPTER

2

SAMPLE SIZE CALCULATIONS FOR SMARTS

A treatment regime is a sequence of functions, one per stage of clinical intervention, that map up-to-date patient information to a recommended treatment. An optimal treatment regime maximizes the mean of some cumulative clinical outcome when applied to select treatments for individuals in a population of interest (Murphy, 2003b; Robins, 2004b). Thus, an optimal treatment regime leads to better overall healthcare by adapting treatment to the evolving health status of each patient; consequently, optimal treatment regimes have become a primary means of operationalizing precision medicine. Optimal treatment regimes have been estimated across a wide range of application domains including breastfeeding (Moodie et al., 2012), bipolar disorder (Wu et al., 2015; Zhang et al., 2017), cancer (Thall

et al., 2000; Zhao et al., 2011; Wang et al., 2012; Zhang et al., 2012, 2015; Murray et al., 2017), cystic fibrosis (Zhou et al., 2017), diabetes (Ertefaie, 2014; Luckett et al., 2016), depression (Zhao et al., 2012), HIV (Moodie et al., 2007; van der Laan et al., 2005; Cain et al., 2010), smoking cessation (Chakraborty et al., 2009), substance abuse (Nahum-Shani et al., 2017) among others.

Sequential Randomized Multiple Assignment Randomized Trials (SMARTs Lavori and Dawson, 2000, 2004; Murphy, 2005a; Kidwell, 2014) are the gold standard for estimating and evaluating treatment regimes (Murphy et al., 2007; Lei et al., 2012; Chakraborty and Moodie, 2013; Kosorok and Moodie, 2015) and are increasingly common design in clinical and intervention science (PSU Methodology Center, 2017b,a). However, sample size calculations for SMARTs are typically based on power calculations for simple comparisons, e.g., comparison of the mean outcome across two pre-specified treatment sequences (Murphy, 2005a; Lei et al., 2012) and estimation of an optimal treatment regime from data collected in a SMART are almost always (we are not aware of an exception) conducted as part of exploratory, hypothesis-generating analyses. This approach is aligned with the *estimate-and-validate* paradigm wherein: (i) an optimal treatment regime is estimated using data collected in a SMART; and (ii) the performance of the estimated optimal regime is validated in a follow-up trial where the estimated regime is compared head-to-head with standard of care (Murphy, 2005a). This approach is appealing in that it avoids a number of nontrivial technical issues associated with estimating and evaluating a treatment regime using the same data (Robins, 2004b; Moodie et al., 2010; Chakraborty et al., 2010; Song et al., 2015a; Chakraborty et al., 2014; Laber et al., 2014); furthermore, sample size formulae for the comparison of fixed treatment sequences or other commonly used criteria to size SMARTs are straightforward in that they resemble those commonly used in non-sequential randomized trials. Another reason that sample size calculations for SMARTs are often based on simple comparisons is the seemingly widely held belief that sizing a trial to guarantee frequentist operating characteristics for an estimated optimal regime, e.g., providing a performance guarantee

for the estimated regime or powering a comparison of the performance of the optimal regime with standard of care, would either: (i) require a prohibitively large sample size; and/or (ii) rely on unrealistic assumptions about the underlying data-generating model. We provide evidence that for a for a large class of generative models neither of these beliefs appear to be well-founded.

We derive sample size procedures for a SMART that ensure sufficient power in comparing the mean outcome under the optimal regime with standard of care and that the estimated optimal regime will be within a given tolerance of the optimal policy with a given probability. The proposed sample sized procedures we develop here are at two possible extremes in terms of modeling assumptions. Our first procedure imposes significant parametric structure on the data-generating model and consequently we are able to derive sample size formulae that resemble the comparison of two means and require elicitation or estimation of a single scalar parameter. Our second procedure imposes structure only on moments and tail behavior of some components of the data generating model and then uses the bootstrap with oversampling to estimate a sufficient sample size; as this procedure imposes less structure the resulting estimator is more variable and consequently the estimated sample size tends to be larger. One reason for considering these two extremes is that they provide a basis for intermediate procedures that impose as much structure, as appropriate for a given application. We leave such intermediate approaches to future work.

We perceive the proposed work as making the following contributions: (i) it provides the first rigorous yet practical sample size procedures for estimating and evaluating optimal treatment regimes using SMARTs; (ii) it generates new knowledge about how much additional data would be needed to estimate a high-quality regime and consequently a provides a sense of how 'underpowered' existing SMARTs are for estimating optimal treatment regimes; and (iii) it provides theoretical guarantees for bootstrap oversampling for sample size calculations that are of independent interest. Furthermore, the proposed criteria used to derive our samples size procedures are closely related to those used in Laber et al. (2015)

to size a single stage two-arm trial to estimate an optimal regime, however, the current procedure applies to multistage trials and provides considerably stronger performance guarantees for the estimated optimal regime.

In Section 2.1, we provide the setup and notation. In Section 2.2, we derive our sample size procedures and state their theoretical properties. In Section 2.3, we evaluate the finite sample performance of the proposed sample size procedures in a series of simulation experiments. A discussion of the proposed methodology and open problems is provided in Section 2.4.

2.1 Setup and notation

We consider choosing the sample size, n, for a two-stage SMART that will produce data, $\mathscr{D}_n = \left\{(\boldsymbol{X}_{1,i}, A_{1,i}, \boldsymbol{X}_{2,i}, A_{2,i}, Y_i)\right\}_{i=1}^{n}$, which comprises $i.i.d.$ trajectories of the form $(\boldsymbol{X}_1, A_1, \boldsymbol{X}_2, A_2, Y)$ where: $\boldsymbol{X}_1 \in \mathbb{R}^{p_1}$ denotes baseline subject information; $A_1 \in \{-1, 1\}$ denotes the first assigned treatment; $\boldsymbol{X}_2 \in \mathbb{R}^{p_2}$ denotes subject information collected during the course of the first treatment; $A_2 \in \{-1, 1\}$ denotes the second assigned treatment; and $Y \in \mathbb{R}$ denotes the outcome, coded so that higher is better. Define $\boldsymbol{H}_1 = \boldsymbol{X}_1$ and $\boldsymbol{H}_2 = (\boldsymbol{X}_1^\intercal, A_1, \boldsymbol{X}_2^\intercal)^\intercal$ so that $\boldsymbol{H}_t$ denotes the history at time $t = 1, 2$. For simplicity, we assume that the trial will employ simple one-to-one randomization so that $P(A_t = a_t \big| \boldsymbol{H}_t) = 1/2$ with probability one for $a_t \in \{-1, 1\}$, $t = 1, 2$; extensions to more complex randomization schemes including those with feasible sets of treatments is straightforward (Schulte et al., 2014). A treatment regime in this context is a pair of functions $\boldsymbol{\pi} = (\pi_1, \pi_2)$ where $\pi_t : \text{dom}\, \boldsymbol{H}_t \rightarrow \text{dom}\, A_t$ so that a decision maker following $\boldsymbol{\pi}$ would recommend treatment $\pi_t(\boldsymbol{h}_t)$ to a patient presenting with $\boldsymbol{H}_t = \boldsymbol{h}_t$ at time $t = 1, 2$. We define a treatment regime as optimal if it leads to maximal mean outcome if applied to the population from which $\mathscr{D}_n$ is drawn (other definitions of optimality are possible, see Kosorok and Moodie, 2015; Linn et al., 2016). To formally define an optimal treatment regime, we use potential outcomes (Rubin, 1978; Splawa-Neyman

et al., 1990).

Let $\boldsymbol{H}_2^*(a_1)$ denote the potential second history under initial treatment a_1 and $Y^*(a_1,a_2)$ the potential outcome under treatment sequence (a_1,a_2). The potential outcome under a regime π is

$$Y^*(\pi)=\sum_{(a_1,a_2)} Y^*(a_1,a_2)1_{\pi_1(\boldsymbol{H}_1)=a_1}1_{\pi_2\{\boldsymbol{H}_2^*(a_1)\}=a_2},$$

where 1_u is an indicator that u is true. For any regime, π, define $V(\pi)=\mathbb{E}Y^*(\pi)$; an optimal regime, π^{opt}, satisfies $V(\pi^{\text{opt}})\geq V(\pi)$ for all π. Our sample size procedures depend on an estimator of π^{opt}, in order to construct such an estimator, we make the following assumptions: (C1) sequential ignorability, $\left\{\boldsymbol{H}_2^*(a_1), Y^*(a_1,a_2):(a_1,a_2)\in\{-1,1\}^2\right\}\perp A_t\Big|\boldsymbol{H}_t$ for $t=1,2$; (C2) positivity, $P(A_t=a_t|\boldsymbol{H}_t)>0$ with probability one for each $a_t\in\{-1,1\}$ for $t=1,2$; and (C3) consistency, $Y=Y^*(A_1,A_2)$ and $\boldsymbol{H}_2=\boldsymbol{H}_2^*(A_1)$. These assumptions are standard in the context of estimating optimal treatment regimes (Robins, 2004b; Schulte et al., 2014; Chakraborty and Moodie, 2013) with (C1) and (C2) holding by design in a SMART.

Under these assumptions, the optimal regime can be characterized in terms of the data-generating model as follows. Define $Q_2(\boldsymbol{h}_2,a_2)=\mathbb{E}(Y|\boldsymbol{H}_2=\boldsymbol{h}_2,A_2=a_2)$ and $Q_1(\boldsymbol{h}_1,a_1)=\mathbb{E}\left\{\max_{a_2}Q_2(\boldsymbol{H}_2,a_2)\Big|\boldsymbol{H}_1=\boldsymbol{h}_1,A_1=a_1\right\}$, then $\pi_t^{\text{opt}}(\boldsymbol{h}_t)=\arg\max_{a_t}Q_t(h_t,a_t)$ for $t=1,2$ (see Murphy, 2005b; Schulte et al., 2014). Furthermore, it can be seen that $V(\boldsymbol{\pi}^{\text{opt}})=\mathbb{E}\max_{a_1}Q_1(\boldsymbol{H}_1,a_1)$. Our sample size procedures are based on constructing estimators of $Q_t(\boldsymbol{h}_1,a_t)$ for $t=1,2$ and subsequently deriving plug-in estimators of π; these procedures vary in the structure we impose on these functions. Before describing specific estimators, we state properties of these estimators that we would like to ensure hold with high-probability provided the sample size is sufficiently large.

Let $\widehat{\pi}_n$ denote an estimator of π^{opt} and let $B_0>0$, $\gamma,\alpha,\eta,\epsilon,\zeta\in(0,1)$ be constants. Our goal is to choose n so that:

(POW) there exists an α-level test of $H_0:V(\boldsymbol{\pi}^{\text{opt}})\leq B_0$ based on $\widehat{\pi}_n$ that has power at least $(1-\gamma)\times 100+o(1)$ provided $V(\pi^{\text{opt}})\geq B_0+\eta$;

(OPT) $P\left[\mathbb{E}\left\{Y^*(\widehat{\boldsymbol{\pi}}_n)\middle|\mathscr{D}_n\right\} \geq V(\boldsymbol{\pi}^{\text{opt}}) - \epsilon\right] \geq 1 - \zeta + o(1)$.

Condition (POW) ensures sufficient power to test the effectiveness of the optimal treatment regime relative to some baseline expected outcome, B_0, e.g., the expected outcome under some standard of care. Condition (OPT) ensures that the expected performance of the estimated optimal regime will be near-optimal with high-probability. These conditions are analogous to those used to size a one-stage clinical trial for estimation of an optimal regime except that (OPT) controls the performance of the estimated optimal regime whereas Laber et al. (2015) control the *estimated performance* of the estimated optimal regime. Furthermore, like the one-stage setting, our sample size procedures depend on approximating the sampling distribution of an estimator of $\mathbb{E}\, Y^*(\pi^{\text{opt}})$; however, as we will later illustrate, constructing a high-quality approximation is markedly more complex in the multistage setting (see also Dawid, 1994; Chakraborty et al., 2009; Moodie et al., 2010; Hirano and Porter, 2012; Chakraborty et al., 2014; Laber et al., 2014; Luedtke and Van Der Laan, 2016).

2.2 Sample size procedures

We derive two sample size procedures. The first procedure imposes more parametric structure on the joint distribution of $(\boldsymbol{X}_1, A_1, \boldsymbol{X}_2, A_2, Y)$ than is typical in Q-learning and thereby avoids (or rather assumes away) some of the complexities associated with non-regularity and exceptional laws (Robins, 2004b; Chakraborty et al., 2009; Moodie et al., 2010; Chakraborty et al., 2013, 2014; Laber et al., 2014; Song et al., 2015b). The second proposed procedure does not impose as much parametric structure but at the expense of a more complex and potentially conservative sample size estimator.

2.2.1 Normality-based sample size procedure

We make the following assumptions about the generative model:

(AN1) $Q_2(\boldsymbol{h}_2, a_2) = \boldsymbol{h}_{2,0}^{\intercal}\beta_{2,0}^* + a_2\boldsymbol{h}_{2,1}^{\intercal}\beta_{2,1}^*$, where $\boldsymbol{h}_{2,0} \in \mathbb{R}^{p_{2,0}}$, $\boldsymbol{h}_{2,1} \in \mathbb{R}^{p_{2,1}}$ are summaries of $\boldsymbol{h}_2$ and $\beta_{2,0}^* \in \mathbb{R}^{p_{2,0}}$, $\beta_{2,1}^* \in \mathbb{R}^{p_{2,1}}$ are unknown parameters;

(AN2) $\mathbb{E}\left(\boldsymbol{H}_{2,0}^{\intercal}\beta_{2,0}^* \middle| \boldsymbol{H}_1 = \boldsymbol{h}_1, A_1 = a_1\right) = \boldsymbol{h}_{1,0}^{\intercal}\xi_{1,0}^* + a_1\boldsymbol{h}_{1,1}^{\intercal}\xi_{1,1}^*$, where $\boldsymbol{h}_{1,0} \in \mathbb{R}^{p_{1,0}}$, $\boldsymbol{h}_{1,1} \in \mathbb{R}^{p_{1,1}}$ are summaries of $\boldsymbol{h}_1$ and $\xi_{1,0}^* \in \mathbb{R}^{p_{1,0}}$, $\xi_{1,1}^* \in \mathbb{R}^{p_{1,1}}$ are unknown parameters;

(AN3) $\boldsymbol{H}_{2,1}^{\intercal}\beta_{2,1}^* = \boldsymbol{H}_{1,2}^{\intercal}\varpi_{1,2}^* + A_1\boldsymbol{H}_{1,3}^{\intercal}\varpi_{1,3}^* + \tau^* Z$, where $\boldsymbol{H}_{1,2} \in \mathbb{R}^{p_{1,2}}$, $\boldsymbol{H}_{1,3} \in \mathbb{R}^{p_{1,3}}$ are summaries of $\boldsymbol{H}_1$, Z is a standard normal random variable which is independent of $\boldsymbol{H}_1, A_1$, and $\tau^* > 0$, $\varpi_{1,2}^* \in \mathbb{R}^{p_{1,2}}$, $\varpi_{1,3}^* \in \mathbb{R}^{p_{1,3}}$, are unknown parameters;

(AN4) $\left(\boldsymbol{H}_{1,0}^{\intercal}\xi_{1,0}^*, \boldsymbol{H}_{1,1}^{\intercal}\xi_{1,1}^*, \boldsymbol{H}_{1,2}^{\intercal}\varpi_{1,2}^*, \boldsymbol{H}_{1,3}^{\intercal}\varpi_{1,3}^*\right)^{\intercal} \sim \text{Normal}(\omega^*, \Omega^*)$, where $\omega^* \in \mathbb{R}^4$ and $\Omega^* \in \mathbb{R}^{4\times 4}$ are unknown parameters.

Assumptions (AN1)-(AN3) are similar to those used in interactive Q-learning (IQ-learning Laber et al., 2014) except that in IQ-learning, (AN3) is replaced with a more general location-scale model than the normal linear model used here. The summaries of the history $\boldsymbol{h}_t$ for $t = 1, 2$ can include basis expansions or other non-linear terms as needed. These assumptions were motivated by a desire to create a generative model that is conceptually consistent with the analysis model used in linear Q-learning which remains the most commonly used method for estimating an optimal treatment regime from SMARTs. Assumption (AN4) is not required by IQ-learning as it conditions on $\boldsymbol{H}_1$. The assumption of joint normality could be relaxed, for example, by using a copula or semi-parametric model, but at the expense of more complex expressions that are less amenable to sample size calculations. Therefore we shall not consider such generalizations further.

The following results, which are proved in Appendix A, will be used to inform the construction of an estimator of the optimal treatment regime (see Schulte et al., 2014; Laber et al., 2014, for related expressions). Let Φ denote the cumulative distribution function of a standard normal random variable.

Lemma 2.2.1. *Assume (AN1)-(AN3). For any* $\beta_1 = (\xi_{1,0}^\intercal, \xi_{1,1}^\intercal, \varpi_{1,2}^\intercal, \varpi_{1,3}^\intercal)^\intercal$ *define*

$$Q_1(\boldsymbol{h}_1, a_1; \beta_1, \tau) = \boldsymbol{h}_{1,0}^\intercal \beta_{1,0} + a_1 \boldsymbol{h}_{1,1}^\intercal \beta_{1,1} + \frac{2\tau}{\sqrt{2\pi}} \exp\left\{ -\frac{\left(\boldsymbol{h}_{1,2}^\intercal \beta_{1,2} + a_1 \boldsymbol{h}_{1,3}^\intercal \beta_{1,3}\right)^2}{2\tau^2} \right\} + \left(\boldsymbol{h}_{1,2}^\intercal \beta_{1,2} + a_1 \boldsymbol{h}_{1,3}^\intercal \beta_{1,3}\right)\left[1 - 2\Phi\left\{ -\frac{\left(\boldsymbol{h}_{1,2}^\intercal \beta_{1,2} + a_1 \boldsymbol{h}_{1,3}^\intercal \beta_{1,3}\right)}{\tau} \right\}\right].$$

Then, $Q_1(\boldsymbol{h}_1, a_1) = Q_1(\boldsymbol{h}_1, a_1; \beta_1^*, \tau^*)$.

Let $W(\boldsymbol{H}_1, \beta_1) = (\boldsymbol{H}_{1,0}^\intercal \xi_{1,0}, \boldsymbol{H}_{1,1}^\intercal \xi_{1,1}, \boldsymbol{H}_{1,2}^\intercal \varpi_{1,2}, \boldsymbol{H}_{1,3}^\intercal \varpi_{1,3})^\intercal$ and define $g : \mathbb{R}^4 \to \mathbb{R}$ as

$$g(\boldsymbol{v}) = \max_{\rho \in \{-1,1\}} \left(v_1 + \rho\, v_2 + \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(v_3 + \rho\, v_4)^2}{2} \right\} + (v_3 + \rho\, v_4)\left[1 - 2\Phi\{-\left(v_3 + \rho\, v_4\right)\}\right] \right);$$

it follows from Lemma (2.2.1) that $\max_{a_1} Q_1(\boldsymbol{H}_1, a_1) = \tau^* g\left\{ W(\boldsymbol{H}_1, \beta_1^*)/\tau^* \right\}$. Let $\psi(\boldsymbol{v}; \omega, \Omega)$ denote the density of a multivariate normal distribution with mean $\omega \in \mathbb{R}^4$ and covariance $\Omega \in \mathbb{R}^{4\times 4}$ and write vech(Σ) to denote the vector-half operator of symmetric matrix Σ (Henderson and Searle, 1979). The following result shows that $V(\boldsymbol{\pi}^{\text{opt}})$ is a smooth function of ω^*, τ^*, and Ω^*.

Corollary 2.2.2. *Assume (AN1)-(AN4) and let* $g : \mathbb{R}^4 \to \mathbb{R}$ *be defined as above. Then*

$$V(\boldsymbol{\pi}^{\text{opt}}) = \nu\{\tau^*, \omega^*, \text{vech}(\Omega^*)\} = \int_{\mathbb{R}^4} \tau^* g\left(\boldsymbol{v}/\tau^*\right) \psi(\boldsymbol{v}; \omega^*, \Omega^*) d\,\boldsymbol{v}.$$

Thus, given estimators $\widehat{\tau}_n$, $\widehat{\omega}_n$, and $\widehat{\Omega}_n$ of τ^*, ω^*, and Ω^*, one can use the preceding result to construct a plugin estimator of $V(\boldsymbol{\pi}^{\text{opt}})$. We next describe how to construct these estimators.

Let $\mathbb{P}_n$ denote the empirical measure. Define $Q_2(\boldsymbol{h}_2, a_2; \beta_2) = \boldsymbol{h}_{2,0}^\intercal \beta_{2,0} + a_2 \boldsymbol{h}_{2,1}^\intercal \beta_{2,1}$ and subsequently define $\widehat{\beta}_{2,n} = \arg\min_{\beta_2} \mathbb{P}_n \left\{ Y - Q_2(\boldsymbol{H}_2, A_2; \beta_2) \right\}^2$. Thus, $\widehat{Q}_{2,n}(\boldsymbol{h}_2, a_2) = Q_2(\boldsymbol{h}_2, a_2; \widehat{\beta}_{2,n})$ and the estimated optimal rule at the second stage is $\widehat{\pi}_{2,n}(\boldsymbol{h}_2) = \arg\max_{a_2} \widehat{Q}_{2,n}(\boldsymbol{h}_2, a_2)$. Define

$$\begin{aligned} \widehat{\xi}_{1,0,n}, \widehat{\xi}_{1,1,n} &= \arg\min_{\xi_{1,0}, \xi_{1,1}} \mathbb{P}_n \left(\boldsymbol{H}_{2,0}^\intercal \widehat{\beta}_{2,0,n} - \boldsymbol{H}_{1,0}^\intercal \xi_{1,0} - A_1 \boldsymbol{H}_{1,1}^\intercal \xi_{1,1} \right)^2 \\ \widehat{\varpi}_{1,2,n}, \widehat{\varpi}_{1,3,n} &= \arg\min_{\varpi_{1,2}, \varpi_{1,3}} \mathbb{P}_n \left(\boldsymbol{H}_{2,1}^\intercal \widehat{\beta}_{2,1,n} - \boldsymbol{H}_{1,2}^\intercal \varpi_{1,2} - A_1 \boldsymbol{H}_{1,3}^\intercal \varpi_{1,3} \right)^2 \end{aligned}$$

so that $\widehat{\beta}_{1,n} = (\widehat{\xi}_{1,0,n}^{\intercal}, \widehat{\xi}_{1,1,n}^{\intercal}, \widehat{\varpi}_{1,2,n}^{\intercal}, \widehat{\varpi}_{1,3,n}^{\intercal})^{\intercal}$ is the least-squares estimator of $\beta_1^* = \big(\xi_{1,0}^{*\intercal}, \xi_{1,1}^{*\intercal}, \varpi_{1,2}^{*\intercal}, \varpi_{1,3}^{*\intercal}\big)^{\intercal}$. In addition, define $\widehat{\tau}_n^2 = \mathbb{P}_n \left\{ \boldsymbol{H}_{2,1}^{\intercal} \widehat{\beta}_{2,1} - \boldsymbol{H}_{1,2}^{\intercal} \widehat{\varpi}_{1,2,n} - A_1 \boldsymbol{H}_{1,3}^{\intercal} \widehat{\varpi}_{1,3,n} \right\}^2$. The plugin estimator of $Q_1(\boldsymbol{h}_1, a_1)$, based on (2.2.1), is $\widehat{Q}_{1,n}(\boldsymbol{h}_1, a_1) = Q_1(\boldsymbol{h}_1, a_1; \widehat{\beta}_{1,n}, \widehat{\tau}_n)$ and the estimated optimal decision rule at the first stage is $\widehat{\pi}_{1,n}(\boldsymbol{h}_1) = \arg\max_{a_1} \widehat{Q}_{1,n}(\boldsymbol{h}_1, a_1)$. Furthermore, define $\widehat{\omega}_n = \mathbb{P}_n W(\boldsymbol{H}_1, \widehat{\beta}_{1,n})$ and $\widehat{\Omega}_n = \mathbb{P}_n \left\{ W(\boldsymbol{H}_1, \widehat{\beta}_{1,n}) - \widehat{\omega}_n \right\} \left\{ W(\boldsymbol{H}_1, \widehat{\beta}_{1,n}) - \widehat{\omega}_n \right\}^{\intercal}$. The plugin estimator of $V(\boldsymbol{\pi}^{\text{opt}})$ is

$$\widehat{V}_n = \nu\big\{\widehat{\tau}_n, \widehat{\omega}_n, \text{vech}\big(\widehat{\Omega}_n\big)\big\} = \int_{\mathbb{R}^4} \widehat{\tau}_n g(\boldsymbol{v}/\widehat{\tau}_n) \psi(\boldsymbol{v}; \widehat{\omega}_n, \widehat{\Omega}_n) d\boldsymbol{v}.$$

To establish consistency and asymptotic normality of $\widehat{V}_n$ we assume:

(AN5) $\{\tau^*, \omega^{*\intercal}, \text{vech}(\Omega^*)\}^{\intercal} \in \Theta \subseteq \mathbb{R}^{15}$, where Θ is compact;

(AN6) $\sqrt{n}\big[\big\{\widehat{\tau}_n, \widehat{\omega}_n^{\intercal}, \text{vech}(\widehat{\Omega}_n)^{\intercal}\big\}^{\intercal} - \{\tau^*, \omega^{*\intercal}, \text{vech}(\Omega^*)^{\intercal}\}^{\intercal}\big] \rightsquigarrow \text{Normal}(0, \Sigma^*)$, where $\Sigma^* \in \mathbb{R}^{15\times 15}$ is positive definite.

Condition (AN6) follows from moment conditions that are common in M-estimation; we provide sufficient conditions for (AN6) in Appendix A.

Lemma 2.2.3. *Assume (C1)-(C3) and (AN1)-(AN6). Then,*

$$\sqrt{n}\left\{\widehat{V}_n - V(\boldsymbol{\pi}^{\text{opt}})\right\} \rightsquigarrow \text{Normal}\big(0, \sigma^{*2}\big),$$

where $\sigma^{*2} = \nabla \nu\{\tau^*, \omega^*, \text{vech}(\Omega^*)\}^{\intercal} \Sigma \nabla \nu\{\tau^*, \omega^*, \text{vech}(\Omega^*)\}$.

Let $\widehat{\sigma}_n^2$ be a consistent estimator of σ^{*2} and let $z_{1-\varrho}$ the $(1-\varrho)$ quantile of a standard normal distribution, then a test that rejects when $\sqrt{n}\left\{\widehat{V}_n - B_0\right\}/\widehat{\sigma}_n \geq z_{1-\alpha}$ is an (asymptotic) α-level test of $H_0 : V(\boldsymbol{\pi}^{\text{opt}}) \leq B_0$ with power exceeding $\Phi(z_\alpha + \sqrt{n}\eta/\sigma^*) + o(1)$ when $V(\boldsymbol{\pi}^{\text{opt}}) \geq B_0 + \eta$. Thus, choosing $n = \left\lceil (\sigma^{*2}/\eta^2)\left\{\Phi^{-1}(1-\gamma) + z_{1-\alpha}\right\}^2 \right\rceil$, satisfies (POW) asymptotically. This expression depends on σ^*, which is unknown in general, thus, a value for σ^* must be elicited from domain experts or estimated from historical data.

The preceding sample size has a familiar form which is unsurprising as it is derived from a test statistic which is asymptotically normal. However, what is perhaps more surprising, is that a similar sample size formula can also be used to ensure that condition (OPT) holds under the following regularity conditions. Define $\Delta Q_j = Q_j(\boldsymbol{H}_j, 1) - Q_j(\boldsymbol{H}_j, -1)$ and $\Delta\widehat{Q}_{j,n} = \widehat{Q}_{j,n}(\boldsymbol{H}_j, 1) - \widehat{Q}_{j,n}(\boldsymbol{H}_j, -1)$ for $j = 1, 2$. To select n so that (OPT) also holds we further assume:

(AN7) there exists positive sequences $\left\{c_{n,j}\right\}_{n\geq 1}$ and $\left\{\ell_{n,j}\right\}_{n\geq 1}$ satisfying $\liminf_{n\to\infty} c_{n,j} \geq c_{0,j} > 0$ and $\liminf_{n\to\infty} \ell_{n,j} \geq \ell_{0,j} > 0$ such that

$$P\left\{\sqrt{n}\left|\Delta\widehat{Q}_j - \Delta Q_j\right| > t\right\} \leq \exp\left(-c_{n,j}\, t^{\ell_{n,j}}\right),$$

for all n and $j = 1, 2$;

(AN8) there exists $M_j, \kappa_j > 0$ such that $P\left(|\Delta Q_j| \leq \epsilon\right) \leq M_j \epsilon^{\kappa_j}$ for $j = 1, 2$ as $\epsilon \to 0$.

The preceding assumptions are relatively mild with (AN7) being weaker than requiring a subexponetial tail; e.g., (AN7) and (AN8) would be satisfied if the histories and outcomes are normally distributed. The following result characterizes the concentration of the marginal mean outcome under $\widehat{\pi}_n$ about π^{opt} which can subsequently be used to choose a sample size n that satisfies (OPT).

Lemma 2.2.4. *Assume (C1)-(C3) and (AN1)-(AN8). Then there exists K and $\delta > 0$ such that*

$$|V(\widehat{\pi}_n) - V(\pi^{\text{opt}})| \leq K n^{-\delta} |\widehat{V}_n - V(\pi^{\text{opt}})| + o_p(1/\sqrt{n}).$$

Corollary 2.2.5. *Assume (C1)-(C3) and (AN1)-(AN8). Then setting*

$$n = \left\lceil \left\{ \frac{\Phi^{-1}(1-\zeta)\sigma^*}{\epsilon} \right\}^2 \right\rceil,$$

satisfies (OPT).

Remark 2.2.6. Given pilot or historical data, one can construct a plug-in estimator of σ^{*2}. In the absence of such data, one can use an elicited value for the variance of Y under standard care as an *ad hoc* surrogate for σ^{*2}. Heuristic justification for this surrogate is as follows. If the variance of the outcome is at least as large under standard care as it is under the optimal regime *and* the parametric estimator $\widehat{V}_n$ is at least as efficient as the sample mean of n observations collected under the optimal policy, then

$$\begin{aligned}
\sigma^{*2} &= \lim_{n\to\infty} \text{Var}\left[\sqrt{n}\left\{\widehat{V}_n - V(\pi^{\text{opt}})\right\}\right] \\
&\leq \lim_{n\to\infty} \text{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{Y_i^*(\pi^{\text{opt}}) - V(\pi^{\text{opt}})\right\}\right] \\
&\leq \lim_{n\to\infty} \text{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{Y_i^*(\pi^{\text{soc}}) - V(\pi^{\text{soc}})\right\}\right] \\
&= \text{Var}\left\{Y^*(\pi^{\text{soc}})\right\},
\end{aligned}$$

where π^{soc} denotes standard of care. If one is unwilling to make the above assumptions, an alternative would be to inflate the elicited value for $\text{Var}\{Y^*(\pi^{\text{soc}})\}$ by a constant factor.

2.2.2 Projection-based sample size procedure

Despite a deluge of new estimators of optimal treatment regimes (see Zhang et al., 2017; Zhou et al., 2017; Laber and Staicu, 2017; Tao and Wang, 2017, and references therein), Q-learning with linear models remains among the most commonly used methods in practice. This popularity can be partly attributed to: (i) the heavy use of linear models in seminal papers on estimation of optimal treatment regimes (Murphy, 2003b; Robins, 2004b; Murphy, 2005b; Qian and Murphy, 2011); (ii) minimal requirements on the joint distribution of the data-generating model; (iii) theoretical tractability (Chakraborty et al., 2009; Moodie et al., 2010; Chakraborty et al., 2014; Laber et al., 2014); and (iv) good empirical performance even under some forms of misspecification (Schulte et al., 2014). Thus, our second sample size procedure is designed for the setting where analysts plan to estimate the optimal

regime using Q-learning with linear models. We do not assume that the analysis model is correctly specified nor do we impose any parametric structure on the generative model. However, unlike the procedure described in the preceding section, the resultant sample size procedure derived here relies on quantities that would be difficult to elicit from domain experts, we therefore require that one has suitable pilot data available; such data could be historical or collected as an internal pilot.

We assume that for $t = 1, 2$ one postulates models of the form $Q_t(\boldsymbol{h}_t, a_t; \mu_t) = \boldsymbol{h}_{t,0}^{\intercal}\mu_{t,0} + a_t\boldsymbol{h}_{t,1}^{\intercal}\mu_{t,1}$, where $\boldsymbol{h}_{t,0}, \boldsymbol{h}_{t,1}$ are summaries of $\boldsymbol{h}_t$ and $\mu_t = (\mu_{t,0}^{\intercal}, \mu_{t,1}^{\intercal})^{\intercal}$ are unknown parameters. Define $\mu_2^* = \arg\min_{\mu_2} P\left\{Y - Q_2(\boldsymbol{H}_2, A_2; \mu_2)\right\}^2$ and $\mu_1^* = \arg\min_{\mu_1} P\left\{\max_{a_2} Q_2(\boldsymbol{H}_2, a_2; \mu_2^*) - Q_1(\boldsymbol{H}_1, A_1; \mu_1)\right\}^2$. If (AN1) holds then, provided the requisite expectations exist, $Q_2(\boldsymbol{h}_2, a_2) = Q_2(\boldsymbol{h}_2, a_2; \mu_2^*)$; however, even if (AN1)-(AN4) hold, it need not follow that $Q_1(\boldsymbol{h}_1, a_1) = Q_1(\boldsymbol{h}_1, a_1; \mu_1^*)$ (Laber et al., 2014). Nevertheless, it is still meaningful to discuss the mean outcome under the estimated optimal regime when these models are misspecified. Define the optimal regime under linear Q-learning with the above class of models as $\boldsymbol{\pi}^{Q,\text{opt}} = (\pi_1^{Q,\text{opt}}, \pi_2^{Q,\text{opt}})$ so that $\pi_t^{Q,\text{opt}}(\boldsymbol{h}_t) = \arg\max_{a_t} Q_t(\boldsymbol{h}_t, a_t; \mu_t^*)$. [1] Define $\widehat{\mu}_{2,n} = \arg\min_{\mu_2} \mathbb{P}_n\left\{Y - Q_2(\boldsymbol{H}_2, A_2; \mu_2)\right\}^2$ and $\widehat{\mu}_{1,n} = \arg\min_{\mu_1}\left\{\max_{a_2} Q_2(\boldsymbol{H}_2, a_2; \widehat{\mu}_{2,n}) - Q_2(\boldsymbol{H}_1, A_1; \mu_1)\right\}^2$. The estimated optimal decision rule at time t is $\widehat{\pi}_{t,n}^{Q}(\boldsymbol{h}_t) = \arg\max_{a_t} Q_t(\boldsymbol{h}_t, a_t; \widehat{\mu}_{t,n})$.

It is well-known that $V(\boldsymbol{\pi}^{\text{opt}})$ is not a smooth functional of the generative model and consequently standard approaches for inference, e.g., the bootstrap or series approximations, will not hold without modification (Robins, 2004b; Moodie et al., 2010; Chakraborty et al., 2009, 2013, 2014; Laber et al., 2014; Luedtke and Van Der Laan, 2016). To derive a test which satisfies (POW) we invert a variant of a projection confidence interval (Berger and Boos, 1994; Robins, 2004b) for $V(\boldsymbol{\pi}^{Q,\text{opt}})$; the interval we propose holds regardless of misspecification of the Q-functions and does not require strong parametric assumptions on the underlying generative model. This approach requires a confidence set for (μ_1^*, μ_2^*)

[1] The notation $\boldsymbol{\pi}^{Q,\text{opt}}$ is a bit misleading in that if the Q-functions are misspecified (which we allow) it need not follow that $V(\boldsymbol{\pi}^{Q,\text{opt}}) \geq V^Q(\mu_1, \mu_2)$ for all $(\mu_1, \mu_2) \in \Theta$; we do not assume that $\boldsymbol{\pi}^{Q,\text{opt}}$ satisfies such an inequality. For additional discussion, see Qian and Murphy (2011) and references therein.

which we construct as follows. Let $\boldsymbol{C}_2$ be as in (AN7) and let

$$\widehat{\mathfrak{W}}_{2,n} = \left\{\mathbb{P}_n \boldsymbol{C}_2 \boldsymbol{C}_2^\intercal\right\}^{-1} \mathbb{P}_n \boldsymbol{C}_2 \boldsymbol{C}_2^\intercal \left(Y - \boldsymbol{C}_2^\intercal \widehat{\mu}_{2,n}\right) \left\{\mathbb{P}_n \boldsymbol{C}_2 \boldsymbol{C}_2^\intercal\right\}^{-1}$$

so that $\mathfrak{Z}_{2,n,\varepsilon} = \left\{\mu_2 : n(\mu_2 - \widehat{\mu}_{2,n})^\intercal \widehat{\mathfrak{W}}_{2,n}^{-1} (\mu_2 - \widehat{\mu}_{2,n}) \leq \chi_{1-\varepsilon,\dim(\boldsymbol{C}_2)}\right\}$ is a Wald-type $(1-\varepsilon)\times 100$ confidence set for μ_2^*, where $\chi^2_{q,k}$ is the qth quantile of a chi-square random variable with k degrees of freedom. For each μ_2 define

$$\mu_1^*(\mu_2) = \arg\min_{\mu_1} \mathbb{E}\left\{\max_{a_2} Q_2(\boldsymbol{H}_2, a_2; \mu_2) - Q_1(\boldsymbol{H}_1, A_1; \mu_1)\right\}^2,$$

so that $\mu_1^*(\mu_2)$ is denotes the population-level for the first-stage Q-function were it known that $\mu_2^* = \mu_2$; thus, $\mu_1^* = \mu_1^*(\mu_2^*)$. Define

$$\widehat{\mu}_{1,n}(\mu_2) = \arg\min_{\mu_1} \mathbb{P}_n\left\{\max_{a_2} Q_2(\boldsymbol{H}_2, a_2; \mu_2) - Q_1(\boldsymbol{H}_1, A_1; \mu_1)\right\}^2,$$

to be the least-squares estimator of $\mu_1^*(\mu_2)$. Let $\boldsymbol{C}_1 = (\boldsymbol{H}_{1,0}^\intercal, A_1 \boldsymbol{H}_{1,1}^\intercal)^\intercal$ and define

$$\mathfrak{W}_{1,n}(\mu_2) = \left\{\mathbb{P}_n \boldsymbol{C}_1 \boldsymbol{C}_1^\intercal\right\}^{-1} \mathbb{P}_n \boldsymbol{C}_1 \boldsymbol{C}_1^\intercal \left\{\max_{a_2} Q_2(\boldsymbol{H}_2, a_2; \mu_2) - \boldsymbol{C}_1^\intercal \widehat{\mu}_{1,n}(\mu_2)\right\} \left\{\mathbb{P}_n \boldsymbol{C}_1 \boldsymbol{C}_1^\intercal\right\}^{-1}.$$

A $(1-\varepsilon)\times 100\%$ Wald-type confidence set for $\mu_1^*(\mu_2)$ is

$$\mathfrak{Z}_{1,n,\varepsilon}(\mu_2) = \left\{\mu_1 : n\left[\mu_1 - \widehat{\mu}_{1,n}(\mu_2)\right]^\intercal \widehat{\mathfrak{W}}_{1,n}^{-1}(\mu_2)\left[\mu_1 - \widehat{\mu}_{1,n}(\mu_2)\right] \leq \chi_{1-\varepsilon,\dim(\boldsymbol{C}_1)}\right\}.$$

Thus, given $\varepsilon_1, \varepsilon_2 \in (0,1)$ with $\vartheta = \varepsilon_1 + \varepsilon_2 \leq 1$, a $(1-\vartheta)\times 100\%$ confidence set for (μ_1^*, μ_2^*) is $\Xi_{n,1-\vartheta} = \left\{(\mu_1, \mu_2) : \mu_2 \in \mathfrak{Z}_{2,n,1-\varepsilon_2} \text{ and } \mu_1 \in \mathfrak{Z}_{1,n,1-\varepsilon_1}(\mu_2)\right\}$.

For each μ_1, μ_2 define

$$\delta(\mu_1,\mu_2) = 4Y 1_{A_1 \boldsymbol{H}_{1,1}^\intercal \mu_{1,1} > 0} 1_{A_2 \boldsymbol{H}_{2,1}^\intercal \mu_{2,1} > 0} + 2\left\{1_{A_1 \boldsymbol{H}_{1,1}^\intercal \mu_{1,1} \le 0} - (1/2)\right\} \max_{a_1} Q_1(\boldsymbol{H}_1, a_1; \mu_1) + 4 1_{A_1 \boldsymbol{H}_{1,1}^\intercal \mu_{1,1} > 0} \left\{1_{A_2 \boldsymbol{H}_{2,1}^\intercal \mu_{2,1} \le 0} - (1/2)\right\} \max_{a_2} Q_2(\boldsymbol{H}_2, a_2; \mu_2),$$

and subsequently define $\widehat{V}_n^Q(\mu_1,\mu_2) = \mathbb{P}_n \delta(\mu_1,\mu_2)$ and its population-level analog $V^Q(\mu_1,\mu_2) = \mathbb{E}\delta(\mu_1,\mu_2)$. Then, $\widehat{V}_n(\widehat{\mu}_{n,1}, \widehat{\mu}_{2,n})$ is the augmented inverse probability weighted estimator of $V(\pi^{Q,\text{opt}})$ (Zhang et al., 2013) and $V^Q(\mu_1^*, \mu_2^*) = V(\pi^{Q,\text{opt}})$ (see also Qian and Murphy, 2011; Zhao et al., 2015). Define

$$\varsigma^2(\mu_1,\mu_2) = \mathbb{E}\left\{\delta(\mu_1,\mu_2) - \mathbb{E}\delta(\mu_1,\mu_2)\right\}^2$$

and

$$\widehat{\varsigma}_n^2 = \mathbb{P}_n\left\{\delta(\mu_1,\mu_2) - \mathbb{P}_n\delta(\mu_1,\mu_2)\right\}^2.$$

For any fixed (μ_1,μ_2), it follows that

$$\sqrt{n}\left\{\widehat{V}_n^Q(\mu_1,\mu_2) - V^Q(\mu_1,\mu_2)\right\} \rightsquigarrow \text{Normal}\left\{0, \varsigma^2(\mu_1,\mu_2)\right\},$$

provided that $\mathbb{E}\delta^2(\mu_1,\mu_2) < \infty$. Choose ϑ_1 and ϑ_2 such that $\vartheta_1 + \vartheta_2 = \alpha$, then the proposed α-level test for (POW) rejects when

$$\inf_{(\mu_1,\mu_2)\in\Xi_{n,1-\vartheta_1}} \left[\widehat{V}_n^Q(\mu_1,\mu_2) - \frac{z_{1-\vartheta_2}\widehat{\varsigma}_n(\mu_1,\mu_2)}{\sqrt{n}}\right] \ge B_0.$$

Under the null, $V(\pi^{Q,\text{opt}}) \leq B_0$, so that the type I error is bounded above by

$$
\begin{aligned}
P\left\{\inf_{(\mu_1,\mu_2)\in\Xi_{n,1-\vartheta_1}}\left[\widehat{V}_n^Q(\mu_1,\mu_2)-\frac{z_{1-\vartheta_2}\widehat{\varsigma}_n(\mu_1,\mu_2)}{\sqrt{n}}\right]\geq V(\pi^{Q,\text{opt}})\right\}& \\
\leq P\left\{\widehat{V}_n^Q(\mu_1^*,\mu_2^*)-\frac{z_{1-\vartheta_2}\widehat{\varsigma}_n(\mu_1^*,\mu_2^*)}{\sqrt{n}}\geq V(\pi^{Q,\text{opt}})\right\}+\vartheta_1+o(1)& \\
=P\left\{\sqrt{n}\left[\frac{\widehat{V}_n^Q(\mu_1^*,\mu_2^*)-V^Q(\mu_1^*,\mu_2^*)}{\widehat{\varsigma}_n(\mu_1^*,\mu_2^*)}\right]\geq z_{1-\vartheta_2}\right\}+\vartheta_1+o(1)& \\
\leq \vartheta_1+\vartheta_2+o(1),&
\end{aligned}
$$

where the first inequality follows from $P\left\{(\mu_1^*,\mu_2^*)\in\Xi_{n,1-\vartheta_1}\right\}\geq 1-\vartheta_1+o(1)$. If $\widehat{\varsigma}_n(\mu_1,\mu_2)>0$ with probability one for all $(\mu_1,\mu_2)\in\Theta$, then it can be seen that the power of the proposed test is

$$
\begin{aligned}
&P\left\{\inf_{(\mu_1,\mu_2)\in\Xi_{n,1-\vartheta_1}}\left[\widehat{V}_n^Q(\mu_1,\mu_2)-\frac{z_{1-\vartheta_2}\widehat{\varsigma}_n(\mu_1,\mu_2)}{\sqrt{n}}\right]\geq B_0\right\} \\
&\quad=P\left\{\inf_{(\mu_1,\mu_2)\in\Xi_{n,1-\vartheta_1}}\left[\frac{\sqrt{n}\left\{\widehat{V}_n^Q(\mu_1,\mu_2)-V^Q(\mu_1,\mu_2)\right\}}{\widehat{\varsigma}_n(\mu_1,\mu_2)}\right.\right. \\
&\qquad\qquad\left.\left.+\frac{\sqrt{n}\left\{V^Q(\mu_1,\mu_2)-B_0\right\}}{\widehat{\varsigma}_n(\mu_1,\mu_2)}\right]\geq z_{1-\vartheta_2}\right\} \qquad (2.1) \\
&\quad\geq P\left\{\inf_{(\mu_1,\mu_2)\in\Xi_{n,1-\vartheta_1}}\left(\frac{\sqrt{n}\left\{\widehat{V}_n^Q(\mu_1,\mu_2)-V^Q(\mu_1,\mu_2)\right\}}{\widehat{\varsigma}_n(\mu_1,\mu_2)}\right.\right. \\
&\qquad\qquad\left.\left.+\frac{\min\left[\sqrt{n}\left\{V^Q(\mu_1,\mu_2)-B_0\right\},\sqrt{n}\eta\right]}{\widehat{\varsigma}_n(\mu_1,\mu_2)}\right)\geq z_{1-\vartheta_2}\right\}. \qquad (2.2)
\end{aligned}
$$

The minimum in (2.2) is analogous to plugging-in the smallest possible difference under the alternative $V(\pi^{Q,\text{opt}})-B_0\geq\eta$; see Remark 2.2.9 for additional discussion. The sampling distribution of the test statistic under the alternative is complex and difficult to approximate using series approximations; thus, to estimate a sample size that will yield the desired power, we use the bootstrap.

2.2.2.1 Bootstrap power calculation

We assume that one has available pilot data $\mathscr{D}_{n_0} = \left\{(\boldsymbol{X}_{1,i}, A_{1,i}, \boldsymbol{X}_{2,i}, A_{2,i}, Y_i)\right\}_{i=1}^{n_0}$ comprising n_0 $i.i.d.$ trajectories from the same population from which trial participants will be drawn. We estimate the power required in (POW) using the bootstrap with a resample size of $n \geq n_0$ and solve for the smallest n such that the estimated power exceeds a given threshold. In our asymptotic analyses, we let both n and n_0 diverge to infinity; however, as we anticipate the trial sample size to be much larger than that of the pilot, we focus on an asymptotics in which n goes to infinity "first." We assume:

(PR1) $\mathbb{E}Y^2||\boldsymbol{C}_2||^2 < \infty$ and $\mathbb{E}||\boldsymbol{C}_1||^2||\boldsymbol{C}_2||^2 < \infty$;

(PR2) $\mathbb{E}\boldsymbol{C}_1\boldsymbol{C}_1^\intercal$ and $\mathbb{E}\boldsymbol{C}_2\boldsymbol{C}_2^\intercal$ are finite and strictly positive definite;

(PR3) $\inf_{(\mu_1,\mu_2)\in\Theta}\mathbb{E}\left\{\delta(\mu_1,\mu_2) - \mathbb{E}\delta(\mu_1,\mu_2)\right\}^2 > 0$ and $\sup_{(\mu_1,\mu_2)\in\Theta}\mathbb{E}\left\{\delta(\mu_1,\mu_2) - \mathbb{E}\delta(\mu_1,\mu_2)\right\}^2 < \infty$;

(PR4) the classes $\mathscr{F}_1 = \left\{\delta(\mu_1,\mu_2) : (\mu_1,\mu_2)\in\Theta\right\}$ and $\mathscr{F}_2 = \left\{\delta^2(\mu_1,\mu_2) : (\mu_1,\mu_2)\in\Theta\right\}$ are Donsker;

(PR5) $\mathbb{E}\delta(\mu_1,\mu_2)$ is uniformly continuous in a neighborhood of (μ_1^*,μ_2^*).

The foregoing assumptions are standard in linear Q-learning and mirror those used in linear regression (Laber et al., 2014).

Let $\mathbb{P}_{n,n_0}^{(b)}$ denote the bootstrap empirical distribution corresponding to a resample size of n. For any functional $Z_n = f(P, \mathbb{P}_{n_0})$ we define its bootstrap analog $Z_{n_0,n}^{(b)} = f\left\{\mathbb{P}_{n_0}, \mathbb{P}_{n,n_0}^{(b)}\right\}$. Let P_B denote probabilities computed with respect the bootstrap distribution conditional on the pilot data. The bootstrap estimator of the sample size required for (POW) is the

positive integer n which solves

$$P_B\left\{\inf_{(\mu_1,\mu_2)\in\Xi^{(b)}_{n_0,n,1-\vartheta_1}}\left(\frac{\sqrt{n}\left\{\widehat{V}^{Q(b)}_{n_0,n}(\mu_1,\mu_2)-\widehat{V}^{Q}_{n_0}(\mu_1,\mu_2)\right\}}{\widehat{\varsigma}^{(b)}_{n_0,n}(\mu_1,\mu_2)}\right.\right.$$
$$\left.\left.+\frac{\min\left[\sqrt{n}\left\{\widehat{V}^{Q}_{n_0}(\mu_1,\mu_2)-B_0\right\},\sqrt{n}\eta\right]}{\widehat{\varsigma}^{(b)}_{n_0,n}(\mu_1,\mu_2)}\right)\geq z_{1-\vartheta_2}\right\}\geq 1-\gamma$$

where $\vartheta_1+\vartheta_2=\alpha$; the probability on the left hand side of the inequality can be computed to desired precision by Monte Carlo methods. The following results establish consistency of the bootstrap as n_0 and n diverge.

Theorem 2.2.7. *Assume (C1)-(C3) and (PR1)-(PR4). Let $\vartheta_1\in(0,1)$ be fixed. Let $\varkappa, K>0$ be arbitrary, then*

$$\lim_{n,n_0\to\infty}P\left(\sup_{|v|\leq K}\left|P_B\left[\inf_{(\mu_1,\mu_2)\in\Xi^{(b)}_{n_0,n,1-\vartheta_1}}\frac{\sqrt{n}\left\{\widehat{V}^{Q(b)}_{n_0,n}(\mu_1,\mu_2)-\widehat{V}^{Q}_{n_0}(\mu_1,\mu_2)\right\}}{\widehat{\varsigma}^{(b)}_{n_0,n}(\mu_1,\mu_2)}\geq v\right]\right.\right.$$
$$\left.\left.-P\left[\inf_{(\mu_1,\mu_2)\in\Xi_{n,1-\vartheta_1}}\frac{\sqrt{n}\left\{\widehat{V}^{Q}_{n}(\mu_1,\mu_2)-V^{Q}(\mu_1,\mu_2)\right\}}{\widehat{\varsigma}_{n}(\mu_1,\mu_2)}\geq v\right]\right|>\varkappa\right)=0.$$

The preceding result does not include $\sqrt{n}\left\{V^Q(\mu_1,\mu_2)-B_0\right\}/\widehat{\varsigma}_n(\mu_1,\mu_2)$ (or its bootstrap analog) because, under the alternative, provided $V(\mu_1,\mu_2)>B_0$ for all (μ_1,μ_2) in a sufficiently small neighborhood of (μ_1^*,μ_2^*), this term (and its bootstrap analog) will diverge to infinity so that the conclusion of the above theorem will hold trivially. The following result characterizes the limiting tail behavior of this term; the factor of $\sqrt{n}$ on the right-hand-side of each probability assignment reflects the fact that, under the alternative, we expect $\sqrt{n}\left\{V^Q(\pi^{Q,\text{opt}})-B_0\right\}$ to diverge at rate $\sqrt{n}$.

Theorem 2.2.8. *Assume (C1)-(C3) and (PR1)-(PR4). Let $\vartheta_1\in(0,1)$ and $\eta\geq 0$ be fixed. In addition, assume that $n_0\to\infty$ as $n\to\infty$ and that there exists $c>0$ so that* $\inf_{(\mu_1,\mu_2)\in\Omega}\widehat{\varsigma}_n(\mu_1,\mu_2)\geq$

c. *Let* $\varkappa, K > 0$ *be arbitrary then*

$$\lim_{n,n_0\to\infty} P\Bigg(\sup_{|v|\le K} \Bigg| P_B \left[\inf_{(\mu_1,\mu_2)\in \Xi^{(b)}_{n_0,n,1-\vartheta_1}} \frac{\min\left[\sqrt{n}\left\{\widehat{V}^{Q,(b)}_{n_0,n}(\mu_1,\mu_2) - B_0\right\}, \sqrt{n}\eta\right]}{\widehat{\varsigma}^{(b)}_{n_0,n}(\mu_1,\mu_2)} \ge \sqrt{n}v \right] - P\left[\inf_{(\mu_1,\mu_2)\in \Xi_{n,1-\vartheta_1}} \frac{\min\left[\sqrt{n}\left\{\widehat{V}^{Q}_{n}(\mu_1,\mu_2) - B_0\right\}, \sqrt{n}\eta\right]}{\widehat{\varsigma}_{n}(\mu_1,\mu_2)} \ge \sqrt{n}v \right] \Bigg| > \varkappa \Bigg) = 0.$$

To choose n so that (OPT) holds asymptotically we make use of the following bound. For any (possibly data-dependent) sequence $(\widetilde{\mu}_{1,n}, \widetilde{\mu}_{2,n}) \in \Xi_{n,1-\vartheta_1}$ such that $\widehat{V}^Q_n(\mu^*_1, \mu^*_2) \le \widehat{V}^Q_n(\widetilde{\mu}_{1,n}, \widetilde{\mu}_{2,n}) + o_P(1/\sqrt{n})$ it follows that

$$P\Bigg[V^Q(\widetilde{\mu}_{1,n}, \widetilde{\mu}_{2,n}) \ge V^Q(\pi^{Q,\text{opt}}) + \inf_{(\mu_1,\mu_2)\in \Xi_{n,1-\vartheta_1}} \left\{\widehat{V}^Q_n(\mu_1,\mu_2) - V^Q(\mu_1,\mu_2)\right\} - \sup_{(\mu_1,\mu_2)\in \Xi_{n,1-\vartheta_1}} \left\{\widehat{V}^Q_n(\mu_1,\mu_2) - V^Q(\mu_1,\mu_2)\right\} \Bigg] \ge 1 - \vartheta_1 + o(1).$$

Thus, if $\mathfrak{Q}_{n,1-\vartheta_2,1-\vartheta_1}$ is the $(1-\vartheta_2)$ quantile of

$$\inf_{(\mu_1,\mu_2)\in \Xi_{n,1-\vartheta_1}} \left\{\widehat{V}^Q_n(\mu_1,\mu_2) - V^Q(\mu_1,\mu_2)\right\} - \sup_{(\mu_1,\mu_2)\in \Xi_{n,1-\vartheta_1}} \left\{\widehat{V}^Q_n(\mu_1,\mu_2) - V^Q(\mu_1,\mu_2)\right\}$$

then choosing $\vartheta_1 + \vartheta_2 \le \zeta$ and n such that $\mathfrak{Q}_{n,1-\vartheta_1,1-\vartheta_2}/\sqrt{n} \le \epsilon$ ensures that (OPT) holds asymptotically. Of course, $\mathfrak{Q}_{n,,1-\vartheta_1,1-\vartheta_2}$ is unknown so we estimtate it using the bootstrap, i.e., we select n so that $\mathfrak{Q}^{(b)}_{n_0,n,1-\vartheta_1,1-\vartheta_2}/\sqrt{n} \le \epsilon$.

Remark 2.2.9. To estimate the power at a given sample size one could use the bootstrap analog of (2.1). Indeed, the preceding theoretical results can be easily modified to hold without the min operation. However, the required sample size derived from (2.1) will be based on an *estimated effect size* rather than the minimal effect size of interest, η. A consequence of using the estimated effect size is that as the true effect size increases the estimated required sample size will decrease keeping the power fixed at (approximately) $(1-\gamma)\times 100$. However, in application, it desirable to have power $(1-\gamma)\times 100$ at effect size η but larger

power if the effect size exceeds η. Taking the minimum, as in (2.2), ensures that the power diverges to one as the true effect size grows large.

2.3 Simulation experiments

We examine the finite sample performance of the proposed sample size procedures using a series of simulation experiments. Performance is measured in terms of the proposed criteria (POW) and (OPT). For each generative model, we also compute the number of samples required to compare the mean outcomes under standard care to that under the fixed regimes $\boldsymbol{\pi}^{i,j}$, $i, j \in \{-1,1\}$, where $\pi_1^{i,j}(\boldsymbol{h}_1) \equiv i$ and $\pi_2^{i,j}(\boldsymbol{h}_2) \equiv j$. This comparison allows us to evaluate how much the sample size must be inflated to estimate and/or evaluate an optimal dynamic treatment regime relative to the comparison of fixed and embedded regimes (Almirall et al., 2012).

We first consider a generative model in which the assumptions (AN1)-(AN8) for the normality-based sample size procedure hold. This generative model is as follows:

$$
\begin{aligned}
&\boldsymbol{X}_1 \sim N_4\{0, \Omega_{AR1}(0.5)\}, && \boldsymbol{H}_{1,0}^T = (1, X_{1,0}),\\
&\boldsymbol{H}_{1,1}^T = (1, X_{1,1}), && \boldsymbol{H}_{1,2}^T = (1, X_{1,2}),\\
&\boldsymbol{H}_{1,3}^T = (1, X_{1,3}), && A_1, A_2 \sim_{i.i.d.} \text{Unif}\{-1,1\},\\
&\phi_1, \phi_2, \upsilon \sim_{i.i.d.} N(0,1), && X_{2,0} = \boldsymbol{H}_{1,0}^T \mu_{1,0}^* + A_1 \boldsymbol{H}_{1,1}^T \mu_{1,1}^* + \phi_1,\\
&X_{2,1} = \boldsymbol{H}_{1,2}^T \mu_{2,0}^* + A_1 \boldsymbol{H}_{1,3}^T \mu_{2,1}^* + \phi_2, && \boldsymbol{H}_{2,0}^T = (1, X_{1,0}, A_1, X_{2,0}),\\
&\boldsymbol{H}_{2,1}^T = (1, X_{1,2}, A_1, X_{2,1}), && Y = \boldsymbol{H}_{2,0}^T \beta_{2,0}^* + A_2 \boldsymbol{H}_{2,1}^T \beta_{2,1}^* + \upsilon,
\end{aligned}
$$

where $\Omega_{AR1}(0.5)$ is an autoregressive covariance matrix such that $\{\Omega_{AR1}(0.5)\}_{ij} = 0.5^{|i-j|}$.

Let $\pi^{\text{fixed,opt}}$ denote the optimal fixed regime such that $\pi^{\text{fixed,opt}} = \pi^{i^*,j^*}$, where $i^*, j^* = \arg\max_{i,j \in \{-1,1\}} V(\boldsymbol{\pi}^{i,j})$. We examine the performance of the proposed methods under parameter values which result in the following relationships between $\boldsymbol{\pi}^{\text{fixed,opt}}$, $\boldsymbol{\pi}^{\text{opt}}$, and B_0:

1. $V(\boldsymbol{\pi}^{\text{opt}}) = V(\boldsymbol{\pi}^{\text{fixed,opt}}) = B_0 + \eta$;

2. $V(\boldsymbol{\pi}^{\text{opt}}) - 0.5\eta = V(\boldsymbol{\pi}^{\text{fixed,opt}}) = B_0 + \eta$;

3. $V(\boldsymbol{\pi}^{\text{opt}}) - \eta = V(\boldsymbol{\pi}^{\text{fixed,opt}}) = B_0 + \eta$;

4. $V(\boldsymbol{\pi}^{\text{opt}}) - 2\eta = V(\boldsymbol{\pi}^{\text{fixed,opt}}) = B_0 + \eta$.

Define $\Delta = \left\{ V(\boldsymbol{\pi}^{\text{opt}}) - V(\boldsymbol{\pi}^{\text{fixed,opt}}) \right\} / \eta$ to be a measure of benefit associated with implementing an optimal dynamic treatment regime relative to the optimal fixed regime. It can be seen that Δ ranges from zero to two across the above scenarios. Parameter settings indexing the generative model which yield these values of Δ when $\eta = 1$ are:

1. $\mu_{1,0} = (-1,1)$, $\mu_{1,1} = (4,1)$, $\mu_{2,0} = (-0.4,-1)$, $\mu_{2,1} = (4,1)$, $\beta^*_{2,0} = (0.5,0.5,-1,1)$ and $\beta^*_{2,1} = (1,0.5,0.5,1)$;

2. $\mu_{1,0} = (-1,1)$, $\mu_{1,1} = (4,1)$, $\mu_{2,0} = (-0.4,-1)$, $\mu_{2,1} = (-4,1)$, $\beta^*_{2,0} = (0.5,0.5,-1,1)$ and $\beta^*_{2,1} = (1,-0.9,0.5,1)$;

3. $\mu_{1,0} = (-1,1)$, $\mu_{1,1} = (4,1)$, $\mu_{2,0} = (-0.4,-1)$, $\mu_{2,1} = (-4,1)$, $\beta^*_{2,0} = (0.5,0.5,-1,1)$ and $\beta^*_{2,1} = (1,-1.75,0.5,1)$;

4. $\mu_{1,0} = (-1,1)$, $\mu_{1,1} = (4,1)$, $\mu_{2,0} = (-0.4,-1)$, $\mu_{2,1} = (-4,1)$, $\beta^*_{2,0} = (0.5,0.5,-1,1)$ and $\beta^*_{2,1} = (1,-3.25,0.5,1)$.

Let $\gamma = 0.1$, $\alpha = 0.05$, $\zeta = 0.1$, and $\epsilon = 0.3$. Thus, if (POW) holds, then an α-level test of $H_0 : V(\boldsymbol{\pi}^{\text{opt}}) \leq B_0$ will have approximately 90% power, and if (OPT) holds, then $P\left[\mathbb{E}\left\{ Y^*(\widehat{\boldsymbol{\pi}}_n) \middle| \mathscr{D}_n \right\} \geq V(\boldsymbol{\pi}^{\text{opt}}) - 0.3\right] \geq 0.9 + o(1)$. Recall that the normality-based sample size procedure requires specification of σ^*. We consider three possibilities: (i) σ^* is known, e.g., correctly elicited from domain experts; (ii) σ^* is estimated using a pilot study of n_0 patients; and (iii) σ^* is estimated using the *ad hoc* procedure presented in Remark 2.2.9 using a sample of n_0 subjects treated under standard care wherein we assume that patients are assigned the optimal treatment 80% of the time and suboptimal treatment the remaining 20% of the time. To form a baseline for comparison, we also compute the sample size

required to power a test of the null $V(\boldsymbol{\pi}^{\text{fixed,opt}}) \leq B_0$ against the alternative $V(\boldsymbol{\pi}^{\text{fixed,opt}}) > B_0$ where it is assumed that $\boldsymbol{\pi}^{\text{fixed,opt}}$ is known *a priori* as is $\text{Var}\left\{ Y^*(\boldsymbol{\pi}^{\text{fixed,opt}}) \right\}$; this reflects the common practice of comparing a fixed regime against standard of care or another fixed regime. All results are based on 500 Monte Carlo replications.

Table 2.1 displays the average estimated sample size and its operating characteristics across the four settings of the proposed generative model and three approaches to selecting σ^* when sizing for just condition (POW). Table 2.2 displays the same results when sizing for condition (OPT). A table of results when sizing for both jointly is contained in Appendix A. In the case where $\Delta = 0$, the optimal treatment regime provides no benefit over the optimal fixed regime, thus, this setting reflects a worst-case in terms of the conservatism of sizing for (POW) and (OPT) rather than simply sizing to identify the optimal fixed regime; in this case, the proposed sample size procedure attains nominal levels for (POW) and (OPT) at the cost of an inflated sample size. However, as Δ increases, so that the benefits of personalizing treatment relative to a fixed regime also increase, it can be seen that the sample size required for (POW) and (OPT) can (perhaps surprisingly) be considerably smaller than required for identifying an optimal embedded regime provided that one has a high-quality estimate for σ^* either through elicitation or a pilot study. Table 2.2 has power 1.0 for all cases considered which is a consequence of using an upper bound on the difference between the value of the estimated regime and the optimal regime. One could potentially explore data-adaptive adjustments, e.g., the double bootstrap, to reduce this excess power.

We also examined the performance of the normality-based sample size when the postulated modeling assumptions are violated. For these simulations, we used the following

Table 2.1 Estimated power (POW) under a correctly specified generative model using the normality-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\widehat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	Method for σ^*	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}$	Med($\hat{n}$)	SD$\hat{n}$
0	known	50	90.0	-	74	130	—	—
0	pilot study	50	87.2	-	74	100.57	99	26.74
0	surrogate	50	98.6	-	74	186.44	185	44.65
0.5	known	50	100	-	111	124	—	—
0.5	pilot study	50	99.8	-	111	132.14	131	29.39
0.5	surrogate	50	100	-	111	164.06	159	46.60
1	known	50	100	-	151	134	—	—
1	pilot study	50	100	-	151	160.43	158	33.11
1	surrogate	50	100	-	151	203.78	199.5	62.16
2	known	50	100	-	251	165	—	—
2	pilot study	50	100	-	251	235.47	233	46.14
2	surrogate	50	100	-	251	280.04	275	82.16

generative model:

$$
\begin{aligned}
&X_1 \sim \mathrm{N}(0,1), && A_1, A_2 \sim_{i.i.d.} \mathrm{Unif}\{-1,1\},\\
&\phi \sim t_3, && X_2 = \mu_0^* + \mu_1^* X_1 + \mu_2^* A_1 + \mu_3^* A_1 X_1 + \mu_4^* X_1^2 + \phi,\\
&\boldsymbol{H}_{2,0} = (1, X_1, A_1, X_1 A_1, X_2), && \boldsymbol{H}_{2,1} = (1, A_1, X_2),\\
&\upsilon \sim N(0,1), && Y = \boldsymbol{H}_{2,0}^T \beta_{2,0}^* + A_2 \boldsymbol{H}_{2,1}^T \beta_{2,1}^* + \upsilon.
\end{aligned}
$$

As previously, we let $V(\pi^{\text{opt}}) - \Delta\eta = V(\pi^{\text{fixed,opt}}) = B_0 + \eta$ and consider $\Delta \in \{0, 0.5, 1, 2\}$. We set $\mu^* = (1, 0.5, 0.5, 0.1, 1)$ and choose β^* as follows:

1. $\Delta = 0$, $\beta_{2,0}^* = (1, 0.5, 0.5, 0.5, 1.5)$, $\beta_{2,1}^* = (-1, -1, 0)$;
2. $\Delta = 0.5$, $\beta_{2,0}^* = (1, 0.5, 0.5, 0.5, 1.5)$, $\beta_{2,1}^* = (-1, -1, 0.55)$;
3. $\Delta = 1$, $\beta_{2,0}^* = (1, 0.5, 0.5, 1, 1.5)$, $\beta_{2,1}^* = (-1, -1, 0.65)$;
4. $\Delta = 2$, $\beta_{2,0}^* = (1, 0.5, 0.5, 2.3, 1.5)$, $\beta_{2,1}^* = (-1, -1, 0.71)$.

Table 2.2 Estimated concentration (OPT) under a correctly specified generative model using the normality-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\widehat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	Method for σ^*	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}$	Med($\hat{n}$)	SD$\hat{n}$
0	known	50	-	100	74	277	—	—
0	pilot study	50	-	100	74	204.51	199.5	60.60
0	surrogate	50	-	100	74	394.63	390	90.04
0.5	known	50	-	100	111	263	—	—
0.5	pilot study	50	-	100	111	269.91	264	62.58
0.5	surrogate	50	-	100	111	353.35	343.5	98.09
1	known	50	-	100	151	285	—	—
1	pilot study	50	-	100	151	335.69	332.5	69.93
1	surrogate	50	-	100	151	431.93	420	130.70
2	known	50	-	100	251	352	—	—
2	pilot study	50	-	100	251	495.52	491	92.61
2	surrogate	50	-	100	251	605.82	578.5	176.10

The average sample size and operating characteristics of the normality-based sample size procedure when sizing for condition (POW) are displayed in Table 2.3 whereas the average sample size and characteristics for condition (OPT) are in Table 2.4. Results for sizing to guarantee both conditions jointly are contained in Appendix A. The proposed method continues to attain nominal levels for $\Delta \geq 1$, but is underpowered when using a pilot study to estimate σ^* and there is little or no benefit to the optimal regime over the optimal embedded regime.

We also applied the projection-based sample size procedure to the two classes of generative models described above. Each Monte Carlo replication consists of the following steps. We first generate a pilot study of size n_0. The bootstrap method described in Section 3.2 is used to calculate the minimum sample size $\hat{n}(\mathscr{D}_{n_0})$ to achieve power $(1-\gamma)\times 100\%$ using 100 bootstrap replications across a grid of potential sample sizes and then using nonlinear least squares to regress the estimated power on the sample sizes. Let $\gamma = 0.1$, $\vartheta_1 = 0.01$, and $\vartheta_2 = 0.04$. Which corresponds to 90% power for a test with 5% significance level based on

Table 2.3 Estimated power (POW) under a model which violated the normality assumptions using the normality-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\widehat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	Method for σ^*	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}$	Med($\hat{n}$)	SD$\hat{n}$
0	known	50	99.2	-	70	275	—	—
0	pilot study	50	72.4	-	70	77.96	59	74.25
0	surrogate	50	92.8	-	70	158.91	148	65.31
0.5	known	50	100	-	35	228	—	—
0.5	pilot study	50	80.0	-	35	65.14	46	79.77
0.5	surrogate	50	98.2	-	35	151.95	131	78.04
1	known	50	100	-	47	296	—	—
1	pilot study	50	89.0	-	47	77.77	56.5	72.61
1	surrogate	50	99.6	-	47	194.20	166	281.04
2	known	50	100	-	103	407	—	—
2	pilot study	50	99.4	-	103	139.88	118.5	90.88
2	surrogate	50	100	-	103	252.01	227.5	110.40

a confidence interval for $V(\boldsymbol{\pi}^{\text{opt}})$ that is being constructed using a 99% confidence set for (μ_1^*, μ_2^*) and a 96% interval for $V(\mu_1, \mu_2)$ for each fixed value of (μ_1, μ_2). Table 2.5 displays the results under the normal generative model when sizing for condition (POW); in some cases the pilot study shows no benefit to tailoring treatment, i.e., $\widehat{V}_{n_0} \leq B_0$, in which case $\widehat{n}(\mathscr{D}_0) = +\infty$. Table 2.6 displays the results when sizing for condition (OPT) under the normal generative model.

Table 2.7 show the results for when the projection-based method when sizing for condition (POW) is applied to the data generating model for which the normality assumptions do not hold. The results of sizing for condition (OPT) for the model which the normality assumptions do not hold is contained in Table 2.8. It can be seen that for $\Delta \geq 0.50$ the proposed procedure attains nominal power for (POW) for both generative models.

Table 2.4 Estimated concentration (OPT) under a model which violated the normality assumptions using the normality-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\widehat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	Method for σ^*	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}$	Med($\hat{n}$)	SD$\hat{n}$
0	known	50	-	100	70	586	—	—
0	pilot study	50	-	89.4	70	161.96	122	147.53
0	surrogate	50	-	100	70	339.32	317.5	134.60
0.5	known	50	-	100	35	485	—	—
0.5	pilot study	50	-	85.6	35	119.20	87	118.01
0.5	surrogate	50	-	100	35	322.16	285.5	190.63
1	known	50	-	100	47	630	—	—
1	pilot study	50	-	0.88	47	186.41	142.5	159.20
1	surrogate	50	-	0.99	47	390.41	344.5	191.12
2	known	50	-	100	103	866	—	—
2	pilot study	50	-	99.6	103	296.34	241	211.56
2	surrogate	50	-	100	103	537.83	490.5	277.45

2.4 Discussion

We proposed two sample size procedures for two-stage SMARTs when the objective is estimation and evaluation of an optimal dynamic treatment regime. These procedurs can be used to design SMARTs or conduct power analyses for observational studies. Furthermore, a comparison of the sample size required for construction of a high-quality estimator of an optimal treatment regime with the sample size required for comparison of fixed treatment sequences (or another simple comparison commonly used to size a SMART) can generate new insights into the cost of precision medicine in a given problem domain.

The proposed procedures were developed under two extremes in terms of the structure imposed on the underlying generative model. At one extreme, we assumed correctly specified parametric models for several functionals of the generative model including the optimal regime; and, at the other extreme, we only imposed moment conditions on a possibly misspecified analysis model. There is large class of intermediate models that could be

Table 2.5 Estimated power (POW) under a model for which the normality assumptions hold using the projection-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\widehat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}(\mathscr{D}_{n_0})$	$\text{Med}\{\hat{n}(\mathscr{D}_{n_0})\}$	$\text{SD}\hat{n}(\mathscr{D}_{n_0})$	$P\{\hat{n}(\mathscr{D}_0)=\infty\}$
0	50	85.17	-	74	381.94	342.5	112.46	0.42
0.5	50	99.13	-	111	316.01	292.5	83.89	0.21
1	50	99.73	-	151	312.39	299	71.96	0.17
2	50	100	-	251	364.95	361.5	62.78	0.06

Table 2.6 Estimated concentration (OPT) under a model for which the normality assumptions hold using the projection-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\widehat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}(\mathscr{D}_{n_0})$	$\text{Med}\{\hat{n}(\mathscr{D}_{n_0})\}$	$\text{SD}\hat{n}(\mathscr{D}_{n_0})$
0	50	-	100	74	114.73	113	9.72
0.5	50	-	100	111	107	106	9.59
1	50	-	99.8	151	112.22	111	10.69
2	50	-	99.4	251	124.3	123	13.22

constructed from these two base approaches. Furthermore, while the proposed approaches focused on regression-based estimators they can be extended to classification-based or direct-search estimators (Orellana et al., 2010; Zhang et al., 2012; Zhao et al., 2012; Zhang et al., 2012, 2013; Zhao et al., 2015; Zhou et al., 2017; Zhao et al., 2015; Laber and Zhao, 2015) which are becoming increasingly popular; we leave the details of this extension to future work.

Table 2.7 Estimated power (POW) under a model which violated the normality assumptions using the projection-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\widehat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}(\mathscr{D}_{n_0})$	$\text{Med}\{\hat{n}(\mathscr{D}_{n_0})\}$	$\text{SD}\hat{n}(\mathscr{D}_{n_0})$	$P\{\hat{n}(\mathscr{D}_0)=\infty\}$
0	50	84.28	-	70	482.18	401	250.04	0.34
0.5	50	92.96	-	35	527.52	491	208.04	0.41
1	50	98.55	-	47	556.37	507.5	227.93	0.28
2	50	99.49	-	103	594.68	536	281.35	0.14

Table 2.8 Estimated concentration (OPT) under a model which violated the normality assumptions using the projection-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\widehat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}(\mathscr{D}_{n_0})$	$\text{Med}\{\hat{n}(\mathscr{D}_{n_0})\}$	$\text{SD}\hat{n}(\mathscr{D}_{n_0})$
0	50	-	96.6	70	84.13	83	11.66
0.5	50	-	93.6	35	88.93	88	12.08
1	50	-	95.6	47	90.29	88	14.16
2	50	-	97.6	103	90.53	88	22.73

CHAPTER

3

Q-LEARNING FOR SURVIVAL ANALYSIS

A dynamic treatment regime is a sequential list of decision rules that map up-to-date patient information to a recommended treatment. An optimal dynamic treatment regime is one that maximizes the mean outcome of interest for patients in the population of interest when assigned treatments suggested by following the regime (Murphy, 2003b; Robins, 2004b). One common approach for the estimation of an optimal regime is Q-learning, a regression-based approximate dynamic programming method (Murphy, 2005b; Schulte et al., 2014). This has been applied in many different contexts such as ADHD (Nahum-Shani et al., 2012), depression (Song et al., 2015a), schizophrenia (Shortreed et al., 2010), smoking cessation (Chakraborty et al., 2010), and dosing for chronic pain (Laber et al., 2018).

In the context of chronic disease research such as cancer, HIV/AIDS, and cardiovascular disease, the outcome of interest is commonly a time to an event. Examples of potential

time-to-events of interest include overall survival time, disease-free survival time, and progression-free survival time. When dealing with time-to-event outcomes we frequently have data in which the event is not observed for some patients and is therefore considered to be censored. This could occur due to administrative censoring in which the event has not occurred by the time the study concludes. Some patients could also drop out of the study or there could be a competing risk that prevents the event of interest from being observed causing censoring. In the context of multiple decision regimes, a patient may experience the event of interest or be censored before reaching all of the possible decision points being considered. These issues will cause bias in a naive application of Q-learning to time-to-event data.

There has been some previous work on the estimation of dynamic treatment regimes for time-to-event outcomes. Bai et al. (2017) and Zhao et al. (2015) proposed using a value search estimator from a classification perspective for a one-stage regime. Hager et al. (2018) use value search for an augmented inverse probability weighted estimator from a classification perspective for estimation of an optimal two-stage regime. Xu et al. (2016) use a Bayesian framework for estimating the value of multi-stage dynamic treatment regimes based on disease status. They do not consider the estimation of an optimal regime though. Jiang et al. (2017) propose a value search estimator using kernel smoothing for multiple decision points. They assume that the timing of the decision points are prespecified and the same for all patients.

Goldberg and Kosorok (2012) proposed adjusting Q-learning by using inverse-probability-of-censoring weighting to correct for the bias. They assumed that censoring was independent of patient history though. Zhao et al. (2011) also use a Q-learning framework with modified support vector regression to adjust for bias due to censoring which is again assumed to be independent of patient history. In this paper we will propose a general framework for correcting for bias in the estimation of an optimal dynamic treatment regime using Q-learning via inverse weighting. We will allow censoring to occur at any point after

the initial treatment and depend on the patient history. The timing of the decision points will also be allowed to vary between patients.

In Section 3.1, we will provide setup and notation. In Section 3.2, we present the inverse weighting Q-learning algorithm. In Section 3.3, we will evaluate the performance of our algorithm with a series of simulation experiments. In Section 3.4, we demonstrate the use of our proposed algorithm on the STAR*D trial which was a study of the effectiveness of depression treatments for patients with a major depressive disorder. In Section 3.5, a discussion of the methodology is provided.

3.1 Setup and notation

We will consider the situation where there is a some fixed maximum number, K, of decision points that any patient could reach. We will assume we have data, $\mathscr{D}_n = \{(\kappa_i, \mathscr{T}_{1,i}, \boldsymbol{X}_{1,i}, A_{1,i}, \ldots, \mathscr{T}_{\kappa_i,i}, \boldsymbol{X}_{\kappa_i,i}, A_{\kappa_i,i}, U_i, \Delta_i)\}_{i=1}^n$ which comprises $i.i.d$ trajectories of the form $(\kappa, \mathscr{T}_1, \boldsymbol{X}_1, A_1, \ldots, \mathscr{T}_\kappa, \boldsymbol{X}_\kappa, A_\kappa, U, \Delta)$ such that κ denotes the number of observed decision points for the patient, $\mathscr{T}_k$ denotes the time of decision point k, $\boldsymbol{X}_1 \in \mathbb{R}^{p_1}$ denotes baseline patient information, $\boldsymbol{X}_k \in \mathbb{R}^{p_k}$ denotes covariates recorded during the course of the $(k-1)^{\text{st}}$ treatment for $k > 1$, $A_k \in \mathscr{A}_k$ denotes the treatment assigned at decision point k, U denotes the time to event or censoring, and Δ is an indicator for whether the event was observed. Let T be the potential time-to-event and C be the potential censoring time of a patient. Then $U = \min(T, C)$ and $\Delta = I(T \leq C)$.

Define $\boldsymbol{H}_k = \{(\mathscr{T}_j, \boldsymbol{X}_j, A_j) I(\kappa \geq j)$ for $j = 1, \ldots, k-1, (\mathscr{T}_k, \boldsymbol{X}_k) I(\kappa \geq k), \kappa I(\kappa < k), (U, \Delta) I(\kappa < k)\}$ for $k = 1, \ldots, K$ to be the history of a patient available to a decision maker at decision point k. Define $\boldsymbol{H}(u)$ to be the patient history up to time u such that $\boldsymbol{H}(u) = \{I(\kappa \geq j, u \geq \mathscr{T}_j), (\mathscr{T}_j, \boldsymbol{X}_j, A_j) I(\kappa \geq j, u \geq \mathscr{T}_j)$ for $j = 1, \ldots, K, I(u > U), (U, \Delta) I(u > U)\}$. A treatment regime is given by a sequence of K functions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ where $\pi_k : \text{dom}\, \boldsymbol{H}_k \rightarrow \text{dom}\, A_k$ so that a decision maker assigning treatments by following $\boldsymbol{\pi}$ would assign treatment $\pi_k(\boldsymbol{h}_k)$

to a patient with history $\boldsymbol{H}_k = \boldsymbol{h}_k$ at time $k = 1, \ldots, K$. We define a treatment regime as optimal if it maximizes the mean outcome when applied to the population of interest. We will use potential outcomes to formally define the value of a regime.

Let $\bar{\boldsymbol{A}}_k = (A_1, \ldots, A_k)$ denote a sequence of the first k treatments. Define $\boldsymbol{H}_k^*(\bar{\boldsymbol{a}}_{k-1})$ to be the potential history at the k^{th} decision point if assigned treatments $\bar{\boldsymbol{A}}_{k-1} = \bar{\boldsymbol{a}}_{k-1}$ previously. Let $\mathscr{T}_k^*(\bar{\boldsymbol{a}}_{k-1})$ be the potential time of the k^{th} decision point. We will let the censoring be considered an external process and will thus not have a potential outcome. Define $\varkappa^*(\bar{\boldsymbol{a}}_K)$ to be the potential number of decision points a patient would reach if they could not be censored and $\varkappa_k^*(\bar{\boldsymbol{a}}_k)$ will denote an indicator for whether a patient reaches the k^{th} decision point without experiencing the event. Let $T^*(\bar{\boldsymbol{a}}_{\varkappa^*(\bar{\boldsymbol{a}}_K)})$ be the potential time-to-event. Let $T^*(\pi)$ be the potential time-to-event for a patient if they were to receive treatments dictated by regime π at each decision point.

Let $S_k^*(\bar{\boldsymbol{a}}_{k-1})$ denote the potential time of the k^{th} decision point or the potential event time if it occurs after the $(k-1)^{\text{st}}$ and before the k^{th} decision point. If the potential event occurs before the $(k-1)^{\text{st}}$ decision point then let $S_k^*(\bar{\boldsymbol{a}}_{k-1}) = 0$. Let $\Upsilon_k^*(\bar{\boldsymbol{a}}_{k-1})$ denote whether the potential event occurs between the $(k-1)^{\text{st}}$ and k^{th} decision point such that

$$
\begin{aligned}
S_k^*(\bar{\boldsymbol{a}}_{k-1}) &= I\{\varkappa_{k-1}^*(\bar{\boldsymbol{a}}_{k-2}) = 1, \varkappa_k^*(\bar{\boldsymbol{a}}_{k-1}) = 0\} T^*(\bar{\boldsymbol{a}}_{k-1}) + I\{\varkappa_k^*(\bar{\boldsymbol{a}}_{k-1}) = 1\} \mathscr{T}_k^*(\bar{\boldsymbol{a}}_{k-1}) \\
\Upsilon_k^*(\bar{\boldsymbol{a}}_{k-1}) &= 1 \quad \text{if } \varkappa_{k-1}^*(\bar{\boldsymbol{a}}_{k-2}) = 1 \text{ and } \varkappa_k^*(\bar{\boldsymbol{a}}_{k-1}) = 0 \\
&= 0 \quad \text{if } \varkappa_{k-1}^*(\bar{\boldsymbol{a}}_{k-2}) = 0 \\
&= -1 \text{ if } \varkappa_k^*(\bar{\boldsymbol{a}}_{k-1}) = 1.
\end{aligned}
$$

Define the set of all possible potential outcomes as $W^* = \{\varkappa^*(\bar{\boldsymbol{a}}), \mathscr{T}_2^*(a_1), X_2^*(a_1), \mathscr{T}_3^*(\bar{\boldsymbol{a}}_2), X_3^*(\bar{\boldsymbol{a}}_2), \ldots, \mathscr{T}_{\varkappa^*(\bar{\boldsymbol{a}})}^*(\bar{\boldsymbol{a}}_{\varkappa^*(\bar{\boldsymbol{a}})-1}), X_{\varkappa^*(\bar{\boldsymbol{a}})}^*(\bar{\boldsymbol{a}}_{\varkappa^*(\bar{\boldsymbol{a}})-1}), T^*(\bar{\boldsymbol{a}}_{\varkappa^*(\bar{\boldsymbol{a}}_K)}),$ for all $\bar{\boldsymbol{a}} \in \bar{\mathscr{A}}\}$. To be able to construct an estimator for the value of a regime from the observed data we will make the following causal assumptions (C1 - C3). (C1) Consistency states that the observed data is equivalent to the potential outcomes for the treatments that were actually assigned.

Therefore

$$\kappa = \varkappa^*(\bar{\boldsymbol{A}}) \text{ and } U = T^*(\bar{\boldsymbol{A}}_\kappa) \text{ when } \Delta = 1$$
$$\kappa = \max_j\{j : \varkappa_j^*(\bar{\boldsymbol{A}}_{j-1}) = 1 \text{ and } U > \mathscr{T}_j^*(\bar{\boldsymbol{A}}_{j-1})\} \text{ when } \Delta = 0$$
$$X_k = X_k^*(\bar{\boldsymbol{A}}_{k-1}),\ \mathscr{T}_k = \mathscr{T}_k^*(\bar{\boldsymbol{A}}_{k-1}) \text{ for } k = 1,\ldots,\kappa.$$

(C2) Sequential ignorability states that the treatment assigned at each decision point is independent of the set of all potential outcomes conditional on the patient history. This can be written as $W^* \perp A_k | H_k, \kappa \geq k$ for $k = 1,\ldots,K$. (C3) Positivity states that there is a non-zero probability of each feasible treatment being assigned for a given patient history and a non-zero probability of not being censored before any decision point. Therefore

$$P(A_k = a_k | H_k = h_k, \kappa \geq k) > 0$$

for all feasible $a_k \in \mathscr{A}_k$ and for all $\boldsymbol{h}_k$ that satisfies $P(\boldsymbol{H}_k = \boldsymbol{h}_k, \kappa \geq k) > 0$. The non-zero probability of not being censored can be expressed formally as

$$P\{U \geq u | S_{k+1}^*(\bar{\boldsymbol{a}}_k) = u, H_k = h_k, A_k = a_k, \kappa \geq k\} > 0$$

for feasible $a_1 \in \mathscr{A}_1$ and history $\boldsymbol{h}_1$ such that $P\{S_{k+1}^*(\bar{\boldsymbol{a}}_k) = u, \boldsymbol{H}_k = \boldsymbol{h}_k, A_k = a_k, \kappa \geq k\} > 0$. Define

$$\lambda_C\{u|\boldsymbol{H}(u)\} = \lim_{du \to 0} du^{-1} P\{u \leq U < u + du, \Delta = 0 | U \geq u, \boldsymbol{H}(u)\}$$

to be the cause-specific hazard function for being censored at time u. We will assume that the censoring is noninformative in that it is independent of the set of potential outcomes conditional on the history. Therefore $W^* \perp C | \boldsymbol{H}_k, A_k, \kappa \geq k$, for $k = 1,\ldots,K$. This implies that $\lambda_C\{u|\boldsymbol{H}(u), W^*\} = \lambda_C\{u|\boldsymbol{H}(u)\}$.

Define $V(\boldsymbol{\pi}) = \mathbb{E}[f\{T^*(\boldsymbol{\pi})\}]$ to be the value of a fixed regime $\boldsymbol{\pi}$ where $f(\cdot)$ is a known,

monotone function. The choice of $f(\cdot)$ depends on the scientific question of interest. Some common quantities of interest for time-to-events are the survival time where $f(t) = t$, the probability of survival until time r where $f(t) = I(t \geq r)$, and the restricted lifetime where $f(t) = \min(t, L)$. Our goal is to estimate a regime π^{opt} such that $V(\pi^{\text{opt}}) \geq V(\pi)$ for all π.

3.2 Method

Q-learning is a regression-based approximate dynamic programming method for estimating an optimal treatment regime. Suppose we have data on patients receiving K treatments without censoring and a final outcome $Y \in \mathbb{R}$ coded such that higher values are better. In general Q-learning works by defining the Q-functions

$$
\begin{aligned}
Q_K(\boldsymbol{h}_K, a_K) &= \mathbb{E}(Y | \boldsymbol{H}_K = \boldsymbol{h}_K, A_K = a_k) \\
Q_k(\boldsymbol{h}_k, a_k) &= \mathbb{E}\Big\{ \max_{a_{k+1}} Q_{k+1}(\boldsymbol{H}_{k+1}, a_{k+1}) | \boldsymbol{H}_k = \boldsymbol{h}_k, A_k = a_k \Big\} \text{ for } k = 1, \ldots, K-1.
\end{aligned}
$$

It can then be shown via dynamic programming that $\pi_k^{\text{opt}}(\boldsymbol{h}_k) = \arg\max_{a_k} Q_k(\boldsymbol{h}_k, a_k)$ (Bellman, 1957). Models $Q_k(\boldsymbol{h}_k, a_k; \beta_k)$ are then posited for each Q-function and fit in a backwards iterative manner. First estimate $\hat{\beta}_K$ by regressing Y on $\boldsymbol{H}_K$ and A_K. For $k = K-1, \ldots, 1$ the Q-function is estimated by regressing $\max_{a_{k+1}} Q_{k+1}(\boldsymbol{H}_{k+1}, a_{k+1}; \hat{\beta}_{k+1})$ on $\boldsymbol{H}_k$ and A_k. This then gives a plug-in estimator for the optimal regimes as $\hat{\pi}_k^{\text{opt}}(\boldsymbol{h}_k) = \arg\max_{a_k} Q_k(\boldsymbol{h}_k, a_k; \hat{\beta}_k)$. Our proposed method for time-to-event outcomes will be based around the same framework with the addition of inverse weighting to correct for bias due to censoring.

For each individual i, let

$$\begin{aligned} S_{ki}^{\dagger} &= I(\kappa_i = k-1)U_i + I(\kappa_i \geq k)\mathscr{T}_{k,i} \\ \Upsilon_{ki}^{\dagger} &= 1 \quad \text{if } \kappa_i = k-1,\ \Delta_i = 1 \\ &= 0 \quad \text{if } \kappa_i = k-1,\ \Delta_i = 0 \\ &= -1 \quad \text{if } \kappa_i \geq k. \end{aligned}$$

Then $S_{ki}^{\dagger}$ is either the time of the k^{th} decision point or the time to event/censoring if it occurs before for a patient who has reached the $(k-1)^{\text{st}}$ decision point and $\Upsilon_{ki}^{\dagger}$ is an indicator for whether a patient observes the event, is censored, or reaches the k^{th} decision point. Define $\lambda_C(u|\boldsymbol{h}_k, a_k)$ to be the hazard function for being censored at time u given $\boldsymbol{H}_k = \boldsymbol{h}_k$ and $A_k = a_k$. We can then define the probability of not being censored until after time u after reaching decision point k as

$$\mathscr{K}_k(u|\boldsymbol{h}_k, a_k) = \exp\left\{-\int_{\tau_k}^{u} \lambda_C(w|\boldsymbol{h}_k, a_k)dw\right\}, \ \ \tau_k \leq u \leq S_{k+1}^{\dagger}$$

where $\mathscr{T}_k = \tau_k$ is contained in $\boldsymbol{H}_k = \boldsymbol{h}_k$ for $\kappa \geq k$. Define

$$Q_K(\boldsymbol{h}_K, a_K) = \mathbb{E}\left\{\frac{\Delta f(U)}{\mathscr{K}_K(U|\boldsymbol{h}_K, a_K)}\middle| \boldsymbol{H}_K = \boldsymbol{h}_K, A_K = a_K, \kappa = K\right\}.$$

Then

$$Q_K(\boldsymbol{h}_K, a_K) = \mathbb{E}\big[f\{T^*(\bar{a}_K)\}|\boldsymbol{H}_K = \boldsymbol{h}_k, A_K = a_K, \kappa = K\big].$$

Define $\Psi(\boldsymbol{h}_k) \subseteq \mathscr{A}_k$ to be the feasible treatment options at decision point k for a patient

with history $\boldsymbol{h}_k$. Therefore $\pi_K^{\text{opt}}(\boldsymbol{h}_K) = \arg\max_{a_k \in \Psi(\boldsymbol{h}_K)} Q_K(\boldsymbol{h}_K, a_K)$. Note that

$$\begin{aligned} Q_K(\boldsymbol{h}_K, a_K) &= \mathbb{E}\big[f\{T^*(\bar{a}_K)\} | \boldsymbol{H}_K = \boldsymbol{h}_K, A_K = a_K, \kappa = K\big] \\ &= \mathbb{E}\big[f(\tau_K) + f\{T^*(\bar{a}_K)\} - f(\tau_K) | \boldsymbol{H}_K = \boldsymbol{h}_K, A_K = a_K, \kappa = K\big] \\ &= f(\tau_K) + \mathbb{E}\big[f\{T^*(\bar{a}_K)\} - f(\tau_K) | \boldsymbol{H}_K = \boldsymbol{h}_K, A_K = a_K, \kappa = K\big]. \end{aligned}$$

Then let

$$Q_K^R(\boldsymbol{h}_K, a_K) = Q_K(\boldsymbol{h}_K, a_K) - f(\tau_K)$$

be the expected remaining time to event after reaching decision point K. Then if we posit a model for $Q_K^R(\boldsymbol{h}_K, a_K)$ that is strictly non-negative we will respect that $f(\cdot)$ is monotone and nondecreasing. Let

$$Q_K^R(\boldsymbol{h}_K, a_K; \beta_K) = \exp\{g_K(\boldsymbol{h}_K, a_K; \beta_K)\}.$$

Define

$$\begin{aligned} V_K(\boldsymbol{h}_K) &= \max_{a_K \in \Psi(h_K)} Q_K(\boldsymbol{h}_K, a_K) \\ &= f(\tau_K) + \max_{a_K \in \Psi(h_K)} Q_K^R(\boldsymbol{h}_K, a_K). \end{aligned}$$

Therefore $\pi_K^{\text{opt}}(\boldsymbol{h}_K) = \arg\max_{a_K \in \Psi(\boldsymbol{h}_K)} Q_K^R(\boldsymbol{h}_K, a_K)$ and

$$V_K(\boldsymbol{h}_K) = \mathbb{E}\big(f\big[T^*\{\bar{a}_{K-1}, \pi_K^{\text{opt}}(h_K)\}\big] | \boldsymbol{H}_K = \boldsymbol{h}_K, \kappa = K\big).$$

The estimated optimal regime at decision point K is then given by

$$\hat{\pi}_{Q,K}^{\text{opt}}(\boldsymbol{h}_K) = \arg \max_{a_K \in \Psi(\boldsymbol{h}_K)} Q_K^R(\boldsymbol{h}_K, a_K; \hat{\beta}_K).$$

To be able to estimate β_K though we also need an estimate for the probability of not being censored, $\mathscr{K}_K(U | \boldsymbol{h}_K, a_K)$. Model $\lambda_C\{u | \boldsymbol{H}(u)\}$ using a proportional hazards model of

the form

$$\lambda_C\{u|\boldsymbol{H}(u);\beta_C\} = \lambda_{C0}(u)\exp[g_C\{u,\boldsymbol{H}(u);\xi_C\}]$$

for an unspecified baseline hazard function $\lambda_{C0}(\cdot)$ and a function $g_C\{u,\boldsymbol{H}(u);\xi_C\}$ that depends on the history and a finite-dimensional parameter ξ_C. Let $\beta_C = \{\lambda_{C0}(t), \xi_C^T\}^T$. For example a posited model for $g_C\{u,\boldsymbol{H}(u);\xi_C\}$ could be given by

$$g_C\{u,\boldsymbol{H}(u);\xi_C\} = \xi_{C,1}^T\boldsymbol{H}_1 + \xi_{C,2}A_1 + \sum_{k=2}^{K} I(\kappa \geq k, \mathscr{T}_k \leq u)(\xi_{C,2k-1}^T\boldsymbol{H}_k + \xi_{C,2k}A_k).$$

We can then fit a model to the data to estimate $\lambda_C\{u|\boldsymbol{H}(u);\hat{\beta}_C\}$ and then substitute the fitted model into our expression for $\mathscr{K}_k(u|\boldsymbol{h}_k,a_k)$ to get

$$\hat{\mathscr{K}}_k(u|\boldsymbol{h}_k,a_k) = \exp\left\{-\int_{\tau_k}^{u}\lambda_C(w|\boldsymbol{h}_k,a_k;\hat{\beta}_C)dw\right\}, \quad k=1\ldots,K.$$

We estimate $\hat{\beta}_K$ by solving the inverse probability of censoring M-estimating equation

$$\sum_{\kappa_i=K}\left[\frac{\Delta_i}{\hat{\mathscr{K}}_K(U_i|\boldsymbol{H}_{Ki},A_{Ki})}\frac{\partial Q_K^R(\boldsymbol{H}_{Ki},A_{Ki};\beta_K)}{\partial\beta_K}\left\{f(U_i)-f(\mathscr{T}_{Ki})-Q_K^R(\boldsymbol{H}_{Ki},A_{Ki};\beta_K)\right\}\right]=0.$$

Proof that the above is an unbiased estimating equation can be found in Appendix B. Now move to decision $K-1$. Let

$$Q_{K-1}(\boldsymbol{h}_{K-1},a_{K-1}) = \mathbb{E}\left\{\frac{I(\Upsilon_K^\dagger=1)f(S_K^\dagger)+I(\Upsilon_K^\dagger=-1)V_K(\boldsymbol{H}_K)}{\mathscr{K}_{K-1}(S_K^\dagger|\boldsymbol{h}_{K-1},a_{K-1})} \right.$$
$$\left. |\boldsymbol{H}_{K-1}=\boldsymbol{h}_{K-1},A_{K-1}=a_{K-1},\kappa\geq K-1\right\}.$$

We then have that

$$Q_{K-1}(\boldsymbol{h}_{K-1},a_{K-1}) = \mathbb{E}\left(f\left[T^*\{\bar{a}_{K-1},\pi_K^{\text{opt}}(\boldsymbol{h}_K)\}\right]\middle|\boldsymbol{H}_{K-1}=\boldsymbol{h}_{K-1},A_{K-1}=a_{K-1},\kappa\geq K-1\right).$$

Form pseudo outcomes for individuals i such that $\kappa_i \geq K-1$ and $\Upsilon^{\dagger}_{K-1,i} \neq 0$,

$$\tilde{V}^R_{Ki} = I(\Upsilon^{\dagger}_{Ki}=1) f(S^{\dagger}_{Ki}) + I(\Upsilon^{\dagger}_{Ki}=-1)[f(S^{\dagger}_{Ki}) + \max_{a_K} Q^R_K\{\boldsymbol{h}_K, a_K; \hat{\beta}_K\}] - f(\mathscr{T}_{K-1,i}).$$

Then estimate $\hat{\beta}_{K-1}$ by solving the unbiased M-estimating equation

$$\sum_{\kappa_i \geq K-1} \left[\frac{I(\Upsilon^{\dagger}_{K,i} \neq 0)}{\hat{\mathscr{K}}_{K-1}(S^{\dagger}_{Ki} | \boldsymbol{H}_{K-1,i}, A_{K-1,i})} \frac{\partial Q^R_{K-1}(\boldsymbol{H}_{K-1,i}, A_{K-1,i}; \beta_{K-1})}{\partial \beta_{K-1}} \right.$$
$$\left. \left\{ \tilde{V}^R_{Ki} - Q^R_{K-1}(\boldsymbol{H}_{K-1,i}, A_{K-1,i}; \beta_{K-1}) \right\} \right] = 0.$$

The estimated optimal regime at decision point $K-1$ is then

$$\hat{\pi}^{\text{opt}}_{Q,K-1}(\boldsymbol{h}_{K-1}) = \arg \max_{a_{K-1} \in \Psi(\boldsymbol{h}_{K-1})} Q^R_{K-1}(\boldsymbol{h}_{K-1}, a_{K-1}; \hat{\beta}_{K-1}).$$

We then continue the same procedure for $k = K-2$ down to $k = 1$. In general for $k = K-1, \ldots, 1$,

$$Q_k(\boldsymbol{h}_k, a_k) = \mathbb{E}\left\{ \frac{I(\Upsilon^{\dagger}_{k+1,i}=1) f(S^{\dagger}_{k+1,i}) + I(\Upsilon^{\dagger}_{k+1,i}=-1) V_{k+1}(\boldsymbol{H}_{k+1})}{\mathscr{K}_k(S^{\dagger}_{k+1} | \boldsymbol{h}_k, a_k)} \right.$$
$$\left. | \boldsymbol{H}_k = \boldsymbol{h}_k, A_k = a_k, \kappa \geq k \right\}.$$

We estimate β_k by solving

$$\sum_{\kappa_i \geq k} \left[\frac{I(\Upsilon^{\dagger}_{k+1,i} \neq 0)}{\hat{\mathscr{K}}_k(S^{\dagger}_{k+1,i} | \boldsymbol{H}_{k,i}, A_{k,i})} \frac{\partial Q^R_k(\boldsymbol{H}_{k,i}, A_{k,i}; \beta_k)}{\partial \beta_k} \left\{ \tilde{V}^R_{k+1,i} - Q^R_k(\boldsymbol{H}_{k,i}, A_{k,i}; \beta_k) \right\} \right] = 0$$

such that for individuals i such that $\kappa_i \geq k$ and $\Upsilon^{\dagger}_{k,i} \neq 0$,

$$\tilde{V}^R_{k+1,i} = I(\Upsilon^{\dagger}_{k+1,i}=1) f(S^{\dagger}_{k+1,i}) + I(\Upsilon^{\dagger}_{k+1,i}=-1)[f(S^{\dagger}_{k+1,i}) + \max_{a_{k+1}} Q^R_{k+1}\{\boldsymbol{h}_{k+1}, a_{k+1}; \hat{\beta}_{k+1}\}] - f(\mathscr{T}_{k,i}).$$

The estimated optimal regime at decision point k is then given by

$$\hat{\pi}^{\text{opt}}_{Q,k}(\boldsymbol{h}_k) = \arg \max_{a_k \in \Psi(\boldsymbol{h}_k)} Q^R_k(\boldsymbol{h}_k, a_k; \hat{\beta}_k).$$

Let $\hat{\pi}^{\text{opt}}_Q = (\hat{\pi}^{\text{opt}}_{Q,1}, \ldots, \hat{\pi}^{\text{opt}}_{Q,K})$. An estimator for the value of the optimal regime is then given by

$$\hat{V}_Q(\pi^{\text{opt}}) = n^{-1} \sum_{i=1}^{n} \tilde{V}^R_{1,i}.$$

3.3 Simulation experiments

We will evaluate the performance of our proposed method using a series of simulation experiments. We will consider two different generative models for the data. The first will be the case in which the survival and censoring time follows a Weibull distribution. The other will be a piecewise constant hazard model for both the survival and censoring time. For each generative model we will compare scenarios with different levels of censoring. We will estimate the optimal treatment regime using three different methods. The first will be the proposed survival Q-learning. We will also compare the results with performing Q-learning by assuming the censoring is independent of the patient history and inverse weighting by using the Kaplan-Meier estimate of the probability of not being censored. The third method will be the naive use of Q-learning for a general outcome using only the patients that were not censored.

We will perform 1000 replications of the simulation study. For each replication an estimated optimal regime will be calculated using each of the three methods and the value of the optimal regime will be estimated. The true value of each estimated regime will then be calculated for each estimated regime using the g-computation algorithm with 1000 Monte-Carlo replications. Performance will be evaluated by examining the mean and standard deviations of the estimated value and true value of the estimated regimes across the 1000 replications. We will let the sample size be equal to 1000 for all of the simulations.

The first generative model we will consider is one in which the survival time, censoring time, and time to the next decision point all have a Weibull distribution. The generative model is as follows:

$$X_1 \sim N(0,1), \quad A_1 \sim \text{Unif}\{-1,1\}, \quad \mathscr{T}_1 = 0$$
$$\boldsymbol{H}_{1,0}^T = \boldsymbol{H}_{1,1}^T = (1, X_1).$$

Generate

$$C_1 = \text{Weibull}\{\text{shape} = \gamma_{C_1}, \text{scale} = \max(1, \boldsymbol{H}_{1,0}^T \zeta_{1,0} + A_1 \boldsymbol{H}_{1,1}^T \zeta_{1,1})\}$$
$$T_1 = \text{Weibull}\{\text{shape} = \gamma_{T_1}, \text{scale} = \max(1, \boldsymbol{H}_{1,0}^T \xi_{1,0} + A_1 \boldsymbol{H}_{1,1}^T \xi_{1,1})\}$$
$$\mathscr{T}_2 = \text{Weibull}\{\text{shape} = \gamma_{\mathscr{T}_2}, \text{scale} = \max(1, \boldsymbol{H}_{1,0}^T \omega_{1,0} + A_1 \boldsymbol{H}_{1,1}^T \omega_{1,1})\}.$$

If $C_1 = \min(C_1, T_1, \mathscr{T}_2)$ then the individual is censored at time C_1. If $T_1 = \min(C_1, T_1, \mathscr{T}_2)$ the individual has the event observed at time T_1. If $\mathscr{T}_2 = \min(C_1, T_1, \mathscr{T}_2)$ then the patient has reached the second decision point. Then

$$X_2 = \boldsymbol{H}_{1,0}^T \mu_{1,0} + A_1 \boldsymbol{H}_{1,1}^T \mu_{1,1} + \phi_1, \quad A_2 \sim \text{Unif}\{-1,1\}$$
$$\boldsymbol{H}_{2,0}^T = \boldsymbol{H}_{2,1}^T = (1, X_1, A_1, X_2)$$

and we generate

$$C_2 = \text{Weibull}\{\text{shape} = \gamma_{C_2}, \text{scale} = \max(1, \boldsymbol{H}_{2,0}^T \zeta_{2,0} + A_2 \boldsymbol{H}_{2,1}^T \zeta_{2,1})\}$$
$$T_2 = \text{Weibull}\{\text{shape} = \gamma_{T_2}, \text{scale} = \max(1, \boldsymbol{H}_{2,0}^T \xi_{2,0} + A_2 \boldsymbol{H}_{2,1}^T \xi_{2,1})\}$$
$$\mathscr{T}_3 = \text{Weibull}\{\text{shape} = \gamma_{\mathscr{T}_3}, \text{scale} = \max(1, \boldsymbol{H}_{2,0}^T \omega_{2,0} + A_2 \boldsymbol{H}_{2,1}^T \omega_{2,1})\}.$$

This identical pattern is continued for all 3 stages of treatments. Refer to Appendix B for

the parameter values used. The results are summarized in table 3.1. Note that the true performance of the estimated regimes for all of the methods are very similar to one another. The estimated value of the regime is close to the actual value for our proposed survival Q-learning method while the other competing methods underestimate the true value. As the degree of censoring increases this underestimation becomes even more severe. Our proposed method does lead to an increase in the variance of the estimated value.

Table 3.1 Estimated value of the estimated optimal regime and the true value of the estimated optimal regime for the case where the time-to-event and censoring follow a Weibull distribution.

Method	Censoring	$\mathbb{E}V(\hat{\pi}^{\text{opt}})$	$\text{SD}\,V(\hat{\pi}^{\text{opt}})$	$\mathbb{E}\hat{V}(\hat{\pi}^{\text{opt}})$	$\text{SD}\,\hat{V}(\hat{\pi}^{\text{opt}})$
Survival Q-learn	21.1%	16.426	0.182	16.844	0.862
KM Q-learn	21.1%	16.485	0.168	15.909	0.578
Naive Q-learn	21.1%	16.405	0.163	15.379	0.571
Survival Q-learn	33.7%	16.468	0.192	16.390	1.075
KM Q-learn	33.7%	16.514	0.170	15.231	0.733
Naive Q-learn	33.7%	16.408	0.176	14.446	0.715
Survival Q-learn	46.5%	16.461	0.202	16.032	1.219
KM Q-learn	46.5%	16.501	0.191	14.875	0.854
Naive Q-learn	46.5%	16.367	0.185	13.810	0.849

For the second generative model we will let the survival time, censoring time, and time to the next decision point all have a piecewise constant hazard model. The generative model is then given by:

$$X_1 \sim N(0,1), \quad A_1 \sim \text{Unif}\{-1,1\}, \quad \mathscr{T}_1 = 0$$

$$\boldsymbol{H}_{1,0}^T = \boldsymbol{H}_{1,1}^T = (1, X_1).$$

Let $(t_1, t_2, \ldots, t_J)$ be time points such that the hazard is constant within each interval (t_{j-1}, t_j).

The hazard function for survival time is given by

$$\lambda_{T1}(u|\boldsymbol{h}_1, a_1) = \begin{cases} j \max(0.5, \boldsymbol{h}_{1,0}^T \xi_{1,0} + a_1 \boldsymbol{h}_{1,1}^T \xi_{1,1})/J & \text{if } t_{j-1} < u \leq t_j \\ \max(0.5, \boldsymbol{h}_{1,0}^T \xi_{1,0} + a_1 \boldsymbol{h}_{1,1}^T \xi_{1,1}) & \text{if } u > t_J. \end{cases}$$

Similarly the cause-specific hazard function for the censoring time and time of the next decision point are given by

$$\lambda_{C1}(u|\boldsymbol{h}_1, a_1) = \begin{cases} j \max(0.5, \boldsymbol{h}_{1,0}^T \zeta_{1,0} + a_1 \boldsymbol{h}_{1,1}^T \zeta_{1,1})/J & \text{if } t_{j-1} < u \leq t_j \\ \max(0.5, \boldsymbol{h}_{1,0}^T \zeta_{1,0} + a_1 \boldsymbol{h}_{1,1}^T \zeta_{1,1}) & \text{if } u > t_J. \end{cases}$$

and

$$\lambda_{\mathscr{T}2}(u|\boldsymbol{h}_1, a_1) = \begin{cases} j \max(0.5, \boldsymbol{h}_{1,0}^T \omega_{1,0} + a_1 \boldsymbol{h}_{1,1}^T \omega_{1,1})/J & \text{if } t_{j-1} < u \leq t_j \\ \max(0.5, \boldsymbol{h}_{1,0}^T \omega_{1,0} + a_1 \boldsymbol{h}_{1,1}^T \omega_{1,1}) & \text{if } u > t_J. \end{cases}$$

We generate a time of event, censoring time, and time of the second decision point using each hazard function. As with the previous generative model if $C_1 = \min(C_1, T_1, \mathscr{T}_2)$ the individual is censored at time C_1. If $T_1 = \min(C_1, T_1, \mathscr{T}_2)$ then the individual has the event observed at time T_1 and if $\mathscr{T}_2 = \min(C_1, T_1, \mathscr{T}_2)$ then the patient has reached the second decision point. If the patient reaches the second stage decision point the intervening information collected and second stage decision are generated as

$$X_2 = \boldsymbol{H}_{1,0}^T \mu_{1,0} + A_1 \boldsymbol{H}_{1,1}^T \mu_{1,1} + \phi_1, \quad A_2 \sim \text{Unif}\{-1, 1\}$$
$$\boldsymbol{H}_{2,0}^T = \boldsymbol{H}_{2,1}^T = (1, X_1, A_1, X_2).$$

The piecewise hazard functions all have the same form as before and the generative model follows the same pattern through 3 stages of treatments. The parameter values used are contained in Appendix B. The results are summarized in table 3.2. As with the first generative

model, the true performance of the estimated regimes for all of the methods are very similar to one another. For this generative model all of the estimated values are less than the true value of the estimated regime. This is expected as the hazard model is not correctly specified which means our method does not guarantee an unbiased estimate of the true value. In this scenario, our proposed method still is the least biased of all the tested methods. For this generative model there is again a small increase in the variance of the estimated value for our proposed survival Q-learning method.

Table 3.2 Estimated value of the estimated optimal regime and the true value of the estimated optimal regime for the case where the time-to-event and censoring have a piecewise hazard model.

Method	Censoring	$\mathbb{E}V(\hat{\pi}^{\text{opt}})$	$\text{SD}\,V(\hat{\pi}^{\text{opt}})$	$\mathbb{E}\hat{V}(\hat{\pi}^{\text{opt}})$	$\text{SD}\,\hat{V}(\hat{\pi}^{\text{opt}})$
Survival Q-learn	29.4%	12.941	0.126	9.294	0.324
KM Q-learn	29.4%	13.020	0.123	8.555	0.196
Naive Q-learn	29.4%	12.781	0.088	6.988	0.137
Survival Q-learn	40.8%	12.986	0.126	8.849	0.305
KM Q-learn	40.8%	13.125	0.106	8.191	0.190
Naive Q-learn	40.8%	12.745	0.090	6.271	0.129

3.4 Analysis of STAR*D trial

We will also demonstrate usage of our method to the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (Rush et al., 2003). STAR*D was a randomized study conducted to compare the effectiveness of different treatment regimes for patients with Major Depressive Disorder (MDD). Effectiveness of treatments was assessed using the *Quick Inventory of Depressive Symptomatology* (QIDS) score which is a measure of the current severity of depression symptoms (Rush et al., 2003). This was used with both clinician and self-reported versions of the QIDS. The scores range from 0 to 27 such that higher values

correspond to worse depression symptoms. A patient with a clinician assessed QIDS score of less than or equal to 5 was considered to be in remission.

The study enrolled 4041 patients, each of which could potentially receive four different stages of treatments each of which could last up to 14 weeks. At any point during the course of a treatment a patient could request to move to the next stage if they could not tolerate the side effects of their current treatment. All of the patients were treated initially with citalopram (CIT) which belongs to the selective serotonin reuptake inhibitor (SSRI) class of antidepressants. If a patient did not enter remission in 14 weeks they progressed to level 2.

In level 2 patients were eligible to receive one of seven different treatment options. We will follow the convention used by Chakraborty and Moodie (2013) to classify the different treatment options into two different categories, SSRI and non-SSRI. In the second stage the potential SSRI options are setraline (SER), CIT+bupropion (BUP), CIT + buspirone (BUS), and CIT + cognitive psycotherapy (CT). The potential non-SSRI options are velafaxine (VEN), BUP, and CT. Patients who were assigned to CIT+CT or CT in level 2 were eligible to move to level 2A in which their level 2 treatment was supplemented with either VEN or BUP. For the analysis we will consider level 2A to be combined with level 2 and any patient who received a SSRI during 2 or 2A will be considered to receive an SSRI in the combined level.

Treatments in level 3 are also classified as either SSRIs or non-SSRIs. The SSRI treatments are an augmentation of any of the level 2 SSRI treatments with lithium (LI) or thyroid hormone (THY). The non-SSRI options are mirtazapine (MIRT), nontriptyline (NTP), or augmenting any of the level 2 non-SSRI treatments with LI or THY. In level 4 the treatment options were tranylcypromine (TCP) or MIRT+VEN. At the beginning of each level, each patient had the option to give a preference to either switch or augment the treatment given in the previous stage. If they gave a preference they were randomized between the options that were consistent with their preference.

For the analysis we will consider level 2 and 2a to be stage 1 for Q-learning since all patients received the same treatment in level 1. Similarly, level 3 and 4 will be stage 2 and 3

respectively. The outcome of interest will be given by the time to remission of depression symptoms which will be denoted by T. There was a significant amount of censoring due to drop out at each stage of the trial. At each of the stages we will include the QIDS score at the beginning of the stage, the slope of the QIDS score over the previous stage, and the patient preference as covariates in the analysis. This follows the analysis conducted in Pineau et al. (2007) and Chakraborty et al. (2013). Let X_{j1} be the beginning QIDS score, X_{j2} be the QIDS slope, and X_{j3} be the patient preference for stages $j = 1,2,3$. The patient preference is coded such that $X_{j3} = 1$ if they prefer to switch from their previous treatment and $X_{j3} = -1$ otherwise. Let the A_j denote the treatment for each stage j such that $A_j = 1$ if the patient was assigned a SSRI and $A_j = -1$ if the patient was assigned a non-SSRI.

Only 92 patients reach the third stage and of those only 16 go into remission before the end of the study. Due to the limited sample size we will restrict the covariates in the model to just depend on the features recorded during the course of the second stage treatment as well as the previous treatments assigned. Our model for the Q-function will then be given by

$$Q_3^R(\boldsymbol{h}_3, a_3; \beta_3, \psi_3) = \exp\{\beta_{30} + \beta_{31}A_1 + \beta_{32}A_2 + \beta_{33}X_{31} + \beta_{34}X_{32} + \beta_{35}X_{33} + (\psi_{30} + \psi_{31}A_1 + \psi_{32}A_2 + \psi_{33}X_{31} + \psi_{34}X_{32})A_3\}$$

and the cause-specific hazard function for censoring will be modeled using a cox proportional hazards model of the form

$$\lambda_C(u|\boldsymbol{h}_3, a_3) = \lambda_{C30}(u)\exp\{\xi_{3,0} + \xi_{3,1}X_{11} + \xi_{3,2}X_{12} + \xi_{3,3}X_{13} + \xi_{3,4}A_1 + \xi_{3,5}X_{21} + \xi_{3,6}X_{22} + \xi_{3,7}X_{23} + \xi_{3,8}A_2 + \xi_{3,9}X_{31} + \xi_{3,10}X_{32} + \xi_{3,11}X_{33} + \xi_{3,12}A_3\}.$$

Since the goal is to minimize the time-to-remission the estimated optimal rule at the third

stage is given

$$\hat{\pi}_3(\boldsymbol{h}_3) = -\text{sign}(\hat{\psi}_{30} + \hat{\psi}_{31}X_{31} + \hat{\psi}_{32}X_{32} + \hat{\psi}_{33}X_{33}).$$

After fitting both models we then form the pseudo outcomes for the third stage. Since we are trying to minimize the time-to-event, the max operator will be replaced by the min resulting in the pseudo outcome being given by

$$\tilde{V}_{3i}^R = I(\Upsilon_{3i}^\dagger = 1)f(S_{3i}^\dagger) + I(\Upsilon_{3i}^\dagger = -1)[f(S_{3i}^\dagger) + \min_{a_3} Q_3^R\{\boldsymbol{h}_3, a_3; \hat{\beta}_3, \hat{\psi}_3\}] - f(\mathscr{T}_{2,i}).$$

For the second and first stage Q-functions we will use the following models:

$$Q_2^R(\boldsymbol{h}_3, a_3; \beta_2, \psi_2) = \exp\{\beta_{20} + \beta_{21}X_{11} + \beta_{22}X_{12} + \beta_{23}X_{13} + \beta_{24}A_1 + \beta_{25}X_{21} + \beta_{26}X_{22} + \beta_{27}X_{23} + (\psi_{20} + \psi_{21}X_{11} + \psi_{22}X_{12} + \psi_{23}X_{13} + \psi_{24}A_1 + \psi_{25}X_{21} + \psi_{26}X_{22} + \psi_{27}X_{23})A_2\}$$

and

$$Q_1^R(\boldsymbol{h}_3, a_3; \beta_1, \psi_1) = \exp\{\beta_{10} + \beta_{11}X_{11} + \beta_{12}X_{12} + \beta_{13}X_{13} + (\psi_{10} + \psi_{11}X_{11} + \psi_{12}X_{12} + \psi_{13}X_{13})A_1\}.$$

We will model the cause specific hazard function for censoring for $\mathscr{T}_2 \leq u < \mathscr{T}_3$ and $u < \mathscr{T}_2$ as

$$\lambda_C(u|\boldsymbol{h}_2, a_2) = \lambda_{C20}(u)\exp\{\xi_{2,0} + \xi_{2,1}X_{11} + \xi_{2,2}X_{12} + \xi_{2,3}X_{13} + \xi_{2,4}A_1 + \xi_{2,5}X_{21} + \xi_{2,6}X_{22} + \xi_{2,7}X_{23} + \xi_{2,8}A_2\}$$

and

$$\lambda_C(u|\boldsymbol{h}_1, a_1) = \lambda_{C10}(u)\exp\{\xi_{1,0} + \xi_{1,1}X_{11} + \xi_{1,2}X_{12} + \xi_{1,3}X_{13}\xi_{1,4}A_1\}.$$

The estimated optimal decision rule at the first two stages are then given by:

$$\hat{\pi}_2(\boldsymbol{h}_2) = -\text{sign}(\hat{\psi}_{20} + \hat{\psi}_{21}X_{11} + \hat{\psi}_{22}X_{12} + \hat{\psi}_{23}X_{13} + \hat{\psi}_{24}A_1 + \hat{\psi}_{25}X_{21} + \hat{\psi}_{26}X_{22} + \hat{\psi}_{27}X_{23})$$

$$\hat{\pi}_1(\boldsymbol{h}_1) = -\text{sign}(\hat{\psi}_{10} + \hat{\psi}_{11}X_{11} + \hat{\psi}_{12}X_{12} + \hat{\psi}_{13}X_{13}).$$

The resulting estimates of the regression coefficient are in tables 3.3 - 3.5. The estimated time to remission of depression symptoms of the estimated optimal treatment regime is given by 82.05 days. For the patients in the clinical trial that were not censored, the mean number of days until remission was 136 days. This demonstrates that tailoring treatments for patients can potentially lead to a significant improvement in patient outcomes.

Table 3.3 Regression Coefficients for Third Stage Q-function for STAR*D trial

Coefficient	Variable	Estimate
β_{30}	Intercept	2.56
β_{31}	Treatment 1	0.28
β_{32}	Treatment 2	0.24
β_{33}	QIDS at beginning of stage 3	0.14
β_{34}	QIDS slope during stage 2	-0.79
β_{35}	Preference at stage 3	-0.30
ψ_{30}	Treatment 3	-0.38
ψ_{31}	Treatment 3 × Treatment 1	0.34
ψ_{32}	Treatment 3 × Treatment 2	-0.05
ψ_{33}	Treatment 3 × QIDS at beginning of stage 3	0.02
ψ_{34}	Treatment 3 × QIDS slope during stage 2	-5.17

3.5 Discussion

We proposed an adaptation of Q-learning for the estimation of an optimal dynamic treatment regime when our outcome of interest is a time-to-event. This procedure is a gener-

Table 3.4 Regression Coefficients for Second Stage Q-function for STAR*D trial

Coefficient	Variable	Estimate
β_{20}	Intercept	3.92
β_{21}	QIDS at beginning of stage 1	0.03
β_{22}	QIDS slope during level 1	-0.67
β_{23}	Preference at stage 2	0.04
β_{24}	Treatment 1	-0.09
β_{25}	QIDS at beginning of stage 2	0.03
β_{26}	QIDS slope during stage 1	-0.89
β_{27}	Preference at stage 2	0.15
ψ_{20}	Treatment 2	0.29
ψ_{21}	Treatment 2 × QIDS at beginning of stage 1	-0.00
ψ_{22}	Treatment 2 × QIDS slope during level 1	-1.36
ψ_{23}	Treatment 2 × Preference at stage 2	0.08
ψ_{24}	Treatment 2 × QIDS at beginning of stage 2	-0.00
ψ_{25}	Treatment 2 × QIDS slope during stage 1	-1.04
ψ_{26}	Treatment 2 × Preference at stage 2	0.16

Table 3.5 Regression Coefficients for First Stage Q-function for STAR*D trial

Coefficient	Variable	Estimate
β_{10}	Intercept	3.49
β_{11}	QIDS at beginning of stage 1	0.07
β_{12}	QIDS slope during level 1	0.07
β_{13}	Preference at stage 2	-0.05
ψ_{10}	Treatment 1	-0.38
ψ_{11}	Treatment 1 × QIDS at beginning of stage 1	0.02
ψ_{12}	Treatment 1 × QIDS slope during level 1	-0.59
ψ_{13}	Treatment 1 × Preference at stage 2	0.04

alization of the method proposed in Goldberg and Kosorok (2012) in which we no longer assume that the censoring is independent of patient history. Here we only assume we have non informative censoring which is a weaker assumption. This method requires correctly specifying the hazard and analysis models though to get an unbiased estimate of the optimal regime and its value.

CHAPTER

4

PARENT MESSAGING STRATEGIES FOR STUDENT ABSENTEEISM

Elementary school attendance is very important to children's development. Chronic absenteeism in young students has been found to be highly correlated with significantly lower academic outcomes (Ehrlich et al., 2014; Ginsburg et al., 2014), future academic success (Chang and Romero, 2008), and increased substance abuse (Henry and Thornberry, 2010). These negative effects have also been found to be even more significant for students from lower socioeconomic classes (Ready, 2010). There is also evidence that there are significant spillover effects and students are negatively impacted academically if their classmates are chronically absent (Gottfried, 2019).

Chronic absenteeism is influenced by many different factors. These include barriers to

attendance such as lack of transportation and illness as well as the student or parent having an aversion to school for reasons such as academic struggles or bullying (Henderson et al., 2014). Families also do not always realize the importance of attendance in school for young children (Chang and Romero, 2008). There is a belief that increased parent engagement could help reduce absenteeism (Ginsburg et al., 2014).

4.1 SMART design

This study was conducted by the Institute of Education Sciences along with the American Institutes for Research to test a parent text messaging system to keep parents informed about their child's attendance as well as the importance of attending school for young students. A Sequential Multiple Assignment Randomized Trial (SMART) was used to compare different interventions and examine whether tailoring interventions based on individual family characteristics could lead to significant reductions in chronic absenteeism.

The study was a two-stage SMART with the first stage running during the fall semester and the second stage during the spring. Families were randomized to three different interventions during the first stage. These were Basic Information Messaging - Positive (BIM-pos), Basic Information Messaging - Negative (BIM-neg), and Business As Usual (BAU). Both of the BIM-pos and BIM-neg interventions include a same-day notification to the parent when their child is absent, a same-day message when the student returns to school, and weekly texts on the importance of attendance in school. The difference is for BIM-pos the messages focus on the positive benefits of regular attendance and BIM-neg focuses on the negative consequences of missing school. BAU is the control group and these families receive no additional support.

The first stage begins in October and runs through December. At the end of the fall semester the response status of each student will be assessed. A student will be considered responsive if they were not chronically absent during the fall semester where chronic absen-

teeism is defined as missing more than 10% of the days they were enrolled. In the second stage the students that were considered responsive during the first stage will continue to receive the same intervention. The students that are non-responsive will be randomized to augment their first stage treatment with Student Staff Outreach (SSO) or Goal Commitment Messaging (GCM). Student Staff Outreach involves a staff member reaching out to the parents regularly to help identify with the parent reasons for the chronic absenteeism as well as suggesting methods to help improve attendance. Goal Commitment Messaging works by sending a text message to the parent at the beginning of the week that prompts them to reply and commit to a goal of having perfect attendance for the week. At the end of the week another message is sent that either praises the parent for achieving their goal or encourages them to improve when the goal is not met. These students will continue to receive the same messages from the first-stage intervention they were assigned to in addition to the augmentation interventions during the second stage. All of the students that were assigned to BAU during the first stage will continue with BAU during the second stage no matter what their response status is. The second stage runs from January through the end of the school year in June. The design of the SMART is shown in figure 4.1.

This design leads to five different adaptive interventions embedded within the trial given by:

AI-1: Assign BIM-pos in the fall. If the student is non-responsive in the fall, augment BIM-pos with SSO in the spring and if they are responsive continue with BIM-pos

AI-2: Assign BIM-pos in the fall. If the student is non-responsive in the fall, augment BIM-pos with GCM in the spring and if they are responsive continue with BIM-pos

AI-3: Assign BIM-neg in the fall. If the student is non-responsive in the fall, augment BIM-neg with SSO in the spring and if they are responsive continue with BIM-neg

AI-4: Assign BIM-neg in the fall. If the student is non-responsive in the fall, augment BIM-neg with GCM in the spring and if they are responsive continue with BIM-neg

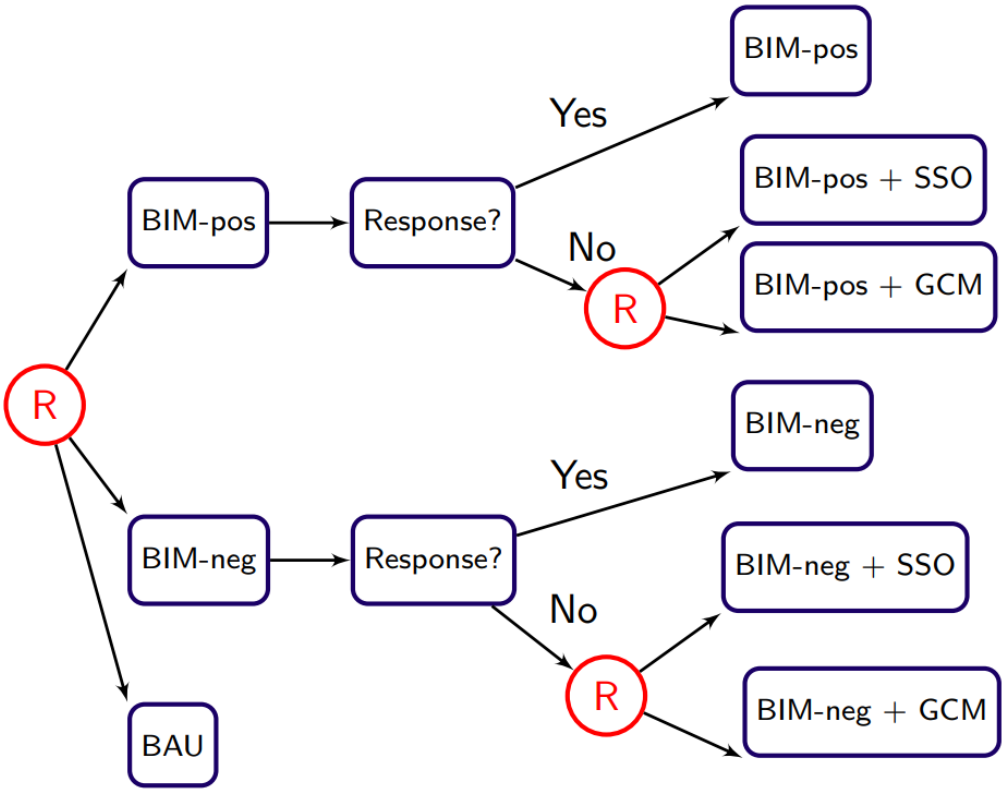

Figure 4.1 SMART diagram for the parent messaging study

BAU: Assign BAU to the family for both semesters.

There are several different outcomes of interest that will be used to assess the effectiveness of the different interventions. These outcomes fall into two different categories, attendance outcomes that assess how much the students are attending school and achievement outcomes that assess the academic performance of the students. The attendance outcomes will include an indicator for whether the student was chronically absent and the number of days the student was absent from school. The achievement outcomes are the students' scores from a standardized test in reading and math administered at the end of the school year.

The study was designed to answer four different research questions on the effectiveness of the interventions being tested in the trial. They are as follows:

- RQ1: Are there significant differences in student attendance and achievement scores

between the first stage interventions during the fall semester?

- RQ2: Is there a significant difference in student outcomes between the five different treatment strategies embedded within the trial?
- RQ3: Is there a significant difference in attendance and achievement scores between assigning SSO and GCM for the families that were non-responsive to the first stage intervention?
- RQ4: Is there a benefit to tailoring the interventions based on individual family and school characteristics?

4.2 Setup and notation

The data collected from this trial will have the following form, $\mathscr{D}_n = \{(\boldsymbol{X}_{ij}^{(1)}, A_{ij}^{(1)}, Y_{ij}^{(1)}, Z_{ij}^{(1)}, R_{ij}, A_{ij}^{(2)}, Y_{ij}^{(2)}, Z_{ij}^{(2)})\}_{i=1,j=1}^{n_j,m}$, which is given by $i.i.d.$ trajectories of the form $(\boldsymbol{X}^{(1)}, A^{(1)}, Y^{(1)}, Z^{(1)}, R, A^{(2)}, Y^{(2)}, Z^{(2)})$. There are $j = 1,\ldots,m$ different schools each of which has n_j different families participating in the trial. Let $n = \sum_{j=1}^{m} n_j$ denote the total number of students in the trial. Let $\boldsymbol{X}^{(1)} \in \mathbb{R}^{p_1}$ denote baseline family and school characteristics which in this case will contain the students race, gender, grade, school ID, disability status, English language learner status, sms availability, whether the student is at-risk for chronic absenteeism based on previous years attendance, percentage of days absent during the one month enrollment period before the study starts, and an indicator for whether they were chronically absent during the enrollment period. $A^{(1)} \in \{-1,0,1\}$ denotes the first stage treatment assigned coded such that $A^{(1)} = 0$ if they were assigned to BAU, $A^{(1)} = -1$ if assigned to BIM-neg, and $A^{(1)} = 1$ if assigned to BIM-pos. Let R denote the responder status of the student at the end of the fall coded such that $R = 1$ if the student responded and $R = 0$ if they didn't. For notational convenience let $R = 1$ for all students assigned to BAU during the first stage. $A^{(2)} \in \{-1,1\}$ denotes the second stage treatment assigned coded such that $A^{(2)} = -1$ if

the student was assigned to GCM and $A^{(2)} = 1$ if the student was assigned to SSO. Note that $A^{(2)}$ is nested within $A^{(1)}$ so $A^{(2)}$ has different meanings for different values of $A^{(1)}$. Let $Y^{(1)}$ denote the number of days the student was absent in the fall and $Y^{(2)}$ denote the number of days absent in the spring. Define $Y = Y^{(1)} + Y^{(2)}$ to be the total number of days the student is absent from school. $Z^{(1)}$ denotes an indicator for whether the student was chronically absent in the fall and $Z^{(2)}$ is a indicator for whether they were chronically absent in the spring such that being chronically absent is defined as missing more than 10% of the possible days of school. Define Z to be an indicator for whether the student was chronically absent for the entire school year.

We will make the same standard causal inference assumptions that were made in chapters 2 and 3. This includes sequential ignorability, positivity, and consistency. Positivity and sequential ignorability hold be design in a SMART. Implicit in the consistency assumption is that there is no interference between students within the trial. That is to say that the outcome of a student is not impacted by the interventions assigned to other students in the trial. This assumption is reasonable in this context though as attendance decisions are determined by the parents for elementary school age children and these decisions are not believed to be impacted by the decisions of other parents in the same school.

4.3 Analysis

4.3.1 Synthetic data

Due to the confidentiality of this study the results of the analysis presented here will use simulated data that has the same form of the actual data. Therefore the conclusions we will make do not reflect the actual results from the study. We will also omit the analysis of the achievement score outcomes, but the methods used are identical to the analysis of the continuous number of days absent outcome. To generate the simulated data we postulate parametric models for the underlying data generating process. We will estimate

the parameters of the model using the actual data and then use the estimated models to generate a new synthetic data set.

For the baseline data we note that the demographic variables such as race, gender, and grade are correlated to one another, but not correlated when conditioned on the school the student attends. Therefore the first step will be to simulate the school for each student which we will assume follows a multinomial distribution with the probability of being from a school given by the sample proportion of students that attended that school. The next step is to simulate the demographic variables which will be given by race, gender, and an indicator for whether the student is in kindergarten. Gender and the kindergarten indicator will be modeled as a binomial random variables with probabilities determined by the within school sample proportions after conditioning on the school the student attends. Race will follow a multinomial distribution again conditional on the school with the probabilities determined by the within school sample proportions.

We will then model the binary student with disability status (SWD) as

$$\text{logit}\left\{P(\text{SWD}_{ij}=1)\right\} = \beta_0 + \beta_1^T X_{ij}^D + \gamma_j$$

such that X_{ij}^D is a vector of the demographic baseline variables and γ_j is a fixed effect for school j. Identical models will be used for the students English language learner status and whether they own a SMS capable phone. We will then use the predicted probabilities from the estimated model to simulate data for each of the variables. The at risk at baseline indicator (RSK) will be modeled as

$$\text{logit}\left\{P(\text{RSK}_{ij}=1)\right\} = \beta_0 + \beta_1^T X_{ij}^D + \beta_2 \text{SWD}_{ij} + \beta_3 \text{ELL}_{ij} + \beta_4 \text{SMS}_{ij} + \gamma_j$$

where SWD is the student with disability indicator, ELL is the English language learner status, and SMS is an indicator for SMS phone capabilities. We again will simulate data using the estimated model that's been fit to the original data. The last baseline variable is

the percentage of days the student was absent during September (PCT) before the study begins at the beginning of October. This will be modeled as

$$\text{PCT}_{ij} = \beta_0 + \beta_1^T X_{ij}^D + \beta_2 \text{SWD}_{ij} + \beta_3 \text{ELL}_{ij} + \beta_4 \text{SMS}_{ij} + \beta_5 \text{RSK}_{ij} + \gamma_j + \epsilon$$

where ϵ is a random error that is assumed to be normally distributed with mean 0 and variance σ^2.

The next step is to model and simulate the number of days absent in the fall for each student. We will split the data into three separate groups based on the assigned first stage intervention and fit a linear model to the baseline covariates with a fixed effect for school. Therefore for each group we have that

$$Y_{ij}^{(1)} = \beta_0 + \xi^T X_{ij}^{(1)} + \gamma_j + \epsilon.$$

The responder status will then be determined by the percentage of days absent in the fall. The non-responsive students will be randomized between Student Staff Outreach (SSO) and Goal Commitment Messaging (GCM). The number of days absent in the spring will also be modeled using a linear model of the same form as the fall with the number of days absent in the fall added as an additional covariate.

4.3.2 Missing data

There is substantial amounts of missing data within this trial due to students dropping out of the study midway through. This was predominantly due to students switching schools during the year causing them to leave the study and therefore have missing outcome data. There is also a small amount of missing baseline data. We will assume that the missing values are missing at random.

For this analysis we will adapt the multiple imputation approach proposed in Shortreed et al. (2014) for handling missing data in SMARTs. Since we are interested in multiple

analyses with this data, by using multiple imputation we will be able to construct the imputed data sets once and use them for all analyses which will ensure consistency across the different analyses. This approach does not use a specification of the full conditional distribution as what is normally used for multiple imputation by chained equations (MICE), but instead uses a forward specification.

The time-ordered nested conditional imputation model used imputes the data sequentially by using only the data that is available at that time point for each imputation. We first impute the missing baseline data using only the baseline data. This is done using a conditional specification of each of the baseline variables. The baseline variables are given by student with disability status, baseline chronic absenteeism, English language learner status, kindergarten status, SMS availability, at risk status, and race. The binary variables are modeled using logistic regression models and the categorical variable, race, is modeled using a polytomous logistic regression model. The values are imputed from simulated draws of the posterior predictive distribution for each variable.

We then impute the number of days absent in the fall semester. This is done by first splitting the semester in half with the first half running from October 1st to November 12th. Any student who was not enrolled for the entire period will be consider missing. The model for the percentage of days absent will contain all of the baseline variables as well as interactions between the baseline variables and the first-stage intervention. The percentage of days absent during the first period is then imputed using predictive mean matching. We then impute the days absent in the second half of the fall in the same manner with the number of days absent in the fall and its interaction with the first stage intervention included in the model.

The response status for the students with missing outcome data is then ascertained based on the imputed values and they are then randomized to student staff outreach or goal commitment messaging if they were non-responsive to the first stage intervention. The number of days absent in the spring is then imputed in an identical manner to the fall

with the semester being split at April 1st.

4.3.3 Research Question 1

The first research question focuses on the main effect of the first stage interventions on attendance during the fall semester. We would like to test the three pairwise comparisons of BIM-pos, BIM-neg, and BAU.

To estimate the effect of assigning BIM-pos or BIM-neg we will fit the following model:

$$Y_{ij}^{(1)} = \zeta_0 + \zeta_1 \text{BIM}_{ij} + \xi^T \boldsymbol{X}_{ij}^{(1)} + \gamma_j + \nu_{ij}$$

where $Y_{ij}^{(1)}$ is the number of days absent in the fall for student i in school j, BIM_{ij} is an indicator for whether they were assigned to either BIM-pos or BIM-neg, $\boldsymbol{X}_{ij}^{(1)}$ is a vector of baseline covariates that are mean centered, γ_j is a fixed effect for school j, and ν_{ij} is a random error for student i in school j that is assumed to be normally distributed with mean 0 and variance σ^2.

When the outcome of interest is an indicator for whether a student is chronically absent the model will be given by:

$$P(Z_{ij}^{(1)} = 1) = \Phi(\zeta_0 + \zeta_1 \text{BIM}_{ij} + \xi^T \boldsymbol{X}_{ij}^{(1)} + \gamma_j)$$

where $\Phi(\cdot)$ is the cumulative distribution function for a standard normal distribution. To test if assigning one of either BIM-pos or BIM-neg has a significant effect for both outcomes, we will test whether ζ_1 is significantly different from zero using a Wald test.

We will also compare assigning BIM-pos, BIM-neg, and BAU. The model for the number of days absent with each of the individual effects will be given by

$$Y_{ij}^{(1)} = \zeta_0 + \zeta_1 \text{POS}_{ij} + \zeta_2 \text{NEG}_{ij} + \xi^T \boldsymbol{X}_{ij}^{(1)} + \gamma_j + \nu_{ij}$$

where $Y_{ij}^{(1)}$ is the number of days absent in the fall for student i in school j, POS_{ij} is an indicator for whether they were assigned to BIM-pos during the fall, NEG_{ij} is an indicator for whether they were assigned to BIM-neg, $\boldsymbol{X}_{ij}^{(1)}$ is a vector of mean centered baseline covariates, γ_j is a fixed effect for school j, and ν_{ij} is a random error for student i in school j that is assumed to be normally distributed with mean 0 and variance σ^2.

Similar to before when the outcome of interest is an indicator for whether a student is chronically absent the model will be given by:

$$P(Z_{ij}^{(1)} = 1) = \Phi(\zeta_0 + \zeta_1 \text{POS}_{ij} + \zeta_2 \text{NEG}_{ij} + \xi^T \boldsymbol{X}_{ij}^{(1)} + \gamma_j)$$

where $\Phi(\cdot)$ is the cumulative distribution function for a standard normal distribution.

To test if assigning BIM-pos has a significant effect for both outcomes, we will test whether ζ_1 is significantly different from zero using a Wald test. Similarly to test whether BIM-neg has a significant effect we will test if ζ_2 is significantly different then zero. To test if there is a difference in the effect of BIM-pos and BIM-neg we will test if $\zeta_1 - \zeta_2$ is significantly different than zero.

Table 4.1 displays the estimated mean number of days absent in the fall semester for those that were assigned to BAU, BIM+, BIM-, or one of either BIM+ or BIM- as well as the results of the test for the pairwise differences between the different intervention options. We use a Bonferroni correction to adjust for multiple comparisons.

Table 4.2 displays the predicted probability of a student being chronically absent in the fall semester if they were assigned to each of the different interventions in the fall. We again use a Bonferroni correction to adjust for multiple comparisons.

We can see that there is a significant reduction in the number of days absent and probability of chronic absenteeism by assigning either BIM-pos or BIM-neg. There was not a significant difference between BIM-pos and BIM-neg for outcomes in the fall semester.

Table 4.1 Estimated effect of first-stage intervention on the number of days absent in the fall semester

Comparison	BIM+ or BIM-	BIM+	BIM-	BAU	Difference	Effect Size	p-value
BIM+ or BIM- vs. BAU	3.160			3.444	0.284	0.119	<0.001
BIM+ vs. BAU		3.155		3.444	0.289	0.122	<0.001
BIM- vs. BAU			3.166	3.444	0.278	0.117	<0.001
BIM+ vs. BIM-		3.155	3.166		0.011	0.005	0.776

Table 4.2 Estimated effect of first-stage intervention on chronic absenteeism in the fall semester

Comparison	BIM+ or BIM-	BIM+	BIM-	BAU	Difference	Effect Size	p-value
BIM+ or BIM- vs. BAU	0.176			0.216	0.040	0.101	<0.001
BIM+ vs. BAU		0.173		0.216	0.043	0.108	<0.001
BIM- vs. BAU			0.178	0.216	0.038	0.095	<0.001
BIM+ vs. BIM-		0.173	0.178		0.005	0.013	0.435

4.3.4 Research Question 2

The second research question examines the difference between the four adaptive regimes embedded in the trial and BAU. We want to test if any of the adaptive regimes are more effective than BAU as well as if there is a difference between any of the adaptive regimes.

Let $\boldsymbol{\pi} = (\pi_1, \pi_2)$ denote a treatment regime such that π_t is a function that maps the available information on a student to one of the possible interventions available to that student. There are five different treatment regimes embedded in the design of the trial which we will denote by:

1. $\boldsymbol{\pi}^{\text{BIM+,SSO}}$ which assigns BIM-pos to all families at the initial stage. During the second stage, those that responded to the first stage intervention continue with BIM-pos and those that are not responsive are given SSO in addition to BIM-pos.

2. $\pi^{\text{BIM+,GCM}}$ which assigns BIM-pos to all families at the initial stage. During the second stage, those that responded to the first stage intervention continue with BIM-pos and those that are not responsive are given GCM in addition to BIM-pos.

3. $\pi^{\text{BIM−,SSO}}$ which assigns BIM-neg to all families at the initial stage. During the second stage, those that responded to the first stage intervention continue with BIM-neg and those that are not responsive are given SSO in addition to BIM-neg.

4. $\pi^{\text{BIM−,GCM}}$ which assigns BIM-neg to all families at the initial stage. During the second stage, those that responded to the first stage intervention continue with BIM-neg and those that are not responsive are given GCM in addition to BIM-neg.

5. π^{BAU} which assigns BAU at all stages.

Consider the students in the trial that received interventions that are consistent with following the treatment regime $\pi^{\text{BIM+,SSO}}$. This includes all the students who received BIM-pos in the first stage that were responsive at the end of the fall and therefore continued with BIM-pos as well as those that received BIM-pos in the fall, were non-responsive and then re-randomized to receive SSO in the spring. The probability of a non-responsive student receiving SSO is equal to 0.5. This causes those students who were considered non-responsive to be underrepresented in the data. We will correct for this by inverse weighting by the probability of receiving the intervention that they were assigned. Let W_{ij} denote the weight for student i in school j. For the students that were responsive to BIM-pos or BIM-neg and those that were assigned to BAU will have $W_{ij} = 3$ since the probability of receiving the initial treatment that they did is equal to 1/3. For the students that received BIM-pos or BIM-neg and were non-responsive will have $W_{ij} = 6$. This is because the probability of receiving their first stage intervention is 1/3 as before and the probability of receiving their second stage intervention is 1/2.

If we wanted to estimate the mean response of a single adaptive intervention we could limit the data set to only those students that were consistent with that adaptive intervention

and fit a model using weighted least squares. If a student was assigned to BIM-pos in the first stage and was responsive at the end of the fall, they are consistent with following the treatment regimes $\pi^{\text{BIM+,SSO}}$ and $\pi^{\text{BIM+,GCM}}$. We therefore need to make further adjustments to properly account for this. We will make a new data set that takes each of the individuals that were responsive to the first stage intervention and creates an identical copy of their observation. For these individuals we will let $A_{ij}^{(2)} = 1$ for one of the observations and $A_{ij}^{(2)} = -1$ for the other.

The model for the number of days absent is then given by

$$Y_{ij} = \beta_0 + \beta_1 \text{BAU}_{ij} + \beta_2 A_{ij}^{(1)} + \beta_3 A_{ij}^{(2)} + \beta_4 A_{ij}^{(1)} A_{ij}^{(2)} + \xi^T \boldsymbol{X}_{ij}^{(1)} + \gamma_j + \nu_{ij}$$

where BAU_{ij} is an indicator for whether student i in school j was assigned to BAU. The regression coefficients are estimated using weighted least squares by minimizing

$$\sum_{j=1}^{m} \sum_{i=1}^{n_j} W_{ij} (Y_{ij} - \beta_0 - \beta_1 \text{BAU}_{ij} - \beta_2 A_{ij}^{(1)} - \beta_3 A_{ij}^{(2)} - \beta_4 A_{ij}^{(1)} A_{ij}^{(2)} - \xi^T \boldsymbol{X}_{ij}^{(1)} - \gamma_j).$$

The expected number of days absent for each of the difference treatment regimes embedded in the trial can then be expressed as a linear combination of the coefficients such that

1. $\mu^{\text{BIM+,SSO}} = \beta_0 + \beta_2 + \beta_3 + \beta_4$
2. $\mu^{\text{BIM+,GCM}} = \beta_0 + \beta_2 - \beta_3 - \beta_4$
3. $\mu^{\text{BIM-,SSO}} = \beta_0 - \beta_2 + \beta_3 - \beta_4$
4. $\mu^{\text{BIM-,GCM}} = \beta_0 - \beta_2 - \beta_3 + \beta_4$
5. $\mu^{\text{BAU}} = \beta_0 + \beta_1$.

We can then test all of the pairwise differences between all of the adaptive treatment regimes. For example, to test if $\mu^{\text{BIM+,SSO}}$ is different than μ^{BAU} is then equivalent to testing

if $\beta_2+\beta_3+\beta_4-\beta_1$ is different than zero. We will use robust sandwich standard errors to properly take into account the fact that some of the observations were replicated.

When the outcome of interest is whether the student was chronically absent the model will be given by

$$P(Z_{ij}=1)=\Phi(\beta_0+\beta_1\text{BAU}_{ij}+\beta_2A_{ij}^{(1)}+\beta_3A_{ij}^{(2)}+\beta_4A_{ij}^{(1)}A_{ij}^{(2)}+\xi^T\boldsymbol{X}_{ij}^{(1)}+\gamma_j)$$

where Φ is the cumulative distribution function of a standard normal random variable.

Table 4.3 and 4.4 display the expected number of days absent for each of the different adaptive interventions embedded in the trial. Table 4.3 displays the results of testing all of the pairwise differences with BAU and table 4.4 displays the tests for the remaining differences. We use a Bonferroni correction to adjust for multiple comparisons.

Table 4.3 Pairwise differences between each adaptive intervention and BAU on the number of days absent during both semesters

Comparison	AI-1	AI-2	AI-3	AI-4	BAU	Diff.	Effect Size	p-value
AI-1 vs. BAU	10.550				11.410	0.860	0.123	<0.001
AI-2 vs. BAU		10.562			11.410	0.847	0.124	<0.001
AI-3 vs. BAU			10.468		11.410	0.941	0.136	<0.001
AI-4 vs. BAU				10.620	11.410	0.790	0.112	<0.001

Table 4.5 and 4.6 display the predicted probability of a student being chronically absent through the entire school year if assigned to each of the different adaptive interventions embedded in the trial. Table 4.5 displays the results of testing all of the pairwise differences with BAU and table 4.6 displays the tests for the remaining differences. We use a Bonferroni correction to adjust for multiple comparisons.

We can see that there is a significant decrease in the number of days absent and the probability of being chronically absent for all of the embedded interventions when compared

Table 4.4 Pairwise differences between each of the adaptive interventions on the number of days absent during both semesters

Comparison	AI-1	AI-2	AI-3	AI-4	Difference	Effect Size	p-value
AI-1 vs. AI-2	10.550	10.562			0.013	0.002	0.895
AI-1 vs. AI-3	10.550		10.468		0.081	0.011	0.507
AI-1 vs. AI-4	10.550			10.620	0.070	0.010	0.572
AI-2 vs. AI-3		10.562	10.468		0.094	0.013	0.435
AI-2 vs. AI-4		10.562		10.620	0.057	0.008	0.638
AI-3 vs. AI-4			10.468	10.620	0.151	0.026	0.126

Table 4.5 Pairwise differences between each adaptive intervention and BAU on chronic absenteeism during both semesters

Comparison	AI-1	AI-2	AI-3	AI-4	BAU	Diff.	Effect Size	p-value
AI-1 vs. BAU	0.167				0.251	0.083	0.206	<0.001
AI-2 vs. BAU		0.170			0.251	0.080	0.198	<0.001
AI-3 vs. BAU			0.167		0.251	0.083	0.206	<0.001
AI-4 vs. BAU				0.179	0.251	0.072	0.175	<0.001

to BAU. There is not a significant difference between any of the four adaptive interventions though.

4.3.5 Research Question 3

The third research question asks, for the students that were non-responsive to the first stage intervention, is it better to assign Student Staff Outreach (SSO) or Goal Commitment Messaging (GCM)? This analysis is based on the results of research question 2. If $\pi^{\text{BIM+,SSO}}$ leads to significantly less missed days than $\pi^{\text{BIM+,GCM}}$ and $\pi^{\text{BIM−,SSO}}$ is also significantly better than $\pi^{\text{BIM−,GCM}}$, we could conclude that SSO is preferable to GCM in the second stage for students that are non-responsive. Likewise if $\pi^{\text{BIM+,GCM}}$ is preferable to $\pi^{\text{BIM+,SSO}}$ and $\pi^{\text{BIM−,GCM}}$ is better than $\pi^{\text{BIM−,SSO}}$ we could conclude that GCM is the best intervention for

Table 4.6 Pairwise differences between each of the adaptive interventions on chronic absenteeism during both semesters

Comparison	AI-1	AI-2	AI-3	AI-4	Difference	Effect Size	p-value
AI-1 vs. AI-2	0.167	0.170			0.003	0.007	0.649
AI-1 vs. AI-3	0.167		0.167		0.000	0.001	0.978
AI-1 vs. AI-4	0.167			0.179	0.012	0.031	0.142
AI-2 vs. AI-3		0.170	0.167		0.003	0.008	0.694
AI-2 vs. AI-4		0.170		0.179	0.009	0.023	0.263
AI-3 vs. AI-4			0.167	0.179	0.012	0.031	0.062

students who were non-responsive after the first stage.

It may not be the case that one of SSO and GCM is always preferable no matter what first stage intervention was assigned. There is also interest in comparing if there is a difference between all regimes that assign SSO to non-responsive students and the ones that assign GCM when averaging over the possible first stage treatments. The model for the number of days absent for this analysis will be identical to the one used in research question 2 so we again have that

$$Y_{ij} = \beta_0 + \beta_1 \text{BAU}_{ij} + \beta_2 A_{ij}^{(1)} + \beta_3 A_{ij}^{(2)} + \beta_4 A_{ij}^{(1)} A_{ij}^{(2)} + \xi^T X_{ij}^{(1)} + \gamma_j + \nu_{ij}$$

and

$$P(Z_{ij} = 1) = \Phi(\beta_0 + \beta_1 \text{BAU}_{ij} + \beta_2 A_{ij}^{(1)} + \beta_3 A_{ij}^{(2)} + \beta_4 A_{ij}^{(1)} A_{ij}^{(2)} + \xi^T \boldsymbol{X}_{ij}^{(1)} + \gamma_j)$$

when we are interested in the probability of a student being chronically absent. We will also fit the model using weighted least squares and duplicating the observations for the responsive students as we did for research question 2.

We then have that the average outcome for regimes that assign School Staff Outreach to

non-responsive students is given by

$$\frac{\mu^{BIM+,SSO} + \mu^{BIM-,SSO}}{2} = \beta_0 + \beta_3$$

and the average outcome for regimes that assign Goal Commitment Messaging to non-responsive students is given by

$$\frac{\mu^{BIM+,SSO} + \mu^{BIM-,SSO}}{2} = \beta_0 - \beta_3.$$

Therefore to test if there is a difference between SSO and GCM we can test if β_3 is significantly different than zero.

Table 4.7 contains the estimated mean number of days absent for regimes that assign Student Staff Outreach or Goal Commitment Messaging to non-responders during the second stage and the results of the test for if there is a significant difference between them. Table 4.8 displays the predicted probabilities of a student being chronically absent for the two different regimes. We can see that there is not a significant difference for both outcomes of interest.

Table 4.7 Differences between second-stage interventions on the number of days absent during both semesters

Comparison	SSO	GCM	Difference	Effect Size	p-value
SSO vs. GCM	10.509	10.591	0.082	0.014	0.234

4.3.6 Research Question 4

The fourth research question examines how to tailor interventions based on individual family and school characteristics. For this analysis we will use Q-learning to estimate an

Table 4.8 Differences between second-stage interventions on chronic absenteeism during both semesters

Comparison	SSO	GCM	Difference	Effect Size	p-value
SSO vs. GCM	0.167	0.174	0.007	0.019	0.100

optimal dynamic treatment regime.

Define $\boldsymbol{H}_{ij}^{(t)}$ to be the history of the student available to the decision maker at stage t. Therefore $\boldsymbol{H}_{ij}^{(1)} = \boldsymbol{X}_{ij}^{(1)}$ and $\boldsymbol{H}_{ij}^{(2)} = (\boldsymbol{X}_{ij}^{(1)}, A_{ij}^{(1)}, Y_{ij}^{(1)}, Z_{ij}^{(1)}, R_{ij})$. As before $\boldsymbol{\pi} = (\pi_1, \pi_2)$ denotes a treatment regime such that $\pi_t : \text{dom } \boldsymbol{H}^{(t)} \rightarrow \text{dom } A^{(t)}$. Define $V(\boldsymbol{\pi}) = \mathbb{E}\, Y^*(\boldsymbol{\pi})$. Our goal is to estimate a regime $\boldsymbol{\pi}^{\text{opt}}$ such that $V(\boldsymbol{\pi}^{\text{opt}}) \leq V(\boldsymbol{\pi})$. Note that in previous chapters we wanted to find the regime, $\boldsymbol{\pi}$, that maximizes $V(\boldsymbol{\pi})$, but in this application the outcome, Y, is the number of days absent for a student which we hope to be as small as possible.

Let $\psi_t(\boldsymbol{h}^{(t)})$ be the feasible intervention options for a student with history $\boldsymbol{h}^{(t)}$. Define $Q_2(\boldsymbol{h}^{(2)}, a^{(2)}) = \mathbb{E}\{Y | \boldsymbol{H}^{(2)} = \boldsymbol{h}^{(2)}, A^{(2)} = a^{(2)}\}$ for all $(\boldsymbol{h}^{(2)}, a^{(2)}) \in \text{dom } \boldsymbol{H}^{(2)} \times \text{dom } A^{(2)}$ such that $a^{(2)} \in \psi_2(\boldsymbol{h}^{(2)})$. Define

$$Q_1\left(\boldsymbol{h}^{(1)}, a^{(1)}\right) = \mathbb{E}\left\{ \min_{a^{(2)} \in \psi_2(\boldsymbol{H}^{(2)})} Q_2\left(\boldsymbol{H}^{(2)}, a^{(2)} | \boldsymbol{H}^{(1)} = \boldsymbol{h}^{(1)}, A^{(1)} = a^{(1)}\right) \right\}$$

for all $(\boldsymbol{h}^{(1)}, a^{(1)}) \in \text{dom } \boldsymbol{H}^{(1)} \times \text{dom } A^{(1)}$ such that $a^{(1)} \in \psi_1(\boldsymbol{h}^{(1)})$. We can show via dynamic programming that $\boldsymbol{\pi}^{\text{opt}} = (\pi_1^{\text{opt}}, \pi_2^{\text{opt}})$ where $\pi_t^{\text{opt}} = \arg\min_{a^{(t)}} Q_t(\boldsymbol{h}^{(t)}, a^{(t)})$ then fulfills $V(\boldsymbol{\pi}^{\text{opt}}) \leq V(\boldsymbol{\pi})$ for all treatment regimes $\boldsymbol{\pi}$.

As discussed in previous chapters, Q-learning is a regression based approximate dynamic programming algorithm that works by constructing estimators for each of the separate Q-functions. This then leads to a plug-in estimate for the optimal treatment regime. For this study, only the non-responsive students are re-randomized in the second stage which leads to some minor adjustments to how the algorithm was presented in Chapters 2 and 3.

We will postulate linear models for the Q-functions. Therefore the model for the second stage Q-function will be given by

$$Q_2(\boldsymbol{h}^{(2)}, a^{(2)}; \beta_2) = \beta_{20} + \beta_{21} x^{(1)} + \beta_{22} a^{(1)} + \beta_{23} x^{(2)} + \beta_{24} a^{(2)} + \beta_{25} x^{(1)} a^{(2)} + \beta_{26} a^{(1)} a^{(2)} + \beta_{27} x^{(2)} a^{(2)}.$$

where $x^{(1)}$ is an indicator for whether the student was considered at risk at baseline and $x^{(2)}$ is the number of days absent in the fall. Then the estimator $\hat{\beta}_2$ is given

$$\hat{\beta}_2 = \arg\min_{\beta_2} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left\{ Y_{ij}^{(2)} - Q_2(\boldsymbol{H}_{ij}^{(2)}, A_{ij}^{(2)}; \beta_2) \right\} (1 - R_{ij}).$$

The estimated optimal rule for non-responsive students during the second stage is then given by

$$\begin{aligned} \hat{\pi}_2(\boldsymbol{h}^{(2)}) &= \arg\min_{a^{(2)}} Q_2(\boldsymbol{h}^{(2)}, a^{(2)}; \hat{\beta}_2) \\ &= -\text{sign}\left(\hat{\beta}_{24} + \hat{\beta}_{25} x^{(1)} + \hat{\beta}_{26} a^{(1)} + \hat{\beta}_{27} x^{(2)}\right). \end{aligned}$$

Define the pseudo outcome, $\tilde{Y}$, to be the predicted number of days missing for a student that follows the estimated optimal regime during the second stage. Therefore

$$\tilde{Y}_{ij} = R_{ij} Y_{ij} + (1 - R_{ij}) \left\{ Y_{ij}^{(1)} + \min_{a^{(2)}} Q_2(\boldsymbol{H}^{(2)}, a^{(2)}; \hat{\beta}_2) \right\}.$$

The model for the first stage Q-function will be given by

$$Q_1(\boldsymbol{h}^{(1)}, a^{(1)}; \beta_1) = \beta_{10} + \beta_{11} x^{(1)} + \beta_{12} a^{(1)} + \beta_{13} x^{(1)} a^{(1)}$$

and we can estimate β_1 by

$$\hat{\beta}_1 = \arg\min_{\beta_1} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \{ \tilde{Y}_{ij} - Q_1(\boldsymbol{H}_{ij}^{(1)}, A_{ij}^{(1)}; \beta_1) \}.$$

We can then estimate the optimal first stage regime by

$$\begin{aligned}\hat{\pi}_1(\boldsymbol{h}_1) &= \arg\min_{a^{(1)}} Q_1(\boldsymbol{h}^{(1)}, a^{(1)}; \hat{\beta}_1) \\ &= -\text{sign}(\hat{\beta}_{12} + \hat{\beta}_{13} x^{(1)}).\end{aligned}$$

Table 4.9 displays the estimated coefficients and 95% confidence intervals for the second stage Q-function. Table 4.10 shows the estimated coefficients and 95% confidence intervals for the first stage Q-function. For the first stage coefficients the confidence intervals are calculated using the m-out-of-n bootstrap proposed in Chakraborty et al. (2013) to obtain valid intervals despite the non-regularity. Note that the confidence intervals for all of the treatment interaction terms contain zero. Therefore there is not significant evidence that tailoring interventions based on these characteristics would provide a decrease in the number of days absent.

Table 4.11 displays inverse probability weighted estimates of the expected number of days absent for the estimated optimal regime as well as the adaptive interventions embedded within the regime. 95% confidence intervals are also provided that were calculated using the non-parametric bootstrap. Note that the bootstrap may perform poorly in this context as the IPW is not a smooth estimator. These results also indicate that there is not enough evidence to conclude that tailoring interventions would lead to a decrease in absenteeism.

Table 4.9 Parameter estimates for the second stage Q-function

Variable	Coefficient Estimate	95% Lower CI	95 % Upper CI
Intercept	2.888	1.648	4.128
Treatment 1	-0.202	-0.456	0.053
At Risk	3.953	3.430	4.476
Fall Absenteeism	1.064	0.889	1.238
Treatment 2	-0.272	-1.512	0.968
Trt 2 * Trt 1	0.020	-0.235	0.274
Trt 2 * At Risk	-0.282	-0.805	0.241
Trt 2 * Fall Absenteeism	0.036	-0.138	0.210

Table 4.10 Parameter estimates for the first stage Q-function

Variable	Coefficient Estimate	95% Lower CI	95 % Upper CI
Intercept	8.586	8.161	8.490
At Risk	7.267	6.635	7.357
Treatment 1	0.058	-0.118	0.249
Trt 1 * At Risk	-0.133	-0.615	0.179

Table 4.11 IPW Estimates of expected outcomes of different regimes

Regime	Mean Days Absent	95% Lower CI	95% Upper CI
$\hat{\pi}$	10.161	9.850	10.496
$\pi^{BIM+,SSO}$	10.561	10.229	10.862
$\pi^{BIM+,GCM}$	10.504	10.182	10.846
$\pi^{BIM-,SSO}$	10.418	10.107	10.737
$\pi^{BIM-,GCM}$	10.663	10.355	11.021

BIBLIOGRAPHY

Almirall, D., S. N. Compton, M. Gunlicks-Stoessel, N. Duan, and S. A. Murphy (2012). Designing a pilot sequential multiple assignment randomized trial for developing an adaptive treatment strategy. *Statistics in medicine 31*(17), 1887–1902.

Bai, X., A. A. Tsiatis, W. Lu, and R. Song (2017). Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime Data Analysis 23*(4), 585–604.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

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

Cain, L. E., J. M. Robins, E. Lanoy, R. Logan, D. Costagliola, and M. A. Hernán (2010). When to start treatment? a systematic approach to the comparison of dynamic regimes using observational data. *The international journal of biostatistics 6*(2).

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.

Chakraborty, B., E. B. Laber, and Y.-Q. Zhao (2014). Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials 11*(4), 408–417.

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

Chakraborty, B., S. Murphy, and V. Strecher (2009). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research 19*(3).

Chakraborty, B., V. Strecher, and S. Murphy (2010). Inference for nonregular parameters

in optimal dynamic treatment regimes. *Statistical Methods in Medical Research 19*(3), 1–27.

Chang, H. N. and M. Romero (2008). *Present, Engaged, and Accounted For: The critical Importance of Addressing Chronic Absence in the Early Grades*. New York, NY: National Center for Children in Poverty.

Dawid, A. (1994). Selection paradoxes of bayesian inference. *Lecture Notes-Monograph Series*, 211–220.

Ehrlich, S. B., J. A. Gwynne, A. S. Pareja, E. M. Allensworth, P. Moore, S. Jagesic, and E. Sorice (2014). *Preschool Attendance in Chicago Public Schools: Relationship with Learning Outcomes and Reasons for Absences*. Chicago: University of Chicago Consortium on Chicago School Research.

Ertefaie, A. (2014). Constructing dynamic treatment regimes in infinite-horizon settings. *arXiv preprint arXiv:1406.0764*, 1–26.

Ginsburg, A., P. Jordan, and H. Chang (2014). *Absences Add Up: How School Attendance Influences Student Success*. San Francisco: Attendance Works.

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

Gottfried, M. A. (2019). Chronic absenteeism in the classroom context: Effects on achievement. *Urban Education 54*(1), 3–34.

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

Henderson, H. V. and S. Searle (1979). Vec and vech operators for matrices, with some uses in jacobians and multivariate statistics. *Canadian Journal of Statistics 7*(1), 65–81.

Henderson, T., C. Hill, and K. Norton (2014). *The Connection Between Missing School and Health: A Review of Chronic Absenteeism and Student Health in Oregon*. Upstream Public Health.

Henry, K. L. and T. P. Thornberry (2010). Truancy and escalation of substance use during adolescence. *Journal of Studies on Alcohol and Drugs 71*(1), 115–124.

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

Jiang, R., W. Lu, R. Song, and M. Davidian (2017). On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society, Series B 79*, 1165–1185.

Kidwell, K. M. (2014). Smart designs in cancer research: Past, present, and future. *Clinical Trials 11*(4), 445–456.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer.

Kosorok, M. R. and E. E. Moodie (2015). *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. SIAM.

Laber, E., K. Linn, and L. Stefanski (2014). Interactive model building for q-learning. *Biometrika 101*(4), 831–847.

Laber, E. and Y. Zhao (2015). Tree-based methods for estimating 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.

Laber, E. B. and A.-M. Staicu (2017). Functional feature construction for individualized treatment regimes. *Journal of the American Statistical Association* (just-accepted).

Laber, E. B., F. Wu, C. Munera, I. Lipkovich, S. Colucci, and S. Ripa (2018). Identifying optimal dosage regimes under safety constraints: An application to long term opioid treatment of chronic pain. *Statistics in Medicine 37*(9), 1407–1418.

Laber, E. B., Y.-Q. Zhao, T. Regh, M. Davidian, A. Tsiatis, J. B. Stanford, D. Zeng, R. Song, and M. R. Kosorok (2015). Using pilot data to size a two-arm randomized trial to find a nearly optimal personalized treatment strategy. *Statistics in medicine*.

Lavori, P. and R. Dawson (2000). A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 163*(1), 29–38.

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

Lei, H., I. Nahum-Shani, K. Lynch, D. Oslin, and S. Murphy (2012). A "smart" design for building individualized treatment sequences. *Annual Review of Clinical Psychology 8*, 21–48.

Linn, K. A., E. B. Laber, and L. A. Stefanski (2016). Interactive q-learning for quantiles. *Journal of the American Statistical Association* (just-accepted), 1–37.

Luckett, D. J., E. B. Laber, A. R. Kahkoska, D. M. Maahs, E. Mayer-Davis, and M. R. Kosorok (2016). Estimating dynamic treatment regimes in mobile health using v-learning. *arXiv preprint arXiv:1611.03531*.

Luedtke, A. R. and M. J. Van Der Laan (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics 44*(2), 713–742.

Moodie, E., T. Richardson, and D. Stephens (2007). Demystifying optimal dynamic treatment regimes. *Biometrics 63*(2), 447–455.

Moodie, E., T. Richardson, and D. Stephens (2010). Estimating optimal dynamic regimes: Correcting bias under the null. *Biometrics 63*(2), 447–455.

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

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

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

Murphy, S. A. (2003b). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B 65*(2), 331–366.

Murphy, S. A. (2005b, Jul). A generalization error for Q-learning. *Journal of Machine Learning Research 6*, 1073–1097.

Murphy, S. A., K. G. Lynch, D. Oslin, J. R. McKay, and T. TenHave (2007). Developing adaptive treatment strategies in substance abuse research. *Drug and alcohol dependence 88*, S24–S30.

Murray, T., Y. Yuan, and P. Thall (2017). A bayesian machine learning approach for optimizing dynamic treatment regimes. *Journal of the American Statistical Association Early view*, 1–37.

Nahum-Shani, I., A. Ertefaie, X. L. Lu, K. G. Lynch, J. R. McKay, D. W. Oslin, and D. Almirall (2017). A smart data analysis method for constructing adaptive treatment strategies for substance use disorders. *Addiction 112*(5), 901–909.

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

Orellana, L., A. Rotnitzky, and J. Robins (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content. *Int. Jrn. of Biostatistics 6*(2), 1–49.

Pineau, J., M. G. Bellemare, A. J. Rush, A. Ghizaru, and S. A. Murphy (2007). Constructing evidence-based treatment strategies using methods from computer science. *Drug Alcohol Dependence 88*, S52–S60.

Praestgaard, J. and J. A. Wellner (1993). Exchangeably weighted bootstraps of the general empirical process. *The Annals of Probability 21*(4), 2053–2086.

PSU Methodology Center, T. (2017a, July). Nih program announcements.

PSU Methodology Center, T. (2017b, July). Smart studies.

Qian, M. and S. Murphy (2011). Performance Guarantees for Individualized Treatment Rules. *The Annals of Statistics 39*(2), 1180–1210.

Qian, M., I. Nahum-Shani, and S. A. Murphy (2013). Dynamic treatment regimes. In *Modern Clinical Trial Analysis*, pp. 127–148. Springer.

Ready, D. D. (2010). Socioecomic disadvantage, school attendance, and early cognitive development: The differential effects of school exposure. *Sociology of Education 83*(4), 271–286.

Robins, J. (2004a). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*.

Robins, J., L. Orellana, and A. Rotnitzky (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine 27*(23), 4678–4721.

Robins, J. M. (2004b). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatitics*, pp. 189–326. Springer.

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics 6*(1), 34–58.

Rush, A., M. Trivedi, and M. Fava (2003). Depression, IV: STAR* D treatment trial for depression. *American Journal of Psychiatry 160*(2), 237.

Rush, A., M. Trivedi, H. Ibrahim, T. Carmody, B. Arnow, D. Klein, J. Markowitz, P. Ninan, S. Kornstein, R. Manber, M. Thase, J. Kocsis, and M. Keller (2003). The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry 54*(5), 573–583.

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

Shortreed, S., E. B. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy (2010). Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine Learning 84*(1–2), 109–136.

Shortreed, S. M., E. Laber, T. S. Stroup, J. Pineau, and S. A. Murphy (2014). A multiple imputation strategy for sequential multiple assignment randomized trials. *Statistics in Medicine 33*(24), 4204–4214.

Song, R., W. Wang, D. Zeng, and M. Kosorok (2015a). Penalized q-learning for dynamic treatment regimes. *Statistica Sinica 25*(3), 901–920.

Song, R., W. Wang, D. Zeng, and M. R. Kosorok (2015b). Penalized q-learning for dynamic treatment regimens. *Statistica Sinica 25*(3), 901.

Splawa-Neyman, J., D. Dabrowska, T. Speed, et al. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472.

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

Thall, P. F., R. E. Millikan, H.-G. Sung, et al. (2000). Evaluating multiple treatment courses in clinical trials. *Statistics in medicine 19*(8), 1011–1028.

van der Laan, M. J., M. L. Petersen, and M. M. Joffe (2005). History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *The International Journal of Biostatistics 1*(1).

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

Watkins, C. (1989). *Learning from Delayed Rewards*. Ph. D. thesis, Cambridge University.

Wu, F., E. B. Laber, I. A. Lipkovich, and E. Severus (2015). Who will benefit from antidepressants in the acute treatment of bipolar depression? a reanalysis of the step-bd study by sachs et al. 2007, using q-learning. *International journal of bipolar disorders 3*(1), 7.

Xu, Y., P. Müller, A. S. Wahed, and P. F. Thall (2016). Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *Journal of the American Statistical Association 111*(515), 921–950.

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

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

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

Zhang, Y., E. Laber, A. Tsiatis, and M. Davidian (2017). List-based treatment regimes. *Journal of the American Statistical Association Early view*, 1–25.

Zhang, Y., E. B. Laber, A. Tsiatis, and M. Davidian (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics 71*(4), 895–904.

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

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

Zhao, Y., D. Zeng, M. A. Socinski, and M. R. Kosorok (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics 67*(4), 1422–1433.

Zhou, X., N. Mayer-Hamblett, U. Khan, and M. R. Kosorok (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association 112*(517), 169–187.

APPENDICES

APPENDIX

A

SAMPLE SIZE CALCULATIONS FOR SMARTS

A.1 Proofs of technical results

Proof of Lemma 3.1. It can be shown by direct calculation that if $W \sim \text{Normal}(\mu, \omega^2)$ then

$$\begin{aligned} \mathbb{E}|W| &= \int \frac{|\mu + \omega w|}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) d\, w \\ &= \mu\left\{1 - \Phi\left(-\frac{\mu}{\omega}\right)\right\} + \frac{\sqrt{2}\omega}{\sqrt{\pi}} \exp\left(-\frac{\mu^2}{2\omega^2}\right). \end{aligned}$$

Write

$$\begin{aligned} Q_1(h_1,a_1) &= \mathbb{E}\left\{\max_{a_2} Q_2(\boldsymbol{H}_2,a_2)\Big|\boldsymbol{H}_1=\boldsymbol{h}_1,A_1=a_1\right\} \\ &= \mathbb{E}\left(\boldsymbol{H}_{2,1}^{\intercal}\beta_{2,0}^{*}+|\boldsymbol{H}_{2,1}^{\intercal}\beta_{2,1}^{*}|\Big|\boldsymbol{H}_1=\boldsymbol{h}_1,A_1=a_1\right) \\ &= \boldsymbol{h}_{1,0}^{\intercal}\xi_{1,0}^{*}+a_1\boldsymbol{h}_{1,1}^{\intercal}\xi_{1,1}^{*} \\ &\quad + \mathbb{E}\left(\left|\boldsymbol{H}_{1,2}^{\intercal}\varpi_{1,2}^{*}+A_1\boldsymbol{H}_{1,3}^{\intercal}\varpi_{1,3}^{*}+\tau^{*}Z\right|\Big|\boldsymbol{H}_1=\boldsymbol{h}_1,A_1=a_1\right) \\ &= \boldsymbol{h}_{1,0}^{\intercal}\xi_{1,0}^{*}+a_1\boldsymbol{h}_{1,1}^{\intercal}\xi_{1,1}^{*} \\ &\quad + \int\frac{|\boldsymbol{h}_{1,2}^{\intercal}\varpi_{1,2}^{*}+a_1\boldsymbol{h}_{1,3}^{\intercal}\varpi_{1,3}^{*}+\tau^{*}z|}{\sqrt{2\pi}}exp\left(-\frac{z^2}{2}\right)dz, \end{aligned}$$

from which the result follows. ■

Proof of Lemma 3.3. The function ν is continuously differentiable in a neighborhood of $\{\tau^{*},\omega^{*\intercal},\text{vech}(\Omega^{*})^{\intercal}\}^{\intercal}$. Thus, the result follows immediately from a Taylor series expansion, (AN6), and Slutsky's theorem; that is,

$$\begin{aligned} \sqrt{n}\left\{\widehat{V}_n-V(\boldsymbol{\pi}^{\text{opt}})\right\} &= \sqrt{n}\left[\nu\left\{\widehat{\tau}_n,\widehat{\omega}_n,\text{vech}\left(\widehat{\Omega}_n\right)\right\}-\nu\{\tau^{*},\omega^{*},\text{vech}(\Omega^{*})\}\right] \\ &= \nabla\nu\left\{\widetilde{\tau}_n,\widetilde{\omega}_n,\text{vech}\left(\widetilde{\Omega}_n\right)\right\}^{\intercal}\sqrt{n}\begin{Bmatrix}\widehat{\tau}_n-\tau^{*}\\ \widehat{\omega}_n-\omega^{*}\\ \text{vech}\left(\widehat{\Omega}_n\right)-\text{vech}(\Omega^{*})\end{Bmatrix} \\ &\rightsquigarrow \text{Normal}(0,\Sigma^{*}). \end{aligned}$$

■

Proof of Lemma 3.4. We show that $\sqrt{n}|V(\widehat{\pi}_n)-V(\pi^{Q,\text{opt}})|=O(n^{-\delta})$. For any regime $\pi=$

(π_1, π_2) write

$$
\begin{aligned}
V(\pi) &= \mathbb{E}\left[\mathbb{E}\left\{\mathbb{E}\left(Y\middle|\boldsymbol{H}_2, A_2\right)\Big|_{A_2=\pi_2(\boldsymbol{H}_2)}\middle|\boldsymbol{H}_1, A_1\right\}\Big|_{A_1=\pi_1(\boldsymbol{H}_1)}\right] \\
&= \mathbb{E}\left(\mathbb{E}\left[Q_2\left\{\boldsymbol{H}_2, \pi_2(\boldsymbol{H}_2)\right\}\middle|\boldsymbol{H}_1, A_1\right]\Big|_{A_1=\pi_1(\boldsymbol{H}_1)}\right) \\
&= \mathbb{E}\left(\mathbb{E}\left\{\max_{a_2} Q_2(\boldsymbol{H}_2, a_2)\middle|\boldsymbol{H}_1, A_1\right\}\Big|_{A_1=\pi_1(\boldsymbol{H}_1)}\right) \\
&\quad -\mathbb{E}\left\{\mathbb{E}\left(|\Delta Q_2| 1_{\pi_2(\boldsymbol{H}_2)\neq\pi_2^{Q,\text{opt}}(\boldsymbol{H}_2)}\middle|\boldsymbol{H}_1, A_1\right)\Big|_{A_1=\pi_1(\boldsymbol{H}_1)}\right\} \\
&= \mathbb{E}[Q_1\left\{\boldsymbol{H}_1, \pi_1(\boldsymbol{H}_1)\right\}] \\
&\quad -\mathbb{E}\left\{\mathbb{E}\left(|\Delta Q_2| 1_{\pi_2(\boldsymbol{H}_2)\neq\pi_2^{Q,\text{opt}}(\boldsymbol{H}_2)}\middle|\boldsymbol{H}_1, A_1\right)\Big|_{A_1=\pi_1(\boldsymbol{H}_1)}\right\} \\
&= \mathbb{E}\left\{\max_{a_1} Q_1(\boldsymbol{H}_1, a_1)\right\} - \mathbb{E}\left(|\Delta Q_1| 1_{\pi_1(\boldsymbol{H}_1)\neq\pi_1^{Q,\text{opt}}(\boldsymbol{H}_1)}\right) \\
&\quad -\mathbb{E}\left\{\mathbb{E}\left(|\Delta Q_2| 1_{\pi_2(\boldsymbol{H}_2)\neq\pi_2^{Q,\text{opt}}(\boldsymbol{H}_2)}\middle|\boldsymbol{H}_1, A_1\right)\Big|_{A_1=\pi_1(\boldsymbol{H}_1)}\right\} \\
&= V(\boldsymbol{\pi}^{Q,\text{opt}}) - R_1(\pi_1) - R_2(\pi_1, \pi_2).
\end{aligned}
$$

Thus, $|V(\widehat{\pi}_n) - V(\boldsymbol{\pi}^{Q,\text{opt}})| \leq R_1(\widehat{\pi}_{1,n}) + R_2(\widehat{\pi}_{1,n}, \widehat{\pi}_{2,n})$. We next derive rates of convergence for these remainder terms.

We will make use of the fact that $R_2(\widehat{\pi}_{1,n}, \widehat{\pi}_{2,n}) \leq 2\mathbb{E}|\Delta Q_2| 1_{\Delta\widehat{Q}_{2,n}\Delta Q_2<0}$ which follows from writing

$$
\begin{aligned}
R_2(\widehat{\pi}_{1,n}, \widehat{\pi}_{2,n}) &= \mathbb{E}\left\{\mathbb{E}\left(|\Delta Q_2| 1_{\Delta\widehat{Q}_{2,n}\Delta Q_2<0}\middle|\boldsymbol{H}_1, A_1\right)\Big|_{A_1=\widehat{\pi}_{1,n}(\boldsymbol{H}_1)}\right\} \\
&\leq \mathbb{E}\left\{\mathbb{E}\left(|\Delta Q_2| 1_{\Delta\widehat{Q}_{2,n}\Delta Q_2<0}\middle|\boldsymbol{H}_1, A_1 = 1\right)\right\} \\
&\quad +\mathbb{E}\left\{\mathbb{E}\left(|\Delta Q_2| 1_{\Delta\widehat{Q}_{2,n}\Delta Q_2<0}\middle|\boldsymbol{H}_1, A_1 = -1\right)\right\} \\
&= 2\mathbb{E}|\Delta Q_2| 1_{\Delta\widehat{Q}_{2,n}\Delta Q_2<0},
\end{aligned}
$$

where we have used $P(A_1 = 1|\boldsymbol{H}_1) = P(A_1 = -1|\boldsymbol{H}_1) = 1/2$ with probability one. Let ϵ_n be any nonnegative sequence diverging to ∞ with n, then

$$\begin{aligned}
2\mathbb{E}|\Delta Q_2|1_{\Delta\widehat{Q}_{2,n}\Delta Q_2<0} &= 2\mathbb{E}|\Delta Q_2|1_{\Delta\widehat{Q}_{2,n}\Delta Q_2<0}1_{|\Delta Q_2|\leq\epsilon_n/\sqrt{n}} \\
&\quad +2\mathbb{E}|\Delta Q_2|1_{\Delta\widehat{Q}_{2,n}\Delta Q_2<0}1_{|\Delta Q_2|>\epsilon_n/\sqrt{n}} \\
&\leq 2\big(\epsilon_n/\sqrt{n}\big)P\big(|\Delta Q_2|\leq\epsilon_n/\sqrt{n}\big) \\
&\quad +2\mathbb{E}|\Delta Q_2|1_{\Delta\widehat{Q}_{2,n}\Delta Q_2<0}1_{|\Delta Q_2|>\epsilon_n/\sqrt{n}} \\
&\leq 2M\epsilon_n^{1+\kappa}/n^{1/2+\kappa/2} \\
&\quad +2\mathbb{E}|\Delta Q_2|1_{\Delta\widehat{Q}_{2,n}\Delta Q_2<0}1_{|\Delta Q_2|>\epsilon_n/\sqrt{n}} \\
&= 2M\epsilon_n^{1+\kappa}/n^{1/2+\kappa/2} \\
&\quad +2\mathbb{E}|\Delta Q_2|1_{(\Delta\widehat{Q}_{2,n}-\Delta Q_2)\mathrm{sgn}(\Delta Q_2)+|\Delta Q_2|<0}1_{|\Delta Q_2|>\epsilon_n/\sqrt{n}} \\
&\leq 2M\epsilon_n^{1+\kappa}/n^{1/2+\kappa/2}+2\sqrt{\mathbb{E}|\Delta Q_2|^2}P\left\{\sqrt{n}\left|\Delta\widehat{Q}_{2,n}-\Delta Q_2\right|>\epsilon_n\right\} \\
&\leq 2M\epsilon_n^{1+\kappa}/n^{1/2+\kappa/2}+2\sqrt{\mathbb{E}|\Delta Q_2|^2}\exp\big(-c_{n,2}\epsilon_n^{\ell_{n,2}}\big),
\end{aligned}$$

where the second inequality follows from (AN8), the third inequality follows from Cauchy-Schwartz, and the fourth from (AN7). Choose $\epsilon_n = \left\{c_{n,2}^{-1}\log\big(n^{1/2+\kappa/2}\big)\right\}^{1/\ell_{n,2}}$ to obtain that $\sqrt{n}R_2(\widehat{\pi}_{1,n},\widehat{\pi}_{n,2}) = O_P(n^{-\kappa/2})$. An identical argument can be used to derive an analogous bound on $\sqrt{n}R_1(\widehat{\pi}_{1,n})$. ■

Proof of Theorem 3.7. Define $\mathscr{F}_1 = \big\{\delta(\mu_1,\mu_2) : (\mu_1,\mu_2)\in\Theta\big\}$ and $\mathscr{F}_2 = \{\delta^2(\mu_1,\mu_2) : (\mu_1,\mu_2)\in\Theta\}$; then both are Donsker by (PR4). By Corollary 2.2 of Praestgaard and Wellner (1993) it follows that $\sqrt{n}(\mathbb{P}^{(b)}_{n_0,n}-\mathbb{P}_{n_0}) \overset{P}{\underset{B}{\rightsquigarrow}} \mathbb{G}$ in $\ell^\infty(\mathscr{F}_1)$ and $\ell^\infty(\mathscr{F}_2)$, where $\mathbb{G}$ is a Brownian bridge with mean zero and covariance $\mathrm{Cov}\big\{\mathbb{G}(f_1),\mathbb{G}(f_2)\big\} = P(f_1-Pf_1)(f_2-Pf_2)$. Similarly, it follows that $\mathbb{P}_{n_0,n} \overset{P}{\underset{B}{\rightsquigarrow}} P$ in $\ell^\infty(\mathscr{F}_1)$ and $\ell^\infty(\mathscr{F}_2)$. Applying Slutsky's theorem for empirical processes (e.g.,

Theorem 7.15 in Kosorok, 2008) it follows that

$$\left\{ \begin{array}{c} \sqrt{n}\left(\mathbb{P}_{n_0,n} - \mathbb{P}_{n_0}\right) \\ \mathbb{P}_{n_0,n} \\ \mathbb{P}_{n_0,n} \end{array} \right\} \overset{P}{\rightsquigarrow}_B \left(\begin{array}{c} \mathbb{G} \\ P_1 \\ P_2 \end{array} \right)$$

in $\ell^\infty(\mathscr{F}_1 \bigcup \mathscr{F}_2)$, where $P_1 = \left\{Pf : f \in \mathscr{F}_1\right\}$ and $P_2 = \left\{Pf : f \in \mathscr{F}_2\right\}$. Define $\mathscr{F}_3 = \{f^2 - (Pf)^2 : f \in \mathscr{F}_1\}$. Application of the bootstrap continuous mapping theorem ensures that

$$\left\{ \begin{array}{c} \sqrt{n}\left(\mathbb{P}_{n_0,n} - \mathbb{P}_{n_0}\right) \\ \mathbb{P}_{n_0,n} \end{array} \right\} \overset{P}{\rightsquigarrow}_B \left(\begin{array}{c} \mathbb{G} \\ P_3 \end{array} \right)$$

in $\ell^\infty(\mathscr{F}_1 \bigcup \mathscr{F}_3)$, where $P_3 = \left\{Pf : f \in \mathscr{F}_3\right\}$. The desired result follows from the bootstrap continuous mapping theorem. ■

Proof of Theorem 3.8. Let Δ denote symmetric set difference, then we have $P_B \widehat{\Theta}^{(b)}_{n_0,n} \Delta \widehat{\theta}_n \to 0$ with probability one. Write $Z^{(b)}_{n_0,n} = O_{P_B}\left\{f(n_0,n)\right\}$ to mean that $\lim_{M\to\infty} P_B(|Z^{(b)}_{n_0,n}/f(n_0,n)| \leq M)$ converges to one in probability. It can be seen that

$$\begin{aligned} \sqrt{n}\left\{\widehat{V}^{Q(b)}_{n_0,n}(\mu_1,\mu_2) - B_0\right\} &= \sqrt{n}\left\{\widehat{V}^{Q(b)}_{n_0,n}(\mu_1,\mu_2) - \widehat{V}^{Q}_{n_0}(\mu_1,\mu_2)\right\} + \sqrt{n}\left\{\widehat{V}_{n_0}(\mu_1,\mu_2) - V^Q(\mu_1,\mu_2)\right\} \\ &\quad + \sqrt{n}\left\{V^Q(\mu_1,\mu_2) - B_0\right\} \\ &= O_{P_B}(1) + O_{P_B}(\sqrt{n}/\sqrt{n_0}) + \sqrt{n}\left\{V^Q(\mu_1,\mu_2) - B_0\right\}. \end{aligned}$$

Thus, provided that $V^Q(\mu_1^*,\mu_2^*) > B_0 + \eta$, it follows that

$$\begin{aligned} &\inf_{(\mu_1,\mu_2)\in\Xi^{(b)}_{n_0,n,1-\vartheta_1}} \frac{\min\left[\sqrt{n}\left\{\widehat{V}^{Q,(b)}_{n_0,n}(\mu_1,\mu_2) - B_0\right\}, \sqrt{n}\eta\right]}{\widehat{\varsigma}^{(b)}_{n_0,n}(\mu_2,\mu_2)} \\ &\quad = \inf_{(\mu_1,\mu_2)\in\Xi^{(b)}_{n_0,n,1-\vartheta_1}} \frac{\min\left[\sqrt{n}\left\{V^Q(\mu_1,\mu_2) - B_0\right\}, \sqrt{n}\eta\right]}{\widehat{\varsigma}^{(b)}_{n_0,n}(\mu_2,\mu_2)} + O_{P_B}\left\{\max(1, \sqrt{n}/\sqrt{n_0})\right\}, \end{aligned}$$

where, by appeal to (PR5), the leading term on the right hand side of the above equality

diverges to $+\infty$ at rate $\sqrt{n}$ and thereby dominates the second term. Similarly, it can be seen that

$$\inf_{(\mu_1,\mu_2)\in\Xi_{n,1-\vartheta_1}} \frac{\min\left[\sqrt{n}\left\{\widehat{V}_n^Q(\mu_1,\mu_2)-B_0\right\},\sqrt{n}\eta\right]}{\widehat{\varsigma}_n(\mu_2,\mu_2)} = \inf_{(\mu_1,\mu_2)\in\Xi_{n,1-\vartheta_1}} \frac{\min\left[\sqrt{n}\left\{V^Q(\mu_1,\mu_2)-B_0\right\},\sqrt{n}\eta\right]}{\widehat{\varsigma}_n(\mu_2,\mu_2)} + O_P(1),$$

where the leading term diverges to $+\infty$ at rate $\sqrt{n}$ and thereby dominates the second term. The desired result follows from Slutsky's theorem. ■

A.2 Simulation results

Table A.1 Estimated power (POW) and concentration (OPT) under a correctly specified generative model using the normality-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\hat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	Method for σ^*	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}$	Med($\hat{n}$)	SD$\hat{n}$
0	known	50	99.8	100	74	277	–	–
0	pilot study	50	97.6	100	74	204.51	199.5	60.60
0	surrogate	50	100	100	74	394.63	390	90.04
0.5	known	50	100	100	111	263	–	–
0.5	pilot study	50	100	100	111	269.91	264	62.58
0.5	surrogate	50	100	100	111	353.35	343.5	98.09
1	known	50	100	100	151	285	–	–
1	pilot study	50	100	100	151	335.69	332.5	69.93
1	surrogate	50	100	100	151	431.93	420	130.70
2	known	50	100	100	251	352	–	–
2	pilot study	50	100	100	251	495.52	491	92.61
2	surrogate	50	100	100	251	605.82	578.5	176.10

Table A.2 Estimated power (POW) and concentration (OPT) under a model which violated the normality assumptions using the normality-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\hat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	Method for σ^*	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}$	Med($\hat{n}$)	SD$\hat{n}$
0	known	50	100	100	70	586	–	–
0	pilot study	50	79.0	86.4	70	153.17	115.5	145.73
0	surrogate	50	99.4	100	70	339.32	317.5	134.60
0.5	known	50	100	100	35	485	–	–
0.5	pilot study	50	90.2	85.0	35	132.75	90.5	145.46
0.5	surrogate	50	99.8	100	35	322.16	285.5	190.63
1	known	50	100	100	47	630	–	–
1	pilot study	50	95.6	88	47	186.41	142.5	159.20
1	surrogate	50	100	99.0	47	390.41	344.5	191.12
2	known	50	100	100	103	866	–	–
2	pilot study	50	100	99.6	103	296.34	241	211.56
2	surrogate	50	100	100	103	537.83	490.5	277.45

Table A.3 Estimated power (POW) and concentration (OPT) under a model for which the normality assumptions hold using the projection-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\hat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}(\mathscr{D}_{n_0})$	Med$\{\hat{n}(\mathscr{D}_{n_0})\}$	SD$\hat{n}(\mathscr{D}_{n_0})$	$P\{\hat{n}(\mathscr{D}_0)=\infty\}$
0	50	84.58	100	74	387.72	346	107.78	0.48
0.5	50	97.54	100	111	304.47	285	82.12	0.24
1	50	100	100	151	314.61	299.5	71.15	0.08
2	50	100	100	251	366.88	364.5	64.13	0.03

Table A.4 Estimated power (POW) and concentration (OPT) under a model which violated the normality assumptions using the projection-based sample size procedure at a nominal level of 90. To form a baseline for comparison, $\hat{n}^{\text{fixed}}$, shows the required sample size to compare the optimal embedded regime with standard of care.

Δ	n_0	(POW)	(OPT)	$\hat{n}^{\text{fixed}}$	$\mathbb{E}\hat{n}(\mathscr{D}_{n_0})$	Med$\{\hat{n}(\mathscr{D}_{n_0})\}$	SD$\hat{n}(\mathscr{D}_{n_0})$	$P\{\hat{n}(\mathscr{D}_0)=\infty\}$
0	50	84.12	100	70	569.31	491.5	270.15	0.41
0.5	50	93.65	100	35	531.89	503	206.85	0.32
1	50	100	100	47	586.21	532.5	254.04	0.21
2	50	100	100	103	581.8	524	251.39	0.03

APPENDIX

B

Q-LEARNING FOR SURVIVAL ANALYSIS

B.1 Proofs of techinical results

Proof that

$$\sum_{\kappa_i=K}\left[\frac{\Delta_i}{\hat{\mathscr{K}}_K(U_i|\boldsymbol{H}_{Ki},A_{Ki})}\frac{\partial Q_K^R(\boldsymbol{H}_{Ki},A_{Ki};\beta_K)}{\partial\beta_K}\left\{f(U_i)-f(\mathscr{T}_{Ki})-Q_K^R(\boldsymbol{H}_{Ki},A_{Ki};\beta_K)\right\}\right]=0.$$

is an unbiased estimating equation.

To show this we will show that

$$\mathbb{E}\left[\frac{I(\kappa=K)\Delta}{\mathscr{K}_K(U|\boldsymbol{H}_K,A_K)}\frac{\partial Q_K^R(\boldsymbol{H}_K,A_K;\beta_K)}{\partial\beta_K}\left\{f(U)-f(\mathscr{T}_K)-Q_K^R(\boldsymbol{H}_K,A_K;\beta_K)\right\}\right]=0.$$

We then have that

$$
\begin{aligned}
&\mathbb{E}\left[\frac{I(\kappa=K)\Delta}{\mathscr{K}_K(U|\boldsymbol{H}_K,A_K)}\frac{\partial Q_K^R(\boldsymbol{H}_K,A_K;\beta_K)}{\partial\beta_K}\left\{f(U)-f(\mathscr{T}_K)-Q_K^R(\boldsymbol{H}_K,A_K;\beta_K)\right\}\right]\\
&=\mathbb{E}\Bigg(\mathbb{E}\Bigg[\frac{I(\kappa=K)\Delta}{\mathscr{K}_K(U|\boldsymbol{H}_K,A_K)}\frac{\partial Q_K^R(\boldsymbol{H}_K,A_K;\beta_K)}{\partial\beta_K}\left\{f(U)-f(\mathscr{T}_K)-Q_K^R(\boldsymbol{H}_K,A_K;\beta_K)\right\}\\
&\qquad\Bigg|\boldsymbol{H}_K,A_K,\kappa=K\Bigg]\Bigg)\\
&=\mathbb{E}\Bigg(I(\kappa=K)\frac{\partial Q_K^R(\boldsymbol{H}_K,A_K;\beta_K)}{\partial\beta_K}\Bigg[\mathbb{E}\left\{\frac{f(U)\Delta}{\mathscr{K}_K(U|\boldsymbol{H}_K,A_K)}\Bigg|\boldsymbol{H}_K,A_k,\kappa=K\right\}\\
&\qquad-\mathbb{E}\left\{\frac{f(\mathscr{T}_K)\Delta+Q_K^R(\boldsymbol{H}_K,A_K;\beta_K)\Delta}{\mathscr{K}_K(U|\boldsymbol{H}_K,A_K)}\Bigg|\boldsymbol{H}_K,A_k,\kappa=K\right\}\Bigg]\Bigg).
\end{aligned}
$$

Note that by definition

$$
\mathbb{E}\left\{\frac{f(U)\Delta}{\mathscr{K}_K(U|\boldsymbol{H}_K,A_K)}\Bigg|\boldsymbol{H}_K,A_k,\kappa=K\right\}=Q_K(\boldsymbol{H}_K,A_K)
$$

and

$$
\mathbb{E}\left\{\frac{\Delta}{\mathscr{K}_K(U|\boldsymbol{H}_K,A_K)}|\boldsymbol{H}_K,A_K,\kappa=K\right\}=1.
$$

Therefore the above becomes

$$
\begin{aligned}
&=\mathbb{E}\left(I(\kappa=K)\frac{\partial Q_K^R(\boldsymbol{H}_K,A_K;\beta_K)}{\partial\beta_K}\left\{Q_K(\boldsymbol{H}_K,A_K)-f(\mathscr{T}_K)-Q_K^R(\boldsymbol{H}_K,A_K;\beta_K)\right\}\right)\\
&=0.
\end{aligned}
$$

B.2 Simulation study parameter values

For the Weibull generative model the following parameter values were used for the case with a low level of censoring:

$$\begin{aligned}
\mu_{1,0} &= (0, 0.25) & \mu_{1,1} &= (0.25, -1) \\
\mu_{2_0} &= (0, 0.25, 0.25, 0.25) & \mu_{2_1} &= (0.25, 0.25, 0.25, -1) \\
\xi_{1,0} &= (7, 1) & \xi_{1,1} &= (-3, 1) \\
\xi_{2,0} &= (7, 1, 1, 1) & \xi_{2,1} &= (0, 1, 1, 1) \\
\xi_{3,0} &= (7, 1, 1, 1, 1, 1) & \xi_{3,1} &= (-3, 1, 1, 1, 2, 2) \\
\zeta_{1,0} &= (7, 1) & \zeta_{1,1} &= (0, 0) \\
\zeta_{2,0} &= (8, 1, 1, 1) & \zeta_{2,1} &= (0, 0, 0, 0) \\
\zeta_{3,0} &= (12, 1, 1, 1, 1, 1) & \zeta_{3,1} &= (0, 0, 0, 0, 0, 0) \\
\omega_{1,0} &= (3, 1) & \omega_{1,1} &= (0, 0) \\
\omega_{2,0} &= (3, 1, 1, 1) & \omega_{2,1} &= (0, 0, 0, 0) \\
\omega_{3,0} &= (3, 1, 1, 1, 1, 1) & \omega_{3,1} &= (0, 0, 0, 0, 0, 0) \\
\gamma_{C_t} &= 5 \quad \gamma_{T_t} = 5 & \gamma_{\mathcal{T}_t} &= 10 \quad \forall\ t.
\end{aligned}$$

To increase the degree of censoring, the intercept in $\zeta_{t,0}$ for all t is decreased while everything else remains the same. For the medium degree of censoring $\zeta_{t,0}$ changes to

$$\begin{aligned}
\zeta_{1,0} &= (5, 1) \\
\zeta_{2,0} &= (6, 1, 1, 1) \\
\zeta_{3,0} &= (10, 1, 1, 1, 1, 1)
\end{aligned}$$

and for the high degree of censoring $\zeta_{t,0}$ changes to

$$\zeta_{1,0} = (4, 1)$$

$$\zeta_{2,0} = (5, 1, 1, 1)$$

$$\zeta_{3,0} = (9, 1, 1, 1, 1, 1).$$

For the piecewise hazard generative model the parameter values for the case of low level censoring is given by:

$$\mu_{1,0} = (0, 0.25) \qquad \mu_{1,1} = (0.25, -1)$$

$$\mu_{2_0} = (0, 0.25, 0.25, 0.25) \qquad \mu_{2_1} = (0.25, 0.25, 0.25, -1)$$

$$\xi_{1,0} = (1, 0.5) \qquad \xi_{1,1} = (-1, 0.3)$$

$$\xi_{2,0} = (1, 0.5, 0.5, 0.5) \qquad \xi_{2,1} = (0, 0.3, 0.3, 0.3)$$

$$\xi_{3,0} = (2, 0.3, 0.3, 0.3, 0.3, 0.3) \qquad \xi_{3,1} = (-2, 0.5, 0.5, 0.5, 1, 1)$$

$$\zeta_{1,0} = (0.2, 0.2) \qquad \zeta_{1,1} = (0, 0)$$

$$\zeta_{2,0} = (0.2, 0.2, 0.2, 0.2) \qquad \zeta_{2,1} = (0, 0, 0, 0)$$

$$\zeta_{3,0} = (0.2, 0.2, 0.2, 0.2, 0.2, 0.2) \qquad \zeta_{3,1} = (0, 0, 0, 0, 0, 0)$$

$$\omega_{1,0} = (9, 3) \qquad \omega_{1,1} = (0, 0)$$

$$\omega_{2,0} = (9, 3, 3, 3) \qquad \omega_{2,1} = (0, 0, 0, 0)$$

$$\omega_{3,0} = (9, 3, 3, 3, 3, 3) \qquad \omega_{3,1} = (0, 0, 0, 0, 0, 0).$$

For the high level censoring case the values of $\zeta_{t,0}$ are changed to

$$\zeta_{1,0} = (1, 0.2)$$

$$\zeta_{2,0} = (1, 0.2, 0.2, 0.2)$$

$$\zeta_{3,0} = (1, 0.2, 0.2, 0.2, 0.2, 0.2).$$