Developing New Statistical Methods for Challenges in Evaluating Dynamic Treatment Regimes

by

Kelly A. Speth

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2021

Doctoral Committee:

Professor Lu Wang, Chair
Research Associate Professor Daniel Almirall
Professor Michael Elliott
Associate Professor Emily Somers
Research Assistant Professor Yilun Sun

Kelly A. Speth

kaspeth@umich.edu

ORCID iD: 0000-0001-8045-8406

ACKNOWLEDGEMENTS

Many faculty members supported me in my career development throughout my tenure at University of Michigan. First and foremost, I would like to express my sincere gratitude to my dissertation advisor, Lu Wang. Lu is a brilliant statistician, a wonderful mentor, and has become a good friend. I would also like to thank and acknowledge my dissertation committee members for their support and contributions: Mike Elliott, Danny Almirall, Emily Somers, as well as Yilun Sun who has also graciously been my mentor for many years, as well. I also wish to acknowledge and thank Jeremy Taylor, Kelley, Kidwell, and Phil Boonstra, for their guidance and support in my early years in the program, as well as other faculty and the staff of the Biostatistics Department.

My collaborators at Michigan Medicine have included physician-investigators from various disciplines, including breast cancer (Dan Hayes and team), plastic and reconstructive surgery (Kevin Chung and team), cardiology (Robert Brook and team), radiation oncology (Whitney Beeler), and rheumatology (Emily Somers).

My applied and research work at the University of Michigan was supported by several grants: NIH CA-83654 (Biostatistics Training in Cancer Research training grant, PI Jeremy Tayor); NIH R01-ES019616 (Environmental Science, PI Robert Brook); Internal UM Funding (deNancrede Professorship in Surgery, Kevin Chung); and NIH R01-AA024150 (Drug Abuse, PI Andrew Quanbeck).

I also would like to acknowledge and thank my professors in the Mathematics and Statistics Department at San Jose State University, including my Master's thesis

committee chair and mentor Martina Bremer, also Tim Hsu, Steve Crunk, Bee Leng Lee, and Andrea Gottlieb. In addition to being outstanding and dedicated educators, they have been kind and supportive mentors.

I would like to thank and acknowledge my friends and colleagues at UM Biostats for their friendship and support, with a particular shout out to Moneyball, and also to Lucius, Charity, and PT.

Finally, I acknowledge and thank my family and friends for their decades of support and friendship. And a special thank you to my partner, Juan Marquez, for his continued support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

ABSTRACT

Personalized medicine is built upon the understanding that patients are uniquely heterogeneous in their existing and emergent comorbidities, as well as their tolerance of, response to, and even preference for different treatments. Dynamic treatment regimes (DTRs) lead to personalized medicine through a series of stage-specific decision rules that map a patient's up-to-date individual characteristics, including treatment history and disease state, to a tailored treatment assignment at each successive treatment stage. In this dissertation we develop new statistical methods to overcome several challenges arising in the field of dynamic treatment regimes.

In Chapter II, we develop Restricted Sub-Tree Learning (ReST-L) to estimate optimal DTRs in a multi-stage multi-treatment setting using observational data while restricting estimated DTRs to include only the set of covariates considered to be meaningful tailoring variables. ReST-L uses a purity measure derived from the augmented inverse probability weighted estimator for the counterfactual mean outcome; it is able to correctly estimate the optimal underlying dynamic treatment regime for a relatively large number of covariates with comparatively small sample sizes. We demonstrate the utility of ReST-L in a study of treatment recommendations for patients presenting to the emergency department with traumatic amputation of digit(s) on the hand.

Chapter III is motivated by a clustered sequential multiple assignment randomized trial (Clustered SMART) designed to improve the clinic-level uptake of evidence-based practices and health outcomes of patients with mood disorders. We develop estimation and inference procedures for Clustered Q-learning to inform the empiri-

cal construction of an optimal clustered adaptive intervention and address the well-known non-regularity challenge that can occur in a multi-stage estimation setting. We show that estimates of model parameters are unbiased and demonstrate near nominal coverage of confidence intervals across two intervention stages when the number of clusters is large and sample sizes within clusters are moderate. We apply Clustered Q-Learning to data from a Clustered SMART to evaluate whether a set of candidate tailoring variables may be used to additionally tailor cluster-level interventions to improve patient-level outcomes of patients with mood disorders.

We develop Penalized Spline-Involved Tree-based (PenSIT) Learning in Chapter IV, which seeks to improve upon existing tree-based approaches to estimating an optimal multi-stage multi-treatment DTR. Instead of using the estimated propensity score to construct the inverse weighting, which may result in unstable estimates when weights are large, we predict missing counterfactual outcomes using regression models that incorporate a penalized spline of pre-transformed propensity scores, as well as other covariates predictive of the outcome. Our simulation results demonstrate good performance of PenSIT Learning across different scenarios, particularly when the level of confounding is high or moderate, or the sample size is small. We apply PenSIT Learning to a retrospectively-collected dataset to estimate a two-stage fluid resuscitation strategy to minimize a measure of organ dysfunction in patients with acute emergent sepsis.

CHAPTER I

Introduction

There are numerous forces motivating the development of tailored interventions in healthcare. First, individuals and organizations are uniquely heterogeneous. There is no one-size-fits-all intervention that can be used for all individuals or organizations across all scenarios. Second, healthcare expenditures in the United States are skyrocketing. In the year 2000, inflation-adjusted outlays for healthcare were estimated to be about $1.8 trillion dollars (*Kamal et al.*, 2019). Less than 20 years later, although the population of the United States has increased by only about 15% (from about 280 million in 2000 to about 325 million in 2017), expenditures for healthcare have doubled (*Kamal et al.*, 2019). Third, there has been an explosion in the past two decades of data collection, storage, and processing capabilities, permitting the investigation of heretofore impossible questions. These factors create the impetus for developing evidence-based interventions that improve outcomes in a measurable way and removing or replacing those interventions that do not.

Another important consideration when devising tailored interventions is the fact that interventions are largely implemented in sequence. In medicine, for example, chronic conditions are managed over a patient's lifetime in response to progression or improvement of disease, and involve a sequence of consecutive evaluations and treatment decisions for improving aspects of a patient's life. If a patient responds

to their initial treatment, they may maintain the current dose or be removed from treatment entirely; if the patient fails to respond to a first stage treatment, however, a higher dose or a more intensive intervention may often be prescribed. Unfortunately, most randomized controlled trials, as well as cohort or other observational studies, are designed to evaluate only a single stage across the continuum of care, which fails to account for synergistic or antagonistic effects that may occur across intervention stages and also fails to reflect long-term treatment goals.

In response to this push for data-driven, evidence-based initiatives to provide multi-stage, tailored interventions, statistical research in the field of dynamic treatment regimes has been advancing. Dynamic treatment regimens (*Chakraborty and Moodie*, 2013; *Murphy*, 2003; *Robins*, 2004), also known commonly as DTRs, individualized treatment rules, or adaptive interventions, represent multi-stage, prescribed treatment sequences that are tailored based on baseline and time-varying characteristics and are a vehicle to operationalize the manner in which interventions are delivered in practice more efficiently (*Chakraborty and Moodie*, 2013). Consider the following example of a DTR for women with early-stage breast cancer: "First perform surgery. If there are cancerous cells present in the lymph nodes following surgery, give aggressive systemic chemotherapy. If cancerous cells are absent from the lymph nodes, continue to monitor the patient over time." Note that this DTR includes the first treatment (surgery) followed by a second treatment (i.e., either chemotherapy or no chemotherapy) where the second treatment depends on a tailoring variable, i.e., the presence or absence of cancerous cells in the lymph nodes. A particular emphasis is on the "dynamic" nature of the DTR, which provides a personalized prescription that changes over time in response to emergent characteristics. This is in contrast to a "treatment sequence", which does not change at any time based on intermediate outcomes or characteristics. In fact, it can be argued that very few interventions are administered as a treatment sequence and not as a DTR in practice because in-

termediate outcomes factor into nearly all healthcare decisions made in a sequential manner. It should be understood that personalized medicine generally, and DTRs specifically, is a search for interactions, as interventions are expected to have different effects in individuals or organizations with different characteristics. DTRs have been of great recent interest in several medical specialities such as oncology, as well as in social and clinical psychology, behavioral health, and implementation science.

DTR methods are grounded in causal inference. Causal inference under Rubin's potential outcomes framework (*Rubin*, 1974) involves a comparison of counterfactual outcomes represented by the set of all potentially observable outcomes under each of the different interventions (or intervention sequences). The challenges of estimating a multi-stage DTR, however, are two-fold. First, as is often described as the fundamental problem of causal inference (*Holland*, 1986), only one of the potential outcomes–the counterfactual outcome consistent with the intervention actually received–can be observed. Thus, in a single stage setting with only two possible treatments, $\frac{1}{2} = 50\%$ of the potential outcomes is observed. Adding a second stage, again with only two possible treatment options at the second stage, it can easily be seen that only $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 25\%$ of the counterfactual outcomes are actually observed, i.e., the outcome associated with only one of the four possible treatment sequence combinations. Adding additional interventions at each stage or considering more than two stages rapidly expands upon this challenge. Secondly, it is often the case in a longitudinal, multi-stage data setting that a variable will be both an intermediate outcome of prior treatment yet a confounder of future treatment, meaning that a variable can be both a mediator yet also confounder, known as "confounding by indication" which, when not addressed, can lead to bias in estimation. In an ideal world every individual would have completed each candidate treatment regime under the exact same circumstances and we would be able to compare outcomes for each patient across regimes and identify covariate-treatment interactions at each stage that

maximize outcomes. Under ordinary circumstances, however, this is not possible. In order to perform causal inference in the absence of all potential outcomes, a series of assumptions must be fulfilled: consistency, positivity, and no unmeasured confounders (NUCA). In a sequential multiple assignment randomized trial (SMART), a trial design defined by the presence of at least two sequential randomizations that is commonly used to evaluate candidate DTRs, these assumptions are fulfilled by the nature of an adequately designed trial and correctly performed randomizations. Often, however, randomized trials are expensive, unethical or infeasible, and/or long in duration. In such cases we may rely on observational data. In analyses of observational data, by contract, treatments are not assigned randomly and actual treatment decisions are likely to be based, either formally or informally, on some known patient characteristics. If we fail to account for the actual treatment assignment mechanism, we cannot make claims of causality and our results are of minimal value.

One primary objective within the field of DTRs is to estimate decision rules spanning the multiple intervention stages such that, if the DTR were applied to the population of interest, outcomes would be optimized. Specifically, this may involve identifying relevant tailoring variable(s) and their respective cutpoints at each stage. Given the great potential value of DTRs in personalized medicine, there has understandably been an abundance of statistical literature related to estimating and evaluating DTRs. Two such classes of methods that are flexible, easy-to-implement, and produce interpretable results include decision tree-type DTR estimation and Q-Learning. Although these methods continue to experience increasing popularity, many unaddressed statistical research questions still exist.

A decision tree, also known as a classification and regression tree (CART; *Breiman et al.*, 1984), is a popular machine learning algorithm used for prediction. Using a set of training observations, the CART algorithm recursively partitions the covariate space until observations within each terminal (leaf) node achieve maximum "purity",

i.e., relative homogeneity of the target variable for observations defined within the partition. A predicted value of the target variable can then be made for a new set of observations based on their distinct covariate values. A purity measure, a term used frequently in the CART literature, is the metric used to determine if and when a binary split of the covariate space will be made. Purity measures commonly used in the CART literature include misclassification rates for binary outcomes or residual sum of squares for continuous outcomes (*Hastie et al.*, 2009). CART algorithms were developed as a prediction tool rather than for the field of causal inference, however. To design a CART algorithm with the goal of estimating a causal relationship, e.g., a DTR, purity measures utilizing estimators of counterfactual outcomes were developed. *Laber and Zhao* (2015) incorporate an inverse probability weighted estimator (IPW) into a purity measure for tree-based learning, whereas *Zhang et al.* (2015) utilize a doubly robust augmented inverse probability weighted (AIPW) estimator within a more restrictive tree-based framework known as a decision list. Tree-based reinforcement learning (T-RL), developed by *Tao et al.* (2018), incorporates the AIPW estimator of the counterfactual mean outcome within a purity measure to estimate an optimal DTR in a multi-stage, multi-treatment setting using a flexible, tree-based framework. Whereas it is reasonable that all variables may be used to estimate the propensity for treatment assignment, as well as the conditional mean outcome models, both of which are needed to derive the AIPW estimate, T-RL dictates that the full set of covariates be considered as candidate tailoring variables, which is often not reasonable in practice. In the example DTR introduced above, it would be unethical to determine whether to treat the patient with chemotherapy following surgery based on her socioeconomic status. Another challenge identified within decision tree-type DTR estimation is the fact that IPW-weighted estimators used within the purity measure may be unstable when weights are large, leading to unstable estimates of counterfactual outcomes. In this dissertation we address both

challenges arising within tree-based DTR estimation, statistical questions which to our knowledge have not yet been explored in the literature.

A second class of methods for estimating DTRs is Q-Learning (*Chakraborty et al.*, 2013; *Moodie et al.*, 2012; *Murphy*, 2005b; *Nahum-Shani et al.*, 2012; *Schulte et al.*, 2014). Q-Learning was originally proposed within the computer science literature and has become a widely-popular method to estimate personalized, multi-stage interventions. Implemented recursively starting with the final stage, Q-Learning utilizes the language of Q-functions, which are defined for each stage k as the expected outcome conditional on covariate and intervention history collected through the k-th stage. A Q-function is said to be "optimal" if the expected counterfactual outcome is maximized (assuming higher values of the outcome are desired). Given their definition, Q-functions can be modeled in many ways; often, however, Q-functions are modeled parametrically using standard linear or generalized linear regression models due to the widespead use and understanding of such methods. Whereas medical research often concerns settings in which the intervention is administered and outcomes are collected at the individual level, much research in the areas of education, behavioral science, and implementation science evaluates interventions administered at the cluster level while the outcome of interest resides at the individual level. In this dissertation we address the challenge of clustered data when using Q-Learning for estimating a multi-stage multi-treatment DTR.

In Chapter II we introduce Restricted Sub-Tree Learning (ReST-L) for estimating an optimal DTR using observational data when restrictions to the set of candidate tailoring variables in the multi-stage decision rules are justified. Given a set of time-varying patient characteristics and treatment history, ReST-L utilizes a decision tree framework to build an estimated DTR based on a sub-tree. We demonstrate that, when clinical knowledge substantiates consideration of only a subset of covariates as candidate tailoring variables but other covariates may define the treatment assign-

ment mechanism or may be related to the outcome, ReST-L provides a flexible and interpretable, semi-parametric analysis approach for estimating the optimal dynamic treatment regime across multiple stages. We show that, when the true, underlying DTR structure is tree-type, ReST-L is a significant improvement over both parametric and nonparametric Q-Learning. Although, asymptotically, performance of ReST-L and T-RL are similar, ReST-L provides more power in the smaller sample sizes common in biomedical research. When the underlying DTR is a complicated nonlinear function, ReST-L also improves upon both parametric and nonparametric Q-Learning across all sample sizes.

In Chapter III we introduce Clustered Q-Learning for use with data from a Clustered SMART–a SMART in which the intervention is applied at the cluster level but the outcome of interest lies at the level of the individual within the cluster–to inform the empirical construction of an optimal clustered adaptive intervention (CAI). This scenario occurs with increasing frequency in the area of implmentation science (*Fernandez et al.*, 2020; *Kilbourne et al.*, 2014; *Kilbourne et al.*, 2018; *Quanbeck et al.*, 2020; *Smith et al.*, 2019; *Zhou et al.*, 2020), a field that seeks to develop data-driven strategies that facilitate uptake of practices–clinical evidence-based practices or standard operating procedures, for example– into regular use. Using Clustered Q-Learning we determine whether a set of pre-specified candidate tailoring variables can be used to tailor multi-stage interventions administered at the cluster level such that, if the CAI were applied to the population of interest, outcomes would be optimized. While estimation of regression parameters is straightforward using a general linear model framework with estimation by either maximum likelihood or generalized estimating equations, a well-known challenge in Q-Learning regression concerns the estimation of confidence intervals under conditions of nonregularity, which is expected to be common. Nonregularity is a phenomenon that will affect the estimation of standard errors for any earlier intervention stage if there is a non-unique treatment effect

(e.g., no treatment effect) at one or more of the later intervention stages. To enable the construction of confidence intervals with nominal coverage rates, we propose the M-out-of-N cluster bootstrap, which is an extension of the m-out-of-n bootstrap for uncorrelated data (*Chakraborty et al.*, 2013), to accommodate data from a Clustered SMART. We show that estimation of model parameters using Clustered Q-Learning is unbiased and demonstrate near nominal coverage of confidence intervals across two intervention stages when the number of clusters is large and the cluster size is moderate.

In Chapter IV we propose a new method to estimate the optimal multi-stage multi-treatment dynamic treatment regime using observational data, PenSIT Learning, which seeks to improve upon existing tree-based approaches that rely on IPW-type estimators for causal inference. While conceptually similar to the implementation of ReST-L, PenSIT Learning makes use of a purity measure derived from the penalized spline of propensity prediction (PSPP) method used for missing data (*Little and An*, 2004; *Zhang and Little*, 2009; *Zhou et al.*, 2019). Specifically, instead of using weights derived from the estimated propensity score as is done with IPW estimators, we predict missing counterfactual outcomes for the treatments not assigned to patients using regression models that incorporate a penalized spline of a function of the propensity to be assigned that treatment and other covariates predictive of the outcome. The estimator retains the property of double robustness against model misspecification. We demonstrate that PenSIT Learning typically outperforms Q-Learning methods across most data generating scenarios and is a viable alternative to T-RL, particularly when the level of confounding is high or moderate and when the sample size is small.

In summary, in this dissertation we address three challenges that exist within the field of optimal DTR estimation. Within the framework of decision tree-type optimal DTR estimation we provide a method that can be selected when only a subset

of variables, based on prior clinical knowledge, may be used as candidate tailoring variables in a multi-stage DTR. Secondly, we adapt Q-Learning to accommodate clustered data arising when interventions are delivered at the cluster level but outcomes of interest lie at the individual level within the cluster. Finally, we develop a novel purity measure based on the penalized spline for propensity prediction method used for missing data, which we incororporate into a decision tree framework for estimating a multi-stage DTR.

CHAPTER II

Restricted Sub-Tree Learning (ReST-L) to Estimate an Optimal Dynamic Treatment Regime Using Observational Data

2.1 Introduction

There is a drive in the healthcare field toward evidenced-based and personalized medicine, which has the potential to both improve patient outcomes, lower costs, and allocate healthcare resources in more efficient ways. In essence, personalized healthcare recognizes that patients are uniquely heterogeneous in their preexisting and emergent comorbidities, as well as their response to or tolerance of–and even preference for–a particular treatment. The overarching goal of personalized medicine, therefore, is to search for treatment interactions that can define which patients will benefit from which treatments. One such avenue to achieve this goal is through dynamic treatment regimes, which have become of great recent interest in several medical specialties including oncology, as well as in social and clinical psychology and behavioral health. Dynamic treatment regimes (*Chakraborty and Moodie*, 2013; *Murphy*, 2003; *Robins*, 2004) also known commonly as DTRs, individualized treatment rules, or adaptive interventions, represent multi-stage, prescribed treatment sequences that are tailored to the individual based on their baseline and time-varying

characteristics and are a vehicle to operationalize the manner in which patient care for chronic diseases is delivered in practice more efficiently (*Chakraborty and Moodie*, 2013).

A primary statistical objective in the field of DTRs is to estimate stage-specific decision rules that will optimize the expected long-term counterfactual outcomes of patients when applied across the population of interest. Given the potential value of DTRs for both improving long-term patient outcomes and optimizing the allocation of resources needed for patient care, there are understandably an abundance of methods that have been developed to estimate optimal DTRs. Existing methods can be classified in a number of ways, one of which relates to its dependence on parametric and semi-parametric, or more flexible nonparametric assumptions (*Hernan et al.*, 2001; *Huang et al.*, 2015; *Murphy*, 2003; *Murphy et al.*, 2001; *Robins*, 1986; *Robins*, 1994; *Robins*, 1997; *Robins*, 2000; *Robins*, 2004; *Schulte et al.*, 2014; *Thall et al.*, 2000; *Thall et al.*, 2002; *Thall et al.*, 2007; *van der Laan and Rubin*, 2006; *Wang et al.*, 2012; *Zhang et al.*, 2013). In recent years, however, due to the increasing capacity for large scale data collection and storage, as well as advancements in computation, flexible methods have become increasingly popular (*Arjas and Saarela*, 2010; *Moodie et al.*, 2013; *Qian and Murphy*, 2011; *Xu et al.*, 2016; *Zhao et al.*, 2009). Tree-based methods that offer flexibility and robustness in estimation are desirable given, often, an abundance of observational data, as well as a high degree of uncertainty with regard to the complex relationships among variables. Additionally, because optimal DTR estimation is exploratory in nature and communication with clinicians is crucial, methods with interpretable results are particularly favored. Tree-based methods informing optimal DTR construction in a single stage and/or with binary treatment decisions (*Laber and Zhao*, 2015; *Zhang et al.*, 2012a; *Zhang et al.*, 2012b; *Zhao et al.*, 2012; *Zhao et al.*, 2015) have been expanded to accommodate a multi-stage and multi-treatment setting (*Tao and Wang*, 2017; *Tao et al.*, 2018; *Zhang et al.*, 2018). *Zhang*

et al. (2018) define a new class of tree-based DTRs known as decision lists, which are expressed as a series of consecutive "if-then" statements. Although the robust nature of the estimator used by *Zhang et al.* (2018), as well as the focus of the methodology on interpretatibility of the estimated optimal DTRs, are important, these list-based DTRs can incorporate only two covariates per rule. Additionally, given the unidirectional growth of this special case decision tree, errors in estimation can accumulate across stages. *Tao et al.* (2018) introduce a more flexible, tree-based method known as tree-based reinforcement learning (T-RL) for multi-stage and multi-treatment estimation of an optimal DTR. T-RL is a semi-parametric approach combining the flexibility of a decision tree (e.g., *Breiman et al.*, 1984) with a purity measure, i.e., a metric used within the decision tree framework to devise binary covariate splits, that is derived from a doubly-robust augmented inverse probability weighted (AIPW) estimator of the counterfactual mean outcome (*Zhang et al.*, 2012b). Although T-RL provides a flexible and robust modeling approach yielding interpretable results, and is able to accommodate multiple treatments across multiple treatment stages, the algorithm requires all observed covariates to be considered as possible tailoring variables in an optimal DTR, which is unlikely to exclusively occur in practice.

For use in real-world settings we desire an estimator that provides interpretable results, so that it can be examined within the clinical community, but is also flexible yet robust against model misspecification. Additionally, it is often necessary to consider only a restricted subset of covariates as candidates for an estimated optimal dynamic treatment regime. For example, envisage the case of a patient presenting to the emergency department with traumatic amputation of digit(s) on the finger(s). Current practice recommends treatment with either replantation, where the digit(s) are replanted at the point at which they were severed, or correction of the amputated stump, also known as "revision amputation". There are de facto treatment decision rules to guide surgeons' choice of treatment; however, rigorous, data-driven guidelines

that optimize long-term patient outcomes have not yet been developed. Because all treatment decisions are ultimately subject to patient choice, conducting a randomized study to evaluate whether replantation or revision amputation of the amputated digits improve long-term patient outcomes would be infeasible. Therefore, use of observational data, e.g., a longitudinal cohort study, is necessary. Given the regional differences in treatment practices, it would be important to incorporate the actual or estimated treatment assignment mechanism (i.e., the propensity model) into an estimator. However, although all measured covariates may be used to define the treatment assignment mechanism, many of the covariates–socioeconomic status or type of insurance coverage (e.g., private insurance versus Medicaid), for example–would, for a variety of reasons, be considered inappropriate to be included in an estimated optimal treatment regime. We therefore seek a flexible and interpretable method that has desirable statistical properties and can be used with observational or randomized data to estimate an optimal multi-stage dynamic treatment regime that is restricted in its prescriptive covariates.

In this manuscript we develop Restricted Sub-Tree Learning (ReST-L) that utilizes a decision tree framework but restricts the covariate space, according to subject-matter knowledge, to build an estimated, optimal DTR based on a sub-tree, a topic which to our knowledge has not yet been explored in the statistical literature. In order to determine the binary splits of the covariate space, we use a purity measure derived from an AIPW estimator of the counterfactual mean outcome, which is doubly robust to model misspecification; however, although all possible tailoring and confounding variables are used to estimate propensity and conditional mean models, both of which are used to derive the AIPW estimator, only a subset of covariates are considered as possible tailoring variables. We input into the algorithm a set of time-varying patient characteristics and treatment history and output a dynamic, personalized, and optimal multi-stage treatment regime. In simulation studies we demonstrate that,

when clinical knowledge substantiates consideration of only a subset of covariates as candidate tailoring variables but other covariates may define the treatment assignment mechanism or may be related to the outcome, ReST-L provides a flexible, semi-parametric analysis approach with interpretable estimates of an optimal, multi-stage dynamic treatment regime.

2.2 Mathematical Formulation and Assumptions

2.2.1 Notation and Formulation

Suppose one of K_j treatments ($k_j = 1, ..., K_j$; $K_j \geq 2$) is administered to every subject $i = 1, ..., n$ at each of $j = 1, ..., J$ stages. The actual treatment received by the i-th patient at stage j is denoted $A_{j,i}$. As is customary, we use the convention of using a capital letter to refer to the unrealized random variable and lowercase to refer to a realized value. For simplicity, we omit the subscript i from future notation when no confusion exists. Let us denote the j-th stage covariates that are observed and available when making the j-th treatment decision as $\mathbf{X}_j$. Assuming that only a subset of covariates among $\mathbf{X}_j$ may be used to define the j-th stage decision rule, we distinguish between $\mathbf{X}_{\text{sub,j}}$ and $\mathbf{X}_{\text{sub,j}}{}^C = \mathbf{X}_j \setminus \mathbf{X}_{\text{sub,j}}$, where $\mathbf{X}_{\text{sub,j}}$ represents a p_j-dimensional vector of multi-scale data ($p_j \geq 1$) corresponding to measured covariates that a clinician would consider as candidates to include in a treatment decision rule at stage j. Conversely, $\mathbf{X}_{\text{sub,j}}{}^C$ is a q_j-dimensional vector of multi-scale data ($q_j \geq 1$) corresponding to the set of measured covariates that may <u>not</u> be included in a clinical treatment decision rule. In the context of our application example to determine an optimal decision rule to treat patients with traumatic amputation of digits on a hand, $\mathbf{X}_{\text{sub}}$ includes variables such as the number of digits amputated, whether the thumb was amputated, and whether the injury occurred to the dominant hand–all of which a clinician would consider as possible tailoring variables–whereas $\mathbf{X}_{\text{sub}}^C$ includes

the patient's type of insurance coverage, socioeconomic status, and other variables that a clinician would not use to assign treatment. After each j-th treatment stage, measurements are made on a set of covariates $\mathbf{X}_{j+1}$, which may also include an intermediate outcome, Y_j. Such an intermediate outcome Y_j may reflect a certain response from previous treatments or may be a function of the previous treatment history and other observed covariates, and may be used to determine treatment for the $(j+1)$-th stage. Thus, Y_j may also be an element of $\mathbf{X}_{\text{sub,(j+1)}}$. Following convention, we use overbar to denote history, i.e., all observations collected at stages on or before the jth stage. For example, $\overline{\mathbf{X}}_j = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_j)$ represents all covariate data collected and available to the clinician prior to the jth treatment decision. Similarly, $\overline{A}_{j-1}$ represents the set of all treatments received prior to the jth treatment decision, i.e., $\overline{A}_{j-1} = (A_1, A_2, ..., A_{j-1})$, and $\overline{Y}_{j-1} = (Y_1, Y_2, ..., Y_{j-1})$ represents the set of all intermediate outcomes observed prior to the jth treatment decision. Suppose the final outcome of interest is $Y = h(Y_1, Y_2, ..., Y_J)$, which may be a function of stage-specific intermediate outcomes $Y_1, Y_2, ..., Y_J$, assumed to be continuous and approximiately normally distributed. Here $h(\cdot)$ represents some clinically-relevant, prespecified function (e.g., sum or last value). The full history prior to the decision at stage j is then expressed as $\mathbf{H}_j = (\overline{A}_{j-1}, \overline{\mathbf{X}}_j)$. $\mathbf{H}_{\text{sub,j}} = (\overline{A}_{j-1}, \overline{\mathbf{X}}_{\text{sub,j}})$ includes the full treatment history prior to the treatment decision at stage j and covariate history only from the subset $\overline{\mathbf{X}}_{\text{sub,j}}$ that may be used in a clinical decision rule. Using a similar convention, $\mathbf{H}^C_{\text{sub,j}} = \overline{\mathbf{X}}^C_{\text{sub,j}}$ includes covariate history through stage j for variables not considered for a treatment regime. Next let $\mathbf{g} = (g_1, g_2, ..., g_J)$ denote a J-stage dynamic treatment regime. Each stage-specific decision rule g_j is a function only of covariates that can be used to make treatment decisions at each stage, i.e., $g_j : \mathbf{H}_{\text{sub, j}} \rightarrow A_j$. Refering to the motivating example introduced previously, an example of an estimated single stage decision rule $g(\mathbf{H}_{\text{sub}})$ that would guide clinicians in determining whether to perform replantation or revision amputation of traumatically amputated digits is

as follows: If the patient has 3 or more fingers amputated, perform digit replantation; otherwise, perform revision amputation.

As we are interested in making a causal claim related to an estimated optimal dynamic treatment regime, we employ Rubin's potential outcome framework (*Rubin*, 1974). At stage J we let $Y^*(A_1, ..., A_{J-1}, a_J)$, or simply $Y^*(a_J)$, denote the counterfactual outcome, also known as a potential outcome, for a patient treated with $a_J \in \mathcal{A}_J$ conditional on prior treatment history $\overline{A}_{J-1}$. Notably, only one counterfactual outcome–the one consistent with the treatment actually received–will be observed. In the context of our estimation problem we can similarly define $Y^*\{\mathbf{g}(\mathbf{H}_{\text{sub}})\}$ as the counterfactual outcome under the multi-stage dynamic treatment regime $\mathbf{g}(\mathbf{H}_{\text{sub}})$. As mentioned above, only one counterfactual outcome will be observed, although in this case the observed counterfactual will be the potential outcome consistent with the DTR followed by the individual. We measure the performance of a multi-stage DTR, $\mathbf{g}(\mathbf{H}_{\text{sub}})$, using the counterfactual mean outcome $E[Y^*\{\mathbf{g}(\mathbf{H}_{\text{sub}})\}]$, the higher the better by convention, and define the optimal DTR $\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})$ as the one that satisfies

$$E[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] \geq E[Y^*\{\mathbf{g}(\mathbf{H}_{\text{sub}})\}]$$

for all $\mathbf{g}(\mathbf{H}_{\text{sub}}) = (g_1, g_2, ..., g_J)^T \in \mathbf{G}_{\text{sub}}$, where $\mathbf{G}_{\text{sub}}$ is the class of all potential regimes constructed using $\mathbf{H}_{\text{sub}}$ only. Our statistical goal, therefore, can be summarized as follows: to estimate an interpretable, optimal, J-stage treatment regime, $\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})$, from observational data such that, if all patients were to be assigned to multi-stage treatment using this regime, the expected counterfactual outcome of our population of interest would be maximized: $\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}}) = \text{argmax}_{\mathbf{g}\in\mathbf{G}_{\text{sub}}} E[Y^*\{\mathbf{g}(\mathbf{H}_{\text{sub}})\}]$.

2.2.2 Link to Observed Data

The above optimization objective features $Y^*\{\mathbf{g}(\mathbf{H}_{\text{sub}})\}$, the counterfactual outcome; however, as is widely known in causal inference, only one of the potential outcomes is observed, making estimation of a causal effect impossible without a series of assumptions. To proceed with estimation of an optimal DTR under Rubin's potential outcomes framework, we make three foundational assumptions: consistency, positivity, and no unmeasured confounders (NUCA).

(1) Consistency. Under an assumption of consistency, the potential outcome under the observed treatment agrees with that of the observed outcome Y. For a single stage treatment, we can express this as: $Y = \sum_{a=1}^{K} Y^*(a)\mathcal{I}(A = a)$, where $\mathcal{I}(\cdot)$ is an indicator function that returns a value of 1 if the argument is true and a value of 0 otherwise. Consistency further assumes that there is no interference between units, which means that one patient's observed and counterfactual outcomes are independent of the treatment(s) of all other patients.

(2) Positivity. $0 < \tau < Pr(A_i = a|\mathbf{H}_i) < 1$ for all $a \in \mathcal{A}, \mathbf{H}_i \in \boldsymbol{\mathcal{H}}$, where τ is a positive constant, i.e., the probability to be assigned to each possible treatment is bounded away from 0.

(3) NUCA. Finally, we assume that there are no unmeasured confounders, i.e., data on all variables associated with both the assignment of treatment A and the outcome Y have been observed. That is, given history $\mathbf{H}$, $Y_1^*(1), ..., Y_K^*(K) \perp\!\!\!\perp A|\mathbf{H}$, where $\perp\!\!\!\perp$ denotes statistical independence.

In contrast to previous work with tree-based reinforcement learning (*Tao and Wang*, 2017; *Tao et al.*, 2018), our interest is to estimate a $\mathbf{g}^{\text{opt}}$ that is based only on covariates in $\mathbf{H}_{\text{sub}}$. Because $Y^*\{g(\mathbf{H}_{\text{sub}})\} = \sum_{a=1}^{K} Y^*(a)\mathcal{I}\{g(\mathbf{H}_{\text{sub}}) = a\}$, the optimal decision rule g for a single stage can be expressed as $g^{\text{opt}}(\mathbf{H}_{\text{sub}}) =$ $\text{argmax}_{g \in \mathcal{G}_{\text{sub}}} E\left[\sum_{a=1}^{K} Y^*(a)\mathcal{I}\{g(\mathbf{H}_{\text{sub}}) = a\}\right]$. After taking an iterated expectation and conditioning on history $\mathbf{H}$, and in accordance with our assumptions of consis-

tency, positivity, and NUCA, we can express the optimal decision rule as follows:

$$g^{\text{opt}}(\mathbf{H}_{\text{sub}}) = \text{argmax}_{g \in \mathcal{G}_{\text{sub}}} E[Y^*\{g(\mathbf{H}_{\text{sub}})\}] = \text{argmax}_{g \in \mathcal{G}_{\text{sub}}} E_{\mathbf{H}}\Big[\sum_{a=1}^{K} E\big\{Y | g(\mathbf{H}_{\text{sub}}) = a, \mathbf{H}\big\}\Big]$$

2.3 Restricted Sub-tree Learning (ReST-L)

2.3.1 Single Treatment Stage

Consider estimation of the optimal decision rule $g^{\text{opt}}(\mathbf{H}_{\text{sub}})$ for a single stage with K possible treatments: $g^{\text{opt}} : \boldsymbol{\mathcal{H}}_{\text{sub}} \rightarrow \{1, 2, ..., K\}$. Define C as a compatibility indicator, with $C = \sum_{a=1}^{K} \mathcal{I}(A = a) \cdot \mathcal{I}\{g(\mathbf{H}_{\text{sub}}) = a\}$, which is equivalent to $\mathcal{I}\{A = g(\mathbf{H}_{\text{sub}})\}$, i.e., that the actual treatment received is consistent with the treatment assigned by rule $g(\mathbf{H}_{\text{sub}})$. Next define $\pi_a(\mathbf{H}) = Pr(A = a|\mathbf{H})$ as the propensity score for treatment assignment, noticing that this potentially depends on all variables in $\mathbf{H}$–not just variables in $\mathbf{H}_{\text{sub}}$. Also, denote $\pi_C(\mathbf{H})$ as the probability of receiving treatment consistent with $g(\mathbf{H}_{\text{sub}})$. Assuming we have observational data, we would posit a propensity model $\pi_a(\mathbf{H}; \gamma)$, e.g., using multinomial logistic regression, to estimate γ. We see that:

$$\pi_C(\mathbf{H}) = Pr(C = 1|\mathbf{H}) = E(C|\mathbf{H}) = E\Big[\sum_{a=1}^{K} \mathcal{I}(A = a) \cdot \mathcal{I}\{g(\mathbf{H}_{\text{sub}}) = a\}|\mathbf{H}\Big] = \sum_{a=1}^{K} \pi_a(\mathbf{H}) \cdot \mathcal{I}\{A = g(\mathbf{H}_{\text{sub}}) = a\}$$

Under the three assumptions in Section 2.2.2, condiser the IPW estimator of $E[Y^*\{g(\mathbf{H}_{\text{sub}})\}]$, i.e., $\hat{E}[Y^*\{g(\mathbf{H}_{\text{sub}})\}] = \mathbb{P}_n\Big\{\frac{C \cdot Y}{\hat{\pi}_C(\mathbf{H}_i; \hat{\gamma})}\Big\}$, where C and $\pi_C(\mathbf{H})$ are defined above, $\mathbb{P}_n(\cdot)$ represents the empirical mean operator evaluated over all patients i, and Y represents our outcome of interest. Under an assumption of consistency and positivity:

$$E\left[\frac{C \cdot Y}{\hat{\pi}_C(\mathbf{H};\hat{\gamma})}\right] = E\left[\frac{C}{\hat{\pi}_C(\mathbf{H};\hat{\gamma})} Y^*\{g(\mathbf{H}_{\text{sub}})\}\right]$$

Taking an iterated expectation conditional on $\mathbf{H}$ and under the assumption of NUCA, the above is equivalent to:

$$E_{\mathbf{H}}\left[E\Big[\frac{C}{\hat{Pr}(C=1|\mathbf{H})} Y^*\{g(\mathbf{H}_{\text{sub}})\}|\mathbf{H}\Big]\right] =$$
$$E_{\mathbf{H}}\left[E\left[\frac{\mathcal{I}\{A=g(\mathbf{H}_{\text{sub}})\}}{\hat{Pr}\{A=g(\mathbf{H}_{\text{sub}})|\mathbf{H}\}}|\mathbf{H}\right] E[Y^*\{g(\mathbf{H}_{\text{sub}})\}|\mathbf{H}]\right]$$

If $\hat{\pi}_C(\mathbf{H})$ is correctly specified, $E\left[\frac{\mathcal{I}\{A=g(\mathbf{H}_{\text{sub}})\}}{\hat{Pr}\{A=g(\mathbf{H}_{\text{sub}})|\mathbf{H}\}}|\mathbf{H}\right] = 1$, which demonstrates that an IPW style estimator is consistent in large samples for estimating the counterfactual mean outcome $E[Y^*\{g(\mathbf{H}_{\text{sub}})\}]$ under a regime $g(\mathbf{H}_{\text{sub}})$:

$$\hat{E}[Y^*\{g(\mathbf{H}_{\text{sub}})\}] = \mathbb{P}_n\left[\frac{C \cdot Y}{\hat{\pi}_C(\mathbf{H};\hat{\gamma})}\right] \to^p E[Y^*\{g(\mathbf{H}_{\text{sub}})\}]$$

However, because the IPW estimator of the counterfactual mean can quickly become unstable as the number of treatment stages increases, estimation can be improved by utilizing a doubly robust estimator of the counterfactual mean, also known as the augmented inverse probability weighted (AIPW) estimator. It has been shown that $\mathbb{P}_n\{\hat{\mu}_a^{\text{AIPW}}(\mathbf{H})\}$ is a consistent estimator for $E\{Y^*(a)\}$ (*Tao et al.*, 2018; and others), where $\hat{\mu}_a^{\text{AIPW}}(\mathbf{H})$ is defined as follows, with $\hat{\mu}_a(\mathbf{H}) = E(Y|A=a,\mathbf{H})$ representing the conditional mean model:

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

Now considering the decision rule $g(\mathbf{H}_{\text{sub}})$, we propose to extend the AIPW estimator to estimate the counterfactual mean under regime $g(\mathbf{H}_{\text{sub}})$, i.e., $\hat{E}[Y^*\{g(\mathbf{H}_{\text{sub}})\}]$, as $\mathbb{P}_n\left[\sum_{a=1}^K \hat{\mu}_a^{\text{AIPW}}(\mathbf{H})\, \mathcal{I}\{g(\mathbf{H}_{\text{sub}})=a\}\right]$, which can also be expressed as:

$$\hat{E}[Y^*\{g(\mathbf{H}_{\text{sub}})\}] = \mathbb{P}_n\left[\frac{C}{\hat{\pi}_C(\mathbf{H})}Y + \left\{1 - \frac{C}{\hat{\pi}_C(\mathbf{H})}\right\}\hat{\mu}_C(\mathbf{H})\right]$$

where both C and $\pi_C(\mathbf{H})$ are as defined above, and $\hat{\mu}_C(\mathbf{H}) = \hat{E}\{Y|A = g(\mathbf{H}_{\text{sub}}) = a, \mathbf{H}\}$. The AIPW estimator of the counterfactual mean outcome for treatment $A = a$ under regime $g(\mathbf{H}_{\text{sub}})$ is then expressed as:

$$\mathbb{P}_n\left[\frac{\mathcal{I}\{A = g(\mathbf{H}_{\text{sub}}) = a\}}{\hat{\pi}_C(\mathbf{H})}Y + \left\{1 - \frac{\mathcal{I}\{A = g(\mathbf{H}_{\text{sub}}) = a\}}{\hat{\pi}_C(\mathbf{H})}\right\}\hat{\mu}_C(\mathbf{H})\right]$$

It is doubly robust in the sense that it will provide a consistent estimator of the counterfactual mean outcome under regime $g(\mathbf{H}_{\text{sub}})$ if either the models of $\hat{\pi}_C(\mathbf{H})$ or the conditional mean outcome model $\hat{E}\{Y|A = g(\mathbf{H}_{\text{sub}}) = a, \mathbf{H}\}$ is correctly specified.

With the above knowledge, we propose ReST-L, a new learning procedure akin to the classification and regression tree (CART) (e.g., *Breiman et al.*, 1984), to estimate $g(\mathbf{H}_{\text{sub}})$. One important feature of CART methods, which will also be an important component of ReST-L, is the "purity measure". A purity measure is used to quantify the degree of similarity–or "purity"–of observations with respect to a target variable. Specifically, the measured purity is used to determine binary covariate splits, mimicking the branching of a tree, such that observations within each "leaf" node are relatively homogeneous with respect to a target variable–and then to use the estimated partition of the covariate space to predict the target variable for a set of new observations. The process of splitting nodes of the tree into binary partitions of the covariate space continues until the pre-specified depth of the tree is achieved or until the improvement in the "purity" falls below a pre-specified level. Examples of purity measures frequently used with CART include entropy, the Gini index, and the sum of squared prediction errors (*Hastie et al.*, 2009). A similarity between CART and ReST-L, as introduced above, includes the fact that a partition of the covariate space is made such that observations within each subset are relatively homogeneous.

A notable difference, however, is the fact that the target of estimation for ReST-L is an optimal decision rule, which determines optimal treatment assignment based on observed covariates but is not directly observed. Furthermore, ReST-L, in contrast to CART methods, rests within the causal inference framework. Therefore, we propose for ReST-L a new purity measure suitable for our goal of estimating a decision rule based on only a subset of covariates while also preserving a causal interpretation. Specifically, we exploit the consistent, large-sample, doubly-robust AIPW estimator of the counterfactual mean outcome for a decision rule based only on a subset of covariates introduced above and define our purity measure represented by the binary partition created by split ω of node Ω, $\mathcal{P}(\Omega, \omega)$, as follows:

$$\mathcal{P}(\Omega,\omega) = \max_{a_1,a_2\in\mathcal{A}} \ \mathbb{P}_n\left[\sum_{a=1}^{K} \hat{\mu}_a^{\text{AIPW}}(\mathbf{H}) \ \mathcal{I}\{A = g_{\omega,a_1,a_2}(\mathbf{H}_{\text{sub}}) = a\} \ \mathcal{I}(\mathbf{H}_{\text{sub}} \in \Omega)\right],$$

where g_{ω,a_1,a_2} denotes a decision rule such that patients in ω are assigned treatment a_1 while patients in the complementary set ω^C are assigned to a_2, with $a_1 \neq a_2$. Using this purity measure, ReST-L is implemented as described in Section 2.3.3.

2.3.2 Multiple Treatment Stages

We now extend ReST-L to a setting with multiple treatment stages, $j = 1, ..., J$, with 2 or more treatment options per stage, i.e., $K_j \geq 2$. Because of the potential for confounding by indication, a problem that can introduce substantial bias in a multi-stage estimation, ReST-L is implemented recursively using backward induction (*Bather*, 2000), beginning with estimation of the final stage and continuing backwards in time through all prior stages.

We first consider estimation of the decision rule for the final, J-th, stage. We perform this estimation in the same manner in which we estimate a single stage decision rule. Following our exposition in section 2.3.1,

$$\hat{\mu}^{\text{AIPW}}_{J,a_J}(\mathbf{H}_J) = \frac{\mathcal{I}(A_J = a_J)}{\hat{\pi}_{J,a_J}(\mathbf{H}_J)} Y + \left\{ 1 - \frac{\mathcal{I}(A_J = a_J)}{\hat{\pi}_{J,a_J}(\mathbf{H}_J)} \right\} \hat{\mu}_{J,a_J}(\mathbf{H}_J)$$

where $\hat{\mu}_{J,a_J}(\mathbf{H}_J) = \hat{E}[Y|A_J = a_J, \mathbf{H}_J]$. The estimator of the J-th stage counterfactual mean outcome for $A_J = a_J$ under regime $g_J(\mathbf{H}_{\text{sub,J}})$ can then be expressed as:

$$\mathbb{P}_n \left(\frac{\mathcal{I}\{A_J = g_J(\mathbf{H}_{\text{sub,J}}) = a_J\}}{\hat{\pi}_{J,C_J}(\mathbf{H}_J)} Y + \left[1 - \frac{\mathcal{I}\{A_J = g_J(\mathbf{H}_{\text{sub,J}}) = a_J\}}{\hat{\pi}_{J,C_J}(\mathbf{H}_J)} \right] \hat{\mu}_{C_J}(\mathbf{H}_J) \right)$$

where $\hat{\mu}_{C_J}(\mathbf{H}_J) = E\{Y|A_J = g_J(\mathbf{H}_{\text{sub,J}}) = a_J, \mathbf{H}_J\}$. Likewise, we define the purity measure for the J-th stage decision rule $g_J(\mathbf{H}_{\text{sub}})$ under a binary split ω (and ω^C) of node Ω as:

$$\mathcal{P}_J(\Omega, \omega) =$$

$$\max_{a_1, a_2 \in \mathcal{A}_J} \ \mathbb{P}_n \left[\sum_{a_J=1}^{K_J} \hat{\mu}^{\text{AIPW}}_{J,a_J}(\mathbf{H}_J) \ \mathcal{I}\{A_J = g_{J,\omega,a_1,a_2}(\mathbf{H}_{\text{sub,J}}) = a_J\} \ \mathcal{I}(\mathbf{H}_{\text{sub,J}} \in \Omega) \right]$$

Having completed estimation of the J-th stage, we generalize estimation now for the j-th stage, each estimated in backward sequence for $j = J-1, ..., 1$. Our goal in a multi-stage setting is to estimate a DTR such that the expected long-term counterfactual outcome is optimized. When estimating the decision rule for the j-th treatment stage, we must also account for the fact that the patient was treated with the optimal treatment at all future stages. Therefore, when performing estimation for any stage prior to the last, it is necessary to have a stage-specific pseudo-outcome $\tilde{Y}_j$ that represents the predicted counterfactual outcome at the j-th stage contingent upon the patient receiving the optimal treatments at all future stages, $j+1, ..., J$. Mathematically this can be expressed as: $\tilde{Y}_j = \hat{E}\{Y^*(A_1, ..., A_j, g^{\text{opt}}_{j+1}, ..., g^{\text{opt}}_J)\}$, as well as in recursive form, $\tilde{Y}_j = \hat{E}\{\tilde{Y}_{j+1}|A_{j+1} = g^{\text{opt}}_{j+1}(\mathbf{H}_{\text{sub,j+1}}), \mathbf{H}_{\text{sub,j+1}}\}$. Denoting $\hat{E}(\tilde{Y}_j|A_j = a_j, \mathbf{H}_{\text{sub,j}})$ as $\tilde{\mu}_{j,a_j}(\mathbf{H}_{\text{sub,j}})$, we can express the j-th stage pseudo-outcome as $\tilde{Y}_j = \tilde{\mu}_{j+1,g^{\text{opt}}_{j+1}}(\mathbf{H}_{\text{sub,j+1}})$. Similar to the delineation above, under the assumptions

of consistency, positivity, and NUCA, we express the optimal decision rule at the j-th stage as a function of the counterfactual pseudo-outcome as follows:

$$g_j^{\text{opt}}(\mathbf{H}_{\text{sub}}) = \text{argmax}_{g_j \in \mathcal{G}_{\text{sub},j}} E[\tilde{Y}_j\{g(\mathbf{H}_{\text{sub}})\}] = \\ \text{argmax}_{g_j \in \mathcal{G}_{\text{sub},j}} E_{\mathbf{H}_j}\left[\sum_{a_j=1}^{K_j} \tilde{\mu}_{j+1,a_{j+1}}(\mathbf{H}_{\text{sub,j}})\mathcal{I}\{A_j = g_j(\mathbf{H}_{\text{sub,j}}) = a_j\}\right]$$

Defining $\tilde{\mu}_{j,a_j}^{\text{AIPW}}(\mathbf{H}_j)$ as:

$$\tilde{\mu}_{j,a_j}^{\text{AIPW}}(\mathbf{H}_j) = \frac{\mathcal{I}(A_j = a_j)}{\hat{\pi}_{j,a_j}(\mathbf{H}_j)}\tilde{Y}_j + \left\{1 - \frac{\mathcal{I}(A_j = a_j)}{\hat{\pi}_{j,a_j}(\mathbf{H}_j)}\right\}\tilde{\mu}_{j,a_j}(\mathbf{H}_j)$$

where $\tilde{\mu}_{j,a_j}(\mathbf{H}_j)$ is $E(\tilde{Y}_j|A_j = a_j, \mathbf{H}_j)$, then the ReST-L purity measure used at the j-th treatment stage is:

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

Implemention of ReST-L is described in the following section.

2.3.3 Implementation

ReST-L is implemented backward recursively, beginning with estimation of the J-th stage. At each stage, several user-defined inputs are needed to implement ReST-L. First, it is necessary to specify a positive value, λ_j, which is used to determine whether or not a binary split of a node identifies a meaningful difference in purity. For example, if $P(\Omega_m)$ represents the purity of node Ω_m in the absence of a binary split and $P(\Omega_m, \omega)$ represents the "new" purity of node Ω_m under a split defined by partition ω (and its complement ω^C), we would expect a split to occur only if a meaningful improvement in purity is achieved, i.e., $P(\Omega_m, \omega) - P(\Omega) > \lambda_j$. Here λ_j

may be selected based on practical or clinical considerations or using cross-validation, as explained in *Tao et al.* (2018). Additional user-defined inputs include the desired minimum size of terminal nodes and the maximum depth of the tree to be estimated, both of which may also vary by stage j. The minimum node size, $n_{0,j}$, reflects the minumum number of observations that can fall into each of the leaf nodes once a split of a parent node is made. The depth of the tree (d_j) refers to the number of times recursive splitting of the root node may occur. A smaller minimum node size and larger tree depth result in more complex tree structures with a possible concern of overfitting whereas the converse may result in underfitting. There is an abundance of literature related to selecting optimal tuning parameters for decision-tree type estimation (e.g., *Boehmke and Greenwell*, 2020; *Hastie et al.*, 2009); in general the choices depend on the desired complexity of the resulting estimated stage j decision rule and often are chosen adaptively from the data. *Mantovani et al.* (2019) suggest that optimal minimum node size for a CART type estimation ranges from 1-20 and a depth of 5 is often a good starting point (*Boehmke and Greenwell*, 2020). Another strategy frequently employed is to grow a large tree and then prune it as needed using a cost metric (*Boehmke and Greenwell*, 2020; *Hastie et al.*, 2009; *Therneau et al.*, 2019), for example.

We briefly summarize the set of criteria for recursive partitioning of the covariate space for each stage $j = J, J-1, ..., 1$. Refer to *Tao et al.* (2018) for additional details. Inputs into the algorithm at the j-th stage include the purity measure $P_j(\Omega_m, \omega)$; the counterfactual pseudo-outcomes calculated via $\hat{\mu}^{\text{AIPW}}_{J,a_J}(\mathbf{H}_J)$ and $\tilde{\mu}^{\text{AIPW}}_{j,a_j}(\mathbf{H}_j)$ for the J-th or j-th stages, respectively; the minimum cut-off level for improvement in purity λ_j; the minimum terminal node $n_{0,j}$; and the maximum tree depth d_j. Beginning with the root node at the j-th stage, a series of recursive, binary splits of the covariate space $\mathbf{H}_{\text{sub},j}$ are made at the level of each node denoted as Ω_m, where the split is identified by ω, if the following criteria are met:

1. The node Ω_m resides at a shallower depth than the maximum, pre-specified tree depth d_j.

2. There are at least $2n_{0,j}$ observations in the node Ω_m and at least $n_{0,j}$ observations in each resulting child node.

3. $P(\Omega_m, \omega) - P(\Omega_m) > \lambda_j$, where $P(\Omega_m)$ refers to the purity in the absence of a split.

If these criteria are met, we compute the estimated optimal split $\hat{\omega}^{\text{opt}} = \text{argmax}_\omega\{P_j(\Omega, \omega)\}$. Recursive partitioning continues for the j-th stage across each node until at least one of the criteria is not met, at which point the node becomes a terminal node. Once all nodes within the j-th stage estimation become terminal, estimation for the j-th stage ends. The optimal j-stage decision rule is then determined by the partition of the covariate space at the j-th stage, with each partition being assigned the optimal treatment that maximizes the mean counterfactual (pseudo)-outcome. Estimation continues backward through all stages from the final stage J to stage 1.

2.4 Simulation Studies

2.4.1 Two-Stage Simulation to Evaluate the Bias of a Naive Implementation of T-RL

Given that the prescriptive variables can be reduced to a smaller set of variables, i.e., $\mathbf{H}_{\text{sub}}$, one may be tempted to input only those variables in $\mathbf{H}_{\text{sub}}$ into the T-RL algorithm. We refer to this method as "Naive T-RL". Assuming a two-stage DTR with three treatment options per stage and a sample size of $n = 1000$, we generate observations under varying conditions, including different levels of covariate correlations (ρ), number of variables in the full covariate history $\mathbf{H}$ and the subset

$\mathbf{H}_{\text{sub}}$ ($|\mathbf{H}|$ and $|\mathbf{H}_{\text{sub}}|$, respectively), and with both underlying tree-type and nontree-type DTRs. Data is generated assuming independent observations. Covariate data with dimension n x $|\mathbf{H}|$ are generated using the multivariate normal distribution with a mean of $\mathbf{0}_{|\mathbf{H}|}$ and an autoregressive (AR1) correlation structure with specified ρ, but with the following modifications: Pairwise correlation between the first variable in $\mathbf{H}_{\text{sub}}$ and the first three variables in $\mathbf{H}^C_{\text{sub}}$ is equal to ρ and pairwise correlations between the first three variables in $\mathbf{H}_{\text{sub}} = 0$. This modification was intended to reflect quasi-real world complexities among covariates but specifying a moderate or high degree of correlation between the variables involved in the optimal DTR and confounding variables. An additional covariate, $Z \in \mathbf{H}^C_{\text{sub}}$, was generated for each observation using a Bernoulli distribution with $p = 0.4$. The actual treatment received A_1 is randomly generated from the multinomial distribution with probabilities $\pi_{10}, \pi_{11}, \pi_{12}$ where $\pi_{10} = 1 - \pi_{11} - \pi_{12}$, $\pi_{11} = \exp(0.5X_{C1} - 0.5X_1 + Z - 0.5)/[1 + \exp(0.5X_{C1} - 0.5X_1 + Z - 0.5) + \exp(0.5X_{C2} + 0.5X_1 - Z)]$ and $\pi_{12} = \exp(0.5X_{C2}+0.5X_1-Z)/[1+\exp(0.5X_{C1}-0.5X_1+Z-0.5)+\exp(0.5X_{C2}+0.5X_1-Z)]$, where X_{C1}, X_{C2} represent the first two covariates in $\mathbf{H}^C_{\text{sub}}$, i.e., confounding variables not considered as candidate tailoring variables. The intermediate outcome following the first stage is defined as $Y_1 = \exp\{1.5+0.3X_{C1}-1.5Z-|1.5X_1-2|\cdot(A-g_1^{\text{opt}})^2\}+\epsilon$, where $\epsilon \sim N(0,1)$. This reflects an unequal penalty dependent on the value of X_1–a variable used in the true optimal treatment regime–if the patient was not treated according to their optimal therapy, which is intended to add an additional degree of complexity into the data generating scenario and reflective of a real world setting. The optimal tree-type decision rule for the first stage is as follows: If $X_1 > -1$ & $X_2 > 0.25$, then $g_1^{\text{opt}} = 2$; if $X_1 > -1$ & $-0.5 < X_2 \leq 0.25$, then $g_1^{\text{opt}} = 1$; otherwise, $g_1^{\text{opt}} = 0$. The optimal first stage, nontree-type decision rule is specified as: $g_1^{\text{opt}} = \mathcal{I}[\{\log_2(|X_1|+1) \leq 2\}$ & $(X_2 < -0.25)] + \mathcal{I}(X_2^2 > 0.35)$. Data for the treatment assignment for the second stage, A_2, are generated randomly also using the multi-

nomial distribution, but with probabilities $\pi_{20}, \pi_{21}, \pi_{22}$, where $\pi_{20} = 1 - \pi_{21} - \pi_{22}$, $\pi_{21} = \exp(0.2Y_1 + 0.5 - Z)/[1.5 + \exp(0.2Y_1 + 0.5 - Z) + \exp(0.5X_{C2} + Z)]$ and $\pi_{12} = \exp(0.5X_{C2} + Z)/[1.5 + \exp(0.2Y_1 + 0.5 - Z) + \exp(0.5X_{C2} + Z)]$. We define the intermediate outcome following the second stage as $Y_2 = \exp\{1.18 + 0.2X_{C2} - 2Z - |1.5X_3 + 2| \cdot (A - g_2^{\text{opt}})^2\} + \epsilon$, and the final outcome Y is defined as the sum of the stage-specific intermediate outcomes, i.e., $Y = Y_1 + Y_2$. The true, second-stage tree-type optimal decision rule is defined as follows: If $Y_1 > 0.5$ & $X_3 > 0$, then $g_2^{\text{opt}} = 2$; if $Y_1 > 0.5$ & $-1 < X_3 \leq 0$, then $g_2^{\text{opt}} = 1$; otherwise, $g_2^{\text{opt}} = 0$. The nontree-type optimal decision rule for the second stage is defined as: $g_2^{\text{opt}} = \mathcal{I}\{(|X_3| > 0.6)$ & $(Y_1 > 1)\} + \mathcal{I}(Y_1^2 > 3)$. In summary, we assume that only variables in $\mathbf{H}_{\text{sub}}$ may be included in an estimated optimal DTR, but that variables from either $\mathbf{H}_{\text{sub}}$ or $\mathbf{H}_{\text{sub}}^C$ may define the intermediate outcomes and the treatment assignment mechanisms. Under optimal treatment allocation $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 5.4$. For the analysis we assume that there is an additive linear relationship between the outcome Y conditional on covariate and treatment history; for ReST-L the assumed model includes all observed covariates in $\mathbf{H}$ and a treatment interaction with the subset of candidate tailoring variables in $\mathbf{H}_{\text{sub}}$ whereas for Naive T-RL the assumed model includes only those observed covariates in $\mathbf{H}_{\text{sub}}$ with a corresponding treatment interaction. We further assume that the propensity models used in Naive T-RL and ReST-L are correctly specified.

Results for this simulation study are presented in Table 2.1 for an underlying tree-type DTR and in Table 2.2 for a nontree-type DTR. It can easily be seen that, under all data generation settings, Naive T-RL will generate a substantial bias in its estimate of the counterfactual mean outcome and a substantially lower percentage of observations correctly classified to their optimal two-stage treatment regime than ReST-L. For a tree-based DTR with 20 covariates and a correlation of $\rho = 0.2$, for example, we observe a relative bias in estimation of the optimal counterfactual mean

outcome of 15.6% for Naive T-RL compared with 5.5% for ReST-L. The corresponding percentage of observations in the test set ($N_{\text{test}} = 1000$) that were correctly classified to their optimal treatment assignment for Naive T-RL and ReST-L are 57.5% and 85.3%, respectively. Refer to Section 2.4.4.4 for additional simulation results reflecting this data generation setting.

Table 2.1: Bias and performance of Naive T-RL and ReST-L in estimating an optimal two-stage dynamic treatment regime (DTR) with three treatments per stage, a sample size of $n = 1000$, and a high degree of confounding with an underlying tree-type structure. Medians (and IQRs) of $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ and % *opt*, as well as absolute and relative bias, are presented. $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; Naive T-RL = Naive Tree-based Reinforcement Learning; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment; IQR = interquartile range; Abs = Absolute Bias; Rel % = relative percent bias; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a DTR estimated using the applicable method. Under optimal treatment allocation $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 5.4$

$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}\text{sub}\rvert$	ρ	Naive T-RL			ReST-L		
		$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)	Abs Bias (Rel %)	%opt (IQR)	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)	Abs Bias (Rel %)	%opt (IQR)
Tree-based DTR							
20/7	0	4.542 (0.588)	0.869 (16.1)	55.0 (20.8)	5.133 (0.269)	0.278 (5.1)	86.5 (10.6)
20/7	0.2	4.569 (0.552)	0.842 (15.6)	57.5 (20.4)	5.114 (0.256)	0.297 (5.5)	85.3 (10.2)
20/7	0.6	4.569 (0.490)	0.842 (15.5)	55.5 (20.1)	5.123 (0.290)	0.288 (5.3)	85.4 (10.5)
50/10	0	4.491 (0.525)	0.920 (17.0)	54.1 (18.6)	5.114 (0.251)	0.297 (5.5)	85.4 (11.4)
50/10	0.2	4.589 (0.523)	0.822 (15.2)	57.2 (21.6)	5.106 (0.265)	0.305 (5.6)	85.2 (12.0)
50/10	0.6	4.525 (0.538)	0.886 (16.4)	54.3 (19.9)	5.093 (0.292)	0.318 (5.9)	84.4 (12.1)
50/35	0	4.489 (0.558)	0.922 (17.0)	53.8 (16.9)	5.056 (0.290)	0.355 (6.6)	82.3 (13.6)
50/35	0.2	4.485 (0.553)	0.926 (17.1)	54.0 (19.0)	5.049 (0.298)	0.362 (6.7)	82.3 (13.2)
50/35	0.6	4.483 (0.541)	0.928 (17.2)	52.3 (20.2)	5.049 (0.267)	0.362 (6.7)	81.1 (11.9)
100/20	0	4.475 (0.587)	0.936 (17.3)	53.4 (19.0)	5.030 (0.319)	0.381 (7.0)	82.0 (13.5)
100/20	0.2	4.560 (0.543)	0.851 (15.7)	56.2 (19.7)	5.030 (0.306)	0.381 (7.0)	82.7 (11.6)
100/20	0.6	4.519 (0.510)	0.892 (16.5)	54.9 (20.1)	5.030 (0.328)	0.381 (7.0)	81.6 (12.9)

Table 2.2: Bias and performance of Naive T-RL and ReST-L in estimating an optimal two-stage dynamic treatment regime (DTR) with three treatments per stage, a sample size $n = 1000$, and a high degree of confounding with an underlying nontree-type structure. Medians (and IQRs) of $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ and % *opt*, as well as absolute and relative bias, are presented. $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; Naive T-RL = Naive Tree-based Reinforcement Learning; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment; IQR = interquartile range; Abs = Absolute Bias; Rel % = relative percent bias; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a DTR estimated using the applicable method. Under optimal treatment allocation $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 5.4$

$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}sub\rvert$	ρ	Naive T-RL			ReST-L		
		$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)	Abs Bias (Rel %)	%opt (IQR)	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)	Abs Bias (Rel %)	%opt (IQR)
nontree-type DTR							
20/7	0	4.797 (0.763)	0.614 (11.3)	66.3 (26.3)	5.132 (0.373)	0.279 (5.2)	82.8 (17.2)
20/7	0.2	4.656 (0.701)	0.755 (14.0)	64.2 (25.7)	5.107 (0.350)	0.304 (5.6)	82.3 (15.9)
20/7	0.6	4.612 (0.666)	0.799 (14.8)	63.1 (19.1)	5.118 (0.364)	0.293 (5.4)	81.4 (15.2)
50/10	0	4.740 (0.737)	0.671 (12.4)	65.7 (25.1)	5.092 (0.428)	0.319 (5.9)	81.6 (20.4)
50/10	0.2	4.646 (0.678)	0.765 (14.1)	64.5 (24.0)	5.118 (0.407)	0.293 (5.4)	82.7 (18.4)
50/10	0.6	4.506 (0.593)	0.905 (16.7)	60.4 (16.2)	5.047 (0.461)	0.364 (6.7)	80.0 (21.6)
50/35	0	4.708 (0.715)	0.703 (13.0)	64.0 (25.1)	5.027 (0.520)	0.384 (7.1)	77.6 (20.5)
50/35	0.2	4.659 (0.715)	0.752 (13.9)	64.3 (24.4)	5.004 (0.466)	0.407 (7.5)	79.4 (19.4)
50/35	0.6	4.465 (0.482)	0.946 (17.5)	58.0 (14.6)	4.971 (0.564)	0.440 (8.1)	76.6 (22.6)
100/20	0	4.706 (0.712)	0.705 (13.0)	65.4 (24.3)	5.002 (0.575)	0.409 (7.6)	78.6 (22.7)
100/20	0.2	4.594 (0.663)	0.817 (15.1)	63.3 (21.6)	5.013 (0.537)	0398 (7.4)	79.0 (22.4)
100/20	0.6	4.479 (0.540)	0.932 (17.2)	59.4 (16.3)	4.988 (0.507)	0.423 (7.8)	76.7 (21.3)

2.4.2 Single Stage Simulation to Evaluate Relative Performance of ReST-L

Next we evaluate the relative performance of ReST-L in a single stage setting with three treatment options. Parameters varied across this simulation study include the sample size (n), the number of covariates in $\mathbf{H}$ and $\mathbf{H}_{\text{sub}}$ (H/Hsub), the correlation ρ used to generate the correlation matrix for covariates in $\mathbf{H}$, and the true, underlying structure of the decision rule (i.e., tree- or nontree-type). Data is generated assuming independent observations. Covariate data $\mathbf{X}$ with dimension n x $|\mathbf{H}|$ are generated using the multivariate normal distribution with a mean of $\mathbf{0}_{|\mathbf{H}|}$ and an autoregressive (AR1) correlation structure with specified ρ, but with the following modifications: Pairwise correlation between the first variable in $\mathbf{H}_{\text{sub}}$ and the first three variables in $\mathbf{H}^C_{\text{sub}}$ is equal to ρ and pairwise correlations between the first three variables in $\mathbf{H}_{\text{sub}} = 0$. This mimics the correlation structure used for the simulation in Section 2.4.1. [Supplemental simulations using a simple exchangeable correlation structure with $\rho = 0.2$ (results not shown) revealed similar results to those presented herein.] The actual treatment received, A, is randomly generated from the multinomial distribution with probabilities π_0, π_1, π_2 where $\pi_0 = 1 - \pi_1 - \pi_2$, $\pi_1 = \exp(0.5X_{C1} + 0.5X_1)/[1 + \exp(0.5X_{C1} + 0.5X_1) + \exp(0.5X_{C2} - 0.5X_1)]$ and $\pi_2 = \exp(0.5X_{C2} - 0.5X_1)/[1 + \exp(0.5X_{C1} + 0.5X_1) + \exp(0.5X_{C2} - 0.5X_1)]$, where X_{C1}, X_{C2} represent the first two covariates in $\mathbf{H}^C_{\text{sub}}$, i.e., confounding variables not considered as candidate tailoring variables. The outcome $Y = \exp\{1.5 + 0.3X_{C1} - |1.5X_1 - 2| \cdot (A - g^{\text{opt}})^2\} + \epsilon$, where $\epsilon \sim N(0, 1)$. The true, underlying tree-type decision rule is defined as follows: If $X_1 > -1$ & $X_2 > 0.5$, then $g^{\text{opt}} = 2$; if $X_1 > -1$ & $-0.5 < X_2 < 0.5$, then $g^{\text{opt}} = 1$; otherwise, $g^{\text{opt}} = 0$. The nontree-type decision rule is defined as: $g^{\text{opt}} = \mathcal{I}\{\log_2(|X_1| + 1) \leq 2 \;\&\; X_2 < 0.25\} + \mathcal{I}\{X_2^2 \leq 0.5\}$. Importantly, the outcome and actual treatment assignment are defined using variables in both $\mathbf{H}_{\text{sub}}$ and $\mathbf{H}^C_{\text{sub}}$. The optimal decision rule, based on the methodologic assumptions of ReST-L, includes

only variables in $\mathbf{H}_{\text{sub}}$. Under optimal treatment allocation $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 4.7$.

We compare the estimated performance of ReST-L with five competing methods: tree-based reinforcement learning (T-RL), standard Q-Learning using linear modeling (Q-L), restricted linear Q-Learning (Q-L-R), Q-Learning using nonparametric modeling (Q-NP), and restricted nonparametric Q-Learning (Q-NP-R). With the exception of T-RL, which represents the unrestricted counterpart to ReST-L, we restrict our comparisons to Q-Learning methods because these are the only existing methods to our knowledge that can accommodate a subset of variables in the estimated treatment regime. For both ReST-L and T-RL we assume that there is an additive linear relationship between the outcome Y conditional on covariate and treatment history that includes all observed covariates, as well as a treatment-interaction with either all observed covariates (T-RL) or with a subset of candidate tailoring variables (ReST-L). We further assume that the propensity model used in ReST-L and T-RL is correcly specified. (Performance results under an incorrectly-specified propensity model are presented for a two-stage simulation in Section 2.4.4.1.) Restricted Q-Learning methods are modifications of the standard Q-Learning models such that only variables in $\mathbf{H}_{\text{sub}}$ are considered as possible candidate tailoring variables (i.e., treatment interactions), which differs from standard Q-Learning in which all variables in $\mathbf{H}$ are possible treatment tailoring variables. Linear Q-Learning assumes a linear relationship between the covariates and the outcome. Nonparametric Q-Learning methods allow a more flexible relationship for the Q-functions, estimated using random forests. Performance is evaluated using two metrics. First, we estimate the optimal treatment regime using data from the training set with sample size n and use a test set ($N_{\text{test}} = 1000$) to determine the percentage of observations correctly classified to their optimal treatment, $\%opt$. Second, using the test set, we estimate $E[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$, the expected counterfactual outcome had everyone in the patient population been treated optimally based on the estimated regime. For each design setting, we tabu-

late the median and interquartile range (IQR) of $\%opt$ and $\hat{E}[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ across all $B = 500$ Monte Carlo iterations.

Estimated performance for tree-type and nontree-type decision rules are displayed in Table 2.3 and Table 2.4, respectively. As observed in the tabulated results, ReST-L selects the optimal treatment decision rule well when the underlying decision rule is either tree-type or nontree-type. Across all data generating settings, for a tree-type decision rule the percent of observations from the test set that are correctly classified to their optimal treatment ranges from about 85% for smaller sample sizes and fewer variables in $\mathbf{H}$ and $\mathbf{H}_{\text{sub}}$ to more than 95% for larger sample sizes (and a correspondingly larger number of variables in $\mathbf{H}$). ReST-L performance improves as the sample size increases, as expected; for example refer to results for sample sizes of $n = 500$ and $n = 750$ when $|\mathbf{H}| = 100$. For the same sample size and number of variables in the covariate history $\mathbf{H}$, performance improves as the proportion of variables in $\mathbf{H}_{\text{sub}}$ relative to $\mathbf{H}$ decreases. For example, for 50 covariates in $\mathbf{H}$, a sample size of $n = 300$ and a correlation $\rho = 0.2$, estimated performance improves from 86.5% to 90.0% correct classification when the number of variables in $\mathbf{H}_{\text{sub}}$ decreases from 35 to 10. ReST-L performance in estimating the optimal decision rule is similar across different degrees of correlation among covariates when all other parameters are held constant. Finally, we observe that the variability for ReST-L in estimating the optimal regime increases when the true, underlying decision rule is nontree-type, and is generally higher with a smaller sample size or as the proportion of variables in $\mathbf{H}_{\text{sub}}$ increases. Furthermore, ReST-L estimates the empirical counterfactual mean outcome under the optimal treatment regime, $\hat{E}[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$, with a high degree of accuracy and relatively low variability across Monte Carlo iterations, particularly as the sample size increases. Assuming a tree-type decision rule, $|\mathbf{H}| = 100$, and $n = 750$, ReST-L estimates the counterfactual mean under the estimated optimal treatment assignment to be 4.6, which is very close to the true empirical counterfactual mean

of 4.7.

ReST-L consistently performs better than all other methods across all one-stage data generating settings presented – for both tree- and nontree-type decision rules. For estimation with an underlying tree-type decision rule, the variability of ReST-L in estimating the optimal regime is smaller than that of T-RL and similar to restricted nonparametric Q-Learning. For a nontree-type decision rule, variability of ReST-L in estimating $\%opt$ is larger than that of restricted nonparametric Q-Learning, but generally remains smaller than that of T-RL overall. Across all simulation settings, restricted Q-Learning methods perform better than their standard Q-Learning counterparts. Both linear Q-learning methods (restricted and unrestricted) perform poorly in all scenarios whether the underlying decision rule is tree-type or nontree-type. With a larger sample size, restricted nonparametric Q-Learning does a reasonable job of estimating the optimal treatment regime for an underlying tree- and nontree-type decision rule; with $n = 500$ and 100/20 variables in $\mathbf{H}/\mathbf{H}_{\text{sub}}$, for example, restricted nonparametric Q-Learning achieves higher than 85% correct classification.

Table 2.3: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal one-stage decision rule with 3 possible treatments based on an underlying, tree-type decision rule. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment rule estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 4.7$

n	H/Hsub	rho	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
			% opt (IQR)					
750	100/20	0.2	97.0 (4.4)	94.6 (5.9)	62.3 (2.3)	56.4 (2.3)	93.8 (3.9)	84.4 (7.0)
750	100/20	0.6	96.6 (4.5)	94.5 (5.3)	62.3 (2.2)	56.4 (2.9)	92.4 (3.9)	82.8 (5.9)
500	100/20	0.2	93.6 (5.8)	84.8 (13.0)	61.4 (2.4)	52.8 (3.1)	88.7 (5.4)	75.8 (7.4)
500	100/20	0.6	93.3 (5.7)	86.2 (9.6)	61.1 (2.3)	52.6 (3.2)	86.5 (5.4)	72.6 (7.1)
300	50/35	0.2	86.5 (11.1)	81.6 (15.6)	56.8 (3.2)	54.2 (3.5)	72.8 (9.0)	66.0 (9.7)
300	50/35	0.6	85.9 (10.5)	82.7 (13.5)	56.9 (3.2)	54.3 (3.6)	70.9 (7.5)	63.4 (8.0)
300	50/10	0.2	90.0 (6.9)	82.4 (14.2)	61.8 (2.4)	54.4 (3.2)	83.4 (6.8)	66.1 (8.9)
300	50/10	0.6	90.4 (7.3)	82.5 (14.1)	61.6 (2.4)	54.3 (3.3)	80.6 (6.9)	62.9 (8.4)
200	20/7	0.2	86.4 (8.5)	83.8 (12.6)	61.9 (2.7)	58.2 (3.4)	78.5 (7.9)	64.3 (9.7)
200	20/7	0.6	86.3 (6.8)	84.4 (10.7)	61.7 (2.7)	58.1 (3.5)	76.2 (7.8)	62.5 (8.0)
			$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ (IQR)					
750	100/20	0.2	4.6 (0.2)	4.5 (0.2)	3.4 (0.1)	3.1 (0.1)	4.5 (0.1)	4.2 (0.3)
750	100/20	0.6	4.6 (0.2)	4.5 (0.2)	3.4 (0.1)	3.1 (0.1)	4.5 (0.1)	4.1 (0.2)
500	100/20	0.2	4.5 (0.2)	4.1 (0.4)	3.3 (0.1)	2.9 (0.1)	4.3 (0.2)	3.8 (0.3)
500	100/20	0.6	4.5 (0.2)	4.2 (0.4)	3.3 (0.1)	2.9 (0.2)	4.2 (0.2)	3.7 (0.3)
300	50/35	0.2	4.2 (0.4)	4.0 (0.6)	3.1 (0.2)	3.0 (0.2)	3.7 (0.3)	3.5 (0.4)
300	50/35	0.6	4.2 (0.4)	4.1 (0.5)	3.1 (0.1)	3.0 (0.2)	3.7 (0.3)	3.4 (0.3)
300	50/10	0.2	4.3 (0.3)	4.0 (0.5)	3.3 (0.1)	3.0 (0.2)	4.1 (0.2)	3.5 (0.3)
300	50/10	0.6	4.3 (0.3)	4.1 (0.5)	3.3 (0.1)	3.0 (0.2)	4.0 (0.2)	3.4 (0.3)
200	20/7	0.2	4.2 (0.3)	4.0 (0.4)	3.3 (0.1)	3.1 (0.2)	4.0 (0.3)	3.4 (0.4)
200	20/7	0.6	4.2 (0.3)	4.1 (0.4)	3.3 (0.1)	3.2 (0.2)	3.9 (0.3)	3.4 (0.3)

Table 2.4: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal one-stage decision rule with 3 possible treatments based on an underlying, nontree-type decision rule. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment rule estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 4.7$

n	H/Hsub	rho	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
			% opt (IQR)					
750	100/20	0.2	98.6 (2.1)	97.8 (6.7)	50.2 (3.9)	40.2 (2.8)	93.0 (3.6)	85.0 (8.9)
750	100/20	0.6	98.4 (2.3)	97.7 (8.5)	50.3 (3.9)	40.4 (2.9)	91.5 (4.1)	77.2 (10.1)
500	100/20	0.2	95.8 (14.1)	78.6 (25.0)	47.3 (4.5)	37.4 (3.1)	87.1 (5.5)	70.5 (11.1)
500	100/20	0.6	94.4 (18.9)	68.8 (24.8)	47.6 (5.0)	37.7 (3.1)	85.5 (5.4)	66.7 (10.4)
300	50/35	0.2	83.9 (29.9)	78.0 (26.5)	40.8 (4.3)	38.7 (4.0)	66.7 (8.8)	61.3 (9.8)
300	50/35	0.6	83.4 (29.7)	77.8 (25.7)	41.2 (4.1)	39.0 (3.7)	67.9 (8.0)	58.6 (8.6)
300	50/10	0.2	94.6 (16.9)	78.6 (26.6)	49.0 (5.6)	38.9 (3.6)	79.6 (6.9)	60.7 (9.3)
300	50/10	0.6	94.0 (16.7)	76.9 (25.9)	49.0 (5.4)	38.9 (3.6)	78.2 (7.0)	58.0 (8.9)
200	20/7	0.2	93.1 (31.6)	82.9 (31.7)	48.9 (7.0)	42.0 (4.7)	74.2 (8.3)	60.5 (9.4)
200	20/7	0.6	92.2 (32.6)	82.2 (32.3)	49.0 (6.3)	42.4 (4.8)	72.8 (7.1)	56.5 (8.5)
			$\hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ (IQR)					
750	100/20	0.2	4.6 (0.1)	4.6 (0.2)	2.8 (0.2)	2.4 (0.1)	4.5 (0.1)	4.2 (0.3)
750	100/20	0.6	4.6 (0.1)	4.6 (0.3)	2.9 (0.2)	2.5 (0.1)	4.5 (0.1)	3.9 (0.4)
500	100/20	0.2	4.5 (0.5)	4.0 (0.9)	2.7 (0.2)	2.3 (0.1)	4.3 (0.2)	3.7 (0.4)
500	100/20	0.6	4.5 (0.6)	3.6 (0.8)	2.8 (0.2)	2.3 (0.2)	4.3 (0.2)	3.5 (0.4)
300	50/35	0.2	4.1 (1.0)	3.9 (1.0)	2.4 (0.2)	2.3 (0.2)	3.6 (0.3)	3.3 (0.4)
300	50/35	0.6	4.1 (1.0)	3.9 (0.9)	2.5 (0.2)	2.4 (0.2)	3.6 (0.3)	3.3 (0.3)
300	50/10	0.2	4.5 (0.6)	3.9 (1.0)	2.8 (0.2)	2.3 (0.2)	4.0 (0.2)	3.3 (0.3)
300	50/10	0.6	4.5 (0.6)	3.9 (0.9)	2.8 (0.3)	2.4 (0.2)	4.0 (0.2)	3.2 (0.3)
200	20/7	0.2	4.4 (1.1)	4.1 (1.1)	2.8 (0.3)	2.5 (0.2)	3.8 (0.3)	3.3 (0.4)
200	20/7	0.6	4.4 (1.1)	4.1 (1.1)	2.8 (0.3)	2.6 (0.2)	3.8 (0.3)	3.2 (0.4)

2.4.3 Two-Stage Simulation to Evaluate Relative Performance of ReST-L

We next evaluate the performance of ReST-L in a two-stage estimation setting with 3 possible treatment options per stage. It can easily be seen that random allocation of one of three treatments in each of two stages would select the optimal two-stage treatment assignment about 1 out of every 3^2 times, which is about 11% of the time. All settings for generating first stage data, including the covariate matrix $\mathbf{X}$, the treatment assignment mechanism for A_1, the intermediate outcome Y_1, and optimal treatment $g_1^{\text{opt}}(\mathbf{H}_{\text{sub}})$, are the same as those used in the single stage setting described above. The second stage treatment A_2 is randomly generated using the multinomial distribution with probabilities $\pi_{20}, \pi_{21}, \pi_{22}$ where $\pi_{20} = 1-\pi_{21}-\pi_{22}$, $\pi_{21} = \{\exp(0.2Y_1-0.5)\}/[1+\{\exp(0.2Y_1-0.5)\}+\{\exp(0.5X_{C2})\}]$, and $\pi_{22} = \{\exp(0.5X_{C2})\}/[1 + \{\exp(0.2Y_1 - 0.5)\} + \{\exp(0.5X_{C2})\}]$. The intermediate outcome is $Y_2 = \exp\{1.18 + 0.2X_{C2} - |1.5X_3 + 2| \cdot (A_2 - g_2^{\text{opt}})^2\} + \epsilon$, where $\epsilon \sim N(0,1)$, and the overall outcome $Y = Y_1 + Y_2$. When a tree-type DTR is assumed, $g_2^{\text{opt}}(\mathbf{H}_{\text{sub}})$ is assigned as follows: If $X_3 > -1$ & $Y_1 > 2$, then $g_2^{\text{opt}} = 2$; if $X_3 > -1$ & $0 < Y_1 \leq 2$, then $g_2^{\text{opt}} = 1$; otherwise, $g_2^{\text{opt}} = 0$. Under an assumed nontree-type DTR: $g_2^{\text{opt}} = \mathcal{I}(|X_3| > 0.6 \;\&\; Y_1 > 0.4) + \mathcal{I}(Y_1^2 > 2.5)$. Both the intermediate outcomes and actual treatment assignments depend on variables in both $\mathbf{H}_{\text{sub}}$ and $\mathbf{H}_{\text{sub}}^C$. However, the optimal DTR are set to include only variables in $\mathbf{H}_{\text{sub}}$. Under optimal treatment allocation and assuming independence across observations, $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 8.0$. Similar to the single stage setting, ReST-L and T-RL assume a correctly-specified propensity model and an incorrectly-specified conditional mean model, which is more likely to happen in practice.

The performance of ReST-L and other competing methods for estimating the optimal two-stage regime with either an underlying tree-type or nontree-type DTR are displayed in Tables 2.5 and 2.6, respectively. Across all sample size and variable settings with a tree-type DTR, ReST-L does a reasonably good job of selecting the

optimal treatment, with correct classification generally between 85-90%. As in a single stage estimation setting, performance improves with sample size, as expected, with an improvement in percent correct classification from 89.6% to 95.2% for sample sizes of $n = 600$ to $n = 1000$ ($\rho = 0.2$). Additionally, performance improves with fewer variables in $\mathbf{H}_{\text{sub}}$ relative to $\mathbf{H}$: The percent correct classification with 50 variables in $\mathbf{H}$ and $\rho = 0.2$ improves from 87.0% to 89.8% as the number of variables in $\mathbf{H}_{\text{sub}}$ is reduced from 35 to 10. With an underlying nontree-type DTR, larger sample sizes are needed to obtain a similar estimated correct classification rate. For example, with a sample size of $n = 600$, $\rho = 0.2$, and $|\mathbf{H}| = 100$ variables, the percent of observations correctly classified to their optimal treatment is just over 70% for the nontree-type DTR compared with nearly 90% for a tree-type DTR; however, with the same specifications but with $n = 1000$, the percent correct classification are similar for tree- and nontree-type DTRs (95.2% and 93.8%, respectively). Variability of estimation of the percent correct treatment allocation of ReST-L is lower for a tree-type DTR than for nontree-type DTR and is larger on average than that observed in a single stage setting. Finally, for ReST-L, the estimated counterfactual mean outcome is closer to the empirical mean when sample size increases; when $n = 1000$, ReST-L achieves an estimated counterfactual mean outcome of 7.8 compared to the empirical mean of 8.0.

For a two-stage, tree-type DTR, ReST-L improves more upon T-RL at lower sample sizes and when the proportion of variables in $\mathbf{H}_{\text{sub}}$ relative to $\mathbf{H}$ decreases. With larger sample sizes, e.g., when $n = 1000$, both ReST-L and T-RL achieve more than 90% correct treatment classification although ReST-L still slightly outperforms T-RL in this case. For a nontree-type DTR, ReST-L improves upon T-RL across all settings, although in particular the benefit of ReST-L is observed with a larger number of covariates. When $n = 1000$ and $\rho = 0.2$, for example, the percent of observations in the test set that were correctly classified to their optimal treatment

is 93.8% for ReST-L compared with 84.8% for T-RL. As in a single stage setting, we observe that the restricted versions of Q-Learning improve upon their unrestricted counterparts, although linear Q-Learning demonstrates poor performance across all settings, never exceeding more than 25% correct treatment classification. Restricted nonparametric Q-Learning, on the other hand, achieves good performance, nearing 90% correct classification for a tree-type DTR with a large sample size.

Table 2.5: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments per stage and based on an underlying, tree-type DTR. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 8.0$.

n	H/Hsub	rho	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
			% opt (IQR)					
1000	100/20	0.2	95.2 (5.5)	92.5 (7.3)	51.0 (2.1)	45.1 (2.6)	88.9 (4.8)	75.0 (8.6)
1000	100/20	0.6	93.8 (5.9)	91.8 (7.2)	50.9 (2.3)	44.8 (2.7)	86.9 (4.7)	73.1 (6.6)
600	100/20	0.2	89.6 (10.4)	77.4 (17.1)	49.4 (2.5)	40.0 (2.9)	79.4 (6.4)	60.1 (9.6)
600	100/20	0.6	89.1 (8.6)	76.8 (15.5)	49.6 (2.3)	39.9 (2.9)	76.2 (7.2)	57.6 (8.2)
500	50/35	0.2	87.0 (11.9)	83.1 (13.5)	46.9 (2.9)	44.8 (2.6)	69.2 (8.8)	60.5 (9.0)
500	50/35	0.6	85.3 (11.1)	82.8 (12.5)	47.1 (2.9)	45.1 (2.9)	67.0 (8.7)	57.6 (8.3)
500	50/10	0.2	89.8 (9.0)	83.5 (13.8)	50.8 (2.4)	44.8 (3.0)	79.8 (6.1)	60.9 (10.1)
500	50/10	0.6	89.2 (8.7)	82.0 (12.2)	50.8 (2.6)	44.7 (3.2)	76.2 (5.7)	57.9 (8.0)
350	20/7	0.2	86.0 (13.8)	80.2 (14.1)	51.0 (2.4)	48.4 (2.8)	75.0 (7.0)	59.6 (9.1)
350	20/7	0.6	85.9 (11.4)	81.6 (13.9)	50.9 (2.6)	48.1 (2.6)	72.0 (6.4)	57.5 (7.8)
			$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)					
1000	100/20	0.2	7.8 (0.2)	7.8 (0.3)	5.9 (0.1)	5.5 (0.2)	7.7 (0.2)	7.1 (0.4)
1000	100/20	0.6	7.8 (0.2)	7.7 (0.3)	5.9 (0.2)	5.6 (0.2)	7.6 (0.2)	7.1 (0.3)
600	100/20	0.2	7.7 (0.3)	7.2 (0.6)	5.8 (0.2)	5.2 (0.2)	7.3 (0.3)	6.4 (0.4
600	100/20	0.6	7.6 (0.3)	7.2 (0.5)	5.9 (0.2)	5.2 (0.2)	7.2 (0.3)	6.3 (0.4)
500	50/35	0.2	7.6 (0.4)	7.5 (0.4)	5.6 (0.2)	5.5 (0.2)	6.9 (0.4)	6.4 (0.4)
500	50/35	0.6	7.6 (0.3)	7.4 (0.4)	5.7 (0.2)	5.6 (0.2)	6.9 (0.3)	6.4 (0.4)
500	50/10	0.2	7.7 (0.3)	7.4 (0.4)	5.9 (0.1)	5.5 (0.2)	7.4 (0.2)	6.5 (0.4)
500	50/10	0.6	7.7 (0.2)	7.4 (0.4)	5.9 (0.2)	5.6 (0.2)	7.2 (0.2)	6.4 (0.4)
350	20/7	0.2	7.5 (0.4)	7.4 (0.5)	5.9 (0.2)	5.7 (0.2)	7.2 (0.3)	6.4 (0.4)
350	20/7	0.6	7.6 (0.3)	7.4 (0.4)	5.9 (0.2)	5.8 (0.2)	7.1 (0.3)	6.4 (0.4)

Table 2.6: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments possible per stage and based on an underlying, nontree-type DTR. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in H; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation , $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 8.0$.

n	H/Hsub	rho	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
			% opt (IQR)					
1000	100/20	0.2	94.0 (33.2)	85.0 (34.4)	24.3 (2.8)	18.3 (2.3)	86.7 (4.6)	69.8 (11.4)
1000	100/20	0.6	93.9 (31.0)	85.9 (32.4)	25.0 (2.6)	18.4 (2.0)	82.0 (5.4)	61.2 (9.3)
600	100/20	0.2	71.4 (32.3)	53.4 (26.4)	22.0 (3.1)	15.5 (2.4)	71.6 (7.4)	45.4 (10.9)
600	100/20	0.6	74.9 (30.7)	55.5 (21.4)	22.5 (3.3)	15.6 (2.3)	66.3 (8.1)	38.4 (10.5)
500	50/35	0.2	65.7 (36.0)	64.3 (34.7)	19.2 (2.8)	18.2 (2.7)	54.4 (9.8)	43.5 (10.9)
500	50/35	0.6	72.0 (28.6)	65.4 (26.0)	19.6 (2.8)	18.3 (2.7)	50.6 (10.3)	37.7 (10.3)
500	50/10	0.2	80.9 (31.4)	64.4 (33.2)	23.9 (3.4)	18.1 (2.8)	72.1 (7.6)	43.6 (10.5)
500	50/10	0.6	78.6 (30.3)	61.5 (27.6)	24.6 (3.3)	18.2 (2.7)	66.3 (7.7)	37.0 (9.4)
350	20/7	0.2	72.2 (31.6)	66.0 (33.6)	23.9 (4.1)	20.8 (3.7)	64.2 (8.1)	41.2 (9.7)
350	20/7	0.6	77.2 (28.3)	68.0 (30.1)	24.3 (4.2)	20.7 (3.6)	60.1 (8.3)	36.8 (8.7)
			$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)					
1000	100/20	0.2	7.8 (1.2)	7.6 (1.2)	4.9 (0.2)	4.4 (0.2)	7.7 (0.2)	7.0 (0.4)
1000	100/20	0.6	7.8 (1.0)	7.6 (1.2)	5.0 (0.2)	4.5 (0.1)	7.5 (0.2)	6.7 (0.4)
600	100/20	0.2	7.3 (1.2)	6.4 (1.0)	4.7 (0.2)	4.2 (0.2)	7.2 (0.3)	6.0 (0.5)
600	100/20	0.6	7.3 (1.2)	6.4 (1.0)	4.8 (0.2)	4.2 (0.2)	7.0 (0.3)	5.8 (0.5)
500	50/35	0.2	7.0 (1.2)	6.9 (1.2)	4.5 (0.2)	4.4 (0.2)	6.5 (0.4)	6.0 (0.5)
500	50/35	0.6	7.2 (1.1)	7.0 (1.1)	4.6 (0.2)	4.5 (0.2)	6.4 (0.4)	5.8 (0.5)
500	50/10	0.2	7.5 (1.2)	6.9 (1.1)	4.9 (0.2)	4.4 (0.2)	7.2 (0.3)	6.0 (0.5)
500	50/10	0.6	7.4 (1.2)	6.7 (1.1)	5.0 (0.2)	4.4 (0.2)	7.0 (0.3)	5.8 (0.4)
350	20/7	0.2	7.3 (1.2)	7.0 (1.1)	4.9 (0.3)	4.6 (0.2)	6.9 (0.3)	5.9 (0.4)
350	20/7	0.6	7.3 (1.1)	7.1 (1.1)	5.0 (0.3)	4.7 (0.2)	6.8 (0.3)	5.7 (0.4)

2.4.4 Supplemental Two-Stage Simulation Experiments

Supplemental simulation studies were conducted to evaluate the performance of ReST-L in a variety of other scenarios. Specifically, we apply ReST-L and T-RL using an incorrectly-specified propensity model (Tables 2.7 and 2.8); or modify the data generating mechanisms to remove confounding of the treatment assignments $\mathbf{A}$ and outcomes $\mathbf{Y}$ by variables in $\mathbf{H}^C_{\text{sub}}$ (Tables 2.9 and 2.10); or modify the data generating mechanisms for $\mathbf{g}$ such that variables defining the true optimal DTR may be in $\mathbf{H}^C_{\text{sub}}$ (Tables 2.11 and 2.12); or evaluate the relative performance of ReST-L compared with other methods under stronger confounding with a binary covariate $Z \in \mathbf{H}^C_{\text{sub}}$ (Tables 2.13 and 2.14), which mimics the data generation model from Section 2.4.1.

2.4.4.1 Simulation Studies to Evaluate the Relative Performance of ReST-L When Using Incorrectly-Specified Propensity Models

We conducted additional simulation experiments in order to demonstrate the sensitivity of our findings when utilizing ReST-L for estimating an optimal DTR when only a subset of the covariates may be considered as candidate tailoring variables. First, we evaluate ReST-L performance in estimation of a two-stage optimal DTR when the propensity model is incorrectly specified. The two-stage data generation specifications for this simulation experiment are the same as those introduced in Section 2.4.3. In contrast to the analyses presented in Tables 2.5 and 2.6 in which we assumed that the variables determining treatment were known, however, here we consider all variables in $\mathbf{H}$ as variables that may be used to define the treatment assignment mechanism at both stages. While we understand that the AIPW estimator is consistent and doubly robust in large samples when either or both the propensity model and the conditional mean model are correctly specified, we believe that this supplemental simulation study in which neither the propensity model nor the condi-

tional mean model are correctly specified will reflect a scenario that is likely to occur frequently in practice and will shed light on the use of ReST-L as an out-of-the-box solution for estimating optimal DTRs.

ReST-L and T-RL performance using incorrectly-specified propensity models is presented in Tables 2.7 and 2.8. Because Q-Learning methods do not rely on a propensity model, performance measures for Q-Learning methods are replicated from Tables 2.5 and 2.6 for ease of comparison with ReST-L and T-RL. For tree-type DTRs, performance of ReST-L is slightly lower overall when the propensity models are incorrectly specified compared with correct specification. For example, with a sample size of $n = 350$, a covariate correlation of $\rho = 0.2$, and $|\mathbf{H}| = 20$, the percent of the test set classified to the correct treatment is 84.8% when the propensity models are incorrectly specified and 86.0% when correctly specified. Similarly, when the sample size and the number of variables in $\mathbf{H}$ are large (i.e., $n = 1000$, $\rho = 0.2$, and $|\mathbf{H}| = 100$), the performance is 93.2% and 95.2% for incorrectly- and correctly- specified propensity models, respectively. The percentage of correctly-treated observations in the test set remains reasonable across all sample sizes and variable settings, hovering above 85% correct classification on average, and maintains an improvement over T-RL across all settings. With an underlying nontree-type DTR, larger sample sizes are necessary to achieve reasonable performance, as was also observed in Tables 2.5 and 2.6. While ReST-L is still likely to be favored when the assumptions of the method are fulfilled, i.e., that the true DTR is defined only in terms of a subset of variables $\mathbf{H}_{\text{sub}}$, the improvement of ReST-L over restricted nonparametric Q-Learning decreases as sample size increases. With $n = 1000$ and a correlation of $\rho = 0.2$, for example, the percent of observations with correct treatment classification is 86.7% for restricted nonparametric Q-Learning compared with 87.0% for ReST-L.

Table 2.7: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments per stage and based on an underlying, tree-type DTR and assuming incorrectly-specified propensity models. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 8.0$.

n	H/Hsub	rho	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
			% opt (IQR)					
1000	100/20	0.2	93.2 (7.0)	90.6 (8.8)	51.0 (2.1)	45.1 (2.6)	88.9 (4.8)	75.0 (8.6)
1000	100/20	0.6	92.4 (6.8)	89.7 (9.6)	50.9 (2.3)	44.8 (2.7)	86.9 (4.7)	73.1 (6.6)
600	100/20	0.2	79.8 (19.3)	67.9 (19.4)	49.4 (2.5)	40.0 (2.9)	79.4 (6.4)	60.1 (9.6)
600	100/20	0.6	81.2 (19.2)	68.7 (18.8)	49.6 (2.3)	39.9 (2.9)	76.2 (7.2)	57.6 (8.2)
500	50/35	0.2	82.0 (15.0)	79.9 (16.1)	46.9 (2.9)	44.8 (2.6)	69.2 (8.8)	60.5 (9.0)
500	50/35	0.6	82.3 (12.5)	78.8 (14.7)	47.1 (2.9)	45.1 (2.9)	67.0 (8.7)	57.6 (8.3)
500	50/10	0.2	86.5 (13.1)	78.2 (16.3)	50.8 (2.4)	44.8 (3.0)	79.8 (6.1)	60.9 (10.1)
500	50/10	0.6	86.5 (12.3)	78.0 (15.0)	50.8 (2.6)	44.7 (3.2)	76.2 (5.7)	57.9 (8.0)
350	20/7	0.2	84.8 (15.2)	79.2 (17.5)	51.0 (2.4)	48.4 (2.8)	75.0 (7.0)	59.6 (9.1)
350	20/7	0.6	84.3 (12.7)	78.8 (15.0)	50.9 (2.6)	48.1 (2.6)	72.0 (6.4)	57.5 (7.8)
			$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)					
1000	100/20	0.2	7.8 (0.2)	7.7 (0.3)	5.9 (0.1)	5.5 (0.2)	7.7 (0.2)	7.1 (0.4)
1000	100/20	0.6	7.8 (0.2)	7.7 (0.3)	5.9 (0.2)	5.6 (0.2)	7.6 (0.2)	7.1 (0.3)
600	100/20	0.2	7.4 (0.6)	6.8 (0.8)	5.8 (0.2)	5.2 (0.2)	7.3 (0.3)	6.4 (0.4)
600	100/20	0.6	7.4 (0.6)	6.9 (0.7)	5.9 (0.2)	5.2 (0.2)	7.2 (0.3)	6.3 (0.4)
500	50/35	0.2	7.4 (0.5)	7.3 (0.5)	5.6 (0.2)	5.5 (0.2)	6.9 (0.4)	6.4 (0.4)
500	50/35	0.6	7.4 (0.4)	7.3 (0.4)	5.7 (0.2)	5.6 (0.2)	6.9 (0.3)	6.4 (0.4)
500	50/10	0.2	7.6 (0.4)	7.3 (0.6)	5.9 (0.1)	5.5 (0.2)	7.4 (0.2)	6.5 (0.4)
500	50/10	0.6	7.6 (0.3)	7.3 (0.5)	5.9 (0.2)	5.6 (0.2)	7.2 (0.2)	6.4 (0.4)
350	20/7	0.2	7.5 (0.4)	7.3 (0.5)	5.9 (0.2)	5.7 (0.2)	7.2 (0.3)	6.4 (0.5)
350	20/7	0.6	7.5 (0.3)	7.4 (0.4)	5.9 (0.2)	5.8 (0.2)	7.1 (0.3)	6.4 (0.4)

Table 2.8: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments possible per stage and based on an underlying nontree-type DTR and assuming incorrectly-specified propensity models. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in H; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 8.0$.

n	H/Hsub	rho	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
			% opt (IQR)					
1000	100/20	0.2	87.0 (34.1)	78.2 (32.0)	24.3 (2.8)	18.3 (2.3)	86.7 (4.6)	69.8 (11.4)
1000	100/20	0.6	88.5 (32.2)	80.7 (29.8)	25.0 (2.6)	18.4 (2.0)	82.0 (5.4)	61.2 (9.3)
600	100/20	0.2	59.3 (33.4)	42.8 (21.6)	22.0 (3.1)	15.5 (2.4)	71.6 (7.4)	45.4 (10.9)
600	100/20	0.6	59.6 (31.5)	46.4 (20.6)	22.5 (3.3)	15.6 (2.3)	66.3 (8.1)	38.4 (10.5)
500	50/35	0.2	62.4 (32.8)	59.1 (30.8)	19.2 (2.8)	18.2 (2.7)	54.4 (9.8)	43.5 (10.9)
500	50/35	0.6	63.6 (28.3)	58.8 (29.2)	19.6 (2.8)	18.3 (2.7)	50.6 (10.3)	37.7 (10.3)
500	50/10	0.2	72.6 (33.2)	58.1 (31.4)	23.9 (3.4)	18.1 (2.8)	72.1 (7.6)	43.6 (10.5)
500	50/10	0.6	74.0 (30.6)	61.0 (26.4)	24.6 (3.3)	18.2 (2.7)	66.3 (7.7)	37.0 (9.4)
350	20/7	0.2	68.4 (33.6)	63.5 (31.0)	23.9 (4.1)	20.8 (3.7)	64.2 (8.1)	41.2 (9.7)
350	20/7	0.6	71.0 (30.0)	64.8 (29.3)	24.3 (4.2)	20.7 (3.6)	60.1 (8.3)	36.8 (8.7)
			$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)					
1000	100/20	0.2	7.7 (1.2)	7.4 (1.3)	4.9 (0.2)	4.4 (0.2)	7.7 (0.1)	7.0 (0.5)
1000	100/20	0.6	7.7 (1.2)	7.4 (1.2)	5.0 (0.2)	4.5 (0.1)	7.5 (0.2)	6.7 (0.4)
600	100/20	0.2	6.6 (1.1)	6.0 (0.9)	4.7 (0.2)	4.2 (0.2)	7.2 (0.3)	6.0 (0.5)
600	100/20	0.6	6.6 (1.2)	6.1 (0.9)	4.8 (0.2)	4.2 (0.2)	7.0 (0.3)	5.8 (0.5)
500	50/35	0.2	6.8 (1.1)	6.7 (1.1)	4.5 (0.2)	4.4 (0.2)	6.5 (0.4)	6.0 (0.5)
500	50/35	0.6	6.9 (1.1)	6.6 (1.1)	4.6 (0.2)	4.5 (0.2)	6.4 (0.4)	5.8 (0.5)
500	50/10	0.2	7.3 (1.2)	6.6 (1.1)	4.9 (0.2)	4.4 (0.2)	7.2 (0.3)	6.0 (0.5)
500	50/10	0.6	7.3 (1.1)	6.7 (1.0)	5.0 (0.2)	4.4 (0.2)	7.0 (0.3)	5.8 (0.4)
350	20/7	0.2	7.1 (1.1)	7.0 (1.1)	4.9 (0.3)	4.6 (0.2)	6.9 (0.3)	5.9 (0.4)
350	20/7	0.6	7.2 (1.1)	6.9 (1.1)	5.0 (0.3)	4.7 (0.2)	6.8 (0.3)	5.7 (0.4)

2.4.4.2 Simulation Studies To Evaluate the Relative Performance of ReST-L in the Absence of Confounding by Variables in $\mathbf{H}^C_{\mathbf{sub}}$

In this supplemental simulation study we evaluate the relative performance of ReST-L in the absence of confounding by variables in $\mathbf{H}^C_{\text{sub}}$, i.e., the true data-generating model for treatment $\mathbf{A}$ and outcomes $\mathbf{Y}$ include only variables in $\mathbf{H}_{\text{sub}}$. This is in contrast to the simulation experiments presented in Section 2.4.3 in which confounding by variables in $\mathbf{H}^C_{\text{sub}}$ exist. Using the data generating mechanisms presented in Section 2.4.3, for this simulation experiment we replace variables X_{C1} and X_{C2}, the first two variables in $\mathbf{H}^C_{\text{sub}}$, with the final two variables in $\mathbf{H}_{\text{sub}}$. As in all previously-reported simulation experiments and by ReST-L assumption, the optimal regimes $\mathbf{g}^{\text{opt}}$ are defined using variables only in $\mathbf{H}_{\text{sub}}$.

As can be seen in Tables 2.9 and 2.10, performance for both ReST-L and T-RL are similar to those reported in Tables 2.5 and 2.6. Performance for Q-Learning methods, with the exception of restricted nonparametric Q-Learning, are also similar. Only restricted nonparametric Q-Learning demonstrates a slightly lower performance in this setting. For a tree-based DTR, for example, we observe 88.9% correct treatment classification in Table 2.5 with a sample size $n = 1000$, $|\mathbf{H}| = 100$ variables and $\rho = 0.2$, compared with 85.5% correct classification when the true treatment allocation $\mathbf{A}$, outcomes $\mathbf{Y}$, and the optimal DTR $\mathbf{g}^{\text{opt}}$ are defined using variables only in $\mathbf{H}_{\text{sub}}$.

Table 2.9: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments per stage and based on an underlying, tree-type DTR with outcomes, treatment assignment, and optimal dynamic treatment regime defined using variables in $\mathbf{H}_{\text{sub}}$. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 8.0$.

n	H/Hsub	rho	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
			% opt (IQR)					
1000	100/20	0.2	94.9 (5.5)	92.5 (6.5)	50.9 (2.4)	45.3 (2.5)	85.5 (5.6)	73.6 (7.6)
1000	100/20	0.6	95.3 (5.6)	92.6 (6.8)	51.1 (2.3)	45.5 (2.6)	82.5 (6.1)	70.0 (8.0)
600	100/20	0.2	89.4 (10.1)	77.0 (18.4)	49.6 (2.5)	40.2 (2.8)	75.4 (7.6)	60.0 (10.8)
600	100/20	0.6	89.2 (10.7)	77.5 (17.5)	49.8 (2.5)	40.7 (3.1)	72.2 (7.7)	54.8 (11.0)
500	50/35	0.2	84.9 (14.1)	83.4 (15.6)	47.2 (2.7)	45.1 (2.7)	65.3 (9.4)	61.2 (9.8)
500	50/35	0.6	85.6 (13.0)	82.7 (15.0)	47.6 (2.9)	45.4 (2.8)	61.5 (8.5)	56.8 (9.7)
500	50/10	0.2	89.9 (10.9)	83.4 (15.5)	50.8 (2.6)	45.1 (3.0)	78.3 (7.7)	61.2 (9.4)
500	50/10	0.6	89.9 (10.3)	83.7 (14.2)	50.7 (2.7)	45.3 (3.0)	74.6 (7.3)	57.0 (9.1)
350	20/7	0.2	85.3 (13.2)	81.3 (15.6)	50.7 (2.4)	48.2 (2.7)	74.5 (7.8)	60.6 (10.2)
350	20/7	0.6	86.7 (13.0)	82.2 (15.5)	51.2 (2.7)	48.8 (3.0)	72.0 (6.7)	58.6 (8.3)
			$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)					
1000	100/20	0.2	7.8 (0.2)	7.7 (0.3)	5.9 (0.2)	5.5 (0.2)	7.5 (0.2)	7.1 (0.3)
1000	100/20	0.6	7.8 (0.2)	7.7 (0.3)	5.9 (0.1)	5.5 (0.2)	7.4 (0.2)	6.9 (0.3)
600	100/20	0.2	7.6 (0.3)	7.2 (0.7)	5.8 (0.2)	5.2 (0.2)	7.1 (0.3)	6.4 (0.5)
600	100/20	0.6	7.6 (0.3)	7.2 (0.6)	5.8 (0.2)	5.2 (0.2)	7.0 (0.4)	6.2 (0.5)
500	50/35	0.2	7.5 (0.4)	7.5 (0.4)	5.6 (0.2)	5.5 (0.2)	6.7 (0.4)	6.5 (0.4)
500	50/35	0.6	7.5 (0.4)	7.4 (0.4)	5.7 (0.2)	5.5 (0.2)	6.5 (0.4)	6.3 (0.4)
500	50/10	0.2	7.7 (0.3)	7.4 (0.5)	5.9 (0.2)	5.5 (0.2)	7.3 (0.3)	6.4 (0.4)
500	50/10	0.6	7.7 (0.3)	7.5 (0.5)	5.9 (0.1)	5.5 (0.2)	7.1 (0.3)	6.3 (0.5)
350	20/7	0.2	7.5 (0.4)	7.4 (0.5)	5.9 (0.2)	5.7 (0.2)	7.1 (0.3)	6.5 (0.5)
350	20/7	0.6	7.6 (0.3)	7.4 (0.4)	5.9 (0.2)	5.7 (0.2)	7.0 (0.3)	6.4 (0.4)

Table 2.10: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments possible per stage and based on an underlying, nontree-type DTR with outcomes, treatment assignment, and optimal dynamic treatment regime defined using variables in $\mathbf{H}_{\text{sub}}$. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in H; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 8.0$.

n	H/Hsub	rho	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
			% opt (IQR)					
1000	100/20	0.2	93.3 (33.5)	85.7 (35.7)	25.2 (2.7)	18.4 (2.0)	81.6 (7.3)	69.4 (12.2)
1000	100/20	0.6	93.6 (35.0)	86.4 (35.3)	25.2 (2.6)	18.3 (2.1)	75.8 (7.5)	60.7 (8.9)
600	100/20	0.2	67.6 (42.6)	52.1 (26.3)	22.7 (3.1)	15.6 (2.4)	64.8 (9.8)	46.7 (12.2)
600	100/20	0.6	77.1 (38.8)	54.8 (28.8)	22.8 (3.0)	15.6 (2.1)	59.8 (8.5)	40.9 (11.4)
500	50/35	0.2	65.7 (37.3)	63.3 (34.8)	19.6 (3.0)	18.2 (2.7)	49.6 (10.5)	44.9 (11.3)
500	50/35	0.6	68.4 (32.3)	64.8 (33.8)	19.9 (2.9)	18.4 (2.9)	46.6 (10.5)	39.9 (10.8)
500	50/10	0.2	78.4 (37.1)	62.6 (35.8)	24.7 (3.6)	18.0 (2.8)	67.5 (7.8)	43.6 (12.3)
500	50/10	0.6	78.2 (33.0)	66.9 (32.5)	24.9 (3.3)	18.5 (2.7)	63.1 (7.3)	40.8 (10.1)
350	20/7	0.2	64.6 (40.0)	59.8 (34.7)	24.4 (3.9)	20.7 (3.2)	61.7 (9.3)	43.0 (9.7)
350	20/7	0.6	65.4 (38.1)	62.5 (33.3)	24.8 (3.9)	21.2 (3.2)	59.0 (8.8)	40.8 (8.7)
			$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)					
1000	100/20	0.2	7.8 (1.2)	7.6 (1.2)	4.9 (0.2)	4.4 (0.1)	7.5 (0.3)	7.0 (0.5)
1000	100/20	0.6	7.8 (1.2)	7.7 (1.2)	4.9 (0.2)	4.4 (0.2)	7.2 (0.3)	6.6 (0.4)
600	100/20	0.2	7.2 (1.2)	6.4 (1.0)	4.7 (0.2)	4.2 (0.2)	6.8 (0.4)	6.1 (0.6)
600	100/20	0.6	7.4 (1.3)	6.5 (1.0)	4.7 (0.2)	4.2 (0.2)	6.7 (0.4)	5.8 (0.6)
500	50/35	0.2	7.0 (1.1)	6.8 (1.1)	4.5 (0.2)	4.4 (0.2)	6.2 (0.5)	6.0 (0.5)
500	50/35	0.6	7.1 (1.1)	7.0 (1.1)	4.5 (0.2)	4.4 (0.2)	6.1 (0.4)	5.8 (0.5)
500	50/10	0.2	7.4 (1.2)	6.8 (1.1)	4.9 (0.3)	4.4 (0.2)	7.0 (0.3)	6.0 (0.5)
500	50/10	0.6	7.5 (1.2)	7.1 (1.1)	4.9 (0.2)	4.4 (0.2)	6.8 (0.3)	5.8 (0.5)
350	20/7	0.2	7.0 (1.2)	6.7 (1.2)	4.9 (0.3)	4.6 (0.2)	6.8 (0.3)	6.0 (0.4)
350	20/7	0.6	7.2 (1.3)	6.9 (1.2)	4.9 (0.3)	4.7 (0.2)	6.7 (0.3)	5.9 (0.4)

2.4.4.3 Simulation Studies to Evaluate the Relative Performance of ReST-L Under Violations of ReST-L Assumptions

ReST-L is an analytic solution that can be used when the investigators are relatively certain about the set of covariates that can be considered in the optimal dynamic treatment regime. For scenarios in which the optimal DTR is actually defined, at least in part, by variables in $\mathbf{H}_{\text{sub}}^{C}$, we would expect the performance of ReST-L to be lower than its unrestricted counterpart, T-RL. Although this will be the case overall, here we present an illustration of performance differences that may be observed under violations of the ReST-L assumption that only variables in $\mathbf{H}_{\text{sub}}$ are involved in an optimal DTR. Specifically, using the data generating mechanisms presented in Section 2.4.3 for both tree- and nontree-type DTRs, we modify g_1^{opt} to be defined using X_{C1} rather than X_1. Additionally, g_2^{opt} is now defined using the final variable in $\mathbf{H}_{\text{sub}}^{C}$ rather than on X_3. (Recall that X_{C1} refers to the first variable in the set of $\mathbf{H}_{\text{sub}}^{C}$.) Otherwise, similar to the data generating mechanisms used in Section 2.4.3, confounding is introduced using variables in both $\mathbf{H}_{\text{sub}}$ and $\mathbf{H}_{\text{sub}}^{C}$.

As can be seen in Table 2.11 for a tree-type DTR, performance of ReST-L is lower than that of T-RL across all data generation settings. With larger sample sizes, performance of ReST-L reaches about 75% of observations correctly classified to their optimal treatment. T-RL, on the other hand, achieves more than 90% correct treatment classification, which is similar to that reported in Table 2.5. With a nontree-type DTR structure, we observe in Table 2.12 that performance of ReST-L is poor across all data settings. Performance for T-RL is mediocre for lower sample sizes, as well, but is similar to the performance presented in Table 2.6.

Table 2.11: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments per stage and based on an underlying, tree-type DTR with outcomes, treatment assignment, and optimal dynamic treatment regime defined using variables in $\mathbf{H}_{\text{sub}}$ and $\mathbf{H}_{\text{sub}}^C$. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 8.0$.

n	H/Hsub	rho	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
			% opt (IQR)					
1000	100/20	0.2	73.6 (3.2)	92.2 (4.7)	47.6 (2.2)	44.8 (2.4)	67.7 (4.3)	64.6 (11.0)
1000	100/20	0.6	73.4 (2.9)	92.0 (6.1)	48.5 (2.3)	44.8 (2.6)	67.3 (4.3)	64.7 (8.0)
600	100/20	0.2	69.8 (8.7)	81.1 (10.7)	46.1 (2.5)	39.8 (2.9)	61.8 (6.9)	50.1 (13.9)
600	100/20	0.6	69.1 (7.4)	79.8 (11.3)	47.0 (2.7)	39.3 (3.0)	60.9 (6.3)	50.1 (9.9)
500	50/35	0.2	65.5 (9.3)	85.0 (8.9)	43.2 (2.9)	44.7 (3.1)	53.4 (10.2)	53.5 (10.7)
500	50/35	0.6	64.8 (10.1)	84.3 (9.1)	44.0 (3.0)	44.6 (3.0)	54.1 (8.2)	52.2 (9.8)
500	50/10	0.2	69.6 (8.8)	84.7 (8.6)	47.2 (2.6)	44.7 (3.0)	63.1 (5.9)	52.6 (12.0)
500	50/10	0.6	69.2 (9.1)	84.2 (9.6)	48.5 (2.8)	44.8 (3.2)	62.4 (5.5)	51.9 (8.8)
350	20/7	0.2	66.7 (10.9)	83.5 (9.5)	47.5 (2.6)	48.0 (2.7)	60.7 (5.7)	56.8 (8.5)
350	20/7	0.6	66.1 (9.9)	83.1 (10.0)	48.6 (2.9)	48.2 (2.9)	60.1 (6.0)	54.5 (7.6)
			$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)					
1000	100/20	0.2	7.2 (0.2)	7.8 (0.2)	5.7 (0.2)	5.4 (0.2)	7.0 (0.2)	6.7 (0.5)
1000	100/20	0.6	7.2 (0.2)	7.8 (0.2)	5.8 (0.1)	5.5 (0.2)	6.9 (0.2)	6.7 (0.3)
600	100/20	0.2	7.1 (0.3)	7.4 (0.5)	5.6 (0.2)	5.1 (0.2)	6.7 (0.3)	6.0 (0.6)
600	100/20	0.6	7.0 (0.3)	7.3 (0.5)	5.7 (0.2)	5.2 (0.2)	6.6 (0.3)	5.9 (0.5)
500	50/35	0.2	6.9 (0.3)	7.5 (0.3)	5.4 (0.2)	5.4 (0.2)	6.2 (0.5)	6.1 (0.5)
500	50/35	0.6	6.9 (0.3)	7.5 (0.4)	5.5 (0.2)	5.5 (0.2)	6.2 (0.4)	6.0 (0.5)
500	50/10	0.2	7.1 (0.3)	7.5 (0.3)	5.7 (0.2)	5.4 (0.2)	6.7 (0.3)	6.1 (0.5)
500	50/10	0.6	7.0 (0.2)	7.5 (0.3)	5.8 (0.2)	5.5 (0.2)	6.7 (0.3)	6.0 (0.4)
350	20/7	0.2	7.0 (0.3)	7.5 (0.4)	5.7 (0.2)	5.6 (0.2)	6.6 (0.3)	6.3 (0.5)
350	20/7	0.6	7.0 (0.3)	7.5 (0.3)	5.8 (0.2)	5.7 (0.2)	6.6 (0.3)	6.2 (0.4)

Table 2.12: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments possible per stage and based on an underlying, nontree-type DTR with outcomes, treatment assignment, and optimal dynamic treatment regime defined using variables in $\mathbf{H}_{\text{sub}}$ and $\mathbf{H}_{\text{sub}}^C$. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in H; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 8.0$.

n	H/Hsub	rho	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
			% opt (IQR)					
1000	100/20	0.2	50.4 (10.6)	64.7 (44.6)	24.8 (2.7)	18.6 (2.2)	46.9 (3.5)	58.3 (17.9)
1000	100/20	0.6	48.2 (10.8)	90.0 (33.9)	25.3 (2.7)	18.7 (2.4)	44.7 (3.4)	48.1 (12.6)
600	100/20	0.2	44.3 (14.2)	43.1 (20.2)	22.3 (3.1)	15.8 (2.3)	40.2 (4.3)	32.1 (11.7)
600	100/20	0.6	43.3 (14.2)	47.8 (24.4)	22.8 (3.4)	15.8 (2.4)	38.3 (5.6)	27.9 (9.3)
500	50/35	0.2	43.0 (14.0)	53.0 (37.3)	19.8 (2.8)	18.8 (2.5)	32.6 (6.0)	32.4 (10.8)
500	50/35	0.6	42.5 (13.7)	62.2 (37.3)	20.1 (3.1)	18.6 (2.8)	31.9 (5.7)	29.2 (9.8)
500	50/10	0.2	45.3 (12.1)	52.4 (37.4)	24.5 (3.3)	18.4 (2.6)	41.1 (4.0)	32.4 (11.7)
500	50/10	0.6	44.0 (13.7)	60.5 (35.9)	25.1 (3.4)	18.4 (2.5)	39.0 (4.3)	28.6 (8.9)
350	20/7	0.2	44.5 (12.5)	54.9 (36.6)	24.6 (4.0)	21.2 (3.3)	38.8 (5.1)	34.2 (10.2)
350	20/7	0.6	43.4 (14.3)	58.8 (36.3)	24.9 (3.7)	21.0 (3.4)	37.2 (4.9)	30.5 (7.7)
			$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)					
1000	100/20	0.2	6.7 (0.3)	6.9 (1.3)	5.0 (0.2)	4.5 (0.2)	6.6 (0.2)	6.6 (0.6)
1000	100/20	0.6	6.6 (0.4)	7.6 (1.3)	5.0 (0.2)	4.5 (0.2)	6.4 (0.2)	6.2 (0.5)
600	100/20	0.2	6.4 (0.9)	6.1 (0.8)	4.8 (0.2)	4.2 (0.2)	6.2 (0.3)	5.5 (0.6)
600	100/20	0.6	6.4 (0.9)	6.2 (0.8)	4.9 (0.2)	4.2 (0.2)	6.1 (0.3)	5.3 (0.5)
500	50/35	0.2	6.4 (0.9)	6.5 (1.1)	4.6 (0.2)	4.5 (0.2)	5.7 (0.3)	5.6 (0.5)
500	50/35	0.6	6.3 (0.9)	6.6 (1.1)	4.6 (0.2)	4.5 (0.2)	5.7 (0.3)	5.4 (0.5)
500	50/10	0.2	6.5 (0.5)	6.5 (1.1)	5.0 (0.2)	4.5 (0.2)	6.2 (0.2)	5.6 (0.5)
500	50/10	0.6	6.5 (0.9)	6.6 (1.1)	5.0 (0.2)	4.5 (0.2)	6.1 (0.2)	5.4 (0.4)
350	20/7	0.2	6.4 (0.7)	6.6 (1.1)	5.0 (0.3)	4.7 (0.2)	6.1 (0.3)	5.6 (0.4)
350	20/7	0.6	6.4 (0.9)	6.6 (1.2)	5.0 (0.3)	4.7 (0.3)	6.0 (0.2)	5.5 (0.3)

2.4.4.4 Simulation Studies to Evaluate the Relative Performance of ReST-L Under a Data Generation Mechanism with a High Degree of Confounding

Here we present comprehensive results for the two-stage simulation experiment introduced in 2.4.1, which demonstrated the bias of Naive T-RL in estimating the counterfactual mean outcome if all patients were to receive treatment according to their estimated optimal DTR. As described previously, the data generating mechanism includes a binary variable Z that has a strong confounding relationship with both the treatment assignment mechanisms for A_1 and A_2. This differs from the simulation experiments presented in Sections 2.4.2 and 2.4.3 in which the confounding relationship is defined using only continuous covariates. Parameters varied across this simulation study include the sample size (n), with fixed sample sizes $n = 500, n = 1000$, and $n = 2000$, the number of covariates in $\mathbf{H}$ and $\mathbf{H}_{\text{sub}}$ (H/Hsub $= 20/7, 50/10, 50/35, 100/20$), the correlation used to generate the correlation matrix for covariates in $\mathbf{H}$ ($\rho = 0, 0.2, 0.6$), and the true, underlying structure of the DTR (i.e., tree or nontree-type). Refer to Section 2.4.1 for a description of the full data generating models. As in the previously-described simulations and consistent with the assumptions required by this method, we assume that only variables in $\mathbf{H}_{\text{sub}}$ may be included in an optimal DTR; variables from either $\mathbf{H}_{\text{sub}}$ or $\mathbf{H}_{\text{sub}}^C$, however, may define the intermediate outcomes and the treatment assignment mechanisms. Under optimal treatment allocation $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 5.4$. It is further assumed in the analysis that propensity models for both ReST-L and T-RL are correctly specified.

Results for the simulation studies are presented in Table 2.13 for a tree-type DTR and in Table 2.14 for a nontree-type DTR. With a tree-type DTR and a strong confounding relationship of covariate Z with the treatment assignment and outcomes, a larger sample size than that needed in the analyses presented in Sections 2.4.2 and 2.4.3 is required to achieve reasonable performance. With a correlation of $\rho = 0.2$ and

$|\mathbf{H} = 20|$, a sample size of $n = 1000$ is required to achieve similar levels of performance to a sample size of $n = 350$ in Table 2.5. With a sample size of $n = 2000$, performance across all data generating settings reaches about 90% correct classification and the ReST-L results for $n = 2000$ are comparable to those of T-RL. For smaller sample sizes, however, we observe that ReST-L improves upon T-RL across all data settings, although the improvement is most apparent with a larger number of covariates and when the proportion of variables in $\mathbf{H}_{\text{sub}}$ relative to $\mathbf{H}$ is lower. Consider, for example, a sample size of $n = 1000$ with $|\mathbf{H}| = 100$ and $\rho = 0.2$ in which we report an estimated 82% correct classification for ReST-L compared with 73% for T-RL. With a nontree-type DTR structure, performance is slightly lower across all settings for both ReST-L and T-RL than with a tree-type DTR, but both ReST-L and T-RL achieve about 88-90% correct classification with $n = 2000$.

2.5 Application to Personalize Hand Injury Treatment Decisions

We illustrate our methods using a de-identified sub-dataset from the FRANCHISE study (*Chung et al.*, 2019), and following the secondary analysis conducted by *Speth et al.* (2020), which includes baseline characteristics and patient-reported and functional outcome measures for 338 consenting adults with traumatic amputation of digits distal to the metacarpophalangeal (MCP) joint who were treated by revision amputation or successful replantation at least one year prior to recruitment. Details on the original study design and enrollment, as well as a comprehensive description of collection methods for functional assessments and patient-reported outcomes, are found in *Chung et al.* (2019).

Four outcome measures are considered in this analysis: hand strength, dexterity, pain, and patient-reported hand quality of life. Refer to *Speth et al.* (2020) for a

Table 2.13: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments based on an underlying, tree-type DTR with $\rho = 0.2$. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Naive T-RL = Naive Tree-based Reinforcement Learning; Q-Linear-R = Restricted Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment rule estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 5.4$

H/Hsub	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
	% opt (IQR)					
n = 500						
20/7	71.6 (18.2)	62.2 (17.3)	41.6 (3.6)	36.4 (3.4)	49.1 (6.3)	45.4 (5.2)
50/10	67.1 (18.7)	55.0 (16.7)	39.8 (3.6)	28.9 (3.0)	45.4 (6.8)	39.8 (5.7)
50/35	57.7 (16.8)	52.8 (17.3)	32.2 (3.4)	29.0 (3.2)	35.0 (8.1)	39.7 (5.0)
100/20	58.4 (16.1)	37.6 (13.6)	35.6 (3.6)	20.5 (2.6)	40.8 (8.2)	35.1 (6.4)
n = 1000						
20/7	85.3 (10.2)	82.1 (12.9)	43.3 (3.9)	40.5 (2.8)	57.3 (4.8)	55.0 (4.8)
50/10	85.2 (12.0)	79.1 (15.6)	42.5 (3.1)	35.1 (2.5)	55.5 (4.9)	50.9 (4.4)
50/35	82.3 (13.2)	79.3 (14.4)	37.7 (2.6)	35.1 (2.6)	49.6 (5.9)	50.8 (4.4)
100/20	82.7 (11.6)	73.9 (16.0)	40.4 (2.8)	29.4 (2.5)	52.5 (5.5)	47.9 (4.5)
n = 2000						
20/7	91.8 (5.0)	91.3 (6.5)	43.8 (3.2)	42.7 (2.8)	66.3 (4.5)	66.0 (4.5)
50/10	91.7 (5.6)	90.6 (6.6)	43.3 (3.4)	39.6 (2.7)	65.6 (4.6)	62.1 (4.5)
50/35	91.4 (6.0)	90.8 (6.7)	41.2 (2.5)	39.6 (2.6)	63.0 (5.4)	62.2 (4.7)
100/20	91.0 (6.2)	89.5 (7.9)	42.8 (3.0)	35.5 (2.3)	65.0 (5.1)	59.7 (4.7)
	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)					
n = 500						
20/7	4.9 (0.5)	4.7 (0.5)	4.1 (0.2)	3.8 (0.2)	4.2 (0.3)	4.3 (0.2)
50/10	4.7 (0.5)	4.4 (0.5)	4.0 (0.2)	3.5 (0.2)	4.1 (0.3)	4.0 (0.3)
50/35	4.5 (0.5)	4.4 (0.5)	3.7 (0.2)	3.5 (0.2)	3.6 (0.4)	4.0 (0.3)
100/20	4.5 (0.5)	3.9 (0.6)	3.8 (0.2)	3.1 (0.2)	3.8 (0.4)	3.8 (0.4)
n = 1000						
20/7	5.1 (0.3)	5.1 (0.3)	4.1 (0.2)	4.0 (0.2)	4.5 (0.2)	4.7 (0.2)
50/10	5.1 (0.3)	5.0 (0.3)	4.1 (0.2)	3.8 (0.2)	4.4 (0.2)	4.5 (0.2)
50/35	5.0 (0.3)	5.0 (0.3)	3.9 (0.2)	3.8 (0.2)	4.2 (0.2)	4.5 (0.2)
100/20	5.0 (0.3)	4.9 (0.4)	4.0 (0.2)	3.5 (0.2)	4.3 (0.2)	4.4 (0.2)
n = 2000						
20/7	5.2 (0.2)	5.2 (0.2)	4.2 (0.2)	4.1 (0.2)	4.8 (0.2)	5.0 (0.2)
50/10	5.2 (0.2)	5.2 (0.2)	4.1 (0.2)	4.0 (0.1)	4.7 (0.2)	4.9 (0.2)
50/35	5.2 (0.2)	5.2 (0.2)	4.0 (0.2)	4.0 (0.2)	4.7 (0.2)	4.9 (0.2)
100/20	5.2 (0.2)	5.2 (0.2)	4.1 (0.2)	3.8 (0.2)	4.7 (0.2)	4.8 (0.2)

Table 2.14: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments based on an underlying, nontree-type DTR with $\rho = 0.2$. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Naive T-RL = Naive Tree-based Reinforcement Learning; Q-Linear-R = Restricted Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment rule estimated using the applicable method; IQR = interquartile range; $\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment. Under optimal treatment allocation $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 5.4$

H/Hsub	ReST-L	T-RL	Q-L-R	Q-L	Q-NP-R	Q-NP
	% opt (IQR)					
n = 500						
20/7	65.2 (23.2)	58.0 (21.7)	32.4 (3.1)	25.0 (3.0)	39.8 (7.6)	32.9 (6.7)
50/10	61.5 (21.0)	49.6 (18.3)	30.9 (3.0)	20.9 (2.4)	35.2 (7.3)	26.3 (5.6)
50/35	52.9 (20.4)	49.2 (19.3)	24.9 (2.8)	20.9 (2.6)	25.3 (5.7)	26.3 (6.0)
100/20	52.1 (17.0)	32.4 (13.2)	27.3 (3.2)	16.2 (2.3)	28.4 (6.1)	22.1 (6.3)
n = 1000						
20/7	82.3 (15.9)	78.5 (19.3)	34.3 (3.4)	28.0 (2.5)	51.0 (7.4)	47.3 (6.4)
50/10	82.7 (18.4)	76.0 (20.8)	33.7 (3.2)	24.3 (2.4)	48.5 (7.6)	40.0 (5.3)
50/35	79.4 (19.4)	75.4 (21.4)	29.2 (2.6)	24.4 (2.4)	38.6 (6.4)	40.6 (5.6)
100/20	79.0 (22.4)	68.2 (21.6)	31.5 (2.8)	20.8 (2.0)	42.6 (7.2)	36.2 (5.3)
n = 2000						
20/7	89.6 (6.5)	89.0 (8.7)	35.0 (3.7)	30.2 (2.5)	62.4 (6.1)	63.3 (5.2)
50/10	89.9 (6.0)	88.6 (10.3)	34.9 (3.2)	27.6 (2.3)	60.9 (6.1)	57.0 (6.7)
50/35	88.7 (8.1)	87.4 (10.8)	32.8 (2.2)	27.6 (2.1)	55.6 (7.6)	57.1 (6.8)
100/20	89.2 (7.4)	87.0 (11.7)	34.0 (2.7)	24.5 (2.1)	58.7 (6.8)	52.4 (6.9)
	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ (IQR)					
n = 500						
20/7	4.7 (0.6)	4.5 (0.6)	3.7 (0.2)	3.5 (0.2)	3.9 (0.3)	3.9 (0.3)
50/10	4.6 (0.6)	4.3 (0.6)	3.6 (0.2)	3.3 (0.2)	3.8 (0.3)	3.6 (0.3)
50/35	4.3 (0.6)	4.3 (0.6)	3.3 (0.2)	3.3 (0.2)	3.3 (0.3)	3.6 (0.3)
100/20	4.3 (0.6)	3.7 (0.5)	3.5 (0.2)	2.9 (0.2)	3.5 (0.3)	3.4 (0.3)
n = 1000						
20/7	5.1 (0.4)	5.1 (0.4)	3.8 (0.2)	3.7 (0.2)	4.3 (0.3)	4.5 (0.2)
50/10	5.1 (0.4)	5.0 (0.5)	3.8 (0.2)	3.5 (0.1)	4.2 (0.3)	4.2 (0.3)
50/35	5.0 (0.5)	4.9 (0.5)	3.6 (0.1)	3.5 (0.2)	3.9 (0.3)	4.2 (0.3)
100/20	5.0 (0.5)	4.8 (0.6)	3.7 (0.2)	3.3 (0.2)	4.0 (0.3)	4.0 (0.3)
n = 2000						
20/7	5.3 (0.2)	5.3 (0.2)	3.8 (0.2)	3.8 (0.1)	4.6 (0.2)	4.9 (0.2)
50/10	5.3 (0.2)	5.2 (0.3)	3.8 (0.2)	3.6 (0.1)	4.6 (0.2)	4.7 (0.3)
50/35	5.2 (0.3)	5.2 (0.3)	3.7 (0.2)	3.6 (0.1)	4.4 (0.3)	4.7 (0.2)
100/20	5.2 (0.3)	5.2 (0.3)	3.8 (0.2)	3.5 (0.2)	4.5 (0.2)	4.6 (0.3)

description of how composite measures were derived. Importantly, all measures were adjusted, as necessary, such that larger values represent better outcomes. Baseline covariates include both patient factors and injury characteristics. Variables considered as candidates for a treatment decision rule include age, number of digits amputated, thumb amputation, dominant hand injury, amputation level, mechanism of injury, and bilateral injury, as these factors have been shown to be clinically relevant to decision making (*Agarwal et al.*, 2010; *Berlin et al.*, 2014; *Boulas*, 1998; *Buntic et al.*, 2008; *Chung and Alderman*, 2002; *Sebastin and Chung*, 2011). In addition to the candidate variables listed above, possible confounding variables included in the propensity model include sex, race, education level, income level, marital status, employment status, occupation group, location of care, work-related injury, health insurance (yes/no), and health insurance type.

We conduct the analysis using a complete, multi-level dataset with missing data singly-imputed using random forest. Using ReST-L estimation, we find that patients for whom hand dexterity is a clinical priority should undergo replantation if they are 58 years or younger and revision amputation otherwise. Alternatively, patients for whom hand-related quality of life is most important should undergo revision amputation if he/she injures the dominant hand and replantation otherwise. Treating all patients with replantation may minimize patient-reported pain long-term compared to revision amputation. Finally, in patients for whom hand strength is paramount, the results of our analysis indicate that age is a principal factor in determining whether patients should undergo revision amputation or replantation. Specifically, our results suggest that middle-aged patients between the ages of 28 and 68 should receive replantation but revision amputation otherwise.

The decision rules related to hand strength and patient-reported pain are slightly different than those identified in *Speth et al.* (2020). We attribute this to the fact that *Speth et al.* (2020) used a naive restricted T-RL method for estimation of treatment

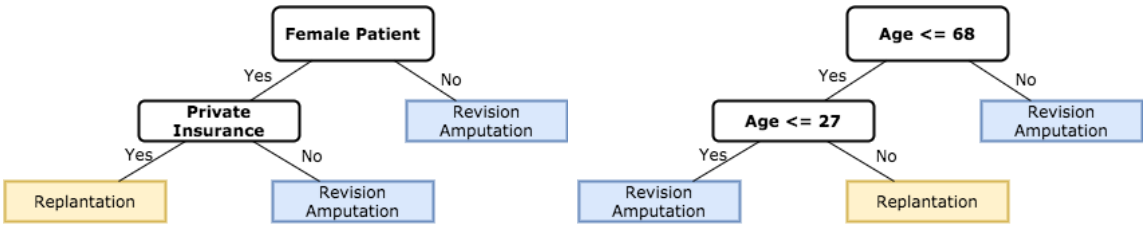

Figure 2.1: Estimated decision rules for the FRANCHISE study using tree-based reinforcement-learning (T-RL, left panel) and Restricted Sub-Tree Learning (ReST-L, right panel) to maximize long-term hand strength. T-RL selects gender and use of private insurance (yes/no) as tailoring variables for an estimated treatment assignment rule. ReST-L, conversely, using only a subset of clinical variables as possible tailoring variables, estimates a treatment decision rule based on age.

assignment rules, where any variable that was deemed inappropriate for a treatment assignment rule was removed post-hoc. Had T-RL been implemented without modification, it would have recommended that, in order to maximize long-term hand strength, female patients with private insurance be treated with replantation whereas men and female without private insurance be treated with revision amputation (Figure 1). Needless to say, making treatment decisions for traumatic amputation patients based on insurance type would be objectionable. With regard to gender, no physiologic differences in hand anatomy between males and females have been identified, suggesting that a decision rule based on a patient's gender would also be inappropriate. Given these results, we view this as additional evidence of the utility of using ReST-L.

2.6 Discussion

Personalized medicine reflects a goal of providing the right treatment to the right person at the right time. ReST-L provides a flexible, data-driven approach grounded in causal inference for estimating an interpretable, optimal multi-stage dynamic treatment regime using observational data when only a subset of covariates, based on clinical or other knowledge, should be considered as candidate tailoring variables.

Importantly, ReST-L addresses a clinical scenario that has not yet been addressed to our knowledge in the literature for tree-based, optimal DTR estimation. We have shown that there is an improvement over other estimation methods when, practically, a clinically-meaningful and ethical treatment decision should be made without certain variables and, given that ReST-L reduces to T-RL when the full set of covariates are considered, this provides an important extension of previous work. ReST-L utilizes a purity measure that is based upon a consistent and doubly robust estimator of the counterfactual mean outcome under a sub-tree regime when either the propensity model or the conditional mean model are correctly specified, resulting in a causal estimator with double protections against model misspecifications. We demonstrate that ReST-L can estimate the optimal DTR in the presence of a moderately large degree of covariates and we base simulation studies on a reasonably complex relationship that is intended to be reflective of data generating scenarios that may be seen in a real world scenario. There are also limitations to our work. We acknowledge that our results reflect a small number of possible data generating scenarios and it is likely that performance estimates would change under different simulation settings. For estimations in a two stage setting, we observe a high degree of variability in the estimated percentage of observations correctly classified to their optimal treatment using ReST-L. While the variability is much lower than the variability observed in T-RL, it is much larger than that estimated using restricted Q-Learning with non-parametric modeling assumptions. However, the median estimated performance is also consistently higher for ReST-L in a two-stage setting than it is for restricted nonparametric Q-Learning, suggesting that a trade of higher variability for higher estimated performance could be warranted. Finally, in our simulation studies we assume that the conditional mean models are incorrectly specified. Although this is useful in order to provide an understanding of performance as an "out of the box" solution for optimal DTR estimation, model selection and diagnostics can be used to

select either the propensity or the conditional mean model, or both. This was not explored, but this may be considered in data applications and/or in future research.

CHAPTER III

Clustered Q-Learning to Inform the Empirical Construction of an Optimal Clustered Adaptive Intervention

3.1 Introduction

In health and education settings, intervention (e.g., treatment) is often provided at the level of a cluster (e.g., clinic or school; *Murray*, 1998; *Raudenbush and Bryk*, 2002; *Raudenbush and Schwartz*, 2020). A clustered adaptive intervention (CAI) is a pre-specified sequence of decision-rules that guides practitioners on how best—and based on which measures—to tailor intervention at the level of a cluster (e.g., clinic or school) with the goal of improving outcomes at the level of individuals within the cluster (e.g., patients within a clinic, or school professionals at a school). In a CAI, intervention is adapted and re-adapted at the level of the cluster across multiple stages of intervention based on pre-specified measures of change in the cluster. These time-varying measures, which both inform subsequent intervention and can be impacted by prior intervention as part of a CAI, are known as tailoring variables.

Consider the following example, two-stage CAI that was designed to improve the uptake of an evidence-based practice for mood disorders, known as Life Goals, at community-based mental health clinics across Colorado and Michigan (*Kilbourne*

et al., 2014; *Necamp et al.*, 2017; *Smith et al.*, 2019): "If a clinic is unsuccessful in implementing Life Goals after 6 months of replicating effective programs support, they should additionally receive external support (EF). After 6 months of EF, they should continue EF if they successfully implement Life Goals but should additionally receive internal support (EF+IF) if they do not." In this example CAI, intervention is provided across two stages at the level of the clinic, but the primary goal of the intervention is to improve the uptake of Life Goals so as to improve mental health quality of life for patients with mood disorders, i.e., an outcome at the level of the individuals within the clinic. In this example, the tailoring variable for the cluster-level intervention is whether the clinic successfully implemented Life Goals following 6 months of external support.

An important scientific question for domain scientists is to better understand whether a particular (set of) covariate(s) ought to be considered for inclusion as a tailoring variable in a CAI that optimizes individual-level outcomes. For example, in the context of the example CAI described above, this question could be posed: "Are there additional clinic-level factors that can be used to further tailor intervention at the clinic-level so as to improve the clinic-level uptake of Life Goals and, correspondingly, the outcomes of patients with mood disorders?" For example, one may ask whether rural clinics have particular needs vis-a-vis clinic-level interventions that urban or suburban clinics do not. This manuscript develops an easy-to-use method to answer such questions when using data obtained from a Clustered SMART.

A Clustered SMART is a multi-stage trial design in which randomization occurs across two or more stages at the level of a cluster, but the primary outcome of interest lies at the level of an individual, or unit, within the cluster (*Almirall et al.*, 2018; *Kilbourne et al.*, 2014; *Necamp et al.*, 2017). Given the broad definition, there are innumerable ways to design a Clustered SMART; three of the most common are displayed in Figure 3.1. Clustered SMARTs can be used to address scientific

questions related to the comparison of CAIs, such as to estimate the relative effect between two or more CAIs (*Almirall et al.*, 2014; *Kosorok and Moodie*, 2016; *Oetting et al.*, 2011). Clustered SMARTs can also be used to identify variables that can be used to additionally tailor multi-stage, cluster-level interventions, such as the question posed in the example above.

Statistical methods supporting single-stage cluster-randomized trials is vast (*Eldridge and Kerry*, 2012; *Hayes and Moulton*, 2017; *Murray et al.*, 2004). So too are methods for the estimation of tailored, multi-stage, adaptive interventions for SMART trials where both randomization and outcome assessment occur at the individual level (*Almirall et al.*, 2014; *Dawson and Lavori*, 2004; *Lavori and Dawson*, 2000; *Lavori and Dawson*, 2004; *Murphy*, 2005b; *Oetting et al.*, 2011; *Wallace et al.*, 2016). One popular method used to estimate personalized, multi-stage adaptive interventions is Q-Learning (*Chakraborty et al.*, 2013; *Murphy*, 2005a; *Nahum-Shani et al.*, 2012; *Schulte et al.*, 2014; etc.). Q-Learning, which was originally proposed within the computer science literature (*Watkins*, 1989), consists of positing a series of stage-specific Q-functions representing the expected outcome conditional on covariate and treatment history. The goal, then, is to identify a multi-stage adaptive intervention as a function of covariate and treatment history such that the Q-functions are optimized. Although the multi-stage Q-functions can be modeled flexibly, parametric linear regression models are often used due to the ease of implementation and widespread use and understanding of linear regression modeling by domain scientists. Methods pertaining to Clustered SMARTs, however, are less well developed (*Raudenbush and Schwartz*, 2020). *Necamp et al.* (2017) proposed a weighted least squares approach to compare the means of individual-level outcomes for two cluster-level adaptive interventions. And sample size formulas needed to achieve a primary goal of comparing mean outcomes for two or more CAIs within a Clustered SMART have been devised for a continuous outcome (*Ghosh et al.*, 2016; *Necamp et al.*, 2017)

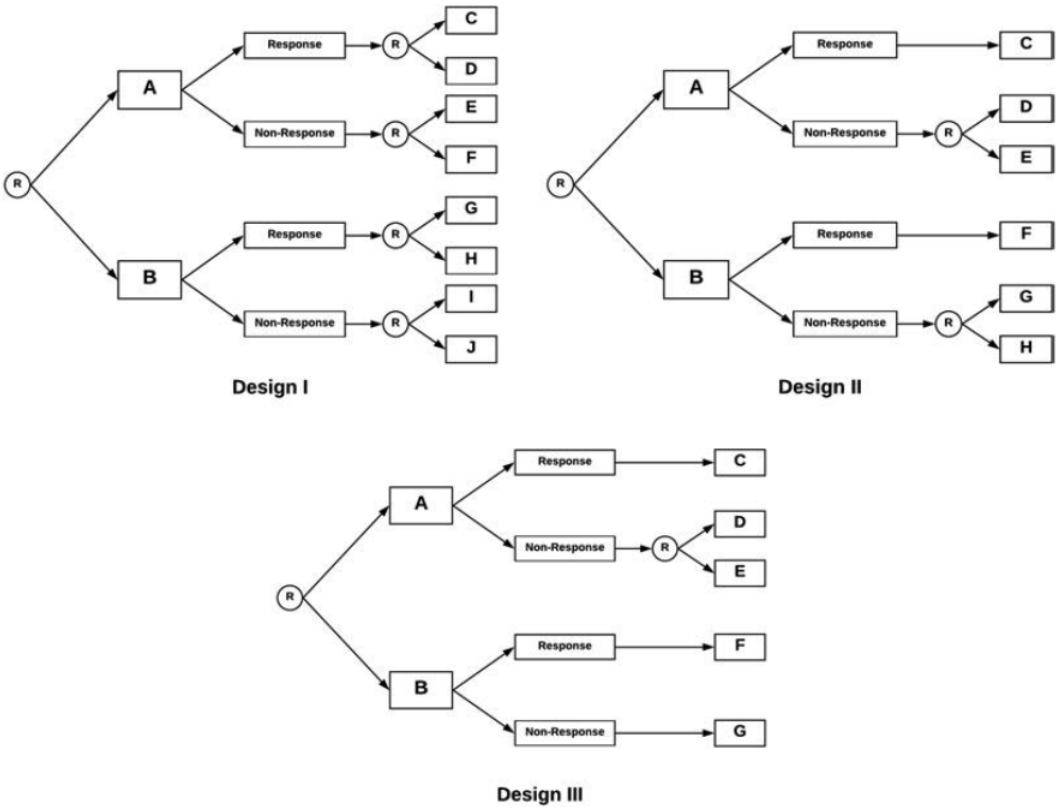

Figure 3.1: Three of the most commonly used Clustered Sequential Multiple Assignment Randomized Trial (Clustered SMART) designs. R denotes a cluster-level randomization and A-J denote cluster-level interventions which need not be unique. Each of the featured designs illustrates two intervention stages with an initial randomization to one of two first-stage interventions followed by an assessment of response. In Design I, all clusters are re-randomized at Stage 2 conditional on both the first-stage intervention and response. Design II features a second-stage randomization only for clusters with a non-response to first-stage intervention whereas Design III includes a second-stage randomization only for clusters non-responsive to initial intervention A. Figure replicated from *Speth and Kidwell* (2019).

and for binary outcomes (*Ghosh et al.*, 2016). To date, however, there are no existing statistical methods to identify cluster-level tailoring variables across multiple stages of intervention that can be used to optimize a patient-level outcome.

In this manuscript we propose Clustered Q-Learning for use with data from a Clustered SMART to inform the empirical construction of a CAI that maximizes the population mean of an outcome. We describe steps to estimate model parameters and we address a well-known challenge in Q-Learning regression concerning the estimation of confidence intervals under conditions of nonregularity, which is expected to be common. Nonregularity is a phenomenon that will affect the estimation of standard errors for any earlier intervention stage if there is a non-unique treatment effect (e.g., no treatment effect) at one or more of the later intervention stages. To enable the construction of confidence intervals with nominal coverage rates, we extend the m-out-of-n bootstrap (*Chakraborty et al.*, 2013) to accommodate data from a Clustered SMART. Using simulation experiments, we show that estimation of model parameters using Clustered Q-Learning is unbiased in estimating Stage 2 model parameters and fully-regular Stage 1 model parameters and exhibits negligible bias in estimating model parameters in nonregular settings when the within-cluster correlation is low. We demonstrate near nominal coverage of estimated confidence intervals across two intervention stages when the number of clusters is large and exhibit slight under-coverage in estimating confidence intervals for Stage 1 model parameters when the number of clusters is small. To illustrate the methods, we use data from ADEPT, the Adaptive Implementation of Effective Programs Trial (*Kilbourne et al.*, 2014), a Clustered SMART, from which we aim to identify variable(s) at each stage that may be used to tailor clinic-level interventions such that clinic-level uptake of evidence-based practices—and, therefore, also the associated outcomes of patients with mood disorders—is improved.

3.2 Methodology

3.2.1 Set up & Notation

Our scientific goal is to evaluate whether a set of candidate tailoring variables may be useful in defining a CAI that will optimize individual-level counterfactual outcomes across the population of interest. To coincide with our motivating example we assume that we have collected multi-stage data for $k = 1, ..., K$ stages from a Clustered SMART with $i = 1, ..., N$ clusters and n_i individuals ($j = 1, ..., n_i$) within each cluster for a total of $\sum_{i=1}^{N} n_i = n$ individuals. Due to the many possible Clustered SMART designs (Figure 3.1), some of which re-randomize only a subset of the clusters randomized at the first stage, we can further denote the number of clusters at each stage as N_k and the total number of individuals within those clusters as n_k. We denote $\mathbf{X}_{k,ij}$ as a p_k-dimensional vector of covariates collected prior to randomization at stage k, i.e., $\mathbf{X}_{1,ij}$ refers to all baseline individual and cluster-level covariates whereas $\mathbf{X}_{2,ij}$ refers to all individual and cluster-level covariates measured at some period of time following the first but prior to the second-stage intervention. Note that $\mathbf{X}_{k,ij}$ may include both individual-level and cluster-level covariates, as well as individual-level covariates aggregated to the cluster level. $A_{k,i}$ refers to the intervention received by cluster i at stage k. Although these methods will accommodate more than two interventions per stage, for simplicity we assume $A_{k,i} \in \{-1, 1\}$. A cluster-level response following intervention $A_{k,i}$ is denoted $R_{k,i}$, often classified as a binary response (or non-response) to intervention. We next define $\mathbf{H}_{k,ij}$ to represent "full history", which includes individual-level and cluster-level covariate data collected starting at baseline through the randomization at stage k, as well as all interventions ($\mathbf{A}$) administered and all responses ($\mathbf{R}$) recorded for all stages $1, ..., k-1$. For example, $\mathbf{H}_{1,ij} = \{\mathbf{X}_{1,ij}\}$, $\mathbf{H}_{2,ij} = \{\mathbf{X}_{1,ij}, \mathbf{X}_{2,ij}, A_{1,i}, R_{1,i}\}$, and so forth. Here we note that, due to the cluster-level randomization, $A_{k,ij} = A_{k,i}$ and $R_{k,ij} = R_{k,i}$ for all

individuals j treated at cluster i. Similarly, all cluster level covariates in $\mathbf{X}_{k,ij}$ are the same for all individuals j within the same cluster i. We can therefore also identify "cluster-level history" defined as $\mathbf{H}_{k,i}$, which includes only cluster-level covariates, interventions and responses collected prior to the stage k intervention decision. Finally, we assume that intermediate, individual-level outcomes $Y_{k,ij}$ may be collected across stages, with $Y_{k,ij} \in \mathbf{X}_{k+1,ij}$. A final, individual-level outcome Y_{ij} is evaluated at the end of the K stages of intervention and may be a function of the stage-specific outcomes, i.e., $Y_{ij} = g(Y_{1,ij}, ..., Y_{K,ij})$ for a known function $g(.)$. For example, $Y_{ij} = Y_{K,ij}$ or $Y_{ij} = \sum_{k=1}^{K} Y_{k,ij}$. For the remainder of our exposition, we assume larger values of Y are desirable such that the goal would be to obtain a maximum value. Let $\mathbf{d}(\mathbf{H}_i) = \{d_1(\mathbf{H}_{1,i}), ..., d_K(\mathbf{H}_{K,i})\}$ be a sequence of decision rules making up a CAI; each stage-specific decision rule $d_k(\mathbf{H}_{k,i})$ is a function only of up-to-date intervention and cluster-level covariate history $\mathbf{H}_{k,i}$ that can be used to make decisions pertaining to interventions at each stage, i.e., $d_k(\mathbf{H}_{k,i}) : \mathbf{H}_{k,i} \rightarrow A_{k,i}$.

Because we are interested in making causal inference pertaining to the importance of a set of candidate prescriptive variables in tailoring intervention, we utilize the potential outcomes framework (*Rubin*, 1974). Let $Y^*(\mathbf{d})$ represent the individual-level counterfactual outcome (also known as a potential outcome) consistent with the CAI $\mathbf{d}$. In a single stage setting with two cluster-level intervention options, e.g., $A \in \{-1, 1\}$, there would be two potential outcomes for each individual, $Y^*(-1)$ and $Y^*(1)$, which represent the individual-level outcome that would be observed had the individual been treated in a cluster with intervention -1 or intervention 1, respectively. Notably, only one of these counterfactual outcomes–the one compatible with the intervention received by the cluster in which the individual was treated–would be observed.

Because the target of our estimation is a counterfactual outcome Y^* but data collected in a Clustered SMART includes observed outcomes Y for only one of the

potential outcomes, various assumptions are needed to link the observed data with the counterfactual outcome. These assumptions include consistency, positivity, and no unmeasured confouders (NUCA). Because our data are collected from a Clustered SMART with randomization events that balance and re-balance baseline and time-varying characteristics, positivity and NUCA are assumed by design and consistency is a reasonable assumption in most cases, although these assumptions can be explored in greater depth (e.g., *Hernan and Robins*, 2020). Under these assumptions, it can be shown for a single stage regimen that $E\{Y^*(d)\} = E_{\mathbf{H}}[\sum_a E(Y|A = a, \mathbf{H})I\{d(\mathbf{H}) = a\}]$, which provides the necessary link between the potential and observed outcomes in order to perform estimation and inference.

3.2.2 Approach: Clustered Q-Learning

We introduce Clustered Q-Learning, which is developed from standard Q-Learning (*Chakraborty et al.*, 2013; *Moodie et al.*, 2012; *Murphy*, 2005a; *Nahum-Shani et al.*, 2012; *Schulte et al.*, 2014), a popular method used to estimate personalized, multi-stage, individual-level interventions. Clustered Q-Learning, in contrast, can estimate a CAI and identify tailoring variables across stages when the outcome is measured at the individual level but the intervention is applied at the level of the cluster. Clustered Q-Learning, similar to standard Q-Learning, utilizes the language of Q-functions, which are defined for each stage k as the expected outcome conditional on covariate and intervention history collected through the k-th stage. The K-stage Q-function, for example, is $Q_{K,ij}(\mathbf{H}_{K,ij}, A_{K,i}) = E(Y_{ij}|\mathbf{H}_{K,ij}, A_{K,i})$. A Q-function is said to be "optimal" if the expected counterfactual outcome is maximized. Due to the possibility of intervention effect confounding that may occur in a multi-stage estimation when conditioning on both the stage-specific intervention(s) of interest as well as intermediate variable(s), estimation of stage-specific decision rules in Q-Learning proceeds in a backwards recursive manner starting with the final stage K. The stage K optimal Q-

function can be represented as $Q^{\text{opt}}_{K,ij}(\mathbf{H}_{K,ij}, A_{K,i}) = \sup_{a_K} E(Y_{ij}|\mathbf{H}_{K,ij}, A_{K,i})$. The optimal Q-functions at all stages prior to the final stage, i.e., $k = 1, ..., K-1$, also known as the k-stage pseudo-outcomes $\tilde{Y}_{k,ij}$, can be understood as the predicted counterfactual outcome at stage k when receiving the optimal intervention(s) at all future stages. The optimal Q-function for the $K-1$ stage, therefore, is $Q^{\text{opt}}_{K-1,ij}(\mathbf{H}_{K-1,ij}, A_{K-1,i}) = E\{\sup_{a_K} Q^{\text{opt}}_{K,ij}(\mathbf{H}_{K,ij}, a_{K,i})|\mathbf{H}_{K-1,ij}, A_{K-1,i}\}$ and the optimal Q-function for the k-th stage generally is $Q^{\text{opt}}_{k,ij}(\mathbf{H}_{k,ij}, A_{k,i}) = E\{\sup_{a_{k+1}} Q^{\text{opt}}_{k+1,ij}(\mathbf{H}_{k+1,ij}, a_{k+1,i})|\mathbf{H}_{k,ij}, A_{k,i}\}$.

3.2.3 Estimation

There are many ways to estimate Q-functions, including linear or nonparametric regression models with or without regularization, splines, neural networks, trees, etc. (*Shortreed et al.*, 2011; *Song et al.*, 2015; *Zhao et al.*, 2009). For the remainder of this manuscript we focus on modeling Q-functions using standard linear regression models given their ease of implementation and interpretability, as well as the broad, general understanding of regression methods across the scientific community.

Assuming the continuous, individual-level outcome Y_{ij} has a distribution that is approximately symmetric and the mean is linear in the regression parameters, we connect the k-stage Q-function with a linear model as follows: $Q_{k,ij}(\mathbf{H}_{k,ij}, A_{k,i}; \beta_k, \Psi_k) = \mathbf{H}_{k0,ij}\boldsymbol{\beta}_k + (\mathbf{H}_{k1,i}\boldsymbol{\Psi}_k)A_{k,i}$ with the error defined as $\epsilon_{k,ij} = Q_{k,ij}(\mathbf{H}_{k,ij}, A_{k,i}; \beta_k, \Psi_k) - \{\mathbf{H}_{k0,ij}\boldsymbol{\beta}_k + (\mathbf{H}_{k1,i}\boldsymbol{\Psi}_k)A_{k,i}\}$. We assume $\epsilon_{k,ij} \perp \epsilon_{k,i'j}$ for all individuals j when $i \neq i'$. Thus we express the within-cluster, k-stage error for cluster i as $\boldsymbol{\epsilon}^T_{k,i} = (\epsilon_{k,i1}, \epsilon_{k,i2}, ..., \epsilon_{k,in_i})$. We assume $E(\boldsymbol{\epsilon}_{k,i}|\mathbf{H}_{k,ij}) = \mathbf{0}$ and that the n_i x n_i dimensional matrix for the k-stage, cluster level covariance is $\text{Cov}(\boldsymbol{\epsilon}_{k,i}|\mathbf{H}_{k,ij}) = \boldsymbol{\Sigma}_{k,i}(\boldsymbol{\theta}_k)$ where $\boldsymbol{\theta}_k$ denotes the set of parameters defining the covariance matrix. For example, $\theta_k = \rho$ under the assumption of an exchangeable correlation structure for outcomes within the same cluster. We assume parameters are common across clusters, i.e., that $\boldsymbol{\theta}_{ki} = \boldsymbol{\theta}_{ki'}$, $\boldsymbol{\beta}_{ki} = \boldsymbol{\beta}_{ki'}$, and $\boldsymbol{\Psi}_{ki} = \boldsymbol{\Psi}_{ki'}$ for all $i \neq i'$. Because our scientific

interest lies in identifying cluster-level tailoring variables which, by definition, will have a qualitative interaction with the intervention A_k, it can be seen that we distinguish between parameters $\boldsymbol{\Psi}_k$ and $\boldsymbol{\beta}_k$. $\boldsymbol{\Psi}_k$ reflect the effects of cluster-level tailoring variables whereas $\boldsymbol{\beta}_k$ may be considered "nuisance" relative to our estimation goals. Variables in $\mathbf{H}_{k1}$ must be variables measured at the cluster-level whereas $\mathbf{H}_{k0}$ may include both individual-level and cluster-level covariates, all of which may be transformed (e.g., mean centered) as needed. Specific decisions about covariates to include in $\mathbf{H}_{k0}$ and $\mathbf{H}_{k1}$ should be made a priori (*Pocock et al.*, 2002) in a similar manner as is conventionally performed with the analysis of randomized controlled trials (RCTs). Although there is less guidance about covariate adjustment in analyzing intervention effects in cluster-randomized trials than for conventional RCTs, it is generally recommended that covariate(s) prognostic of the outcome, including those used to stratify randomization, should be included in $\mathbf{H}_{k0}$ (*European Agency for the Evaluation of Medicinal Products*, 2003; *ICH E Expert Working Group*, 1999; *Pocock et al.*, 2002; *Raab et al.*, 2000), although we note that, in a linear regression model, both adjusted and unadjusted intervention effect estimates lead to unbiased estimation. It should also be considered whether to include effects of an individual-level covariate at both the individual and cluster level as these may not coincide and variation of a covariate is likely to exist at both the individual level and cluster level (*Wright*, 2015). Our methodology provides freedom to make modeling choices for the Q-functions and it should be understood that different scenarios necessitate different modeling decisions (e.g., refer to *European Agency for the Evaluation of Medicinal Products*, 2003; *Pocock et al.*, 2002; *Raab et al.*, 2000; *Wright et al.*, 2015).

Due to the backward recursive manner in which the optimal, stage-specific Q-functions are defined, it is reasonable that estimation in the Q-Learning context is also performed recursively, beginning with estimation of the final stage K and continuing through all prior stages $K-1, ..., 1$, until the multi-stage estimation is

complete. Under the above assumptions then, we perform estimation as follows:

1. Estimate parameters for Stage K: Using standard regression techniques to minimize residual sum of squares between the overall outcome Y_{ij} and the final stage Q-function $Q_K(\mathbf{H}_K, A_K)$, estimate: $(\hat{\boldsymbol{\beta}}_K, \hat{\boldsymbol{\Psi}}_K) = \text{argmin}_{\boldsymbol{\beta}_K, \boldsymbol{\Psi}_K} \sum_{i=1}^{N_2} \Big[\mathbf{Y}_i - \{\mathbf{H}_{K0,i}\boldsymbol{\beta}_K + (\mathbf{H}_{K1,i}\boldsymbol{\Psi}_K)A_{K,i}\}\Big]^T \boldsymbol{\Sigma}_{K,i}^{-1}(\boldsymbol{\theta}_K)\Big[\mathbf{Y}_i - \{\mathbf{H}_{K0,i}\boldsymbol{\beta}_K + (\mathbf{H}_{K1,i}\boldsymbol{\Psi}_K)A_{K,i}\}\Big]$, for clusters $i = 1, ..., N_2$, where $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})^T$, $\boldsymbol{\beta}_K = (\beta_1, ..., \beta_{p_{K0}})^T$, $\mathbf{H}_{K0,i} = (\mathbf{H}_{K0,i1}^T, ..., \mathbf{H}_{K0,in_i}^T)^T$ and $\mathbf{H}_{K0,ij} = (H_{K0,ij1}, ..., H_{K0,ijp_{K0}})$ for p_{K0} distinct stage K predictive covariates. Similarly, $\boldsymbol{\Psi}_K = (\Psi_1, ..., \Psi_{p_{K1}})^T$, $\mathbf{H}_{K1,i} = (\mathbf{H}_{K1,i1}^T, ..., \mathbf{H}_{K1,in_i}^T)^T$ and $\mathbf{H}_{K1,ij} = (H_{K1,ij1}, ..., H_{K1,ijp_{K1}})$ for p_{K1} distinct stage K prescriptive covariates. When $\boldsymbol{\Sigma}_{K,i}(\boldsymbol{\theta}_K) = \text{Cov}(\boldsymbol{\epsilon}_{K,i}|\mathbf{H}_{K,i})$ is unknown, the parameters $\boldsymbol{\theta}_K$ defining $\boldsymbol{\Sigma}_{K,i}(\boldsymbol{\theta}_K)$ can be estimated empirically using restricted maximum likelihood (REML), for example, based on an assumed working covariance structure (e.g., exchangeable).

2. For Stages $k = K-1, ..., 1$:

   (a) Calculate the estimated Stage k pseudo-outcome $\tilde{Y}_{k,ij}$ for all individuals j within cluster i, noting that the pseudo-outcome will depend on the function $g(\cdot)$ that defines the relationship between the overall outcome Y_{ij} and intermediate outcomes $Y_{1,ij}, ..., Y_{K,ij}$. When $Y_{ij} = Y_{K,ij}$, we have: $\tilde{Y}_{k,ij} = \max_{a_{k+1,i}}\{Q_{k+1,ij}(\mathbf{h}_{k+1,ij}, a_{k+1,i}; \hat{\boldsymbol{\beta}}_{k+1}, \hat{\boldsymbol{\Psi}}_{k+1})\} = \max_{a_{k+1,i}}\{\mathbf{h}_{k+1,0,ij}\,\hat{\boldsymbol{\beta}}_{k+1} + (\mathbf{h}_{k+1,1,i}\,\hat{\boldsymbol{\Psi}}_{k+1})a_{k+1,i}\}$.

   (b) Estimate parameters for Stage k: Using standard regression techniques to minimize residual sum of squares between the k stage pseudo-outcome $\tilde{Y}_{k,ij}$ and the Stage k Q-function, $Q_{k,ij}(\mathbf{H}_{k,ij}, A_{k,ij})$, estimate: $(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\Psi}}_k) = \text{argmin}_{\boldsymbol{\beta}_k, \boldsymbol{\Psi}_k} \sum_{i=1}^{N_1} \Big[\tilde{\boldsymbol{Y}}_{k,i} - \{\mathbf{H}_{k0,i}\boldsymbol{\beta}_k + (\mathbf{H}_{k1,i}\boldsymbol{\Psi}_k)A_{k,i}\}\Big]^T \boldsymbol{\Sigma}_{k,i}^{-1}(\boldsymbol{\theta}_k)\Big(\tilde{\boldsymbol{Y}}_{k,i} - (\mathbf{H}_{k0,i}\boldsymbol{\beta}_k + (\mathbf{H}_{k1,i}\boldsymbol{\Psi}_k)A_{k,i})\Big)$ for clusters $1, ..., N_1$, and vectors $\tilde{\mathbf{Y}}_{k,i} = (\tilde{Y}_{k,i1}, ..., \tilde{Y}_{k,in_i})^T$, with $\boldsymbol{\beta}_k$, $\mathbf{H}_{k0,i}$, $\mathbf{H}_{k0,ij}$, $\boldsymbol{\Psi}_k$, $\mathbf{H}_{k1,i}$, and $\mathbf{H}_{k1,ij}$ defined as in

Step 1 above but adjusted to represent the k-th (not the K-th) stage. As above, the parameters $\boldsymbol{\theta}_k$ defining $\boldsymbol{\Sigma}_{k,i}(\boldsymbol{\theta}_k) = \text{Cov}(\boldsymbol{\epsilon}_{k,i}|\mathbf{H}_{k,i})$ can be estimated empirically using REML based on an assumed working covariance structure.

3.2.4 Inference

Conducting inference in a single stage estimation with clustered data is straightforward; standard regression techniques that account for correlation of outcomes within clusters are used for estimation and inference. Estimation in a multi-stage setting, however, can become problematic. Specifically, due to the fact that estimation proceeds in a backward manner and, in the Q-Learning approach, involves regression on a pseudo-outcome that accounts for the optimal intervention received at all future stages, likelihood-based estimation at all stages prior to the last may involve maximization of a non-smooth function.

This problem of nonregularity is well-described in the literature (*Chakraborty and Moodie*, 2013; *Chakraborty et al.*, 2010; *Laber et al.*, 2014; *Robins*, 2004). Formally, in a multi-stage intervention setting, nonregularity exists in the estimation of parameters indexing the optimal adaptive intervention for any stage prior to the last stage when there is positive probability that the final stage intervention effect is zero, i.e., $P(\mathbf{H}_{K,i} : \boldsymbol{\Psi}_K^T \mathbf{H}_{K1,i} = 0) > 0$. This will occur when there is a non-unique intervention effect for at least some subset of clusters at Stage K–or any stage following the k-th stage being estimated–because the limiting distributions may vary depending on the intervention effect at the $k+1$ and future stages and, consequently, standard asymptotic theory may not apply. We exhibit this problem in the context of a 2-stage Clustered SMART with two intervention options at each stage. Using the notation introduced above, the Stage 2 Q-function is defined as $Q_{2,ij}(\mathbf{H}_{2,ij}, A_{2,i}) = E(Y_{ij}|\mathbf{H}_{2,ij}, A_{2,i})$. If we assume that our overall outcome of inter-

est is equal to the outcome observed following stage 2, i.e., $Y = Y_2$, the first stage Q-function is $Q_{1,ij}(\mathbf{H}_{1,ij}, A_{1,i}) = E(\max_{a_2} Q^{\text{opt}}_{2,ij}(\mathbf{H}_{2,ij}, a_{2,i})|\mathbf{H}_{1,ij}, A_{1,i})$. Given our assumed parametric model for the Q-functions, we can further express the Stage 1 Q-function as $Q_{1,ij}(\mathbf{H}_{1,ij}, A_{1,i}) = E(\hat{\boldsymbol{\beta}}_2^T \mathbf{H}_{20,ij} + |\hat{\boldsymbol{\Psi}}_2^T \mathbf{H}_{21,i}| \,|\, \mathbf{H}_{1,ij}, A_{1,i})$. Notably, the Stage 1 estimation now contains a non-smooth function, i.e., the absolute value function that maximizes the linear combination of the Stage 2 Q-function corresponding to the prescriptive cluster-level variables. Thus, the Stage 1 Q-function is non-smooth and its smoothness depends on $\hat{\boldsymbol{\Psi}}_2^T \mathbf{H}_{21}$. It can be seen that if $\hat{\boldsymbol{\Psi}}_2^T \mathbf{H}_{21}$ is far away from zero for all clusters, standard asympotitic theory will ensure consistent estimation of the Stage 1 Q-function. If $\hat{\boldsymbol{\Psi}}_2^T \mathbf{H}_{21}$ is zero or near zero for at least some subset of clusters, the non-smoothness of the Stage 1 Q-function may impede the use of standard asymptotics in finite samples.

Multiple solutions that have been developed to account for nonregular estimators. Several authors have proposed a threshold estimator (*Chakraborty et al.*, 2010; *Moodie and Richardson*, 2010) that can be used in conjunction with a standard nonparametric bootstrap procedure (*Efron*, 1979) to estimate the standard error of nonregular parameters. The threshold estimators are used as an attempt to "regularize" a nonregular estimator, i.e., by shrinking the effect of the non-smooth function toward zero and thereby reducing the degree of nonregularity (*Chakraborty et al.*, 2010; *Moodie and Richardson*, 2010). While these methods do not demonstrate asymptotic regularity in limiting distribution, they demonstrate reasonable empirical finite sample performance in simulation experiments. An alternate approach is that of adaptive confidence intervals (ACI) which, under local asymptotic theory and by taking a supremum of the non-smooth functional, demonstrates a limiting distribution that is regular and asymptotically normal (*Laber et al.*, 2014). Although ACI is currently the only known method that demonstrates regularity and asymptotic normality in limiting distribution, it is known to be quite complicated both theoretically and com-

putationally. Another approach to estimating confidence intervals for nonregular parameters with Q-Learning is the m-out-of-n bootstrap (*Chakraborty et al.*, 2013), originating from a technique used for the estimation of confidence sets of non-smooth functionals (*Bickel et al.*, 1997; *Shao*, 1994). As is generally well understood, the standard nonparametric bootstrap is often used as an alternative to estimate standard errors of an estimator that has an uncertain parametric distribution or when calculation of standard errors is challenging. In the standard bootstrap, the empirical distribution function $F_N(x)$ converges to the true generative distribution $F(x)$ as $n \to \infty$. The m-out-of-n bootstrap, however, uses a resample size m, which is of a smaller order of magnitude than n. It has been shown that the empirical distribution converges to the true generative distribution at a faster rate than the bootstrap empirical distribution converges to the empirical distribution. Given this unique construct, the bootstrap resamples are reflective of the true generative distribution. *Chakraborty et al.* (2013) demonstrated that this method produces consistent confidence sets under fixed alternatives and performs well in the estimation of confidence intervals for nonregular parameters indexing the optimal adaptive intervention in standard Q-Learning.

Herein we propose the M-out-of-N cluster bootstrap for clustered data, adapted from the work of *Chakraborty et al.* (2013), for estimating confidence intervals for parameters indexing the optimal cluster-level adaptive intervention in the setting of parametric Clustered Q-learning. The cluster bootstrap performs resampling at the cluster level rather than the individual level, which is critical when estimating the degree of variability of an estimator in the presence of correlated data (*Hox*, 2010). Assuming that the number of clusters is large (*Field and Welsh*, 2007), model errors are uncorrelated across clusters but correlated within clusters, clusters are exchangeable (*Bouwmeester et al.*, 2013), and the empirical distribution $F_N(x)$ is a reasonable approximation to the underlying distribution $F(x)$, the cluster bootstrap is

asympotically consistent. The choice of M clusters, with $M \leq N$, reflects the degree of nonregularity in the underlying data. The degree of nonregularity in the k-th stage estimation is assessed using estimated model parameters of the $k+1$ stage. In the case of a 2-stage estimation, for example, $\hat{\mathbf{\Psi}}_2$, the estimated parameters associated with the Stage 2 tailoring variables, are used to estimate $p_2 = P(\mathbf{H}_{21} : \mathbf{H}_{21}\mathbf{\Psi}_2 = 0)$, the degree of nonregularity corresponding to the linear combination of the prescriptive cluster-level covariates at Stage 2. Extended from *Chakraborty et al.* (2013), to estimate the resample size M_k in the k-th stage estimation ($1 \leq k < K$), a T-statistic is calculated for each of N_{k+1} clusters as $T_{k,i} = \dfrac{\mathbf{H}_{k+1,1i}\hat{\mathbf{\Psi}}_{k+1}}{\hat{\text{se}}(\mathbf{H}_{k+1,1i}\hat{\mathbf{\Psi}}_{k+1})}$, with the standard errors derived using the sandwich variance estimator. Using a pre-defined threshold η, the proportion of clusters with an absolute T-statistic below η is calculated: $\hat{p}_{k+1} = \dfrac{1}{N_{k+1}} \displaystyle\sum_{i=1}^{N_{k+1}} I(|T_{k,i}| \leq \eta)$. Using $\hat{p}_{k+1}$ as the estimated degree of nonregularity in the Stage k estimation, a value M_k, the number of resamples from N_k clusters to be used in estimation of the Stage k parameters, is selected as $\hat{M}_k = N_k^{f(p_{k+1})}$. As suggested by *Chakraborty et al.* (2013), $f(p_{k+1})$ can be modeled by a simple function that is monotone decreasing in p_{k+1}, takes values in $(0, 1]$, $f(0) = 1$, and is continuous with a bounded first derivative. Using $f(p_{k+1}) = \frac{1+\chi(1-p_{k+1})}{1+\chi}$ with tuning parameter χ, $\hat{M}_k = N_k^{(1+\chi-\chi\hat{p}_{k+1})(1+\chi)^{-1}}$. In a highly regular setting, i.e., when there is a strong stage $k+1$ intervention effect for all clusters, $p_{k+1} = 0$ and, consequently, $M_k = N_k$, which represents the standard cluster bootstrap. With increasingly higher degrees of nonregularity observed in estimation at the $k+1$ stage, the value of M_k decreases relative to N_k.

3.3 Implementation

We implement this method using a modification of the `Q-Learn` package (*Chakraborty et al.*, 2013) in `R` (*R Core Team*, 2018) to accommodate clustered data. For simplicity we demonstrate implementation using data from a two-stage Clustered SMART with

a continuous outcome that is approximately normally distributed.

Estimate Stage 1 and Stage 2 parameters:

1. Using only those (n_2) observations from (N_2) clusters treated at Stage 2, estimate $\boldsymbol{\Psi}_2$, the prescriptive parameters in the second stage Q-function, using a linear regression model that accommodates continuous outcome data (e.g., `lm` or `geeglm` function in R) by regressing the overall outcome Y on Stage 2 covariates.

2. Estimate the Stage 1 pseudo-outcome $\tilde{Y}_{1,ij} = \hat{\boldsymbol{\beta}}_2^T \mathbf{H}_{20,ij} + |\hat{\boldsymbol{\Psi}}_2^T \mathbf{H}_{21,i}|$.

3. Using all (n_1) observations from $N_1 = N$ clusters treated at Stage 1, estimate $\boldsymbol{\Psi}_1$, the prescriptive model parameters in the first stage Q-function using a linear regression model that accommodates continuous outcome data (e.g., `lm` or `geeglm` function in R) by regressing Stage 1 pseudo-outcome $\tilde{Y}$ on Stage 1 covariates.

Estimate confidence intervals for Stage 2 parameters:

4. Draw B independent cluster bootstrap samples of size N_2 drawn from N_2 clusters with replacement so that the probability of selecting any of the N_2 clusters is $1/N_2$. Construct the b-th bootstrap sample as the set of $N_2 \cdot \sum_{i \in b} n_i$ observations, where i represents the index for the cluster selected in the b-th bootstrap sample. All observations contained within each cluster selected in the b-th bbootstrap sample are included in the b-th bootstrap sample.

5. For $b = 1$ to B, estimate the prescriptive parameters for each of the B bootstrap replicates in the regression model, i.e., $\hat{\boldsymbol{\Psi}}_{\mathbf{2}}^{(b)}$.

6. For each parameter, $1, .., p_{2,1}$, in $\hat{\boldsymbol{\Psi}}_{\mathbf{2}}$ and given a pre-specified significance level α reflecting the maximum desired Type I error, use the B bootstrap estimates

$(\hat{\mathbf{\Psi}}_{\mathbf{2}}{}^{(1)}, ..., \hat{\mathbf{\Psi}}_{\mathbf{2}}{}^{(B)})$ to estimate the desired lower and upper bounds of the empirical distribution using the $(\frac{\alpha}{2}, 1-\frac{\alpha}{2})$ sample quantiles of the reverse percentile bootstrap estimates.

Estimate confidence intervals for Stage 1 parameters:

7. Using $\hat{\mathbf{\Psi}}_2$ calculated in Step 1 above, estimate $p_2 = P[\mathbf{H}_{21} : \mathbf{H}_{21}\mathbf{\Psi}_2 = 0]$, the degree of nonregularity corresponding to the estimated linear combination of the Stage 2 intervention effect of the cluster-level covariates. Calculate a T-statistic for each of N_2 clusters as $T_{1,i} = \dfrac{\mathbf{H}_{21,i}\hat{\mathbf{\Psi}}_2}{\hat{\text{se}}(\mathbf{H}_{21,i}\hat{\mathbf{\Psi}}_2)}$, with the standard errors derived using the sandwich variance estimator. Using a pre-defined threshold η, calculate the proportion with a T-statistic below η: $\hat{p}_2 = \dfrac{1}{N_2}\sum_{i=1}^{N_2} I(|T_{1,i}| \leq \eta)$. Using $f(p) = \frac{1+\chi(1-p)}{1+\chi}$ with tuning parameter χ, select a value M_1, the number of resamples from N_1 clusters, with $M_1 \leq N_1$, as $\hat{M}_1 = N_1^{(1+\chi-\hat{p_2}\chi)(1+\chi)^{-1}}$.

8. Select B independent cluster bootstrap samples of size M_1 drawn from N_1 clusters with replacement so that the probability of selecting any of the N_1 clusters is $1/N_1$. Construct the b-th bootstrap sample as the set of $M_1 \cdot \sum_{i \in b} n_i$ observations, where i represents the index for the cluster selected in the b-th bootstrap sample. All observations within each cluster selected in the b-th bootstrap sample are included in the b-th bootstrap sample.

9. Following Steps 1-3 above: For each bootstrap resample $b = 1, ..., B$, estimate $\hat{\mathbf{\Psi}}_2^{(b)}$. Use $\hat{\mathbf{\Psi}}_2^{(b)}$ to construct the Stage 1 pseudo-outcome $\tilde{Y}_{1,ij}^{(b)}$. Estimate $\hat{\mathbf{\Psi}}_1^{(b)}$.

10. For each parameter, $1, .., p_{1,1}$, in $\hat{\mathbf{\Psi}}_{\mathbf{1}}$ and given a pre-specified significance level α, use the B bootstrap estimates $(\hat{\mathbf{\Psi}}_{\mathbf{1}}{}^{(1)}, ..., \hat{\mathbf{\Psi}}_{\mathbf{1}}{}^{(B)})$ to estimate the desired lower and upper bounds of the empirical distribution using the $(\frac{\alpha}{2}, 1-\frac{\alpha}{2})$ sample quantiles of the reverse percentile bootstrap estimates.

3.3.1 Tuning Parameter Selection for M-out-of-N Cluster Bootstrap

Two tuning parameters are involved in the selection of resample size M: η, which represents the threshold at which nonregularity is detected for the linear combination of the Stage 2 intervention effects and is used in the estimation of the degree of nonregularity $\hat{p}$; and χ, used within the function $f(\hat{p})$ to calculate the number of resamples M. Choice of η is straightforward as it represents a quantile from the t-distribution with ν degrees of freedom and a pre-specified significance level α corrected for multiple hypothesis testing. For example, given $\alpha = 0.08$ and assuming $N = 80$ clusters, we desire a maximum Bonferroni-adjusted Type I error of 0.001, which corresponds to $t_{\nu,1-\alpha/(N*2)} = 3.42$. The tuning parameter $\chi \in \{0.025, 0.05, ..., 1\}$, for example, can be selected in a data-driven manner using a double bootstrap algorithm (*Chakraborty et al.*, 2013). Empirically, we observed that lower values of χ, i.e., $\chi \in \{0.025, 0.05\}$ tended to provide estimated coverage closest to nominal with larger values of χ (e.g., $\chi = 0.10$) tending to overcoverage.

3.4 Simulation Experiments

3.4.1 Simulation Setup

To evaluate whether our proposed method is able to estimate the regression parameters corresonding to the multi-stage candidate prescriptive variables with a low degree of bias and their associated confidence intervals with near nominal coverage in settings representing varying degrees of nonregularity, we generate data for a two-stage Clustered SMART. In Simulation 1, we explore performance with a large number of clusters and a fixed number of individuals per cluster ($N = 80; n_i = 20$). In Simulation 2 we evaluate performance with a smaller number of clusters and a larger, fixed number of individuals per cluster (i.e., $N = 20, n_i = 80$). Finally, in Simulation 3 we investigate performance differences with a larger number of clusters

($N = 80$) but a variable number of individuals per cluster.

Data is generated for all simulations as follows. One binary candidate cluster-level tailoring variable X_1 is generated at baseline. Randomization to the first stage intervention A_1 occurs in a ratio of 1:1, with approximately half of the clusters assigned to one of two interventions. An intermediate response and candidate Stage 2 tailoring variable, i.e., X_2, which is also binary, depends on the baseline cluster-level covariate X_1 and the first stage intervention A_1. Intervention A_2 may depend on an intermediate response, e.g., X_2, although we assume equal probability of random assignment also to the second stage intervention A_2. The final, individual-level outcome observed following the second stage, Y, is assumed to be continuous and approximiately normally distributed, with larger values desired. Correlation of within-cluster outcomes is achieved using an intraclass correlation coefficient ρ. One particular advantage of this data generating schema is the fact that it allows the true stage 1 parameters, $\boldsymbol{\Psi}_1$ and $\boldsymbol{\beta}_1$, to exist in closed form, permitting straightforward estimates of bias and coverage. The exact specifications used in the simulation experiements are provided below.

- $X_1 \sim \text{Bern}(p = 0.5)$, $X_1 \in \{-1, 1\}$
- $A_1 \sim \text{Bern}(p = 0.5)$, $A_1 \in \{-1, 1\}$
- $P[X_2 = 1|X_1, A_1] = 1 - P[X_2 = 0|X_1, A_1] = \text{expit}(\delta_1 X_1 + \delta_2 A_1)$, where $\text{expit}(x) = \exp(x)/\{1 + \exp(x)\}$, $X_2 \in \{-1, 1\}$
- $A_2|X_2 \sim \text{Bern}(p = p_{a_1,x_2})$, $A_2 \in \{-1, 1\}$, $p_{a_1,x_2} \in (0, 1)$, $p_{r_{a_1,x_2}} \perp p_{s_{a_1,x_2}}$ for $r \neq s$, indicating that the second-stage randomizations may have different probability of random assignment depending on the intervention A_1 received and the value of X_2 observed. Although the randomization probability can be modified, in our simulation studies we assume $P[A_2|X_2] = 0.5$.

- $\mathbf{Y}_i = \gamma_1 + \gamma_2 X_{1,i} + \gamma_3 A_{1,i} + \gamma_4 X_{1,i} A_{1,i} + \gamma_5 A_{2,i} + \gamma_6 X_{2,i} A_{2,i} + \gamma_7 A_{1,i} A_{2,i} + \boldsymbol{\epsilon}_i$; $\boldsymbol{\epsilon}_i \sim \mathbf{N}_{n_i}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is defined with an exchangeable correlation structure using a prespecified $\text{Cov}(\epsilon_{ij}, \epsilon_{ij'}) = \rho$ for all individuals within cluster i with $j \neq j'$.

Following *Chakraborty et al.* (2010) and *Laber et al.* (2014), we evaluate nine distinct data generating mechanisms that reflect different underlying clinical assumptions and varying degrees of nonregularity. We present results that illustrate fully "regular" settings, in which no problems in confidence interval coverage rates should be observed at either stage, "nonregular" settings in which estimated coverage at the first stage may be adversely affected if not properly accounted for methodologically, and "near-nonregular" settings that would technically be classified as regular settings but may be nearly indistinguishable from a nonregular setting. We induce varying degrees of nonregularity in our estimation by exploiting different combinations of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ specified in the data generating mechanisms above (*Chakraborty et al.*, 2010; *Laber et al.*, 2014). Specifically, we can generate nonregular settings when the linear combination $\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1 = 0$ with positive probability, i.e., $[P(\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1) = 0] > 0$. Thus, we define p and ζ as follows, both of which represent two dimensions of the nonregularity phenomenon, as described in *Chakraborty et al.* (2010): $p = P[\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1 = 0]$, and the "standardized effect size" is defined as $\zeta = [E[\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1] / \sqrt{Var[\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1]}]$. Refer to Table 3.1 for a list of the parameter specifications of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, and corresponding values of p and ζ, used within these simulation experiments. The closed form true values for $\boldsymbol{\beta}_1$ and $\boldsymbol{\Psi}_1$ are also provided (Table 3.1).

As introduced above, the nine data generating settings with varying choices of $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ reflect different underlying, clinical settings (*Chakraborty et al.*, 2010; *Laber et al.*, 2014). Example 1 simulates a scenario in which no cluster experiences any effect of intervention at either stage, which is considered a fully nonregular scenario (all

$\boldsymbol{\gamma} = 0$). Example 2 simulates a very weak effect of Stage 2 intervention ($\gamma_5 = 0.01$), but this can be classified as "near-nonregular" in that the weak effect may be masked by the degree of noise in the data generating process. Example 3 presents a scenario in which there is a reasonably large effect of Stage 2 intervention for about half of the clusters, but an effect is absent for the other half ($\gamma_3 = -0.5, \gamma_5 = \gamma_7 = 0.5$). Example 4 builds upon Example 3 in that half of the clusters retain their relatively large Stage 2 intervention effect while the remainder now have a small Stage 2 intervention effect ($\gamma_3 = -0.5, \gamma_5 = 0.5, \gamma_7 = 0.49$). Example 5 modifies the proportion of clusters with a relatively large Stage 2 intervention effect while the remaining clusters have a very small Stage 2 intervention effect ($\gamma_3 = -0.5, \gamma_5 = 1.0, \gamma_6 = \gamma_7 = 0.50$). Example 6 reflects a "regular" setting in which Stage 2 intervention effects are generated for all clusters ($\gamma_3 = -0.5, \gamma_5 = 0.25, \gamma_6 = \gamma_7 = 0.5$). Example A also simulates a strongly regular setting ($\gamma_3 = -0.25, \gamma_5 = 0.75, \gamma_6 = \gamma_7 = 0.5$). Example B simulates a non-regular setting in which the nonregularity depends on the Stage 1 intervention: Those clusters receiving Intervention A at Stage 1 will have no Stage 2 intervention effect while those receiving Intervention B at Stage 1 have a moderate Stage 2 intervention effect ($\gamma_3 = 0, \gamma_5 = \gamma_7 = 0.25$). Similarly, Example C extends Example B to simulate a weak Stage 2 intervention effect for those clusters receiving Intervention A at Stage 1 ($\gamma_3 = 0, \gamma_5 = 0.25, \gamma_7 = 0.24$).

Table 3.1: Data generating mechanisms for Examples (Ex) 1-9 and A-C. $\boldsymbol{\gamma}$ refers to parameters used to specify the outcome model; $\boldsymbol{\delta}$ refers to parameters used to specify the X_2 variable assignment model (Refer also to Section 4.1). $p = P[\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1 = 0]$ refers to the probability that the linear combination of the Stage 2 prescriptive variables equals 0; $\hat{p} > 0$ represent nonregular settings; $\zeta = [E[\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1]/\sqrt{Var[\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1]}]$ represents the standardized effect size of the linear combination of the Stage 2 prescriptive variables. $\boldsymbol{\beta}_1, \boldsymbol{\Psi}_1 = (\beta_{01}, \beta_{11}, \psi_{10}, \psi_{11})$, which refers to the effects of predictive parameters $\boldsymbol{\beta}$ and prescriptive parameters $\boldsymbol{\Psi}$ consistent with the Stage 1 Q-function $Q_1(\mathbf{H}_1, A_1) = \beta_{10} + \beta_{11} X_1 + (\psi_{10} + \psi_{11} X_1) A_1$.

Ex.	$\boldsymbol{\gamma}$	$\boldsymbol{\delta}$	Type	Regularity Measures	$\boldsymbol{\beta}_1, \boldsymbol{\Psi}_1$
1	$(0,0,0,0,0,0,0)^T$	$(0.5,0.5)^T$	nonregular	$p = 1, \zeta = 0/0$	$(0,0,0,0)$
2	$(0,0,0,0,0.01,0,0)^T$	$(0.5,0.5)^T$	near-nonregular	$p = 0, \zeta = \infty$	$(0.01,0,0,0)$
3	$(0,0,-0.5,0,0.5,0,0.5)^T$	$(0.5,0.5)^T$	nonregular	$p = 0.5, \zeta = 1.00$	$(0.5,0,\sim 0,0)$
4	$(0,0,-0.5,0,0.5,0,0.49)^T$	$(0.5,0.5)^T$	near-nonregular	$p = 0, \zeta = 1.02$	$(0.5,\sim 0,-0.01,\sim 0)$
5	$(0,0,-0.5,0,1.0,0.5,0.5)^T$	$(1.0,0.0)^T$	nonregular	$p = 0.25, \zeta = 1.41$	$(1,0.23,\sim 0,0)$
6	$(0,0,-0.5,0,0.25,0.5,0.5)^T$	$(0.1,0.1)^T$	regular	$p = 0, \zeta = 0.35$	$(0.64,0.01,-0.37,0.02)$
A	$(0,0,-0.25,0,0.75,0.5,0.5)^T$	$(0.1,0.1)^T$	regular	$p = 0, \zeta = 1.035$	$(0.88,0.02,0.14,0.01)$
B	$(0,0,0,0,0.25,0,0.25)^T$	$(0,0)^T$	nonregular	$p = 0.5, \zeta = 1.00$	$(0.25,0,0.25,0)$
C	$(0,0,0,0,0.25,0,0.24)^T$	$(0,0)^T$	near-nonregular	$p = 0, \zeta = 1.03$	$(0.25,0,0.24,0)$

For each of the 9 data generating settings presented within each simulation experiment, we perform Clustered Q-Learning assuming correctly specified models for the Stage 1 and Stage 2 Q-functions: $Q_2(H_2, A_2) = \beta_{20} + \beta_{21}X_1 + \beta_{22}A_1 + \beta_{23}A_1X_1 + (\psi_{20} + \psi_{21}X_2 + \psi_{22}A_1)A_2$ and $Q_1(H_1, A_1) = \beta_{10} + \beta_{11}X_1 + (\psi_{10} + \psi_{11}X_1)A_1$. Estimates of bias and coverage rates are presented for each setting using fixed values of $\chi = 0.025$ and $\eta = 0.001$ with 1000 bootstrap samples within each of $B = 500$ Monte Carlo iterations. Parameter estimates, including bias and coverage, are presented in Tables 3.2 - 3.4 for ψ_{21}, ψ_{10}, and ψ_{11}; simulation results for parameters ψ_{20} and ψ_{22} (not shown) revealed similar trends. Estimated coverage rates are presented for the M-out-of-N cluster bootstrap (MN), as well as the standard m-out-of-n-cluster bootstrap (mn).

3.4.2 Simulation 1: Large Number of Clusters ($N = 80$)

We conduct this simulation with an assumed $N = 80$ clusters and $n_i = 20$ individuals per cluster. Estimates for the parameter indexing the Stage 2 interaction of intervention A_2 with the covariate X_2, i.e., ψ_{21}, are unbiased and coverage is near nominal across all regularity settings and levels of within-cluster correlation (Table 3.2). This result is expected given that nonregularity will not be observed in the Stage 2 estimation and the cluster bootstrap (with $M_2 = N_2$) is appropriately able to account for correlation within clusters. Estimates of parameter and coverage for the Stage 1 X_1A_1 interaction effect, ψ_{11}, reveal generally low bias (Table 3.3); only the nonregular setting Example 5 exhibits bias of about 0.01 across all levels of correlation. Coverage rates estimated for ψ_{11} are generally near nominal or slightly higher, and we observe a higher range of estimated coverage rates across regularity settings with lower within-cluster correlation: Estimated coverage rates range from 94.4% (Example 5) to 97.0% (Example 2) for $\rho = 0.10$ compared with a range of 94.8% (Example 6) to 96.4% (Examples 1 and 2) when $\rho = 0.40$. With regard to estimates

of the main effect of the Stage 1 intervention, ψ_{10}, there is a slight negative bias observed for the data generating settings defined by Examples 3 and 4, as well as Examples 8 and 9, the magnitude of which increases with increasing correlation to a maximum estimated bias of about 0.05 (Table 3.4). As previously described, Examples 4 and 9 illustrate settings in which there is only a very small true treatment effect at Stage 2 for some subset of the population; the similarity of these estimates to those observed for their nonregular counterparts in Examples 3 and 8, respectively, suggest that similar challenges in estimation are faced even in a near-nonregular setting. Even with the slight negative bias, however, estimated coverage rates hover around or slightly above nominal, although overcoverage may be anticipated when there is a higher degree of within-cluster correlation in a setting similar to that of Example 2, which represents a case in which all clusters experience a very weak effect of Stage 2 intervention. Coverage estimated without accounting for the clustering inherent in the data generation process (i.e., the "mn" column in Tables 3.2 - 3.4), is much lower than nominal across all parameter estimates and regularity settings, with coverage estimates ranging from about 50% for data generated using a within-cluster correlation of 0.40 to about 75% for data generated using a correlation level of 0.10. This result is expected even in the case of a fully regular setting due to the fact that the m-out-of-n standard bootstrap fails to account for the clustering within the data generation models.

Table 3.2: Estimates of bias and 95% confidence interval coverage for the X_2A_2 interaction effect, ψ_{21}, estimated in the second stage estimation for $N = 80$ clusters with $n_i = 20$ individuals per cluster. ρ refers to the intra-cluster correlation used to generate the data; Est = estimated value; MN = 95% confidence interval coverage estimated using M-out-of-N cluster bootstrap; mn = 95% confidence interval coverage estimated using m-out-of-n standard bootstrap; nr = nonregular setting; nnr = near-nonregular setting; r = regular setting.

Example	Truth	$\rho = 0.10$				$\rho = 0.25$				$\rho = 0.40$			
		Est	Bias	MN	mn	Est	Bias	MN	mn	Est	Bias	MN	mn
Ex 1-nr	0	0.001	0.001	94.2	75.6	-0.001	-0.001	94.8	60.4	0.001	0.001	94.4	50.2
Ex 2-nnr	0	0.001	0.001	93.8	75.8	-0.001	-0.001	94.6	60.4	0.001	0.001	94.2	50.2
Ex 3-nr	0	0.001	0.001	94.6	75.6	-0.001	-0.001	94.4	59.8	0.001	0.001	94.6	50.2
Ex 4-nnr	0	0.001	0.001	94.8	76.0	-0.001	-0.001	94.8	59.0	0.001	0.001	94.4	49.8
Ex 5-nr	0.5	0.499	-0.001	95.2	76.6	0.501	0.001	95.0	59.8	0.498	-0.002	95.0	49.2
Ex 6-r	0.5	0.501	0.001	95.0	76.2	0.498	-0.002	94.8	60.0	0.502	0.002	94.8	51.2
Ex 7-r	0.5	0.501	0.001	95.0	77.0	0.498	-0.002	95.2	60.8	0.502	0.002	94.6	50.0
Ex 8-nr	0	-0.000	-0.000	94.2	75.2	0.001	0.001	94.6	56.4	-0.001	-0.001	95.2	50.4
Ex 9-nnr	0	-0.000	-0.000	94.6	75.4	0.001	0.001	94.6	56.2	-0.001	-0.001	95.2	50.0

Table 3.3: Estimates of bias and 95% confidence interval coverage for the X_1A_1 interaction effect, ψ_{11}, estimated in the first stage estimation for $N = 80$ clusters with $n_i = 20$ individuals per cluster. ρ refers to the intra-cluster correlation used to generate the data; Est = estimated value; MN = 95% confidence interval coverage estimated using M-out-of-N cluster bootstrap; mn = 95% confidence interval coverage estimated using m-out-of-n standard bootstrap; nr = nonregular setting; nnr = near-nonregular setting; r = regular setting.

Example	Truth	$\rho = 0.10$				$\rho = 0.25$				$\rho = 0.40$			
		Est	Bias	MN	mn	Est	Bias	MN	mn	Est	Bias	MN	mn
Ex 1-nr	0	-0.001	-0.001	96.4	78.4	-0.003	-0.003	96.8	60.6	-0.002	-0.002	96.4	50.6
Ex 2-nnr	0	-0.001	-0.001	97.0	78.2	-0.003	-0.003	96.6	59.8	-0.002	-0.002	96.4	51.2
Ex 3-nr	0	-0.001	-0.001	95.8	74.6	-0.002	-0.002	96.4	59.4	-0.001	-0.001	96.2	50.0
Ex 4-nnr	0	-0.001	-0.001	95.4	75.2	-0.002	-0.002	96.4	59.4	-0.001	-0.001	96.2	50.2
Ex 5-nr	0	0.008	0.008	94.4	62.4	0.011	0.011	95.0	52.0	0.013	0.013	95.4	46.0
Ex 6-r	0.02	0.021	0.001	95.0	60.4	0.021	0.001	95.2	49.4	0.022	0.002	94.8	44.0
Ex 7-r	0.01	0.007	-0.003	95.2	59.6	0.007	-0.003	95.2	48.0	0.008	-0.002	96.0	43.4
Ex 8-nr	0	0.000	0.000	96.4	74.8	-0.001	-0.001	95.6	60.2	0.000	0.002	95.6	49.4
Ex 9-nnr	0	0.000	0.000	95.6	75.2	-0.001	-0.001	95.6	59.0	0.000	0.000	95.6	49.6

Table 3.4: Estimates of bias and 95% confidence interval coverage for the A_1 main effect, ψ_{10}, estimated in the first stage estimation for $N = 80$ clusters with $n_i = 20$ individuals per cluster. ρ refers to the intra-cluster correlation used to generate the data; Est = estimated value; MN = 95% confidence interval coverage estimated using M-out-of-N cluster bootstrap; mn = 95% confidence interval coverage estimated using m-out-of-n standard bootstrap; nr = nonregular setting; nnr = near-nonregular setting; r = regular setting.

Example	Truth	$\rho = 0.10$				$\rho = 0.25$				$\rho = 0.40$			
		Est	Bias	MN	mn	Est	Bias	MN	mn	Est	Bias	MN	mn
Ex 1-nr	0	0.001	0.001	96.6	83.6	-0.001	-0.001	97.4	65.6	0.002	0.002	96.6	56.8
Ex 2-nnr	0	0.001	0.001	96.6	82.6	-0.001	-0.001	97.8	66.4	0.002	0.002	97.4	56.6
Ex 3-nr	0	-0.030	-0.030	94.4	74.4	-0.042	-0.042	95.0	57.8	-0.052	-0.052	94.8	48.6
Ex 4-nnr	-0.01	-0.035	-0.025	95.0	75.4	-0.047	-0.037	95.6	60.8	-0.057	-0.047	95.2	49.0
Ex 5-nr	0	-0.012	-0.012	95.6	65.8	-0.018	-0.018	94.4	54.6	-0.024	-0.024	97.2	47.0
Ex 6-r	-0.37	-0.368	0.002	96.2	65.6	-0.367	0.003	95.0	55.4	-0.368	0.002	95.0	46.2
Ex 7-r	0.14	0.145	0.005	96.6	68.4	0.146	0.006	95.4	55.8	0.144	0.004	95.4	49.8
Ex 8-nr	0.25	0.219	-0.031	95.0	73.8	0.209	-0.041	95.0	57.4	0.197	-0.053	94.6	47.8
Ex 9-nnr	0.24	0.214	-0.026	95.0	75.0	0.204	-0.036	95.4	60.0	0.192	-0.048	94.8	49.8

3.4.3 Simulation 2: Small Number of Clusters ($N = 20$)

Using the same data generation mechanisms introduced in Section 3.4.1 but with the exception that now we simulate $N = 20$ clusters with $n_i = 80$ individuals within each cluster, we evaluate the performance of our proposed method with a smaller number of clusters and a larger number of individuals with each cluster. Compared with the performance reported in Section 3.4.2 with $N = 80$ clusters and $n_i = 20$ individuals, with fewer clusters we observe slightly larger bias overall in each respective parameter estimate (Tables 3.5 - 3.7). However, similar to previous results reported in Section 3.4.2, absolute bias tends to increase with increasing correlation among covariates. For the parameter associated with the X_2A_2 interaction effect (Table 3.5), for example, we observe a maximum absolute bias of 0.007 associated with a correlation of 0.10 compared with 0.013 for $\rho = 0.40$. Interestingly, absolute bias is largest in the regular data generating settings (Examples 6 and 7). Confidence interval coverage estimated using the M-out-of-N cluster bootstrap is near-nominal across all specified values of ρ (Table 3.5), generally ranging from about 93.5% for Example 5 (regular setting) to about 97% in Example 9 (near-nonregular setting); however, variability of coverage estimates across the regularity settings and pre-specified values of ρ is larger with $N = 20$ clusters compared with $N = 80$ clusters. When evaluating estimates of the X_1A_1 interaction effect with a smaller number of clusters (Table 3.6), we observe a negligible degree of absolute bias. Although the absolute bias is slightly higher than that reported in Section 3.4.2 with a large number of clusters, bias estimates do not exceed 0.02 across any regularity setting and degree of correlation. Estimated coverage rates for the ψ_{11} parameter do exhibit undercoverage when the number of clusters is smaller, however, with estimates at/near 90%, which is substantially below the nominal level of 95%. The lowest estimated coverage rates across all degrees of correlation are reported for Examples 6 and 7, both of which reflect fully regular settings. In Table 3.7 we observe here also that absolute bias of the A_1 main effect is

about twice that observed with a larger cluster size in Table 3.4. The largest degree of bias is found in the nonregular and near-nonregular settings Examples 3 and 4, as well as Examples 8 and 9, with absolute bias estimates associated with $\rho = 0.40$ exceeding 0.1. The M-out-of-N cluster bootstrap estimates of coverage for the A_1 main effect are lower than nominal and average around 90% across all regularity settings and levels of covariate correlation. Additionally, estimated coverage appears to decrease slightly as the correlation increases. Finally, as shown in the estimation of all regression parameters, without taking clustering into consideration, the standard m-out-of-n bootstrap estimates of coverage when $N = 20$ are equivalent to those assuming $N = 80$, as expected; these estimates decrease with increasing ρ and remain substantially lower than nominal across all data generating settings.

Table 3.5: Estimates of bias and 95% confidence interval coverage for the X_2A_2 interaction effect, ψ_{21}, estimated in the second stage estimation for $N = 20$ clusters with $n_i = 80$ individuals per cluster. ρ refers to the intra-cluster correlation used to generate the data; Est = estimated value; MN = 95% confidence interval coverage estimated using M-out-of-N cluster bootstrap; mn = 95% confidence interval coverage estimated using m-out-of-n standard bootstrap; nr = nonregular setting; nnr = near-nonregular setting; r = regular setting.

Example	Truth	$\rho = 0.10$				$\rho = 0.25$				$\rho = 0.40$			
		Est	Bias	MN	mn	Est	Bias	MN	mn	Est	Bias	MN	mn
Ex 1-nr	0	-0.003	-0.003	94.6	75.6	-0.005	-0.005	94.6	60.4	-0.007	-0.007	94.6	50.2
Ex 2-nnr	0	-0.003	-0.003	94.6	75.8	-0.005	-0.005	95.0	60.4	-0.006	-0.006	94.6	50.2
Ex 3-nr	0	-0.003	-0.003	95.2	75.6	-0.005	-0.005	94.8	59.8	-0.007	-0.007	95.2	50.2
Ex 4-nnr	0	-0.003	-0.003	95.0	76.0	-0.005	-0.005	95.0	59.0	-0.006	-0.006	95.6	49.8
Ex 5-nr	0.5	0.497	-0.003	93.8	76.6	0.496	-0.004	93.6	59.8	0.493	-0.007	93.4	49.2
Ex 6-r	0.5	0.493	-0.007	95.9	76.2	0.490	-0.010	95.9	60.0	0.487	-0.013	96.3	51.2
Ex 7-r	0.5	0.493	-0.007	96.1	77.0	0.489	-0.011	96.1	60.8	0.487	-0.013	95.9	50.0
Ex 8-nr	0	-0.000	-0.000	96.1	75.2	-0.001	-0.001	96.9	56.4	-0.001	-0.001	96.5	50.4
Ex 9-nnr	0	-0.000	-0.000	97.1	75.4	-0.000	-0.000	96.7	56.2	-0.000	-0.000	97.1	50.0

Table 3.6: Estimates of bias and 95% confidence interval coverage for the X_1A_1 interaction effect, ψ_{11}, estimated in the first stage estimation for $N = 20$ clusters with $n_i = 80$ individuals per cluster. ρ refers to the intra-cluster correlation used to generate the data; Est = estimated value; MN = 95% confidence interval coverage estimated using M-out-of-N cluster bootstrap; mn = 95% confidence interval coverage estimated using m-out-of-n standard bootstrap; nr = nonregular setting; nnr = near-nonregular setting; r = regular setting.

Example	Truth	$\rho = 0.10$				$\rho = 0.25$				$\rho = 0.40$			
		Est	Bias	MN	mn	Est	Bias	MN	mn	Est	Bias	MN	mn
Ex 1-nr	0	-0.005	-0.005	91.5	78.4	-0.008	-0.008	91.3	60.6	-0.010	-0.010	91.5	50.6
Ex 2-nnr	0	-0.005	-0.005	91.3	78.2	-0.008	-0.008	91.3	59.8	-0.010	-0.010	91.3	51.2
Ex 3-nr	0	-0.004	-0.004	91.9	74.6	-0.006	-0.006	92.3	59.4	-0.008	-0.008	92.3	50.0
Ex 4-nnr	0	-0.004	-0.004	91.7	75.2	-0.006	-0.006	91.9	59.4	-0.008	-0.008	92.1	50.2
Ex 5-nr	0	0.010	0.010	90.7	62.4	0.016	0.016	91.9	52.0	0.019	0.019	91.1	46.0
Ex 6-r	0.02	0.016	-0.004	89.0	60.4	0.014	-0.006	89.0	49.4	0.013	-0.007	89.4	44.0
Ex 7-r	0.01	0.000	-0.010	89.0	59.6	0.000	-0.010	90.1	48.0	-0.000	-0.01	90.7	43.4
Ex 8-nr	0	0.001	0.001	92.1	74.8	0.001	0.001	92.3	60.2	0.000	0.000	92.5	49.4
Ex 9-nnr	0	0.001	0.001	92.1	75.2	0.001	0.001	92.6	59.0	0.001	0.001	93.2	49.6

Table 3.7: Estimates of bias and 95% confidence interval coverage for the A_1 main effect, ψ_{10}, estimated in the first stage estimation for $N = 20$ clusters with $n_i = 80$ individuals per cluster. ρ refers to the intra-cluster correlation used to generate the data; Est = estimated value; MN = 95% confidence interval coverage estimated using M-out-of-N cluster bootstrap; mn = 95% confidence interval coverage estimated using m-out-of-n standard bootstrap; nr = nonregular setting; nnr = near-nonregular setting; r = regular setting.

Example	Truth	$\rho = 0.10$				$\rho = 0.25$				$\rho = 0.40$			
		Est	Bias	MN	mn	Est	Bias	MN	mn	Est	Bias	MN	mn
Ex 1-nr	0	-0.006	-0.006	90.1	83.6	-0.010	-0.010	89.2	65.6	-0.012	-0.012	89.4	56.8
Ex 2-nnr	0	-0.007	-0.007	90.0	82.6	-0.010	-0.010	90.3	66.4	-0.013	-0.013	90.0	56.6
Ex 3-nr	0	-0.068	-0.068	90.3	74.4	-0.103	-0.103	90.5	57.8	-0.129	-0.129	90.3	48.6
Ex 4-nnr	-0.01	-0.073	-0.073	90.7	75.4	-0.109	-0.099	90.5	60.8	-0.134	-0.124	90.0	49.0
Ex 5-nr	0	-0.038	-0.038	89.9	65.8	-0.056	-0.056	90.1	54.6	-0.069	-0.069	89.3	47.0
Ex 6-r	-0.37	-0.373	-0.003	89.2	XX	-0.374	-0.004	89.0	55.4	-0.374	-0.004	89.0	46.2
Ex 7-r	0.14	0.133	-0.007	89.4	68.4	0.124	-0.016	88.4	55.8	0.114	-0.026	88.8	49.8
Ex 8-nr	0.25	0.186	-0.064	91.1	73.8	0.155	-0.095	90.0	57.4	0.135	-0.115	88.8	47.8
Ex 9-nnr	0.24	0.181	-0.059	91.3	75.0	0.149	-0.091	90.7	60.0	0.129	-0.111	88.8	49.8

3.4.4 Simulation 3: Large Number of Clusters with Variable Number of Individuals per Cluster

In Simulation 3 we generate data using the same mechanisms described in Section 3.4.1 with $N = 80$ clusters, but instead of a fixed number of individuals per cluster, we assume that the number of individuals within each cluster is randomly generated from a normal distribution with a mean of 20 individuals and a standard deviation of 5, rounded to the nearest integer. Under this mechanism, the number of individuals across clusters ranges from about 7 to 33 individuals per cluster, which we believe represents a more realistic use case.

We report and compare performance to fixed cluster sizes shown in Section 3.4.2. Estimates of the Stage 2 parameter representing the X_2A_2 interaction effect exhibit a slightly higher but still negligible absolute bias, not exceeding 0.006 at the highest level of covariate correlation (Table 3.8). Coverage estimated using the M-out-of-N cluster bootstrap is slightly lower than nominal across all levels of ρ, averaging about 93% for nonregular and near-nonregular settings when $\rho = 0.10$ and about 92% when $\rho = 0.40$. These estimates demonstrate some undercoverage with highly variable sample sizes across clusters compared with a fixed and moderately-sized sample. When estimating the parameter associated with the Stage 1 interaction effect X_1A_1 (Table 3.9), absolute bias remains low across all data generating settings and levels of correlation; only the nonregular Example 5 reaches a bias larger than 0.01 for $\rho = 0.40$. Interestingly, although estimates of bias across all levels of correlation are low, absolute bias appears slightly lower with $\rho = 0.40$. Coverage estimates using the M-out-of-N cluster bootstrap are nominal or slightly above nominal, with estimates of about 95%-96% across all regularity settings and degrees of correlation. These findings are consistent with those reported in Section 3.4.2, although it is somewhat surprising that estimated coverage does not suffer from the variability introduced into the cluster sizes as was observed for the Stage 2 X_2A_2 interaction effect in Table 3.8.

As seen in Table 3.10, estimation of the A_1 main effect, ψ_{10}, exhibits bias similar to that shown with a fixed number of individuals per clusters, reaching about 0.05 for Examples 3, 4, 8, and 9. M-out-of-N cluster bootstrap coverage estimates for the A_1 main effect in Stage 1 are at/near nominal for $\rho = 0.10$ – with the possible exception of regular setting Example 6, which shows slight undercoverage at 93.8%. For $\rho = 0.40$, coverage estimates appear slightly lower than nominal, typically between 93%-94% across most data generating settings, with the exception of Examples 1 and 2, which exhibit nominal estimated coverage. Compared with a fixed sample size, coverage is slightly lower, but remains at or slightly below nominal, as described. Finally, as observed across all simulation experiments with $N = 80$ clusters and a variable number of individuals sampled within each cluster, estimated coverage percentage using the m-out-of-n standard bootstrap fails to reach 80% for $\rho = 0.10$ and is lower than 50% with $\rho = 0.40$, suggesting again that the m-out-of-n standard bootstrap is inadequate to estimate confidence intervals when data are inherently clustered.

Table 3.8: Estimates of bias and 95% confidence interval coverage for the X_2A_2 interaction effect, ψ_{21}, estimated in the second stage estimation for $N = 80$ clusters with variable individuals per cluster: $n_i \sim N(20, \sigma = 5)$. σ refers to the standard deviation; ρ refers to the intra-cluster correlation used to generate the data; Est = estimated value; MN = 95% confidence interval coverage estimated using M-out-of-N cluster bootstrap; mn = 95% confidence interval coverage estimated using m-out-of-n standard bootstrap; nr = nonregular setting; nnr = near-nonregular setting; r = regular setting.

Example	Truth	$\rho = 0.10$				$\rho = 0.25$				$\rho = 0.40$			
		Est	Bias	MN	mn	Est	Bias	MN	mn	Est	Bias	MN	mn
Ex 1-nr	0	0.003	0.003	92.6	69.8	-0.001	-0.001	94.6	55.6	0.006	0.006	92.0	44.2
Ex 2-nnr	0	0.003	0.003	92.8	69.6	-0.001	-0.001	95.0	55.0	0.006	0.006	91.8	44.0
Ex 3-nr	0	0.003	0.003	93.0	69.8	-0.001	-0.001	94.8	54.8	0.006	0.006	91.8	43.8
Ex 4-nnr	0	0.003	0.003	93.2	69.0	-0.001	-0.001	94.8	55.6	0.006	0.006	92.0	43.6
Ex 5-nr	0.5	0.498	-0.002	92.6	70.8	0.495	-0.005	94.0	56.0	0.501	0.001	95.2	44.8
Ex 6-r	0.5	0.503	0.003	94.4	73.2	0.499	-0.001	95.0	58.2	0.506	0.006	91.8	46.2
Ex 7-r	0.5	0.503	0.003	93.8	72.8	0.499	-0.001	94.6	58.2	0.506	0.006	90.6	45.8
Ex 8-nr	0	0.003	0.003	92.0	75.2	0.001	0.001	95.2	53.8	0.004	0.004	93.0	48.6
Ex 9-nnr	0	0.003	0.003	92.2	75.2	0.001	0.001	95.2	54.0	0.004	0.004	92.8	48.4

Table 3.9: Estimates of bias and 95% confidence interval coverage for the X_1A_1 interaction effect, ψ_{11}, estimated in the first stage estimation for $N = 80$ clusters with variable individuals per cluster: $n_i \sim N(20, \sigma = 5)$. σ refers to the standard deviation; ρ refers to the intra-cluster correlation used to generate the data; Est = estimated value; MN = 95% confidence interval coverage estimated using M-out-of-N cluster bootstrap; mn = 95% confidence interval coverage estimated using m-out-of-n standard bootstrap; nr = nonregular setting; nnr = near-nonregular setting; r = regular setting.

Example	Truth	$\rho = 0.10$				$\rho = 0.25$				$\rho = 0.40$			
		Est	Bias	MN	mn	Est	Bias	MN	mn	Est	Bias	MN	mn
Ex 1-nr	0	-0.005	-0.005	96.0	78.2	-0.010	-0.010	96.4	63.6	-0.004	-0.004	96.2	49.0
Ex 2-nnr	0	-0.005	-0.005	96.0	79.2	-0.010	-0.010	96.2	63.4	-0.004	-0.004	96.8	48.8
Ex 3-nr	0	-0.004	-0.004	95.6	76.4	-0.009	-0.009	96.0	62.6	-0.002	-0.002	95.2	47.0
Ex 4-nnr	0	-0.004	-0.004	95.6	76.2	-0.009	-0.009	96.6	62.8	-0.002	-0.002	95.4	46.4
Ex 5-nr	0	0.004	0.004	96.2	64.8	0.001	0.001	95.0	54.0	0.011	0.011	96.0	46.4
Ex 6-r	0.02	0.017	-0.003	95.6	62.2	0.013	-0.007	95.8	54.0	0.020	0.000	94.4	45.8
Ex 7-r	0.01	0.003	-0.007	96.0	58.6	-0.001	-0.011	95.6	53.2	0.006	-0.004	95.0	45.0
Ex 8-nr	0	-0.005	-0.005	96.0	76.6	-0.008	-0.008	96.6	62.6	-0.002	-0.002	95.4	47.6
Ex 9-nnr	0	-0.005	-0.005	95.6	76.0	-0.008	-0.008	97.2	62.4	-0.002	-0.002	95.2	47.0

Table 3.10: Estimates of bias and 95% confidence interval coverage for the A_1 main effect, ψ_{10}, estimated in the first stage estimation for $N = 80$ clusters with variable individuals per cluster: $n_i \sim N(20, \sigma = 5)$. σ refers to the standard deviation; ρ refers to the intra-cluster correlation used to generate the data; Est = estimated value; MN = 95% confidence interval coverage estimated using M-out-of-N cluster bootstrap; mn = 95% confidence interval coverage estimated using m-out-of-n standard bootstrap; nr = nonregular setting; nnr = near-nonregular setting; r = regular setting.

Example	Truth	$\rho = 0.10$				$\rho = 0.25$				$\rho = 0.40$			
		Est	Bias	MN	mn	Est	Bias	MN	mn	Est	Bias	MN	mn
Ex 1-nr	0	0.002	0.002	96.4	83.2	0.004	0.004	96.4	62.8	0.001	0.001	96.0	54.6
Ex 2-nnr	0	0.002	0.002	96.4	83.0	0.004	0.004	96.4	62.8	0.001	0.001	95.4	54.8
Ex 3-nr	0	-0.031	-0.031	95.2	73.2	-0.043	-0.043	93.6	53.6	-0.053	-0.053	93.0	46.6
Ex 4-nnr	-0.01	-0.036	-0.026	95.2	74.6	-0.048	-0.038	94.6	54.4	-0.057	-0.047	93.8	49.4
Ex 5-nr	0	-0.011	-0.011	95.2	63.6	-0.017	-0.017	95.0	54.4	-0.019	-0.019	93.4	45.8
Ex 6-r	-0.37	-0.369	0.001	93.8	68.4	-0.366	0.004	92.6	52.8	-0.368	0.002	93.4	46.0
Ex 7-r	0.14	0.144	0.004	94.2	67.2	0.146	0.006	94.0	55.2	0.144	0.004	94.2	44.6
Ex 8-nr	0.25	0.218	-0.032	94.8	74.4	0.207	-0.043	93.8	52.4	0.198	-0.052	93.8	47.0
Ex 9-nnr	0.24	0.213	-0.027	95.6	74.8	0.202	-0.038	94.0	55.4	0.193	-0.047	94.0	48.6

3.5 Data Analysis

One important objective in the mental health community is to increase the use at treatment centers of evidence-based practices (EBPs) (*Guyatt et al.*, 1992), which have been shown to improve patient-level outcomes for individuals with anxiety or depression, post-traumatic stress disorder, autism, and others (*Badamgarav et al.*, 2003; *Drake et al.*, 2003). The desire would be to provide the minimum support necessary to effectively achieve this goal. However, given what can be a large degree of heterogeneity across clinics, we would expect that distinct clinics may respond differently to varying levels of implementation support. Thus, determining whether there are any clinic-specific factors that may be used to tailor cluster-level interventions will be an important goal.

The Adaptive Implementation of Effective Programs Trial (ADEPT) is a Clustered SMART mirroring Design III in Figure 3.1. It was conducted at community-based, outpatient clinics in Michigan and Colorado and was designed to determine how best to support nonresponsive clinics in implementing EBPs (*Kilbourne et al.*, 2014). One of the stated objectives of this trial was to identify clinic-level factors at each intervention stage, if any, that could be used to tailor the level of intervention necessary to ensure the clinic would successfully implement EBPs across their practice. At the outset of the study, all participating clinics were offered training in replicating effective programs, a system designed to help them implement evidence-based practices. As can be seen in Figure 3.2, the first randomization event included only those clinics who failed to effectively implement the EBPs. These clinics were randomized 1:1 to receive one of two different intervention support systems: external support alone (EF) or both external and internal support (EF/IF). Refer to *Kilbourne et al.* (2014) and *Smith et al.* (2019) for additional information about REP and the interventions EF and IF. Following the Stage 1 response assessment, which occurred 6 months after the first randomization event, clinics who received EF at Stage 1 were withdrawn from

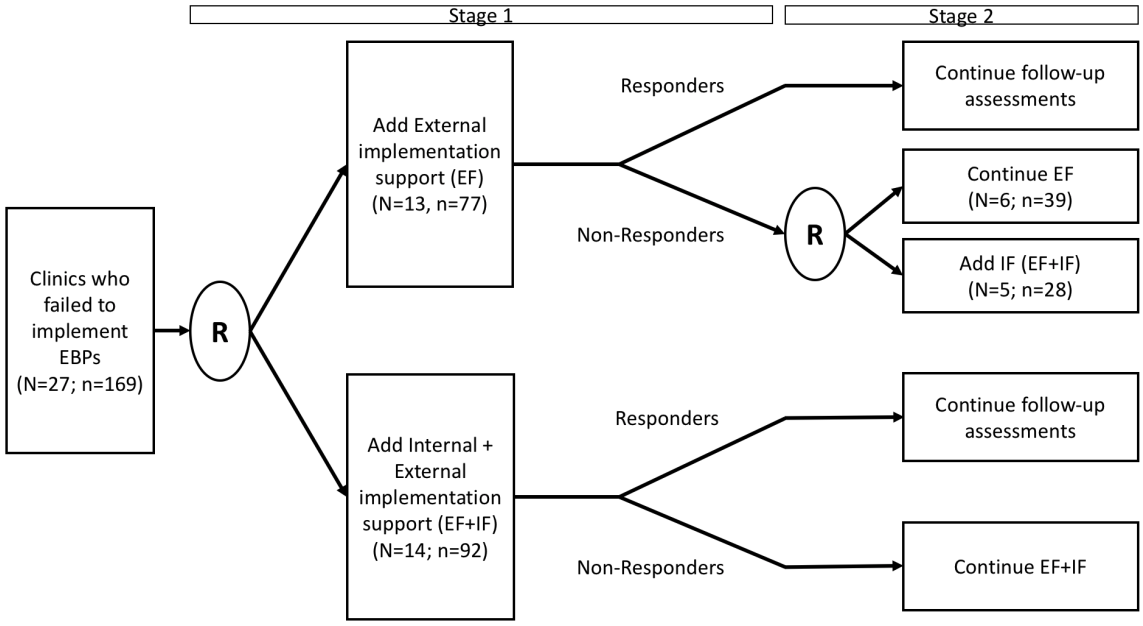

Figure 3.2: Clustered Sequential Multiple Assignment Randomized Trial (Clustered SMART) designed to evaluate use of Internal (IF) and/or External (EF) implementation support for primary and mental health clinics who failed to implement evidence-based practices (EBPs) after a 6 month run-in period. R indicates 1:1 randomization performed. N = number of clinics; n = number of patients within the N clinics. (Figure adapted from *Kilbourne et al.* (2014) and *Smith et al.* (2019).)

EF if they had effectively implemented the EBPs. On the other hand, EF-clinics who had not been successful in implementing EBPs were re-randomized to either continue EF or to add internal support (i.e., EF/IF). All clinics who received EF/IF at Stage 1 either stopped (or continued) EF/IF if they had (or had not) effectively implemented the EBPs. Patient-level outcomes were collected at baseline and after the first and second stages.

We apply Clustered Q-Learning to data collected for the ADEPT study to determine whether clinic-level factors can be used to tailor implementation strategies for EBPs at Stage 1 or Stage 2 with the ultimate goal of improving mental health outcomes for patients with mood disorders. The primary outcome is mental health quality of life (MHQOL) collected for each patient following the Stage 2 intervention. MHQOL is assessed using the Short Form-12 (SF-12) (*Vilagut et al.*, 2013; *Ware Jr et al.*, 1996), which is scored from 0-100 with higher scores indicating better MHQOL.

Interventions at both stages (denoted $A1$ and $A2$, respectively) include EF+IF (versus EF alone), both stages coded as 1 (or -1). Cluster-level covariates collected prior to the first randomization include rural (or urban) clinic $(1/-1)$, clinic located in Michigan (or Colorado) $(1/-1)$, and higher (or lower) than average site-aggregated MHQOL stratum $(1/-1)$. Covariates collected prior to Stage 2 include a factor indicating a higher (or lower) than average site-aggregated SF-12 stratum $(1/-1)$. At Stage 1 we consider two candidate tailoring variables: site-aggregated mean MHQOL level preceding the first randomization (M6-MH) and the state in which the clinic was located; high MHQOL strata and the state of Michigan (MI) were used as reference categories. At Stage 2 we evaluate site-aggregated mean MHQOL level immediately preceding the second randomization (M12-MH). Variables used to stratify randomization are also included in both stage-specific regression models.

We conduct this analysis using the implementation guidelines described in Section 3. The Stage 2 Q-function is specified as follows: $Q_2(\mathbf{H}_2, A_2) = \beta_{20} + \beta_{21}(\text{Rural}) + \beta_{22}(\text{M12-MH}) + \big\{\psi_{20} + \psi_{21}(\text{M12-MH})\big\} * A2$. For patients treated at clinics re-randomized at Stage 2, the Stage 1 pseudo-outcome is calculated as $\tilde{Y}_{MHj} = \hat{\beta}_{20} + \hat{\beta}_{21}(\text{Rural})_{ij} + \hat{\beta}_{22}(\text{M12-MH})_{ij} + \big|\hat{\psi}_{20} + \hat{\psi}_{21}(\text{M12-MH})_{ij}\big|$. The Stage 1 Q-function is modeled using: $Q_1(\mathbf{H}, A_1) = \beta_{10} + \beta_{11}(\text{Rural}) + \beta_{12}(\text{MI}) + \beta_{13}(\text{M6-MH}) + \big\{\psi_{10} + \psi_{11}(\text{MI}) + \psi_{12}(\text{M6-MH})\big\} * A1$. We perform M-out-of-N cluster bootstrap resampling at both stages with $B = 2500$ iterations, but remove any bootstrap resample that fails to generate estimates due to singularity and, given the exploratory nature of our analysis, estimate confidence intervals based on a pre-specified significance level of $\alpha = 0.10$. Due the high degree of missingness of the overall MHQOL for patients treated at clinics who were re-randomized at Stage 2, as well as the composition of the sites re-randomized at Stage 2 (i.e., no urban clinics), we utilize multiply imputed datasets (*Smith et al.*, 2019) with appropriate combining rules (*Little and Rubin*, 2019).

Thirteen clinics were randomized at the first stage to receive EF and 14 clinics were randomized to receive EF+IF (Figure 2). Of the clinics randomized to EF at Stage 1, six clinics were randomized at the second stage to continue EF alone and five clinics were randomized to also receive IF support. The intraclass correlation of outcomes within each clinic was estimated to be 0.23. Refer also to *Smith et al.* (2019) for summary statistics describing the patient cohort and results of the primary analysis.

Estimated regression coefficents and associated 90% M-out-of-N cluster bootstrap confidence intervals are shown in Table 3.11. In order to determine whether the set of candidate variables may be useful in tailoring a CAI that will optimize individual-level counterfactual outcomes across the population of interest, we examine the interaction effects of cluster-level covariates with the intervention EF+IF at both stages (Table 3.11, Rows 5, 6, 10). At Stage 1, the estimated confidence intervals for the Stage 1 EF+IF interventions with state and high clinic mean Month 6 MHQOL are $(-3.11, 2.89)$ and $(-1.44, 3.24)$, respectively, both of which include zero. Similarly, at Stage 2, the estimated 90% confidence interval for the EF+IF interaction with high mean Month 12 MHQOL is $(-3.34, 2.25)$, which also includes 0, suggesting there is insufficient evidence in our data to conclude any of these candidate variables would be useful in additionally tailoring a CAI to support implementation of EBPs. As this was an exploratory analysis and the study was not powered based on this statistical objective, is it possible either that these tailoring variables should not be used to further refine clinic-level interventions to improve implementation of EBPs at primary care and mental health clinics located in Michigan and Colorado, or that there is insufficient power in our dataset to identify these effects.

Table 3.11: Estimated Stage 1 and Stage 2 regression coefficients and associated 90% M-out-of-N cluster bootstrap confidence intervals (CI). Outcome of interest is patient-level Month 18 Mental Health Quality of Life (MHQOL). M6 = Month 6 (prior to first randomization); M12 = Month 12 (prior to second randomization); Interventions at both Stage 1 and Stage 2 include EF+IF (external and internal implementation support) versus EF alone. All covariates are measured at the cluster level.

	Stage 1 Variable	Estimate	90% CI
1	Rural (vs. Not Rural)	-6.80	(-15.2, -7.0)
2	Michigan (vs. Colorado)	0.64	(-2.65, 3.16)
3	High Mean M6 MHQOL (vs. Low)	0.40	(-2.11, 2.42)
4	EF+IF (vs. EF alone)	-3.74	(-4.34, 0.97)
5	(EF+IF):(Michigan)	1.31	(-3.11, 2.89)
6	(EF+IF):(High Mean M6 MHQOL)	0.82	(-1.44, 3.24)
	Stage 2 Variable	Estimate	90% CI
7	Rural (vs. Not Rural)	-5.94	(-14.8, 1.2)
8	High Mean M12 MHQOL (vs. Low)	-1.53	(-3.9, 0.82)
9	EF+IF (vs. EF alone)	-0.19	(-2.65, 2.30)
10	(EF+IF):(High Mean M12 MHQOL)	-0.60	(-3.34, 2.25)

3.6 Discussion

We propose Clustered Q-Learning for evaluating whether candidate tailoring variables may be useful in further tailoring multi-stage interventions delivered at the cluster level with the goal of improving outcomes at the level of the individual within the cluster. We demonstrate that with an asymptotically large number of clusters and a moderate number of individuals within each cluster, estimates of parameters and their associated confidence intervals have a low degree of bias and near or slightly higher than nominal coverage. Given these results, Clustered Q-Learning can be selected as the appropriate analysis method for a Clustered SMART when the planned number of clusters is large and the number of individuals within each cluster is moderate. While our simulation results demonstrate minimal bias and reasonable–although slightly lower than nominal–coverage when the number of individuals per cluster is variable, care should be taken to encourage clusters to recruit the expected number

of individuals into the study. When either the planned number of clusters is small, i.e. < 30, or when the expected number of individuals recruited into each cluster is small–for example, fewer than 20 on average, alternative approaches for confidence interval estimation should be considered.

We have proposed Clustered Q-Learning with a parametric regression framework. There are two major reasons why we believe this is advantageous. First, parametric, linear regression models enjoy widespread use and general understanding among domain scientists. Secondly, we employ Clustered Q-Learning to answer a well-defined research question using data from a Clustered SMART. In this context we believe parametric modeling of the Q-functions will deliver unbiased estimates with greater precision than semi- or nonparametric alternatives.

We note four straightforward extensions of Clustered Q-Learning. First, Clustered Q-Learning is easily extended to a setting with more than two interventions per stage and more than two stages per CAI. Although Clustered SMARTs to evaluate implementation science initiatives are not likely to exceed two stages, Clustered Q-Learning applied to the setting of mobile health and micro-randomized trials, where there could be an indefinite number of stages with many possible interventions per stage, may be of great interest. Additionally, here we consider an outcome that is continuous and approximately normally distributed. Given our use of the generalized estimating equations framework for estimation of the multi-stage Q-functions, however, Clustered Q-Learning is easily extended to non-continuous outcomes using the generalized linear model framework with appropriate choices for the outcome distribution and the link function. Third, although we apply Clustered Q-Learning to analyze data from a Clustered SMART, literature pertaining to the use of standard Q-Learning to estimate causal effects highlights the potential of Clustered Q-Learning also in multi-stage estimation with observational data sources (*Chakraborty and Moodie*, 2013; *Moodie et al.*, 2012; *Schulte et al.*, 2014). In this case, appropri-

ate use of propensity adjustment, as well application of more flexible models for the Q-functions (*Moodie et al.*, 2013; *Qian and Murphy*, 2011; *Zhao et al.*, 2011), may be explored. Finally, while our goal was to determine whether a set of candidate tailoring variables can be used to further tailor a multi-stage, cluster-level intervention, the methods of Clustered Q-Learning can also be used to estimate predictive effects associated with both stages, as well as to identify a specific CAI that may can be used to tailor intervention overall.

There are two limitations to our approach. First, we explore a relatively simple data generation mechansism, with one binary baseline covariate and one binary time-varying covariate. Although these data generating mechanisms are purposefully designed to reflect and evaluate varying degrees of nonregularity induced by different underlying clinical settings and represent the clustered analog of the settings used by *Chakraborty et al.* (2010) and *Laber et al.* (2014) to evaluate performance under nonregularity, these provide a limited understanding of performance among more complex data generating settings. Secondly, although Q-functions are modeled parametrically, estimation of confidence intervals using the M-out-of-N cluster bootstrap is dependent upon selection of two tuning parameters, χ and η, which induces a degree variability – as well as flexibility. In various supplemental simulation experiments (results not shown) we find generally that a range of $\chi \in (0.025, 0.10)$ is reasonable and that larger values of χ are associated with a smaller resample size M and generate larger estimated confidence intervals. Although we found that a pre-specified value of $\chi = 0.025$ demonstrated estimated coverage closest to nominal in simulation experiments, there is room for additional investigation of the specification of χ. A second tuning parameter needed when applying the M-out-of-N cluster bootstrap is η, which represents the value consistent with the desired quantile of the T-distribution used to evaluate the estimated degree of nonregularity at the second stage, $\hat{p}_2$. We selected η as the value associated with the 0.001-th quantile of the T-distribution

with N_2 degrees of freedom which, after applying a Bonferroni correction for multiple (i.e., $N = 80$) distinct hypothesis tests, reflects a maximum desired Type I error of $\alpha = 0.08$. However, changes to the pre-specified value of η can also lead to additional variability in estimation of confidence intervals for the parameters indexing the optimal first-stage CAI.

The ADEPT study, unfortunately, failed to achieve its enrollment goals of 80 centers with 20 patients per center, and our analysis did not reveal any tailoring variables. The Clustered SMART study design has become increasingly popular over the last few years, however, with several Clustered SMARTs now in the field or soon to begin study enrollment (e.g., *Fernandez et al.*, 2020; *Kilbourne et al.*, 2018; *Quanbeck et al.*, 2020; *Zhou et al.*, 2020). Clustered Q-Learning with the M-out-of-N cluster bootstrap, introduced in this manuscript, provides a simple yet effective and easy-to-implement solution to identify multi-stage tailoring variables and can provide insights to improve both cluster-level implementation efforts and individual-level outcomes across a host of domains.

CHAPTER IV

Penalized Spline-Involved Tree-based (PenSIT) Learning for Estimating an Optimal Dynamic Treatment Regime Using Observational Data

4.1 Introduction

Given the increasing prevalence of chronic health conditions, as well as the rapid increase in healthcare expenditures overall, large scale initiatives to deliver personalized medicine are underway. Personalized medicine is built upon the understanding that patients are uniquely heterogeneous in their existing and emergent comorbidities, as well as their tolerance of, response to, and even preference for different treatments. As such, one of its goals is to identify distinct variables that have an interaction with treatment, which can help define which patients will benefit from certain treatments or treatment sequences. One such avenue to advance personalized medicine is through dynamic treatment regimes (DTRs; *Chakraborty and Moodie*, 2013; *Murphy*, 2003), the statistical methods of which are grounded in causal inference. DTRs, also known as adaptive interventions, are a series of stage-specific decision rules that map a patient's measured baseline and time-varying characteristics to a treatment assignment at each successive stage. One particular objective within the field is to estimate an optimal DTR such that, if the population of interest were to receive treatment

consistent with this regime, overall patient-level outcomes would be optimized.

DTR estimation methods can be classified as model-based (*Huang et al.*, 2015; *Murphy*, 2003; *van der Laan and Rubin*, 2006; *Wang et al.*, 2012; *Zhang et al.*, 2013), including parametric and likelihood-based methods, and semiparametric or nonparametric methods (*Arjas and Saarela*, 2010; *Moodie et al.*, 2013; *Qian and Murphy*, 2011; *Xu et al.*, 2016). With the abundance of observational data available to us, it is now generally accepted that more flexible and robust estimation methods, which are able to account for what is expected to be a complex relationship among variables of interest, are desired. Additionally, because optimal DTR estimation is largely an exploratory process and collaboration with clinician scientists is critical, the need for interpretability in an estimated optimal DTR is paramount. As a result, flexible and robust methods that yield interpretable results – for example, those with a decision tree-type structure – have been enjoying much popularity. Over the past decade tree-based methods have evolved from the ability to handle a single stage and/or binary treatment setting (*Laber and Zhao*, 2015; *Zhang et al.*, 2015; *Zhao et al.*, 2015) to a multi-stage setting with multiple treatment options per stage (*Sun and Wang*, 2020; *Tao et al.*, 2018; *Zhang et al.*, 2018), which better reflects how care for chronic health conditions is delivered in practice. *Zhang et al.* (2018) estimate an optimal multi-stage DTR using a decision list; however, computational demands restrict each statement to a maximum of two covariates and, additionally, the unidirectional growth of decision lists precludes correction of estimation error(s) that may have occurred at previous steps. *Tao et al.* (2018) develop tree-based reinforcement learning (T-RL), a direct method of estimating an optimal multi-stage DTR, which cleverly embeds into the decision tree framework (CART; *Breiman et al.*, 1984) a purity measure devised from the augmented inverse probability weighted (AIPW) estimate of the counterfactual mean outcome. Although the AIPW estimator is consistent and doubly robust for the counterfactual mean outcome, however, it has been well established that IPW-style

estimators are unstable when weights are highly variable (e.g., *Kang and Schafer*, 2007), which will often be the case with low propensity of treatment assignment and/or as the number of stages increases.

In this manuscript we propose Penalized Spline-Involved Tree-based (PenSIT) Learning, which seeks to improve upon existing tree-based approaches for estimating an optimal multi-stage multi-treatment DTR. While conceptually similar to the implementation of T-RL, PenSIT Learning makes use of a different purity measure – one that uses a Penalized Spline-Involved (PenSI) estimator of counterfactual outcomes developed from the penalized spline of propensity prediction method used for missing data (*Little and An*, 2004; *Zhang and Little*, 2009; *Zhou et al.*, 2019). Specifically, we predict missing counterfactual outcomes for the treatments not assigned to patients using regression models that incorporate a penalized spline of a function of the propensity to be assigned that treatment and other covariates predictive of the outcome. The PenSI estimator of the counterfactual mean outcome, like the AIPW estimator, is consistent and retains the property of double robustness against model misspecifications, which may lend more stability and provide improved performance (e.g., higher percentage of observations correctly classified to their optimal DTR) under certain data generating mechanisms. PenSIT Learning estimates stage-specific optimal decision rules using backward induction (*Bather*, 2000), beginning with the final stage, to remove bias arising from confounding by indication. PenSIT Learning is a viable alternative to T-RL for estimation of an optimal multi-stage DTR and may be advantageous in correctly identifying the optimal multi-stage treatment sequence when the underlying DTR is tree-based, sample sizes are small, and/or when the level of confounding is high. Additionally, due to the modeling flexibility afforded by PenSIT Learning, it may be preferred to T-RL in many situations.

In Section 4.2 we introduce relevant notation and formulation, followed by PenSIT Learning methodology in Section 4.3. Implementation of PenSIT Learning is

described in detail in Section 4.4. We present simulation results in Section 4.5 and, in Section 4.6, describe the application of PenSIT Learning to data obtained from the Medical Information Mart for Intensive Care (MIMIC-III) Clinical Database to estimate an optimal, two-stage DTR reflecting restrictive or liberal fluid resuscitation strategies designed to minimize a measure of multi-organ dysfunction.

4.2 Notation and Formulation

4.2.1 Notation

Suppose we are estimating a J-stage treatment regime ($j = 1, ..., J$) in which one of K_j treatments ($k_j = 1, ..., K_j$; $K_j \geq 2$) is administered to every subject $i = 1, ..., n$. Treatment received by the i-th individual at the j-th stage is denoted $A_{j,i} \in \mathcal{A}_j$, with A_j assumed to be categorical. As is customary, a capital letter denotes the random variable with a lower case letter denoting the realized value. We suppress the patient-level indicator i when it can be safely omitted without confusion. Variables collected and available when making the j-th treatment decision are denoted $\mathbf{X}_j$. Following the j-th stage treatment A_j, measurements are made on a set of covariates $\mathbf{X}_{j+1}$, which may also include an intermediate reward outcome, Y_j. We denote the full covariate and treatment history prior to the decision at stage j as $\mathbf{H}_j = (A_1, ..., A_{j-1}, \mathbf{X}_1, ..., \mathbf{X}_j)$. A final outcome of interest $Y = h(Y_1, Y_2, ..., Y_J)$ is a clinically-relevant, prespecified function $h(\cdot)$ of observed, stage-specific intermediate reward outcomes, the higher the better by convention. Common functions for $h(\cdot)$, for example, include the sum or last value. Our fully-observed data, then, represents the collection of independent and identically distributed multivariate observations from subjects $i = 1, ..., n$ in our population of interest and is summarized as follows: $\{A_{1i}, ..., A_{Ji}, \mathbf{X}_{1i}, ..., \mathbf{X}_{Ji}, Y_i\}_{i=1}^n$. Now, further to our estimation goal, we let $\mathbf{g} = (g_1, g_2, ..., g_J)$ denote a J-stage DTR. Each stage-specific decision rule g_j maps history

to the j-th treatment decision, i.e., $g_j : \mathbf{H}_j \rightarrow A_j$. Thus, we can more specifically express $\mathbf{g} = (g_1, g_2, ..., g_J)$ as $\mathbf{g}(\mathbf{H}) = \{g_1(\mathbf{H}_1), g_2(\mathbf{H}_2), ..., g_J(\mathbf{H}_J)\}$.

Following Rubin's potential outcomes framework (*Rubin*, 1974), we use $Y^*(A_1, ..., A_{J-1}, a_J)$, or simply $Y^*(a_J)$, to denote the counterfactual outcome, known interchangeably as a "potential outcome", for a patient treated with $A_j = a_J \in \mathcal{A}_J$ conditional on prior treatment history, $A_1, ..., A_{J-1}$. Similarly, $Y^*(a_1, a_2, ..., a_J)$ identifies the counterfactual outcome under the actual treatment regime $\mathbf{A}$ and $Y^*\{\mathbf{g}(\mathbf{H})\}$ denotes the counterfactual outcome under regime $\mathbf{g}(\mathbf{H})$. Using the counterfactual mean outcome, $E[Y^*\{\mathbf{g}(\mathbf{H})\}]$, to evaluate performance, the optimal DTR, $\mathbf{g}^{\text{opt}}(\mathbf{H})$, is the one that satisfies

$$E[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] \geq E[Y^*\{\mathbf{g}(\mathbf{H})\}]$$

for all $\mathbf{g}(\mathbf{H}) = (g_1, g_2, ..., g_J) \in \mathbf{G}$, where $\mathbf{G}$ is the class of all potential regimes. Our statistical goal, therefore, is: to estimate an interpretable, optimal, J-stage DTR, $\mathbf{g}^{\text{opt}}(\mathbf{H})$, using observational data such that, if all patients were to be assigned to multi-stage treatment using this regime, the expected counterfactual outcome of our population of interest would be maximized: $\mathbf{g}^{\text{opt}}(\mathbf{H}) = \text{argmax}_{\mathbf{g} \in \mathbf{G}} E[Y^*\{\mathbf{g}(\mathbf{H})\}]$.

4.2.2 Link to Observed Data

As mentioned previously, only one of the potential outcomes is observed, making estimation of $\mathbf{g}^{\text{opt}}(\mathbf{H})$ impossible without a series of assumptions. Therefore, we make the following three foundational assumptions: consistency, positivity, and ignorability (*Robins and Hernan*, 2009).

(1) Consistency: The potential outcome under the observed treatment agrees with that of the observed outcome. For example, in the final stage J, we can express this as: $Y_J|A_1, ..., A_{J-1} = \sum_{a_J=1}^{K_J} Y_J^*(A_1, ..., A_{J-1}, a_J)\, I(A_J = a_J)$, where $I(\cdot)$ is an indicator

function that returns a value of 1 if the argument is true and a value of 0 otherwise. Consistency further assumes that there is no interference between units, which means that one patient's observed and counterfactual outcomes are independent of the treatments of all other patients.

(2) Positivity: An assumption of positivity is fulfilled if there is a positive probability for each subject of being assigned $A_j = a_j \in \mathcal{A}_j$ at the j-th stage treatment decision conditional on history $\mathbf{H}_{j,i}$. Expressed mathematically, under the positivity assumption $0 < \tau < Pr(A_{j,i} = a_{j,i}|\mathbf{H}_{j,i}) = \pi_{a_j,i}(\mathbf{H}_{j,i}) < 1$ for all $a_j \in \mathcal{A}_j, \mathbf{H}_j \in \boldsymbol{\mathcal{H}_j}$ and for all $j = 1, ..., J$, where $\pi_{a_j,i}(\mathbf{H}_{j,i})$ represents the propensity score for subject i and τ represents a positive constant.

(3) Ignorability: Also known as the assumption of no unmeasured confounders (NUCA), ignorability implies that data on all variables that are associated both with the assignment of A and the outcome Y have been observed. Furthermore, ignorability implies that the counterfactual outcomes are independent of the treatment given the propensity score. For example, $Y_J^*(1), ..., Y_J^*(K) \perp\!\!\!\perp A_J|\pi_J(\mathbf{H}_J)$ (*Rosenbaum and Rubin*, 1983; *Zhang and Little*, 2009; *Zhou et al.*, 2019).

Under the assumptions of consistency, positivity, and ignorability, it can be shown that $E(Y|A_J = a_J, \mathbf{H}_J)$ is a consistent estimator of $E\{Y^*(a_J)\}$. Following the derivation for a single stage in *Tao and Wang* (2017), we can express the optimal decision rule for the final stage as:

$$g_J^{\text{opt}}(\mathbf{H}_J) = \text{argmax}_{g_J \in \mathcal{G}_J} E_{\mathbf{H}_J}\left[\sum_{a_J=1}^{K_J} E\big(Y|A_J = a_J, \mathbf{H}_J\big)\, I\{g_J(\mathbf{H}_J) = a_J\}\right]. \qquad (4.1)$$

Further, following the theory of *Rosenbaum and Rubin* (1983), conditioning on all the full history, $\mathbf{H}_J$, is mathematically the same as conditioning on the propensity score.

Therefore, we also have:

$$g_J^{\text{opt}}(\mathbf{H}_J) = \text{argmax}_{g_J \in \mathcal{G}_J} E_{\mathbf{H}_J}\left[\sum_{a_J=1}^{K_J} E\{Y|A_J = a_J, \pi_{a_J}(\mathbf{H}_J)\}\, I\{g_J(\mathbf{H}_J) = a_J\}\right]. \tag{4.2}$$

Note that Equations (4.1) and (4.2) are solvable using observed data. Based on the above, we formally introduce our proposed PenSIT Learning in the following section.

4.3 Penalized Spline-Involved Tree-based (PenSIT) Learning

4.3.1 Tree-based Learning

As is well known in the CART literature (*Breiman et al.*, 1984; and others) and also as discussed in tree-based optimal DTR estimation literature (*Laber and Zhao*, 2015; *Sun and Wang*, 2020; *Tao et al.*, 2018), a decision tree is a widely-used machine learning technique constructed by recursive partitioning of the covariate space that is used to identify features and their associated cutpoints that are best able to describe the relative homogeneity of another variable or outcome. The result of a decision tree analysis can be represented in a tree-like structure with nodes representing distinct features and leaves capturing those observations identifed as most similar or "pure". A purity metric, $P(\Omega_m, \omega)$, where Ω_m refers to a parent node indexed by m and ω (with its complement ω^C) refers to a specific partition applied to Ω_m, is a criterion used to determine which of these binary splits ω will be applied at Ω_m. Purity measures used previously in the CART literature, for example, include misclassification rates for binary outcomes or residual sum of squares for continuous outcomes (*Hastie et al.*, 2009). In addition to a purity measure, estimation in the tree-based learning context relies upon a set of criteria needed to establish whether a parent node will or will not be partitioned. These criteria, often termed "stopping criteria", are determined by

pre-specified, user-defined inputs and are discussed in Section 4.4.3.

Here it is important to note the differences between estimation using CART versus estimation of an optimal DTR using tree-based methods. First, CART is a supervised learner whereas estimation of an optimal DTR is unsupervised; in CART, the object of estimation, which is often an outcome of interest, is directly observed whereas, for optimal DTR estimation, the object of estimation is an optimal treatment sequence that is not directly observed. Secondly, while CART is used for prediction of a target variable based on observed covariates, the goal of optimal DTR estimation represents the crux of personalized medicine: to estimate stage-specific and dynamic decision rules for assigning treatment to individuals based on their unique demographic or disease-specific characteristics such that overall (counterfactual) outcomes for the population will be maximized. Given that optimal DTR estimation has a causal goal, purity measures used previously in this context are derived from estimators of the counterfactual mean outcome; these include an IPW-based estimator (*Laber and Zhao*, 2015), an AIPW-based estimator (*Tao et al.*, 2018), and others. As will be shown in the following sections, we replace the expression $E\big\{Y|A_J = a_J, \mathbf{H}_J\big\}$ identified in IPW-based estimators with the expression $E\big\{Y|A_J = a_J, \pi_{a_J}(\mathbf{H}_J)\big\}$ introduced in Equation (4.2).

4.3.2 PenSIT Estimation for Final Stage J

For simplicity, denote $\mu_{J,a_J}(\mathbf{H}_J) = E\big\{Y|A_J = a_J, \pi_{a_J}(\mathbf{H}_J)\big\}$. We propose to model $\mu_{J,a_J}(\mathbf{H}_J)$ as:

$$\hat{\mu}_{J,a_J}(\mathbf{H}_J) = s[\, l\{\hat{\pi}_{J,a_J}(\mathbf{H}_J)\}\, ; \boldsymbol{\theta}_{J,a_J}] + r_{J,a_J}(\mathbf{H}_J; \boldsymbol{\beta}_{J,a_J}), \tag{4.3}$$

where $l(\cdot)$ is a pre-specified transformation of the propensity score, e.g., logit or identity; $s(\cdot)$ denotes a penalized spline with fixed knots (*Eilers and Marz*, 1996; *Wand*,

2003) indexed using the parameters $\boldsymbol{\theta}_{J,a_J}$; and $r_{J,a_J}(\cdot)$ refers to a parametric function of other covariates in $\mathbf{H}_J$, indexed by the parameters $\boldsymbol{\beta}_{J,a_J}$. Note in Equation (4.3) that the conditional mean outcome is modeled semiparametrically using covariates predictive of the outcome and a penalized spline of a function of the propensity score, which makes the model more flexible and easier to achieve correct model specification. We then propose to estimate the counterfactual mean outcome under treatment a_J, i.e., $E\{Y^*(a_J)\}$, as $\mathbb{P}_n\{\hat{\mu}^{\text{PenSI}}_{J,a_J}(\mathbf{H}_J)\}$, where we introduce the PenSI estimator as $\hat{\mu}^{\text{PenSI}}_{J,a_J}(\mathbf{H}_J) = I(A_J = a_J)Y + I(A_J \neq a_J)\cdot\hat{\mu}_{J,a_J}(\mathbf{H}_J)$, and $\mathbb{P}_n(\cdot)$ refers to the empirical mean operator. Following *Zhou et al.* (2019), we can show that $\mathbb{P}_n\{\hat{\mu}^{\text{PenSI}}_{J,a_J}(\mathbf{H}_J)\}$ is consistent and doubly robust for $E\{Y^*(a_J)\}$.

Proposition IV.1. *Assuming observations* $\{\boldsymbol{H}_{Ji}, A_{Ji}, Y_i\}_{i=1}^n$ *are independent and identically distributed across all individuals* i *and following some unspecified multivariate distribution* $\boldsymbol{p}$, $\mathbb{P}_n\{\hat{\mu}^{PenSI}_{J,a_J}(\boldsymbol{H}_J)\}$ *is consistent and doubly robust for* $E\{Y^*(a_J)\}$ *if either of the following conditions are met:*

(i) The conditional mean model $\hat{\mu}_{J,a_J}(\boldsymbol{H}_J)$ *consisting of* $s[\,l\{\hat{\pi}_{J,a_J}(\boldsymbol{H}_J)\}\,|\boldsymbol{\theta}_{J,a_J}]$ *and* $r_{J,a_J}(\boldsymbol{H}_J|\boldsymbol{\beta}_{J,a_J})$ *is correctly specified.*

(ii) The parametric component of the conditional mean model, $r_{J,a_J}(\boldsymbol{H}_J|\boldsymbol{\beta}_{J,a_J})$*, is misspecified but the propensity models* $\pi_{J,a_J}(\boldsymbol{H}_J)$ *and the relationship between the outcome and the function of the propensity score* $s[\,l\{\hat{\pi}_{H,a_J}(\boldsymbol{H}_J)\}\,|\boldsymbol{\theta}_{J,a_J}]$ *are correctly specified.*

Proof. $\mathbb{P}_n\{\hat{\mu}^{\text{PenSI}}_{J,a_J}(\mathbf{H}_J)\} = \mathbb{P}_n\{I(A_J = a_J)Y + I(A_J \neq a_J)\hat{\mu}_{J,a_J}(\mathbf{H}_J)\}$. Let n_1 be the number of observations with $A_J = a_J$ and n_0 the number of observations with $A_J \neq a_J$. Then $\mathbb{P}_n\{I(A_J = a_J)Y\} = \dfrac{n_1}{n}\cdot\dfrac{1}{n_1}\displaystyle\sum_{n_1:A_J=a_J} Y$ and $\mathbb{P}_n\{I(A_J \neq a_J)\hat{\mu}_{J,a_J}(\mathbf{H}_J)\} = \dfrac{n_0}{n}\cdot\dfrac{1}{n_0}\displaystyle\sum_{n_0:A_J\neq a_J}\hat{\mu}_{J,a_J}(\mathbf{H}_J)$. Under the assumptions of consistency, positivity, and ignorability, and following the proof in the supplementary material Section 1.1 of *Zhou et al.* (2019), it can be shown that $\mathbb{P}_n\{\hat{\mu}^{\text{PenSI}}_{J,a_J}(\mathbf{H}_J)\}$ is doubly robust. □

The consistency of $\mathbb{P}_n\{\hat{\mu}^{\text{PenSI}}_{J,a_J}(\mathbf{H}_J)\}$ offers a valid large sample estimate of the counterfactual mean outcome under treatment $A_J = a_J$ conditional on prior treatment $A_1, ..., A_{J-1}$; additionally, the double robustness of the the PenSI estimator provides two opportunities for consistent estimation. Furthermore, *Zhang and Little* (2009) maintain that, by regressing the outcome Y on a spline of the logit of the propensity score, condition (ii) in Proposition IV.1 is met under relatively weak conditions due to the modeling flexibility of a spline, which requires minimal assumptions about the relationship between the outcome Y and the propensity for treatment assignment $\pi_{J,a_J}(\mathbf{H}_J)$.

Given our goal to estimate a treatment regime $g_J(\mathbf{H}_J)$ that maximizes the counterfactual mean outcome following regime $g_J(\mathbf{H}_J)$, we estimate $E[Y^*\{g_J(\mathbf{H}_J)\}]$ using our proposed PenSI estimator as follows:

$$\hat{E}[Y^*\{g_J(\mathbf{H}_J)\}] = \mathbb{P}_n\left[\sum_{a_J=1}^{K_J} \hat{\mu}^{\text{PenSI}}_{J,a_J}(\mathbf{H}_J) I\{g_J(\mathbf{H}_J) = a_J\}\right]$$

Based on this formulation we then propose a purity measure, $P_J^{\text{PenSI}}(\Omega_m, \omega)$, suitable for constructing a tree when estimating an optimal treatment rule at the final, J-th stage:

$$P_J^{\text{PenSI}}(\Omega_m, \omega) = \max_{a_1, a_2 \in \mathcal{A}_J} \mathbb{P}_n\left[\sum_{a_J=1}^{K_J} \hat{\mu}^{\text{PenSI}}_{J,a_J}(\mathbf{H}_J) I\{g_{J,\omega,a_1,a_2}(\mathbf{H}_J) = a_J\} I(\mathbf{H}_J \in \Omega_m)\right] \tag{4.4}$$

Specifically, $P_J^{\text{PenSI}}(\Omega_m, \omega)$ refers to the maximum empirical version of the expected counterfactual outcome under decision rule g_{J,ω,a_1,a_2} when node Ω_m is split according to partition ω such that patients in subset ω are assigned treatment a_1 while patients in the complementary set ω^C are assigned to a_2, for $a_1 \neq a_2$, and where subscripts 1 and 2 need not refer to treatment $A_J = 1$ and $A_J = 2$.

4.3.3 PenSIT Estimation for Stages $1, ..., J-1$

Similar to other tree-based optimal DTR estimation methods (*Sun and Wang*, 2020; *Tao et al.*, 2018), estimation proceeds in a backwards recursive manner (*Bather*, 2000), beginning with estimation of the J-th stage decision rule. This is important to account for time-varying confounding by indication and delayed effects related to treatments received at earlier stages, both of which can result in biased estimation. Furthermore, in the context of estimation of an optimal multi-stage DTR, the optimal j-stage decision rule relies upon the patient receiving the optimal treatment at all future stages.

Following our exposition in the previous sections, $g_J^{\text{opt}}(\mathbf{H}_J)$ is estimated within the tree-based construct using the PenSI purity measure for the J-th stage, $P_J^{\text{PenSI}}(\Omega_m, \omega)$, introduced in Equation (4.4). In order to generalize for estimation of the j-th stage decision rule, for $j = 1, ..., J-1$, we now introduce additional notation pertaining to estimation for the j-th stage. Let $\tilde{Y}_j(a_j)$ refer to the predicted pseudo-outcome at stage j under treatment a_j, which is never actually observed. The assumption under an optimal, multi-stage treatment assignment regime is that the long-term outcome Y is maximized. Therefore, when estimating the decision rule for the j-th treatment stage, we must account for the fact that the patient was treated with the optimal treatment at all future stages. To this end we construct a stage-specific counterfactual pseudo-outcome $\tilde{Y}_j$ for any stage prior to the last, which represents the predicted counterfactual outcome at the j-th stage contingent upon the patient receiving the optimal treatment at all future stages. Mathematically this can be expressed as: $\tilde{Y}_j = \hat{E}\{Y^*(A_1, ..., A_j, g_{j+1}^{\text{opt}}, ..., g_J^{\text{opt}})\}$. Under an assumption of consistency, $\tilde{Y}_j = \sum_{a_j=1}^{K_j} \tilde{Y}_j(a_j) I(A_j = a_j)$. Positivity in the multi-stage setting was introduced in Section 4.2.2 and the assumption of ignorability can be expressed for the j-th stage estimation as $\{\tilde{Y}_j(1), ..., \tilde{Y}_j(K_j)\} \perp\!\!\!\perp A_j | \mathbf{H}_j = \{\tilde{Y}_j(1), ..., \tilde{Y}_j(K_j)\} \perp\!\!\!\perp A_j | \boldsymbol{\pi}_{a_j}(\mathbf{H}_j)$. Therefore, similar to that introduced in Equation (4.2), the optimal decision rule at

the j-th stage can be expressed as a function of the predicted pseudo-outcome, with $\hat{E}\{\tilde{Y}_j|A_j = a_j, \pi_{j,a_j}(\mathbf{H}_j)\} = \tilde{\mu}_{j,a_j}(\mathbf{H}_j)$:

$$g_j^{\text{opt}}(\mathbf{H}_j) = \text{argmax}_{g_j \in \mathcal{G}_j} E_{\mathbf{H}_j}\left[\sum_{a_j=1}^{K_j} \tilde{\mu}_{j,a_j}(\mathbf{H}_j) I\{g_j(\mathbf{H}_j) = a_j\}\right]$$

We define $\tilde{\mu}_{j,a_j}^{\text{PenSI}}(\mathbf{H}_j) = I(A_j = a_j) \cdot \tilde{Y}_j + I(A_j \neq a_j) \cdot \tilde{\mu}_{j,a_j}(\mathbf{H}_j)$ and propose to estimate the mean pseudo-outcome under treatment a_j as $\mathbb{P}_n\{\tilde{\mu}_{j,a_j}^{\text{PenSI}}(\mathbf{H}_j)\}$, where, using the notation and modeling choices introduced in Equation (4.3):

$$\tilde{\mu}_{j,a_j}(\mathbf{H}_j) = s[\, l\{\hat{\pi}_{j,a_j}(\mathbf{H}_j)\}\, |\boldsymbol{\theta}_{j,a_j}] + r_{j,a_j}(\mathbf{H}_j|\boldsymbol{\beta}_{j,a_j})$$

Assuming consistent estimation at all future stages through backward induction and following Proposition IV.1 and *Zhou et al.* (2019), $\mathbb{P}_n\{\tilde{\mu}_{j,a_j}^{\text{PenSI}}(\mathbf{H}_j)\}$ is a consistent and doubly robust estimator for $E\{\tilde{Y}_j(a_j)\}$. The associated PenSI purity measure used at the j-th treatment stage, i.e., $P_j^{\text{PenSI}}(\Omega_m, \omega)$, can then be defined as follows, where node Ω_m is split by the partition identified by ω based on rule g_{j,ω,a_1,a_2}, which assigns treatment $A_j = a_1$ to patients in the set defined by ω and assigns $A_j = a_2$ to those in ω^C, for $a_1 \neq a_2$:

$$P_j^{\text{PenSI}}(\Omega_m, \omega) = \max_{a_1,a_2 \in \mathcal{A}_j} \;\; \mathbb{P}_n\left[\sum_{a_j=1}^{K_j} \tilde{\mu}_{j,a_j}^{\text{PenSI}}(\mathbf{H}_j) \;\; I\{g_{j,\omega,a_1,a_2}(\mathbf{H}_j) = a_j\} \;\; I(\mathbf{H}_j \in \Omega_m)\right]$$

4.4 Implementation

4.4.1 Estimation of $\tilde{\mu}_{a_j}^{\text{PenSI}}(\mathbf{H}_j)$

Stage-specific PenSI estimates are a key component of the PenSI purity measure at each stage, $j = 1, ..., J$. We refer to $\tilde{\mu}_{a_j}^{\text{PenSI}}(\mathbf{H}_j)$ for the generic j-th stage throughout, but note in the case of estimation for the final J-th stage that $\tilde{\mu}_{a_j}^{\text{PenSI}}(\mathbf{H}_j) = \hat{\mu}_{a_J}^{\text{PenSI}}(\mathbf{H}_J)$. Furthermore, although we refer to modeling of the pseudo-outcome $\tilde{Y}_j$

throughout, it is understood that, for the final stage J, $\tilde{Y}_j = Y_J$. We assume outcomes are continuous and approximately normally distributed; however, outcomes under other distributional assumptions can be accommodated under a generalized linear modeling framework with a link function appropriate for the outcome of interest. We assume (pseudo-) outcomes for all future stages, i.e., $j = j+1, ..., J$, have been consistently estimated. For simplicity in the implementation guide below we default to including all covariates in $\mathbf{H}_j$; however, more sophisicated models that include interactions and/or transformations or include only a subset of covariates can and should be considered.

1. Estimate the propensity model for treatment assignment at the j-th stage using the full observed dataset.

   (a) Estimate model parameters ($\hat{\boldsymbol{\gamma}}_{a_j}$) for the propensity score, i.e., the probability to be assigned treatment $A_j = a_j$ given history $\mathbf{H}_j$: $\hat{\pi}_{a_j}(\mathbf{H}_j) = \hat{Pr}(A_j = a_j|\mathbf{H}_j; \hat{\boldsymbol{\gamma}}_{a_j})$. Select a parametric regression model suitable for the scale of the treatment variable A_j, e.g., logistic regression for binary treatment using the `glm` function in `R` (*R Core Team*, 2018) or multinomial logistic regression for an ordinal treatment using the function `multinom`.

   (b) Using the estimates of $\hat{\boldsymbol{\gamma}}_{a_j}$ obtained in 1(a) and the observed $\mathbf{h}_{j,i}$, calculate for all i the predicted probabilities of assignment to $A_j = a_j$ for $\forall A_j \in \mathcal{A}_j$. Then calculate for all i the estimated propensity of receiving the observed treatment $A_{j,i} = a_{j,i}$. For example, when $A_j \in \{0, 1\}$ with the reference category assigned as $A_j = 1$, the propensity model is $\hat{Pr}(A_{j,i} = 1|\mathbf{H}_{j,i}, \hat{\boldsymbol{\gamma}}_{a_j}) = \hat{\pi}_{1,i}(\mathbf{H}_{j,i}; \hat{\boldsymbol{\gamma}}_{a_j})$ and the propensity of receiving the observed treatment $A_{j,i} = a_{j,i}$ can be computed as: $I(A_{j,i} = 1) \cdot \hat{\pi}_{1,i}(\mathbf{H}_{j,i}) + I(A_{j,i} = 0) \cdot \{1 - \hat{\pi}_{1,i}(\mathbf{H}_{j,i})\}$.

   (c) We define $l\{\, \hat{\pi}_{a_j}(\mathbf{H}_j)\,\} = \text{logit}\{\, \hat{\pi}_{a_j}(\mathbf{H}_j)\,\} = \log\left[\hat{\pi}_{a_j}(\mathbf{H}_j)/\{1 - \hat{\pi}_{a_j}(\mathbf{H}_j)\}\right]$

for consistency with *Zhou et al.* (2019), but other choices for $l(\cdot)$, e.g., the identity function, may be used.

2. For each $A_j = a_j \in \mathcal{A}_j$: Using only those observations assigned to treatment group $A_j = a_j$, estimate prediction models for $\tilde{Y}_j(a_j)$ (or, in the case of the final stage J, for $Y^*(a_J)$).

   (a) For a desired number of knots D and using $l\{\,\hat{\pi}_{a_j}(\mathbf{H}_j)\,\}$ estimated in Step 1(c), determine the D locations of the knots for the spline model. We set the default to $D = \min(35, 1/4 \cdot n_{a_j})$ (*Rupert*, 2002; *Wand*, 2003) with equal knot spacing, where n_{a_j} refers to the sample size for observations with $A_j = a_j$.

   (b) Create a basis matrix for the spline with fixed knots $B_1, ..., B_D$. We default to a truncated linear basis (*Wand*, 2003): $[\,l\{\hat{\pi}_{a_j,i}(\mathbf{H}_j)\,\} - B_d]_+ = \max[l\{\hat{\pi}_{a_j,i}(\mathbf{H}_j)\} - B_d, 0]$; however, alternate basis specifications are possible. The dimension of the basis matrix will be n_{a_j} x D.

   (c) Fit a regression model for the mean stage j pseudo-outcome $\tilde{Y}_j$ as follows:

$$\hat{E}\{\tilde{Y}_j|\mathbf{H}_j, A_j; \boldsymbol{\theta}_{a_j}, \boldsymbol{\beta}_{a_j}\} = \tilde{\mu}_{j,a_j}(\mathbf{H}_j) = s[\,l\{\hat{\pi}_{a_j}(\mathbf{H}_j)\}\,; \boldsymbol{\theta}_{a_j}\,] + r_{a_j}(\mathbf{H}_j; \boldsymbol{\beta}_{a_j})$$

   where $s[\,l\{\hat{\pi}_{a_j}(\mathbf{H}_j)\}\,; \boldsymbol{\theta}_{a_j}\,] = \theta_{j0,a_j} + \theta_{j1,a_j} l\{\hat{\pi}_{a_j}(\mathbf{H}_j)\,\} + \sum_{k=1}^{K} \theta_{j1k,a_j}[\,l\{\hat{\pi}_{a_j}(\mathbf{H}_j)\,\} - B_d]_+$. $r_{a_j}(\cdot)$ represents a parametric function of other covariates predictive of the outcome, indexed by parameters $\boldsymbol{\beta}_{a_j}$. Due to the stratified nature of model estimation used for $\tilde{\mu}_{a_j}(\mathbf{H}_j)$, different variables can be selected to model the pseudo-outcomes for each treatment $A_j \in \mathcal{A}_j$. We can express the spline model as a linear mixed model (*Wand*, 2003; *Zhou et al.*, 2019), where $\mathbf{C}_1 = [1, \mathbf{H}_j, l\{\hat{\pi}_{a_j}(\mathbf{H}_j)\}]$ represents a design matrix for the fixed effects with model parameters $\boldsymbol{\beta}_{a_j}$

and $\mathbf{C}_2 = ([l\{\hat{\pi}_{a_j}(\mathbf{H}_j)\} - B_1]_+, ..., [l\{\hat{\pi}_{a_j}(\mathbf{H}_j)\} - B_D]_+)$ is the design matrix for the truncated linear basis matrix obtained above with parameters $\boldsymbol{\theta}_{a_j}$, which are included as random effects. Linear mixed models can be easily fit with standard software (e.g., `lmer` function in `R`) using restricted maximum likelihood estimates of the parameters $\boldsymbol{\theta}_{a_j}$ and $\boldsymbol{\beta}_{a_j}$.

(d) Predict $\hat{Pr}(A_{j,i} = a_j|\mathbf{H}_{j,i} = \mathbf{h}_{j,i})$ using the fully observed $\mathbf{h}_j$ and the propensity model estimated in Step 1(a); then apply $l(\cdot)$ to obtain the transformation of the estimated propensity to receive treatment $A_{j,i} = a_j$ for all i.

(e) Using the estimated models in Steps 2(c) and 2(d) above, predict the individual-level counterfactual (pseudo-) outcomes $\tilde{Y}_{j,i}(a_j)$ under $A_j = a_j$.

(f) Obtain $\tilde{\mu}^{\text{PenSI}}_{j,a_j,i}(\mathbf{H}_{j,i}) = I(A_{j,i} = a_j) \cdot \tilde{Y}_{j,i} + I(A_{j,i} \neq a_j) \cdot \tilde{\mu}_{j,a_j,i}(\mathbf{H}_{j,i})$ for all individuals i.

(g) Repeat Steps 2(a) to 2(f) for all $a_j \in \mathcal{A}_j$ and obtain a n x $|A_j|$ matrix representing the estimated counterfactual outcomes under each treatment $a_j \in \mathcal{A}_j$.

4.4.2 Selection of tuning parameters for tree-based estimation

Several user-defined inputs are needed to implement PenSIT Learning. First, a positive value, λ_j, must be specified in order to determine whether a potential split of node Ω_m by partition ω identifies a meaningful difference in purity, i.e., $P^{\text{PenSI}}(\Omega_m, \omega) - P^{\text{PenSI}}(\Omega_m) > \lambda_j$. We recommend that λ be selected to represent a level of clinical or practical significance determined based on clinical knowledge or practical rationale, although λ may also be chosen adaptively from the data (*Hastie et al.*, 2009; *Tao et al.*, 2018). One strategy, for example, grows a full tree by selecting a small value of λ_j and then prunes the tree using a cost-complexity measure (*Boehmke and Greenwell*, 2020; *Breiman et al.*, 1984; *Hastie et al.*, 2009; *Therneau*

et al., 2019). Following *Tao et al.* (2018), another approach is to evaluate the estimated counterfactual mean outcome across a reasonable range of values for λ and select the value of λ using 10-fold cross validation that maximizes the estimated counterfactual mean.

Two other user-specified tuning parameters are also necessary at each stage to perform PenSIT Learning: the minimum number of observations that can fall into each of the terminal nodes, $n_{0,j}$, and a maximum depth to which the tree is allowed to grow, d_j. Generally, the smaller the minimum node size and the larger the depth, the more complex the estimated optimal decision rule will be, leading to the potential of overfitting (*Sun and Wang*, 2020). An optimal range for the minimum node size in a CART-type analysis between 1-20 has been suggested (*Mantovani et al.*, 2019). Similarly, a depth of 5 if often considered a good starting point (*Boehmke and Greenwell*, 2020).

4.4.3 PenSIT Learning Node Splitting and Stopping Rules

As discussed above, tree-based estimation is performed using backward induction, commencing with the J-th stage and ending with Stage 1. Input for the tree-based partitioning at each stage j include estimated counterfactual outcomes $\tilde{\mu}_{a_j}^{\text{PenSI}}(\mathbf{H}_j)$ for all $A_j = a_j \in \mathcal{A}_j$ and user pre-specified λ_j, $n_{0,j}$, d_j, as mentioned above. Using the framework of *Tao et al.* (2018), the following terminal criteria determine when a node $\Omega_{j,m}$ becomes a terminal node in the j-th stage estimation:

- If the size of a node $\Omega_{j,m}$ is less than twice the minimum node size $2n_{0,j}$, i.e., $|\Omega_{j,m}| < 2n_{0,j}$, $\Omega_{j,m}$ becomes a terminal node.

- If all possible splits of $\Omega_{j,m}$ result in child nodes with fewer than $n_{0,j}$ observations, i.e., $|\omega_{j,m}| < n_{0,j}$, for all possible partitions $\omega_{j,m}$, then $\Omega_{j,m}$ becomes a terminal node.

- If the tree depth reaches the pre-specified depth d_j, all nodes $\Omega_{j,m}$ at depth d_j become terminal nodes.

The process of recursively splitting the tree into successively smaller partitions is conducted as follows. Begin with root node $\Omega_{j,m}$, for $m = 1$.

1. At node $\Omega_{j,m}$, evaluate the three terminal criteria above.

   - If at least one termination criterion is satisfied, no splits of the node are carried out. Assign a single best treatment to all subjects in $\Omega_{j,m}$: $\text{argmax}_{a_j \in \mathcal{A}_j} \mathbb{P}_n\{P_j^{\text{PenSI}}(\Omega_{j,m})\}$, where $P_j^{\text{PenSI}}(\Omega_{j,m})$ refers to the purity in the absence of a split.

   - If no termination criteria are met, determine the best split as follows: $\hat{\omega}_{j,m}^{\text{opt}} = \text{argmax}_{\omega_{j,m}}\{P_j^{\text{PenSI}}(\Omega_{j,m}, \omega_{j,m})\}$.

     – If $P_j^{\text{PenSI}}(\Omega_{j,m}, \omega_{j,m}) - P_j^{\text{PenSI}}(\Omega_{j,m}) \leq \lambda_j$, no split of the node is carried out. Assign a single best treatment to all subjects in $\Omega_{j,m}$: $\text{argmax}_{a_j \in \mathcal{A}_j} \mathbb{P}_n\{P_j^{\text{PenSI}}(\Omega_{j,m})\}$

     – If $P_j^{\text{PenSI}}(\Omega_{j,m}, \omega_{j,m}) - P_j^{\text{PenSI}}(\Omega_{j,m}) > \lambda_j$, split $\Omega_{j,m}$ into child nodes $\Omega_{j,2m}$ and $\Omega_{j,2m+1}$ as determined by $\hat{\omega}_{j,m}^{\text{opt}}$.

2. If all nodes are terminal nodes, stop. If not, set $m = m+1$ and repeat Step (a).

4.5 Simulation Studies

We consider an observational study for a two-stage DTR with 2 treatment options per stage, with independent individuals. We evaluate sample sizes of $n = 300, 500, 1000$. For each individual we generate a set of $20, 50$, or 100 baseline covariates, $\mathbf{X}_1$, from a multivariate normal distribution with a mean of $\mathbf{0}_{|X_1|}$ and an exchangeable correlation structure defined by correlation coefficient $\rho = 0.2$, where $|\mathbf{X}_1|$ refers to the cardinality, or size, of the vector $\mathbf{X}_1$ for each individual. First stage

treatment $A_1 \in \{0, 1\}$ is generated from a binomial distribution, with the data generating mechanisms reflecting varying degrees of confounding. Specifically, we evaluate a low, moderate, and high degree of confounding of the outcome using the following specifications, with $\pi_{10} = 1 - \pi_{11}$:

- Low: $\pi_{11} = \exp(0.5X_4 + 0.5X_1)/\{1 + \exp(0.5X_4 + 0.5X_1)\}$
- Moderate: $\pi_{11} = \exp(0.5X_4 + 1.5X_1)/\{1 + \exp(0.5X_4 + 1.5X_1)\}$
- High: $\pi_{11} = \exp(0.5X_4 + 3.5X_1)/\{1 + \exp(0.5X_4 + 3.5X_1)\}$

The stage 1 optimal decision rule assuming an underlying tree-type DTR is $g_1(\mathbf{H}_1) = I(X_1 > -0.5 \,\&\, X_2 > -0.5)$ and $g_1(\mathbf{H}) = I(\sqrt{|X_2 + 5|} + 0.5X_1 < 2.2)$ when an underlying non-tree-type DTR structure is assumed. The intermediate reward outcome following Stage 1 is generated as $Y_1 = \exp\{1.5 + 0.3X_1 - |1.5X_6 - 2| \cdot I(A_1 \neq g_1^{\text{opt}}(\mathbf{H}_1))\} + \epsilon_1$, where $\epsilon_1 \sim N(0, 1^2)$. This reflects an unequal penalty dependent on one of the observed covariate values if the patient was not treated according to their optimal therapy; this is intended to add an additional degree of complexity into the data generating scenario and be more reflective of data that may be encountered in a real world setting. Second stage treatment $A_2 \in \{0, 1\}$ is generated from a binomial distribution and is also adapted to reflect varying degrees of confounding, as follows, with $\pi_{20} = 1 - \pi_{21}$:

- Low: $\pi_{21} = \exp(0.2Y_1 + 0.5X_5 - 0.5)/\{1 + \exp(0.2Y_1 + 0.5X_5 - 0.5)\}$
- Moderate: $\pi_{21} = \exp(0.2Y_1 + 1.5X_5 - 0.5)/\{1 + \exp(0.2Y_1 + 1.5X_5 - 0.5)\}$
- High: $\pi_{21} = \exp(0.2Y_1 + 3.5X_5 - 0.5)/\{1 + \exp(0.2Y_1 + 3.5X_5 - 0.5)\}$

The stage 2 optimal decision rule assuming an underlying tree-type DTR is $g_2(\mathbf{H}_2) = I(X_3 > -1 \,\&\, Y_1 > 2)$ and, when an underlying nontree-type DTR structure is assumed, $g_2(\mathbf{H}_2) = I\{\log_2(|Y_1| + 1) + 3X_3 \leq 1.8\}$. The final outcome $Y = Y_1 + Y_2$

where $Y_2 = \exp\{1.18 + 0.2X_5 - |1.5X_7 + 2| \cdot I(A_2 \neq g_2^{\text{opt}}(\mathbf{H}_2))\} + \epsilon_2$, with $\epsilon_2 \sim N(0, 1^2)$. Under optimal treatment allocation and assuming independence across observations, $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] = 8.0$.

We compare the performance of PenSIT Learning to tree-based reinforcement learning (T-RL), Q-Learning using linear modeling (Q-Linear), and Q-Learning using nonparametric modeling (Q-NP). T-RL is a tree-based DTR estimation method that uses a purity measure constructed using the AIPW estimator of the counterfactual mean outcome (Tao et al., 2018). Q-Learning using linear regression modeling assumes a linear and additive relationship between the covariates and the expected outcome (using the `lm` function in `R`). Q-Learning with nonparametric modeling allows a more flexible relationship for the Q-functions, which are estimated using random forest prediction (using `randomForest` in `R`). Both T-RL and PenSIT Learning use random forest prediction to generate stage 1 pseudo-outcomes. Across all simulation studies we assume that the parametric component of the conditional mean model in PenSIT Learning, $r_{j,a_j}(\mathbf{H}_j|\boldsymbol{\beta}_{j,a_j})$, is incorrectly specified for all $a_j \in \mathcal{A}_j$ in order to evaluate performance under a scenario more reflective of the real world. Performance of PenSIT Learning and T-RL are evaluated under both a correctly- and incorrectly-specified propensity model. Although consistent estimation of the counterfactual mean outcome is not ensured when the propensity model is misspecified, we present results for an incorrectly-specified propensity model in order to demonstrate performance when this assumption is not met, which may be likely to occur in practice. With a test set of size 1000 ($N_{\text{test}} = 1000$), performance is evaluated using two metrics: 1) the percentage of observations correctly classified to their optimal DTR, $\%opt$, and 2) the estimated counterfactual mean outcome $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$, which reflects the expected counterfactual outcome had everyone in the patient population of interest been treated "optimally" based on the regime estimated using each respective method. For each simulation design setting we tabulate the median and interquartile

range (IQR) of $\%opt$ and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ across all $B = 500$ Monte Carlo iterations. For all stages within each simulation experiment, λ is set as a 5% improvement although we find in supplemental simulations that PenSIT Learning appears to be more sensitive to perturbations of λ than does T-RL, particularly with a low degree of confounding. For all simulations we set the minimum node size and tree-depth at 20 and 5, respectively. We tabulate results based on the logit transformation of the estimated propensity score; we also explored the use of the identity transformation and found performance differences of the identity transformation compared to the logit transformation to be negligible (results not shown). We also investigated performance using different knot sizes, from 5-35 knots, and also using stepwise variable selection based on improvement of AIC; no meaningful differences based on these modifications were found (results not shown), suggesting robustness of this method to the choices of knot size and knot placement.

As can be seen generally across all out simulation results (Tables 4.1 - 4.9), PenSIT Learning performance improves with sample size when the number of baseline covariates are held fixed, as expected. When the sample size is fixed, performance generally worsens as the number of baseline covariates increases, although this is most apparent with smaller sample sizes. When the underlying DTR structure is tree-type, PenSIT Learning performs well across all data generating settings (Tables 4.1 - 4.3), although performance is best when the level of confounding is moderate (Table 4.2) or high (Table 4.3). When $n = 500$ and $|\mathbf{H}| = 20$, for example, PenSIT Learning is able to correctly identify the optimal DTR more than 95% of the time when confounding is moderate or high, but this falls to around 85% correct classification of the optimal DTR with a low degree of confounding. Additionally, the percentage of correctly-classified treatments for PenSIT Learning is similar–typically within 1%–when the propensity model is either correctly or incorrectly specified across all covariate cardinality (Tables 4.1 - 4.3). When confounding is low, we observe a

high degree of variability for PenSIT Learning in both the estimated percentage of correctly-classified observations and in the estimated counterfactual mean outcome. When $n = 500$ and $|\mathbf{H}| = 20$, for example, the interquartile range of the correctly-classified optimal DTR is more than 13%, compared with less than half that value when confounding is moderate or high. With a sample size of $n = 300$ (top panels, Tables 4.7 - 4.9), PenSIT Learning under an assumed tree-type DTR reveals similar performance to that discussed above, exceeding 90% correct classification with moderate and high levels of confounding across all settings, including a correct or incorrectly-specified propensity model. When the underlying DTR is nontree-type, PenSIT Learning performance is modest across all sample sizes (Tables 4.4 - 4.6 and Tables 4.7 - 4.9), correctly estimating the optimal DTR between 75-80% of the time, with performance improving both as the level of confounding decreases and as the number of covariates decreases.

Under a tree-type DTR, T-RL achieves very good performance with larger sample sizes across all levels of confounding when the propensity model is correctly specified, regardless of the number of covariates (Tables 4.1 - 4.3), achieving correct classification of more than 90% in all settings. If the propensity model is incorrectly specified, however, performance of T-RL deteriorates as the level of confounding increases– particularly as the covariate cardinality increases relative to the sample size. With a moderate degree of confounding and an incorrectly-specified propensity model, for example, T-RL correctly identifies the optimal DTR 87% of the time when $n = 500$ and $|\mathbf{H} = 100|$, compared with nearly 95% correct classification with a correctly-specified propensity model. When the assumed DTR is nontree-type (Tables 4.4 - 4.6), T-RL correctly classifies the optimal DTR between 70-80% of the time, with performance improving as the sample size increases, the level of confounding decreases, as the number of covariates decreases, and/or when the propensity model is correctly specified. With a small sample size (Tables 4.7 - 4.9), estimated correct classification using T-

RL for a tree-type DTR ranges from more than 95% when the level of confounding is low and the propensity model is correctly specified to about 86% when the propensity model is incorrectly specified and the level of confounding is either high or moderate with a larger number of covariates. For a nontree-type DTR, T-RL's performance in small samples mimics those trends observed with a larger sample size.

Q-Learning implemented with few model assumptions (i.e., Q-NP) performs better than linear Q-Learning across all data generating settings for a tree-type DTR, although we see larger performance gains of Q-NP relative to Q-Linear as the sample size increases (Tables 4.1 - 4.3). As the level of confounding increases, little difference in performance is observed for Q-Linear, which is in contrast to Q-NP, which tends to perform best with a moderate degree of confounding and worst with high confounding (Tables 4.2 - 4.3). With $n = 1000$ and $|\mathbf{H} = 100|$, for example, Q-NP correctly classifies the optimal DTR at least 95% of the time when confounding is moderate or low, but only about 82% under high confounding. When the DTR is nontree-type (Tables 4.4 - 4.6), Q-Linear performance improves relative to estimation of a tree-type DTR for the same sample sizes and number of covariates, e.g., with about 80% correct classification with $|\mathbf{H} = 20|$ for lower confounding compared with about 70% for the same settings with a tree-type DTR. The performance of Q-NP, however, is worse than in the tree-type setting, all other parameters being equal. For a nontree-type setting Q-Linear and Q-NP perform similarly when $|\mathbf{H} = 20|$ across larger sample sizes (Tables 4.4 - 4.6), but Q-Linear performance quickly deteriorates as the cardinality of covariates increases, particularly for $n = 500$. For a small sample size (Tables 4.7 - 4.9), Q-Linear and Q-NP generally perform poorly across all levels of confounding, generally ranging from about 60% to 70% correct classification with $|\mathbf{H} = 50|$ for both a tree-type and nontree-type DTR.

For a fixed λ and an true, underlying tree-type DTR, PenSIT Learning is preferred to competing methods when the level of confounding is high. Furthermore,

the improvement of PenSIT Learning over other methods is most pronounced when the covariate cardinality is large or when the sample size is small. With moderate confounding, PenSIT Learning performance is comparable to T-RL across all settings and is comparable to Q-NP when $n = 1000$, although PenSIT Learning achieves a clear advantage over T-RL when the sample size is small and the propensity model is incorrectly specfied. When the level of confounding is low, for a fixed λ T-RL is preferred over PenSIT Learning across all sample sizes. With a nontree-type DTR structure, PenSIT Learning is generally preferred across all data generating scenarios, including both correctly- and incorrectly- specified propensity models and levels of confounding, although the improvement over T-RL is modest, with both methods correctly classifying the optimal treatment regime between 75-80% of the time. Under low confounding and low covariate cardinality with a nontree-type DTR, Q-NP would be preferred to both PenSIT Learning and T-RL. Relative to PenSIT Learning, Q-Linear exhibits inferior performance to PenSIT Learning across all sample sizes and levels of confounding when the true DTR structure is tree-type. When the sample size is small and the DTR is nontree-type, PenSIT Learning achieves slightly better estimated correct classification rates than does Q-NP and Q-Linear, particularly when the number of covariates increases.

Table 4.1: Performance summary [% *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 2 possible treatments per stage, assuming an underlying tree-type DTR structure with a lower degree of confounding for larger sample sizes of $n = 500$ and $n = 1000$. Generated with specified training dataset sample size (n) and $N_2 = 1000$ test dataset size; No.Var.H = number of variables in covariate history $\mathbf{H}$; Propensity model $\pi_a(\mathbf{H})$ is generated using either "correct" or "incorrect" specification; $\mathbf{H}$ generated using multivariate normal distribution with using exchangeable correlation structure and $\rho = 0.20$; IQR = interquartile range; PenSIT = Penalized Spline-Involved Tree-based Learning; T-RL = Tree-based Reinforcement Learning; Q-Linear = Linear Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set classified to its optimal treatment using a treatment rule estimated using the applicable method; $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ refers to the estimated counterfactual mean outcome under the estimated optimal DTR. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] = 8.0$.

$\pi_a(\mathbf{H})$	Method	No.Var.H = 20		No.Var.H = 50		No.Var.H = 100	
		% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
Lower Degree of Confounding							
$n = 500$							
	Q-Linear	69.6 (2.9)	7.0 (0.14)	64.0 (2.9)	6.8 (0.15)	55.9 (2.9)	6.5 (0.16)
	Q-NP	87.4 (5.5)	7.6 (0.18)	81.6 (7.1)	7.5 (0.24)	76.2 (7.9)	7.3 (0.29)
Correct	T-RL	98.1 (2.3)	7.9 (0.12)	97.9 (2.9)	7.9 (0.14)	97.4 (4.4)	7.9 (0.18)
	PenSIT	84.8 (13.6)	7.7 (0.29)	85.3 (14.2)	7.7 (0.32)	85.3 (13.6)	7.7 (0.30)
Incorrect	T-RL	98.1 (2.4)	7.9 (0.14)	97.4 (4.4)	7.9 (0.16)	93.8 (12.5)	7.8 (0.30)
	PenSIT	85.0 (13.8)	7.7 (0.30)	86.0 (13.4)	7.7 (0.30)	83.7 (4.5)	7.6 (0.20)
$n = 1000$							
	Q-Linear	71.7 (2.1)	7.1 (0.12)	68.8 (2.4)	7.0 (0.14)	64.2 (2.3)	6.8 (0.13)
	Q-NP	96.1 (2.2)	7.9 (0.12)	94.3 (3.1)	7.8 (0.13)	92.3 (4.1)	7.8 (0.14)
Correct	T-RL	98.8 (1.1)	8.0 (0.11)	98.9 (0.9)	8.0 (0.11)	98.6 (1.4)	8.0 (0.11)
	PenSIT	85.8 (14.2)	7.7 (0.30)	85.8 (14.4)	7.7 (0.33)	85.6 (14.0)	7.7 (0.30)
Incorrect	T-RL	98.7 (1.2)	8.0 (0.11)	98.7 (1.4)	7.9 (0.11)	98.3 (1.9)	7.9 (0.12)
	PenSIT	86.0 (14.2)	7.7 (0.30)	86.1 (14.5)	7.7 (0.34)	86.7 (14.4)	7.8 (0.31)

Table 4.2: Performance summary [% *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 2 possible treatments per stage, assuming an underlying tree-type DTR structure with a moderate degree of confounding for larger sample sizes of $n = 500$ and $n = 1000$. Generated with specified training dataset sample size (n) and $N_2 = 1000$ test dataset size; No.Var.H = number of variables in covariate history $\mathbf{H}$; Propensity model $\pi_a(\mathbf{H})$ is generated using either "correct" or "incorrect" specification; $\mathbf{H}$ generated using multivariate normal distribution with using exchangeable correlation structure and $\rho = 0.20$; IQR = interquartile range; PenSIT = Penalized Spline-Involved Tree-based Learning; T-RL = Tree-based Reinforcement Learning; Q-Linear = Linear Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set classified to its optimal treatment using a treatment rule estimated using the applicable method; $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ refers to the estimated counterfactual mean outcome under the estimated optimal DTR. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] = 8.0$.

$\pi_a(\mathbf{H})$	Method	No.Var.H = 20		No.Var.H = 50		No.Var.H = 100	
		% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
Moderate Degree of Confounding							
$n = 500$							
	Q-Linear	70.0 (2.6)	7.0 (0.14)	64.5 (3.2)	6.8 (0.14)	56.6 (3.2)	6.5 (0.15)
	Q-NP	89.1 (6.3)	7.7 (0.18)	84.3 (6.8)	7.6 (0.19)	80.5 (6.5)	7.5 (0.22)
Correct	T-RL	96.6 (6.3)	7.9 (0.22)	96.0 (5.8)	7.9 (0.22)	94.8 (7.1)	7.8 (0.24)
	PenSIT	96.9 (5.5)	7.9 (0.21)	96.6 (5.1)	7.9 (0.20)	96.0 (6.8)	7.9 (0.24)
Incorrect	T-RL	96.1 (6.3)	7.9 (0.23)	94.4 (12.7)	7.8 (0.33)	87.1 (16.2)	7.7 (0.45)
	PenSIT	96.9 (5.1)	7.9 (0.20)	96.6 (5.1)	7.9 (0.20)	95.9 (7.5)	7.9 (0.26)
$n = 1000$							
	Q-Linear	72.0 (2.2)	7.1 (0.12)	69.2 (2.1)	7.0 (0.13)	64.9 (2.6)	6.8 (0.14)
	Q-NP	97.0 (2.0)	7.9 (0.10)	96.2 (2.9)	7.9 (0.12)	95.1 (4.1)	7.9 (0.14)
Correct	T-RL	98.0 (2.5)	7.9 (0.12)	97.5 (2.8)	7.9 (0.14)	97.3 (2.9)	7.9 (0.13)
	PenSIT	98.4 (1.9)	8.0 (0.14)	98.3 (1.9)	7.9 (0.14)	98.3 (1.8)	7.9 (0.13)
Incorrect	T-RL	97.8 (2.6)	7.9 (0.13)	97.1 (3.0)	7.9 (0.14)	96.5 (4.6)	7.9 (0.13)
	PenSIT	98.4 (1.9)	8.0 (0.12)	98.3 (1.8)	7.9 (0.14)	98.3 (1.7)	7.9 (0.13)

Table 4.3: Performance summary [% *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 2 possible treatments per stage, assuming an underlying tree-type DTR structure with a higher degree of confounding for larger sample sizes of $n = 500$ and $n = 1000$. Generated with specified training dataset sample size (n) and $N_2 = 1000$ test dataset size; No.Var.H = number of variables in covariate history $\mathbf{H}$; Propensity model $\pi_a(\mathbf{H})$ is generated using either "correct" or "incorrect" specification; $\mathbf{H}$ generated using multivariate normal distribution with using exchangeable correlation structure and $\rho = 0.20$; IQR = interquartile range; PenSIT = Penalized Spline-Involved Tree-based Learning; T-RL = Tree-based Reinforcement Learning; Q-Linear = Linear Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set classified to its optimal treatment using a treatment rule estimated using the applicable method; $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ refers to the estimated counterfactual mean outcome under the estimated optimal DTR. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] = 8.0$.

$\pi_a(\mathbf{H})$	Method	No.Var.H = 20		No.Var.H = 50		No.Var.H = 100	
		% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
Higher Degree of Confounding							
$n = 500$							
	Q-Linear	69.6 (2.7)	7.0 (0.14)	64.7 (3.1)	6.8 (0.15)	56.7 (3.3)	6.5 (0.16)
	Q-NP	79.7 (4.7)	7.5 (0.18)	77.6 (4.4)	7.4 (0.18)	76.3 (4.6)	7.4 (0.17)
Correct	T-RL	92.1 (15.4)	7.8 (0.42)	92.9 (13.9)	7.8 (0.38)	91.1 (14.8)	7.7 (0.37)
	PenSIT	96.9 (4.5)	7.9 (0.18)	96.6 (5.4)	7.9 (0.21)	95.6 (7.7)	7.9 (0.24)
Incorrect	T-RL	91.0 (16.5)	7.7 (0.42)	88.0 (15.1)	7.7 (0.38)	78.6 (1.2)	7.3 (0.30)
	PenSIT	96.9 (4.7)	7.9 (0.18)	97.0 (4.7)	7.9 (0.18)	95.1 (9.6)	7.8 (0.29)
$n = 1000$							
	Q-Linear	71.6 (2.3)	7.1 (0.12)	69.1 (2.3)	7.0 (0.13)	64.7 (2.6)	6.8 (0.14)
	Q-NP	89.4 (9.2)	7.8 (0.25)	85.1 (8.8)	7.6 (0.26)	82.1 (5.3)	7.6 (0.18)
Correct	T-RL	95.4 (13.0)	7.8 (0.28)	95.5 (12.9)	7.9 (0.32)	95.7 (9.3)	7.9 (0.27)
	PenSIT	98.5 (1.7)	7.9 (0.12)	98.2 (2.2)	7.9 (0.12)	98.4 (2.0)	7.9 (0.14)
Incorrect	T-RL	95.9 (11.2)	7.9 (0.28)	93.9 (16.6)	7.8 (0.41)	85.5 (18.6)	7.6 (0.48)
	PenSIT	98.5 (1.7)	7.9 (0.12)	98.3 (2.0)	7.9 (0.12)	98.3 (2.1)	7.9 (0.12)

Table 4.4: Performance summary [% *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 2 possible treatments per stage, assuming an underlying nontree-type DTR structure with a lower degree of confounding for larger sample sizes of $n = 500$ and $n = 1000$. Generated with specified training dataset sample size (n) and $N_2 = 1000$ test dataset size; No.Var.H = number of variables in covariate history $\mathbf{H}$; Propensity model $\pi_a(\mathbf{H})$ is generated using either "correct" or "incorrect" specification; $\mathbf{H}$ generated using multivariate normal distribution with using exchangeable correlation structure and $\rho = 0.20$; IQR = interquartile range; PenSIT = Penalized Spline-Involved Tree-based Learning; T-RL = Tree-based Reinforcement Learning; Q-Linear = Linear Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set classified to its optimal treatment using a treatment rule estimated using the applicable method; $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ refers to the estimated counterfactual mean outcome under the estimated optimal DTR. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] = 8.0$.

$\pi_a(\mathbf{H})$	Method	No.Var.H = 20		No.Var.H = 50		No.Var.H = 100	
		% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
Lower Degree of Confounding							
$n = 500$							
	Q-Linear	79.2 (3.9)	7.3 (0.17)	71.0 (3.4)	7.1 (0.16)	60.4 (4.0)	6.7 (0.18)
	Q-NP	82.7 (4.2)	7.5 (0.15)	78.5 (5.4)	7.4 (0.16)	74.0 (6.9)	7.2 (0.21)
Correct	T-RL	79.2 (5.3)	7.3 (0.16)	79.6 (5.1)	7.3 (0.16)	79.2 (6.1)	7.3 (0.18)
	PenSIT	80.4 (4.6)	7.3 (0.14)	80.2 (4.3)	7.3 (0.16)	80.2 (5.2)	7.3 (0.15)
Incorrect	T-RL	79.6 (5.9)	7.3 (0.17)	78.8 (5.4)	7.3 (0.17)	77.7 (6.1)	7.3 (0.21)
	PenSIT	80.3 (4.3)	7.3 (0.15)	80.4 (4.4)	7.3 (0.15)	80.0 (5.5)	7.3 (0.17)
$n = 1000$							
	Q-Linear	83.1 (2.9)	7.5 (0.14)	78.0 (3.0)	7.3 (0.14)	71.2 (2.7)	7.1 (0.14)
	Q-NP	86.3 (3.1)	7.6 (0.12)	83.2 (3.5)	7.5 (0.13)	81.6 (3.9)	7.5 (0.13)
Correct	T-RL	80.1 (4.3)	7.4 (0.14)	80.0 (4.6)	7.3 (0.14)	79.7 (4.4)	7.3 (0.14)
	PenSIT	81.0 (3.7)	7.4 (0.12)	81.0 (3.9)	7.4 (0.13)	80.6 (3.6)	7.4 (0.13)
Incorrect	T-RL	80.1 (4.3)	7.4 (0.13)	79.9 (4.7)	7.4 (0.14)	79.4 (4.6)	7.3 (0.15)
	PenSIT	81.0 (3.7)	7.4 (0.12)	80.9 (4.0)	7.4 (0.13)	80.9 (3.5)	7.4 (0.13)

Table 4.5: Performance summary [% *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 2 possible treatments per stage, assuming an underlying nontree-type DTR structure with a moderate degree of confounding for larger sample sizes of $n = 500$ and $n = 1000$. Generated with specified training dataset sample size (n) and $N_2 = 1000$ test dataset size; No.Var.H = number of variables in covariate history $\mathbf{H}$; Propensity model $\pi_a(\mathbf{H})$ is generated using either "correct" or "incorrect" specification; $\mathbf{H}$ generated using multivariate normal distribution with using exchangeable correlation structure and $\rho = 0.20$; IQR = interquartile range; PenSIT = Penalized Spline-Involved Tree-based Learning; T-RL = Tree-based Reinforcement Learning; Q-Linear = Linear Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set classified to its optimal treatment using a treatment rule estimated using the applicable method; $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ refers to the estimated counterfactual mean outcome under the estimated optimal DTR. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] = 8.0$.

$\pi_a(\mathbf{H})$	Method	No.Var.H = 20		No.Var.H = 50		No.Var.H = 100	
		% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
Moderate Degree of Confounding							
$n = 500$							
	Q-Linear	78.0 (4.0)	7.3 (0.17)	69.2 (3.9)	7.0 (0.17)	58.7 (3.7)	6.6 (0.18)
	Q-NP	80.0 (4.6)	7.4 (0.16)	75.8 (4.9)	7.3 (0.15)	73.1 (5.0)	7.2 (0.17)
Correct	T-RL	76.4 (6.1)	7.2 (0.20)	76.0 (6.8)	7.2 (0.21)	75.4 (5.8)	7.2 (0.21)
	PenSIT	78.0 (4.6)	7.3 (0.16)	78.2 (4.9)	7.3 (0.16)	77.5 (5.1)	7.3 (0.17)
Incorrect	T-RL	75.8 (6.5)	7.2 (0.21)	75.0 (6.7)	7.2 (0.22)	73.1 (8.1)	7.1 (0.27)
	PenSIT	78.0 (4.6)	7.3 (0.16)	78.1 (4.8)	7.3 (0.16)	77.6 (5.2)	7.3 (0.17)
$n = 1000$							
	Q-Linear	82.4 (3.2)	7.4 (0.13)	76.5 (3.0)	7.3 (0.16)	69.6 (2.9)	7.0 (0.15)
	Q-NP	83.8 (2.9)	7.5 (0.12)	81.1 (3.2)	7.5 (0.14)	79.1 (3.5)	7.4 (0.15)
Correct	T-RL	77.1 (5.3)	7.3 (0.15)	76.9 (5.0)	7.3 (0.16)	77.2 (5.2)	7.3 (0.16)
	PenSIT	78.7 (4.0)	7.3 (0.14)	78.2 (3.7)	7.3 (0.15)	78.3 (4.6)	7.3 (0.14)
Incorrect	T-RL	77.2 (5.1)	7.3 (0.15)	77.0 (5.3)	7.3 (0.17)	76.6 (5.6)	7.2 (0.17)
	PenSIT	78.7 (4.1)	7.3 (0.14)	78.3 (3.8)	7.3 (0.15)	78.4 (4.6)	7.3 (0.14)

Table 4.6: Performance summary [% *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 2 possible treatments per stage, assuming an underlying nontree-type DTR structure with a higher degree of confounding for larger sample sizes of $n = 500$ and $n = 1000$. Generated with specified training dataset sample size (n) and $N_2 = 1000$ test dataset size; No.Var.H = number of variables in covariate history $\mathbf{H}$; Propensity model $\pi_a(\mathbf{H})$ is generated using either "correct" or "incorrect" specification; $\mathbf{H}$ generated using multivariate normal distribution with using exchangeable correlation structure and $\rho = 0.20$; IQR = interquartile range; PenSIT = Penalized Spline-Involved Tree-based Learning; T-RL = Tree-based Reinforcement Learning; Q-Linear = Linear Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set classified to its optimal treatment using a treatment rule estimated using the applicable method; $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ refers to the estimated counterfactual mean outcome under the estimated optimal DTR. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] = 8.0$.

$\pi_a(\mathbf{H})$	Method	No.Var.H = 20		No.Var.H = 50		No.Var.H = 100	
		% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
Higher Degree of Confounding							
$n = 500$							
	Q-Linear	73.6 (5.5)	7.2 (0.21)	65.3 (5.2)	6.9 (0.21)	55.4 (4.5)	6.5 (0.21)
	Q-NP	75.2 (5.4)	7.3 (0.19)	70.9 (5.5)	7.2 (0.20)	67.1 (7.1)	7.0 (0.26)
Correct	T-RL	73.2 (6.7)	7.2 (0.23)	72.8 (7.8)	7.1 (0.27)	71.7 (7.7)	7.1 (0.29)
	PenSIT	75.7 (5.5)	7.2 (0.16)	75.4 (5.2)	7.2 (0.16)	75.2 (6.0)	7.2 (0.19)
Incorrect	T-RL	72.6 (7.4)	7.1 (0.27)	72.2 (8.5)	7.1 (0.26)	70.4 (9.5)	7.1 (0.31)
	PenSIT	75.4 (5.4)	7.2 (0.15)	75.6 (5.4)	7.2 (0.17)	74.5 (6.2)	7.2 (0.21)
$n = 1000$							
	Q-Linear	78.7 (4.0)	7.3 (0.16)	73.1 (3.8)	7.1 (0.16)	66.1 (3.9)	6.9 (0.18)
	Q-NP	80.2 (3.5)	7.5 (0.13)	77.4 (3.9)	7.4 (0.14)	75.0 (3.9)	7.3 (0.15)
Correct	T-RL	73.7 (6.3)	7.2 (0.20)	74.2 (6.2)	7.2 (0.19)	74.4 (5.4)	7.2 (0.19)
	PenSIT	76.0 (4.1)	7.3 (0.15)	76.4 (4.4)	7.3 (0.14)	76.1 (4.6)	7.3 (0.15)
Incorrect	T-RL	74.6 (6.3)	7.2 (0.19)	74.3 (6.2)	7.2 (0.18)	73.2 (5.5)	7.1 (0.24)
	PenSIT	76.0 (4.2)	7.3 (0.15)	76.2 (4.4)	7.3 (0.14)	75.9 (4.6)	7.3 (0.15)

4.6 Application of PenSIT Learning to MIMIC-III Data

Sepsis is a clinical syndrome characterized by systemic inflammation and infection and is associated with one of the highest rates of mortality among conditions commonly treated in emergency departments (EDs) and intensive care units (ICUs) (*Marino*, 2014). Due to the large degree of heterogeneity in presentation, which may include varying degrees of organ dysfunction, sepsis is a difficult condition to diagnose and even more difficult to successfully treat. Sepsis is routinely treated using fluid resuscitation, antibiotics, and may also include treatment with vasopressors, mechanical ventilation, and others. The established clinical guidelines for treating sepsis, known as the "Surviving Sepsis Campaign" (*Rhodes et al.*, 2017), strongly recommends that resuscitation of at least 30 mL/kg of IV fluid be given within the first 3 hours. However, this recommendation is given with a stated "low quality of evidence" due to the fact that results across studies have been inconsistent with indirect evidence, imprecise results, and a likelihood of bias.

Due to the paucity of strong evidence as to the most beneficial fluid resuscitation strategy in the early hours of treatment, we estimate an optimal two-stage DTR in adult septic patients admitted to the intensive care unit after presenting to the ED (Figure 4.1). We evaluate whether treatment with fluid restrictive or fluid liberal strategies can be further tailored in order to minimize organ dsyfunction scores measured by the Sequential Organ Failure Assessment (SOFA; *Vincent et al.*, 1996) at 24 hours following admission. This analysis is conducted using electronic medical record and administrative data from the Medical Information Mart for Intensive Care III (MIMIC-III; *Johnson et al.*, 2016; *Johnson et al.*, 2017; *Pollard and Johnson*, 2016), a retrospectively-collected and freely-available database accessible through PhysioNet (*Goldberger et al.*, 2000) that contains de-identified and anonymized data for more than 45,000 patients treated in an ICU at Beth Israel Deaconness Medical Center in Boston, Massachusetts.

Table 4.7: Performance summary [% *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 2 possible treatments per stage to evaluate performance in smaller samples ($n = 300$) for both tree- and nontree-type DTRs with a lower degree of confounding. Generated with training dataset sample of size $n = 300$ with $N_2 = 1000$ test dataset size; No.Var.H = number of variables in covariate history $\mathbf{H}$; Propensity model $\pi_a(\mathbf{H})$ is generated using either "correct" or "incorrect" specification; $\mathbf{H}$ generated using multivariate normal distribution with using exchangeable correlation structure and $\rho = 0.20$; IQR = interquartile range; PenSIT = Penalized Spline-Involved Tree-based Learning; T-RL = Tree-based Reinforcement Learning; Q-Linear = Linear Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set classified to its optimal treatment using a treatment rule estimated using the applicable method; $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ refers to the estimated counterfactual mean outcome under the estimated optimal DTR. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] = 8.0$.

$\pi_a(\mathbf{H})$	Method	No.Var.H = 20		No.Var.H = 50	
		% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
Lower Degree of Confounding					
Tree-type DTR					
	Q-Linear	66.8 (3.2)	6.9 (0.14)	58.4 (3.8)	6.6 (0.16)
	Q-NP	76.8 (6.3)	7.3 (0.24)	69.3 (7.8)	7.0 (0.27)
Correct	T-RL	96.6 (7.5)	7.9 (0.21)	95.6 (8.2)	7.8 (0.23)
	PenSIT	85.3 (13.3)	7.7 (0.31)	84.5 (12.5)	7.7 (0.31)
Incorrect	T-RL	96.0 (8.3)	7.9 (0.24)	93.3 (12.9)	7.8 (0.31)
	PenSIT	84.8 (13.7)	7.7 (0.32)	87.0 (12.3)	7.7 (0.23)
Nontree-type DTR					
	Q-Linear	75.1 (4.5)	7.2 (0.17)	63.5 (4.3)	6.8 (0.19)
	Q-NP	78.4 (6.1)	7.4 (0.17)	71.7 (7.9)	7.2 (0.25)
Correct	T-RL	78.4 (5.8)	7.3 (0.17)	77.6 (7.0)	7.3 (0.22)
	PenSIT	79.6 (4.6)	7.3 (0.15)	79.3 (5.7)	7.3 (0.18)
Incorrect	T-RL	78.2 (6.3)	7.3 (0.18)	76.8 (7.4)	7.2 (0.23)
	PenSIT	79.4 (4.9)	7.3 (0.14)	78.6 (4.1)	7.3 (0.15)

Table 4.8: Performance summary [% *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 2 possible treatments per stage to evaluate performance in smaller samples ($n = 300$) for both tree- and nontree-type DTRs with a moderate degree of confounding. Generated with training dataset sample of size $n = 300$ with $N_2 = 1000$ test dataset size; No.Var.H = number of variables in covariate history $\mathbf{H}$; Propensity model $\pi_a(\mathbf{H})$ is generated using either "correct" or "incorrect" specification; $\mathbf{H}$ generated using multivariate normal distribution with using exchangeable correlation structure and $\rho = 0.20$; IQR = interquartile range; PenSIT = Penalized Spline-Involved Tree-based Learning; T-RL = Tree-based Reinforcement Learning; Q-Linear = Linear Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set classified to its optimal treatment using a treatment rule estimated using the applicable method; $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ refers to the estimated counterfactual mean outcome under the estimated optimal DTR. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] = 8.0$.

$\pi_a(\mathbf{H})$	Method	No.Var.H = 20		No.Var.H = 50	
		% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
Moderate Degree of Confounding					
Tree-type DTR					
	Q-Linear	67.5 (3.3)	6.9 (0.15)	58.9 (4.0)	6.6 (0.16)
	Q-NP	78.0 (7.4)	7.4 (0.23)	72.6 (7.0)	7.2 (0.26)
Correct	T-RL	94.6 (12.0)	7.8 (0.34)	93.1 (13.0)	7.8 (0.34)
	PenSIT	94.7 (11.2)	7.8 (0.27)	93.2 (12.7)	7.8 (0.31)
Incorrect	T-RL	93.4 (14.8)	7.8 (0.38)	86.8 (17.2)	7.6 (0.46)
	PenSIT	94.7 (11.4)	7.8 (0.28)	93.2 (12.9)	7.8 (0.31)
Nontree-type DTR					
	Q-Linear	73.0 (4.8)	7.1 (0.20)	61.7 (5.0)	6.7 (0.22)
	Q-NP	75.4 (5.6)	7.3 (0.18)	69.4 (7.3)	7.1 (0.24)
Correct	T-RL	74.7 (7.8)	7.2 (0.23)	74.5 (8.2)	7.2 (0.25)
	PenSIT	77.2 (6.1)	7.3 (0.17)	77.0 (6.3)	7.3 (0.19)
Incorrect	T-RL	74.3 (8.0)	7.2 (0.25)	72.0 (9.0)	7.1 (0.31)
	PenSIT	77.0 (6.3)	7.3 (0.17)	76.5 (6.3)	7.2 (0.19)

Table 4.9: Performance summary [% *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 2 possible treatments per stage to evaluate performance in smaller samples ($n = 300$) for both tree- and nontree-type DTRs with a higher degree of confounding. Generated with training dataset sample of size $n = 300$ with $N_2 = 1000$ test dataset size; No.Var.H = number of variables in covariate history $\mathbf{H}$; Propensity model $\pi_a(\mathbf{H})$ is generated using either "correct" or "incorrect" specification; $\mathbf{H}$ generated using multivariate normal distribution with using exchangeable correlation structure and $\rho = 0.20$; IQR = interquartile range; PenSIT = Penalized Spline-Involved Tree-based Learning; T-RL = Tree-based Reinforcement Learning; Q-Linear = Linear Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set classified to its optimal treatment using a treatment rule estimated using the applicable method; $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H})\}]$ refers to the estimated counterfactual mean outcome under the estimated optimal DTR. Under optimal treatment allocation $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H})\}] = 8.0$.

$\pi_a(\mathbf{H})$	Method	No.Var.H = 20		No.Var.H = 50	
		% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$	% opt	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$
Higher Degree of Confounding					
Tree-type DTR					
	Q-Linear	67.1 (3.7)	6.9 (0.17)	58.9 (4.4)	6.6 (0.16)
	Q-NP	73.8 (5.8)	7.3 (0.20)	70.0 (6.0)	7.2 (0.22)
Correct	T-RL	89.4 (15.5)	7.7 (0.43)	86.6 (18.2)	7.6 (0.47)
	PenSIT	94.6 (8.5)	7.8 (0.25)	91.7 (11.7)	7.7 (0.32)
Incorrect	T-RL	86.1 (19.8)	7.7 (0.50)	86.4 (11.3)	7.6 (0.25)
	PenSIT	94.4 (8.6)	7.8 (0.27)	92.4 (11.4)	7.8 (0.30)
Nontree-type DTR					
	Q-Linear	68.6 (6.1)	7.0 (0.22)	57.6 (6.0)	6.6 (0.25)
	Q-NP	69.9 (7.8)	7.1 (0.27)	62.8 (8.9)	6.9 (0.36)
Correct	T-RL	71.1 (9.3)	7.1 (0.33)	70.3 (10.1)	7.0 (0.34)
	PenSIT	75.0 (6.4)	7.2 (0.18)	74.3 (7.1)	7.2 (0.21)
Incorrect	T-RL	71.7 (11.3)	7.1 (0.36)	61.6 (22.8)	6.8 (0.71)
	PenSIT	75.1 (6.4)	7.2 (0.19)	74.4 (7.9)	7.2 (0.24)

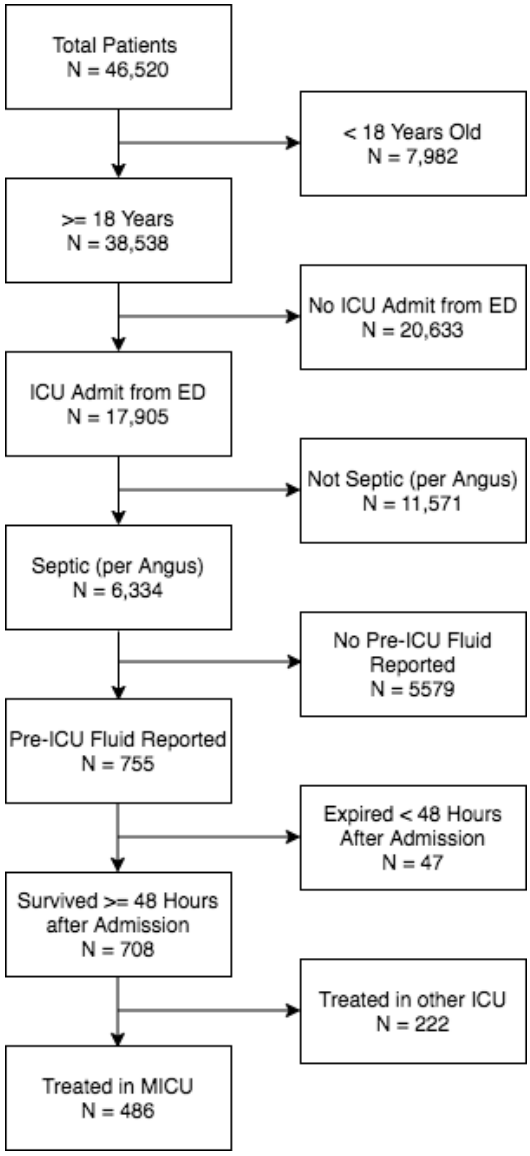

Figure 4.1: Analysis eligibility criteria and patient population. A total of 486 patients were included in this analysis. Inclusion criteria included being ≥ 18 years old at the time of medical intensive care unit (MICU) admission from the emergency department, having a diagnosis of suspected sepsis (*Angus et al.*, 2001; *Horng et al.*, 2017; *Iwashyna et al.*, 2014), receiving documented pre-MICU fluids, and surviving at least 48 hours after MICU admission.

Baseline covariates considered as candidate tailoring variables for treatment strategies included demographics such as gender, age, weight, and racial/ethnic groups, Elixhauser comorbidity score (*Elixhauser et al.*, 1998; *van Walraven et al.*, 2009), and the time of year in which the patient was treated. Stage 1 treatment was defined as either a fluid restrictive (< 30 mL/kg) or a fluid liberal (≥ 30 mL/kg) strategy within the first three hours after admission to the MICU. Intermediate variables collected prior to Stage 2 treatment included indicators of treatment with mechanical ventilation and vasopressors within the first three hour time period, as well as the patient's SOFA score evaluated at three hours post-admission. Stage 2 treatment was defined as either a fluid restrictive (< 30 mL/kg) or a fluid liberal (≥ 30 mL/kg) strategy between 3-24 hours after MICU admission. The final outcome of interest is the SOFA score evaluated at 24 hours post-admission. Because SOFA scores typically exhibit a right-skewed distribution with values ranging from 0 to 18 and lower scores indicate better prognosis, values are log-transformed and then inverted (as $6 - x$) so that larger values represent better outcomes; the resulting transformed values are approximately symmetric and normally distributed. All models needed to estimate stage-specific counterfactual outcomes using the PenSI estimator, including the propensity for assignment to a fluid liberal strategy and the conditional mean model for the transformed SOFA score, assume an additive linear relationship with the log odds of assignment or with the outcome, respectively. Selection of variables to include in the stage-specific propensity models and the conditional mean models estimated for each treatment within each stage was performed using stepwise variable selection with the `MASS` package in `R`. For implementation of tree-based learning, at both stages we specify the need for a 2.5% improvement in the counterfactual outcomes mean across treatments in order to perform a covariate split; after accounting for the transformation of the outcome used in the regression models, this represents a SOFA improvement of roughly 0.5, which we consider to be clinically meaningful.

Additionally, we specify a depth of 3 and a minimum of 20 observations per node.

Four hundred eighty-six (486) patients were included in the analysis cohort. The average patient was a 69 year old white (76%) male (52%) with 0 reported Elixhauser comorbidities (Table 4.10). The median length of hospital stay was 7.8 days with an interquartile range (IQR) of 5.0-13.8. The median fluid input received within 0-3 hours and 3-24 hours post-admission is 41.4 mL/kg (IQR: 22.8-60.9) and 20.2 mL/kg (IQR: 3.6-52.4), respectively. Summary statistics stratified by treatment stage (i.e., 0-3 hours and 3-24 hours post-MICU admission) demonstrate covariate imbalance for age, gender, and weight across fluid resuscitation strategies in the first treatment stage, for race/ethnicity at the second stage, and for the use of mechanical ventilation and vasopressors across fluid resuscitation strategies for both stages, suggesting that confounding is an issue that must be addressed in our analysis in order to make causal inference.

As can be observed in Figure 4.2, it is recommended that all patients should receive liberal fluid resuscitation ($\geq$ 30 mL/kg) within the first three hours following admission to the MICU for treatment of acute emergent sepsis. If the patient has received the liberal fluid resuscitation by 3 hours post-admission in accordance with this estimated decision rule, the patient should receive restrictive fluid resuscitation ($<$ 30 mL/kg) to follow. If the patient was not given liberal fluid resuscitation within the first three hours following admission, the patient should receive liberal fluid resuscitation within 3-24 hours post-admission. Notably, no tailoring variables were identified at the first treatment stage that would result in a meaningful improvement in outcomes overall. Second stage treatment, however, can be tailored based on the patient's first-stage treatment in order to optimize counterfactual outcomes overall.

Although the question of how to optimally treat septic patients is complex and multi-faceted, we applied a robust and flexible causal method with interpretable results to determine whether tailoring of fluid resuscitation strategies at each of two

Table 4.10: Characteristics of the analysis cohort. Summary statistics of demographics, treatment, and outcomes for MIMIC-III analysis cohort are included. n = sample size. Stage 1 = 0-3 hours post-admission. Stage 2 = 3-24 hours post-admission. R = restrictive fluid resuscitation (< 30 mL/kg); L = liberal fluid resuscitation (≥ 30 mL/kg); IQR = interquartile range; kg = kilogram; LOS = length of hospital stay; Mech Vent = Mechanical Ventilation; Vasos = Vasopressors; L = liters; mL/kg = milliliters per kilogram; SOFA = sequential organ failure assessment; hrs = hours. Median (IQR) are presented for continuous variables; frequency (percentage) are provided for categorical variables.

	Overall (n=486)	Stage 1		Stage 2	
		Restrictive (n=163)	Liberal (n=323)	Restrictive (n=289)	Liberal (n=197)
Patient Characteristics					
Age (years)	69 [54-82]	71 [57-81]	68 [53-82]	69 [55-82]	69 [54-82]
Gender					
Male	252 (52)	94 (58)	158 (49)	149 (52)	103 (52)
Female	234 (48)	69 (42)	165 (51)	140 (48)	96 (48)
Race/Ethnicity					
White	371 (76)	123 (76)	248 (77)	226 (78)	145 (74)
Nonwhite	115 (24)	40 (24)	75 (23)	63 (22)	52 (26)
Weight (kg)	77 [65-91]	83 [69-98]	74 [62-87]	80 [68-94]	72 [62-85]
LOS (days)	7.8 [5.0-13.8]	8.5 [5.0-15.7]	7.7 [5.0-12.7]	7.3 [4.8-12.0]	8.2 [5.8-14.8]
0-3 hours post-Admission					
Use of Mech Vent	111 (23)	43 (26)	68 (21)	74 (26)	37 (19)
Use of Vasos	89 (18)	16 (10)	73 (23)	50 (17)	39 (20)
Total Input (L)	3.0 [2.0-5.0]	1.2 [0.9-2.0]	4.0 [3.0-5.2]	2.9 [1.5-4.0]	4.0 [2.5-5.3]
Total Input (mL/kg)	41.4 [22.8-60.9]	16.7 [11.2-22.9]	53.8 [41.4-71.4]	35.2 [17.4-51.9]	53.3 [34.0-74.4]
SOFA (3 hours)	4 [2-6]	4 [2-6]	4 [2-6]	5 [3-6]	4 [2-6]
3-24 hours post-Admission					
Use of Mech Vent	197 (41)	69 (42)	128 (40)	99 (34)	98 (50)
Use of Vasos	189 (39)	49 (30)	140 (43)	80 (28)	109 (55)
Total Input (L)	2.5 [1.0-4.5]	1.5 [1.0-2.7]	3.2 [1.3-5.5]	1.0 [0.7-1.6]	4.5 [3.3-6.0]
Total Input (mL/kg)	20.2 [3.6-52.4]	10.7 [0.0-25.2]	30.3 [9.0-62.7]	7.2 [0.0-16.9]	58.9 [44.0-87.3]
SOFA (24 hours)	5 [3-7]	5 [3-7]	5 [3-8]	5 [3-6]	6 [3-9]

stages within the first 24 hours after MICU admission can be used to improve outcomes overall. Consistent with the Surviving Sepsis Campaign best practice recommendations, we find that liberal fluid resuscitation should be given as early as possible in this patient population in order to reduce early indicators of organ dysfunction.

4.7 Discussion

PenSIT Learning retrofits a decision tree with a novel PenSI purity measure that incorporates the estimated propensity for treatment assignment as a spline predictor rather than a weight (*Zhou et al.*, 2019). Not only does PenSIT Learning retain the flexibility of T-RL and other tree-based optimal DTR estimation methods, but it provides added robustness under conditions of a high degree of confounding. Additionally, the PenSI estimator of the counterfactual mean utilized within the PenSI purity measure fulfills the properties of consistency and double robustness in asymptotia under standard regularity conditions.

There are several distinct advantages of PenSIT Learning. First, the PenSI estimator of counterfactual outcomes is derived using standard regression models for the treatment assignments and the conditional outcomes at all stages, suggesting that the standard knowledge base surrounding regression models, including model building and selection, model fit diagnostics, etc., can and should be applied liberally. Moreover, although our simulation experiments were designed such that the same variables were used to define both counterfactual outcomes within each stage, PenSIT Learning allows modeling choices for the counterfactual outcomes within a stage to differ, which makes practical sense as there is no reason why we would expect the true mechanisms to be the same. Finally, although we focused on a continuous outcome that is approximately symmetric, the simplicity of PenSIT Learning makes it straightforward to make adjustments to regression models based on the scale of the outcome, e.g., using a generalized linear mixed model framework.

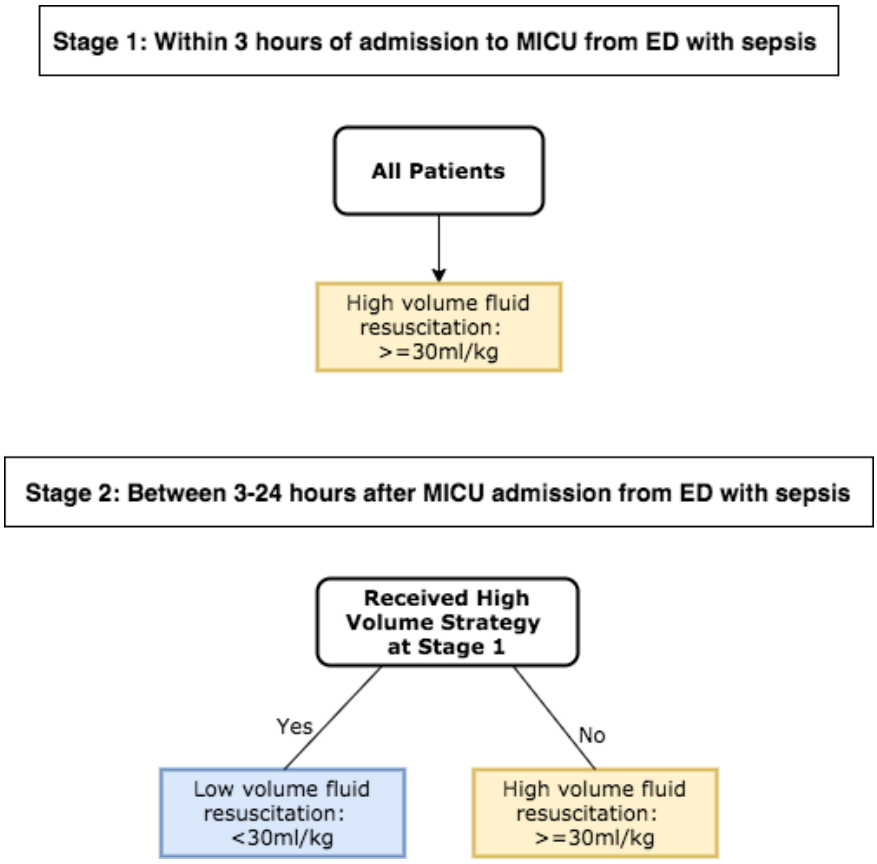

Figure 4.2: Two stage treatment strategy to optimize the patient-level Sequential Organ Failure Assessment (SOFA) score evaluated at 24 hours following admission. We estimate that all patients should receive a high volume fluid resuscitation strategy ($\geq$ 30 mL/kg) within three hours after admission to the Medical Intensive Care Unit (MICU). If the patient receives high volume fluid resuscitation within the first three hours following admission in accordance with this strategy, they should receive low volume fluid resuscitation ($<$ 30 mL/kg) between 3-24 hours following MICU admission. If they did not receive an initial high volume resuscitation strategy in accordance with the estimated guideline, however, they should receive high volume ($\geq$ 30 mL/kg) fluid resuscitation between 3-24 hours following MICU admission.

Several cautions should be heeded when applying PenSIT Learning. First, PenSIT Learning involves a spline function of the estimated propensity score in the conditional mean model. Whereas this allows for a flexible relationship between the propensity score and the outcome, the propensity model should, in theory, be correctly specified. In practice, however, simulation results suggest that the assumption of a correctly-specified propensity model is perhaps less critical. Second, our method is based on fulfilling the assumptions of consistency, positivity, and ignorability. The consistency assumption is generally reasonable in many experimental settings, although a priori assessment is of course required. Positivity can also generally be evaluated using content knowledge and by covariate balance diagnostics conditional on the propensity score; however, in the setting with a large number of covariates and/or multiple treatments and stages, some accommodations to ensure positivity may be needed (*Crump et al.*, 2009; *Gutman and Rubin*, 2015; *Ho et al.*, 2007; *Rosenbaum*, 2012). Ignorability, conversely, may not always be reasonable in an observational data setting, although investigators may be willing to proceed under an assumption of ignorability due to the fact that optimal DTR estimation is largely an exploratory pursuit that should be challenged in confirmatory studies. Lastly, supplemental simulation studies reveal that PenSIT Learning may be particularly sensitive to the choice of λ, especially when there is a lower degree of confounding of the relationship between the treatment and the outcome. However, we do generally expect a moderate to high degree of confounding when using observational data to evaluate causal effects in a medical setting and we maintain the view that the choice of λ and other tuning parameters should be guided by scientific knowledge.

There are several extensions to PenSIT Learning that we believe would be of interest to the research community. The first would be to explore the performance of PenSIT Learning using more flexible modeling approaches for the treatment assignment and/or the conditional counterfactual outcome models (e.g., using BART, random

forests, kernel-based methods, etc.). It is possible that these methods may improve performance under more complex data generation settings, and may be more desirable for accommodating an abundance of data, e.g., from electronic medical records, from which there are few a priori patterns or insights. Second, although we strongly believe that the choice of λ should be driven based on scientific knowledge, in the absence of information about a suitable λ, data-driven approaches for selecting the tuning parameters could be explored. Finally, extensions of PenSIT Learning to account for potential overfitting inherent in decision tree-type constructs, for example, with stochastic tree search or incorporating a lookahead procedure, can be considered.

CHAPTER V

Summary and Future Work

In this dissertation we have addressed three open challenges in the estimation of optimal multi-stage multi-treatment dynamic treatment regimes with the goal of improving the delivery of healthcare. We provide the means to evaluate previously unaddressed problems commonly encountered in the medical and healthcare sphere and we envision extensions of these methods to tackle new challenges.

Chapter II introduced ReST-L, which builds an estimated DTR using a sub-tree defined by a subset of candidate tailoring variables that are determined based on prior knowledge to be clinically or scientifically meaningful. Although an existing method, T-RL, also incorporates an AIPW-style estimator into a decision tree construct, we provide the theoretical justification to include only a subset of variables in an estimated decision rule at each stage, but to include all possible variables in the AIPW estimator for the counterfactual outcomes. While straightforward in principle, it provides a channel for DTR estimation under a scenario that is commonly encountered in medical research.

Next we consider data from a Clustered SMART, in which interventions are applied at the level of the cluster but the outcome of interest lies at the level of the individual within the cluster, a trial design that is becoming increasingly popular in mental health, education, and implementation science research. In Chapter III we

introduced Clustered Q-Learning to estimate model parameters of the Q-functions when clustering occurs by nature of the study design. We further propose the M-out-of-N Cluster Bootstrap, a method extending the m-out-of-n standard bootstrap, for estimating confidence intervals for parameters defining the Q-functions under conditions of nonregularity. Although Clustered Q-Learning and the M-out-of-N Cluster Bootstrap are straightforward extensions of existing methods to the clustered data setting, this is an important contribution to the statistical literature.

Finally, in Chapter IV we introduced PenSIT Learning, which incorporates into a decision-tree type framework a novel purity measure derived from an estimator previously proposed within the missing data literature. PenSIT Learning, while still maintaining the flexible and interpretable structure of the decision tree framework, diverges from the use of IPW-style estimators of the counterfactual outcomes as these may become unstable when weights are large, a phenomenon that may occur in practice with higher levels of confounding or as the number of stages and treatments within each stage increase. PenSIT Learning provides another method for statisticians and clinicians to estimate a multi-stage multi-treatment DTR that may add robustness in scenarios of higher confounding, which is likely to occur with observational data collected in a medical setting.

There are several extensions of our work that may be considered. First, ReST-L and PenSIT Learning are ripe for use with outcomes and treatments on alternate scales. In the medical community, for example, time-to-event outcomes subject to censoring mechanisms are common across many clinical specialties, and could be a target for future work. PenSIT Learning, for example, could be a prime candidate for use with a time-to-event outcome as it can easily be adapted for use within a Cox proportional hazards model. In our estimation of tailored fluid resuscitation strategies for patients with acute sepsis, for example, outcomes of ICU- and hospital- length of stay, as well as survival overall, could be investigated. Additionally, whereas we

approach both ReST-L and PenSIT Learning from the perspective of discrete treatments, which is of course reasonable and common for medical treatments, we could extend these methods to acccount for continuous treatments. In our analysis of patients with acute sepsis, for example, although we evaluate restrictive versus liberal fluid resuscitation using a documented treatment strategy (i.e., using a cut point of 30mL/kg), an estimated DTR based on a continuous treatment scale could be of great interest. Finally, whereas we estimate propensity models and conditional mean models parametrically for ReST-L and we utilize parametric and semi-parametric models within PenSIT Learning, the use of more flexible methods to estimate counterfactual outcomes could be particularly desirable given the abundance of observational data often available to answer a research question.

With regard to Clustered Q-Learning, we effectively address a clinical scenario in which there are a relatively large number of clusters–each with a moderate cluster size. Although we reveal low bias and near nominal coverage across two intervention stages under these conditions, we find that both bias and coverage suffer when the number of clusters is small and, to a smaller extent, when the number of clusters is large but the cluster sizes are small. Therefore, methods that accommodate the clustering mechanism but also address constraints of a smaller cluster size would be desirable. Secondly, just as an abundance of novel modeling choices has been proposed to date for standard Q-Learning, we expect these developments are well-positioned to take hold within Clustered Q-Learning, as well. Finally, although our Clustered Q-Learning was motivated by data collected from a Clustered SMART, a straightforward use of Clustered Q-Learning with observational data with the necessary model adjustments to account for the confounding inherent in these settings is also reasonable.

BIBLIOGRAPHY

BIBLIOGRAPHY

Agarwal, J., M. Trovato, S. Agarwal, P. Hopkins, D. Brooks, and G. Buncke (2010), Selected outcomes of thumb replantation after isolated thumb amputation injury, *J Hand Surg Am.*, *35*(9), 1485–1490.

Almirall, D., I. Nahum-Shani, N. Sherwood, and S. Murphy (2014), Introduction to SMART Designs for the Development of Adaptive Interventions: With Application to Weight Loss Research, *Translational Behavioral Medicine*, *4*(3), 260–274.

Almirall, D., I. Nahum-Shani, L. Wang, and C. Kasari (2018), *Experimental Designs for Research on Adaptive Interventions: Singly and Sequentially Randomized Trials In L. Collins (Ed.), Optimization of Behavioral, Biobehavioral, and Biomedical Interventions: Advanced Topics*, 89-120 pp., Springer International Publishing, Cham, Switzerland.

Angus, D., W. Linde-Zwirble, J. Lidicker, et al. (2001), Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care, *Critical Care Medicine*, *29*(7), 1303–1310.

Arjas, E., and O. Saarela (2010), Optimal Dynamic Regimes: Presenting a Case for Predictive Inference, *International Journal of Biostatistics*, *6*(2), 10, doi:10.2202/1557-4679.1204.

Badamgarav, E., S. Weingarten, K. Henning, et al. (2003), Effectiviness of Disease Management Programs in Depression: A Systematic Review, *American Journal of Psychiatry*, *160*, 2080–90.

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

Berlin, N., C. Tuggle, J. Thomson, and A. Au (2014), Digit replantation in children: a nationwide analysis of outcomes and trends of 455 pediatric patients, *Hand*, *9*(2), 244–252.

Bickel, P., F. Goetze, and W. Zwet (1997), Resampling fewer than n observations: Gains, losses and remedies for losses, *Statistica Sinica*, *7*, 1–31.

Boehmke, B., and B. Greenwell (2020), *Decision Trees In: Hands-On Machine Learning with R*, Boca Raton, FL.

Boulas, H. (1998), Amputations of the fingers and hand: indications for replantation, *J Am Acad Orthop Surg*, *6*(2), 100–105.

Bouwmeester, W., K. Moons, T. Kappen, W. van Klei, J. Twisk, M. Eijkemans, and Y. Vergouwe (2013), Internal Validation of Risk Models in Clustered Data: A Comparison of Bootstrap Schemes, *American Journal of Epidemiology*, *177*(11), 1209–1217, doi:10.1093/aje/kws396.

Breiman, L., J. Freidman, R. Olshen, and C. Stone (1984), *Classification and regression trees*, Wadsworth, Belmont, CA.

Buntic, R., D. Brooks, and G. Buncke (2008), Index finger salvage with replantation and revascularization: revisiting conventional wisdom, *Microsurgery*, *28*(8), 612–616.

Chakraborty, B., and E. Moodie (2013), *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine. Statistics for Biology and Health.*, Springer, New York, NY.

Chakraborty, B., S. Murphy, and V. Strecher (2010), Inference for Nonregular Parameters in Optimal Dynamic Treatment Regimes, *Stat Methods Med Res*, *19*(3), 317–343, doi:10.1177/0962280209105013.

Chakraborty, B., E. Laber, and Y. Zhao (2013), Inference for Optimal Dynamic Treatment Regimes using an Adaptive m-out-of-n Bootstrap Scheme, *Biometrics*, *69*(3), doi:10.1111/biom.12052.

Chung, K., and A. Alderman (2002), Replantation of the upper extremity: Indications and outcomes, *Journal of the American Society for Surgery of the Hand*, *2*(2), 78–94.

Chung, K., A. Yoon, S. Malay, M. Shauver, L. Wang, S. Kaur, and FRANCHISE Group (2019), Patient reported and functional outcomes after revision amputation and replantation of digit amputations: The FRANCHISE multicenter international retrospective cohort study, *JAMA Surgery*, *154*(7), 637–646.

Crump, R., V. Hotz, G. Imbens, and O. Mitnik (2009), Dealing with limited overlap in estimation of average treatment effects, *Biometrika*, *96*(1), 187–199.

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

Drake, R., S. Rosenberg, G. Teague, S. Bartels, and W. Torrey (2003), Fundamental principles of evidence-based medicine applied to mental health care, *Psychiatric Clinics of North America*, *26*, 811–820.

Efron, B. (1979), Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, *7*(1), 1–26.

Eilers, P., and B. Marz (1996), Flexible Smoothing with B-splines and Penalties, *Statistical Science*, *11*(2), 89–102.

Eldridge, S., and S. Kerry (2012), *A Practical Guide to Cluster Randomized Trials in Health Services Research*, John Wiley and Sons Ltd., West Sussex, UK.

Elixhauser, A., C. Steiner, D. Harris, et al. (1998), Comorbidity measures for use with administrative data, *Medical Care*, *36*, 8–27.

Fernandez, M., C. Schlechter, G. Del Fiol, et al. (2020), QuitSMART Utah: an implementation study protocol for a cluster-randomized, multi-level Sequential Multiple Assignment Randomized Trial to increase Reach and Impact of tobacco cessation treatment in Community Health Centers, *Implementation Science*, *15*, doi: 10.1186/s13012-020-0967-2.

Field, C., and A. Welsh (2007), Bootstrapping Clustered Data, *Journal of the Royal Statistical Society Series B*, *69*(3), 369–390.

for the Evaluation of Medicinal Products, E. A. (2003), Points to Consider on Adjustment for Baseline Covariates.

Ghosh, P., Y. Cheung, and B. Chakraborty (2016), *Sample size calculations for clustered SMART designs. In E. M. Moodie (Ed.), Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, SIAM, Alexandria, VA.

Goldberger, A., L. Amaral, L. Glass, J. Hausdorff, et al. (2000), PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, *Circulation*, *101*(23), e215–e220.

Gutman, R., and D. Rubin (2015), Estimation of causal effects of binary treatments in unconfounded studies, *Statistics in Medicine*, *34*, 3381–3398.

Guyatt, G., J. Cairns, D. Churchill, et al. (1992), Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine, *JAMA*, *268*.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer, New York, NY.

Hayes, R., and L. Moulton (2017), *Cluster Randomized Trials: Second Edition*, CRC Press, Boca Raton, FL.

Hernan, M., and J. Robins (2020), *Causal Inference: What If*, Chapman and Hall/CRC, Boca Raton, FL.

Hernan, M., B. Brumback, and J. Robins (2001), Marginal structural models to estimate the joint causal effect of non-randomized treatments, *Journal of the American Statistical Association*, *96*, 440–448.

Ho, D., K. Imai, G. King, and E. Stuart (2007), Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference, *Political Analysis*, *15*(3), 199–236.

Holland, P. (1986), Statistics and Causal Inference, *J. Amer. Statist. Assoc*, *81*(396), 945–960, doi:10.1080/01621459.1986.10478354.

Horng, S., L. Nathanson, D. Sontag, and N. Shapiro (2017), Evaluation of the Angus ICD9-CM Sepsis Abstraction Criteria, *bioRXiv*, doi:10.1101/124289.

Hox, J. (2010), *Multilevel Analysis: Techniques and Applications*, Routledge, Taylor Francis, Hove, UK.

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

ICH E Expert Working Group (1999), ICH harmonised tripartite guideline: Statistical principles for clinical trials, *Statistics in Medicine*, *18*(15), 1905–1942.

Iwashyna, T., A. Odden, J. Rohde, et al. (2014), Identifying patients with severe sepsis using administrative claims: patient-level validation of the Angus implementation of the International Consensus Conference Definition of Severe Sepsis, *Medical Care*, *52*(6), e39–43.

Johnson, A., T. Pollard, L. Shen, et al. (2016), MIMIC-III, a freely accessible critical care database, *Scientific data*, *3*, 160,035.

Johnson, A., D. Stone, L. Celi, and T. Pollard (2017), The MIMIC Code Repository: enabling reproducibility in critical care research, *Journal of the American Medical Informatics Association*.

Kamal, R., D. McDermott, and C. Cox (2019), How has U.S. spending on healthcare changed over time?, https://www.healthsystemtracker.org/chart-collection/u-s-spending-healthcare-changed-time/item-start.

Kang, J., and J. Schafer (2007), Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data (with discussion), *Statistical Science*, *22*, 523–539.

Kilbourne, A., D. Almirall, D. Eisenberg, J. Waxmonsky, D. Goodrich, J. Fortney, J. Kirchner, et al. (2014), Protocol: Adaptive Implementation of Effective Programs Trial (ADEPT): Cluster Randomized SMART Trial Comparing a Standard versus Enhanced Implementation Strategy to Improve Outcomes of a Mood Disorders Program, *Implementation Science*, *9*(132).

Kilbourne, A., S. Smith, S. Choi, et al. (2018), Adaptive School-based Implementation of CBT (ASIC): clustered-SMART for building an optimized adaptive implementation intervention to improve uptake of mental health interventions in schools, *Implementation Science*, *13*(119), doi:10.1186/s13012-018-0808-8.

Kosorok, M., and E. Moodie (2016), *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, SIAM, Alexandria, VA.

Laber, E., and Y. Zhao (2015), Tree-based methods for individualized treatment regimes, *Biometrika*, *102*(3), 501–514, doi:10.1093/biomet/asv028.

Laber, E., D. Lizotte, M. Qian, W. Pelham, and S. Murphy (2014), Dynamic treatment regimes: Technical challenges and applications, *Electron J Stat*, *8*, 1225–1272, doi:10.1214/14-EJS920.

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*, *163*(1), 29–38.

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

Little, R., and H. An (2004), Robust Likelihood-based Analysis of Multivariate Data with Missing Values, *Statistica Sinica*, *14*, 949–968.

Little, R., and D. Rubin (2019), *Statistical Analysis with Missing Data, 3rd Edition*, John Wiley Sons, Inc., Hoboken, NJ.

Mantovani, R., T. Horvath, R. Cerri, S. Junior, J. Vanschoren, and A. de Carvalho (2019), An empirical study on hyperparameter tuning of decision trees, *arXiv*.

Marino, P. (2014), *Marino's The ICU Book, Fourth Edition*, Wolters Kluver Health/Lippincott Williams and Wilkins, Philadelphia, PA.

Moodie, E., and T. Richardson (2010), Estimating Optimal Dynamic Regimes: Correcting Bias under the Null, *Scandinavian Journal of Statistics*, *37*, 126–146, doi: 10.1111/j.1467-9469.2009.00661.x.

Moodie, E., B. Chakraborty, and M. Kramer (2012), Q-learning for estimating optimal dynamic treatment rules from observational data, *Canadian Journal of Statistics*, *40*(4), 629–645, doi:10.1002/cjs.11162.

Moodie, E., N. Dean, and Y. Sun (2013), Q-Learning: Flexible Learning About Useful Utilities, *Stat Biosci*, *6*, 223–243, doi:10.1007/s12561-013-9103-z.

Murphy, S. (2003), Optimal Dynamic Treatment Regimes, *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *65*(2), 331–355.

Murphy, S. (2005a), A Generalization Error for Q-Learning, *Journal of Machine Learning Research*, *6*, 1073–1097.

Murphy, S. (2005b), An experimental design for the developmentof adaptive treatment strategies, *Statistics in Medicine*, *24*, 1455–1481.

Murphy, S., M. van der Laan, J. Robins, and CPPRG (2001), Marginal Mean Models for Dynamic Regimes, *Journal of the American Statistical Association*, *96*(456), 1410–1423.

Murray, D. (1998), *Design and Analysis of Group-Randomized Trials, First Edition*, Oxford University Press, New York, NY.

Murray, D., S. Varnell, and J. Blitstein (2004), Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments, *American Journal of Public Health*, *94*(3), 423–432.

Nahum-Shani, I., M. Qian, D. Almirall, W. Pelham, B. Gnagy, G. Fabiano, J. Waxmonsky, J. Yu, and S. Murphy (2012), Q-Learning: A Data Analysis Method for Constructing Adaptive Interventions, *Psychological Methods*, *17*(4), 478–494, doi: 10.1037/a0029373.

Necamp, T., A. Kilbourne, and D. Almirall (2017), Comparing Cluster-Level Dynamic Treatment Regimens Using Sequential, Multiple Assignment, Randomized Trials: Regression Estimation and Sample Size Considerations, *Statistical Methods in Medical Research*, *26*(4), 1572–1589, doi:10.1177/0962280217708654.

Oetting, A., J. Levy, R. Weiss, and S. Murphy (2011), *Statistical Methodology for a SMART Design in the Development of Adaptive Treatment Strategies. In Causality and Psychopathology: Finding the Determinants of Disorders and Their Cures*, American Psychiatric Publishing, Inc, Arlington, VA.

Pocock, S., S. Assmann, L. Enos, and L. Kasten (2002), Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, *Statistics in Medicine*, *21*(19), 2917–2930.

Pollard, T. J., and A. E. Johnson (2016), The MIMIC-III Clinical Database, http://dx.doi.org/10.13026/C2XW26, doi:10.13026/C2XW26.

Qian, M., and S. Murphy (2011), Performance Guarantees for Individualized Treatment Rules, *Annals of Statistics*, *39*(2), 1180–1210, doi:10.1214/10-AOS864.

Quanbeck, A., D. Almirall, N. Jacobson, et al. (2020), The Balanced Opioid Initiative: protocol for a clustered, sequential, multiple-assignment randomized trial to construct an adaptive implementation strategy to improve guideline-concordant opioid prescribing in primary care, *Implementation Science*, *15*(26), doi:10.1186/s13012-020-00990-4.

R Core Team (2018), *R: A language and environment for statistical computing.*, R Foundation for Statistical Computing, Vienna, Austria.

Raab, G., S. Day, and J. Sales (2000), How to select covariates to include in the analysis of a clinical trial, *Controlled Clinical Trials*, *21*(4), 330–342.

Raudenbush, S., and A. Bryk (2002), *Hierarchical Linear Models, Second Edition*, Sage Publications, Thousand Oaks, CA.

Raudenbush, S., and D. Schwartz (2020), Randomized Experiments in Education, with Implications for Multilevel Causal Inference, *Annual Review of Statistics and its Application*, *7*, 117–208.

Rhodes, A., L. Evans, A. Waleed, et al. (2017), Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016, *Critical Care Medicine*, *45*(3), 486–552.

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

Robins, J. (1994), Correcting for non-compliance in randomized trials using structural nested mean models, *Communications in Statistics-Theory and methods*, *23*, 2379–2412.

Robins, J. (1997), *Causal inference from complex longitudinal data. In Latent Variable Modeling and Applications to Causality*, 69-117 pp., Springer, New York, NY.

Robins, J. (2000), Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference, *Statistical Models in Epidemiology: The Environment and Clinical Trials*, *116*, 95–134.

Robins, J. (2004), Optimal Structural Nested Models for Optimal Sequential Decisions, in *Lecture Notes in Statistics*, Proceedings of the Second Seattle Symposium in Biostatistics, pp. 189–326, Springer, New York, NY.

Robins, J., and M. Hernan (2009), *Estimation of the causal effects of time-varying exposures. In: Longitudinal Data Analysis (eds Fitzmaurice, Davidian, et al.*, Chapman and Hall/CRC, Boca Raton, FL.

Rosenbaum, P. (2012), Optimal Matching of an Optimally Chosen Subset in Observational Studies, *Journal of Computational and Graphical Statistics*, *21*(1), 57–71.

Rosenbaum, P., and D. Rubin (1983), The central role of the propensity score in observational studies for causal effects, *Biometrics*, *70*(1), 41–55.

Rubin, D. (1974), Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology*, *66*, 688–701.

Rupert, D. (2002), Selecting the Number of Knots for Penalized Splines, *Journal of Computational and Graphical Statistics*, *11*(4), 735–757, doi:10.1198/106186002853.

Schulte, P., A. Tsiatis, E. Laber, and M. Davidian (2014), Q- and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes, *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *29*(4), 640–661.

Sebastin, S., and K. Chung (2011), A systematic review of the outcomes of replantation of distal digital amputation, *Plast Reconstr Surg*, *128*(3), 723–737.

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

Shortreed, S., E. Laber, D. Lizotte, T. Stroup, J. Pineau, and S. Murphy (2011), Informing sequential clinical decision-making through reinforcement learning: an empirical study, *Machine Learning*, *84*(1-2), 109–136, doi:10.1007/s10994-010-5229-0.

Smith, S., D. Almirall, K. Prenovost, C. Liebrecht, J. Kyle, D. Eisenberg, M. Bauer, and A. Kilbourne (2019), Change in Patient Outcomes After Augmenting a Low-level Implementation Strategy in Community Practices That Are Slow to Adopt a Collaborative Chronic Care Model: A Cluster Randomized Implementation Trial, *Medical Care*, *57*(7), 503–511.

Song, R., W. Wang, D. Zeng, and M. Kosorok (2015), Penalized q-learning for dynamic treatment regimes, *Statistica Sinica*, *25*(3), 901–920, doi:10.5707/ss.2012.364.

Speth, K., and K. Kidwell (2019), *Sequential, Multiple Assignment, Randomized Trials In S. Michiels (Ed.), Textbook of Clinical Trials in Oncology: A Statistical Perspective*, 399 pp., CRC Press, Boca Raton, FL.

Speth, K., A. Yoon, L. Wang, and K. Chung (2020), Tree-based Statistical Learning to Recommend the Optimal Treatment Decision Rules for Traumatic Finger Amputations, *JAMA Network Open*, *3*(2), doi:10.1001/jamanetworkopen.2019.21626.

Sun, Y., and L. Wang (2020), Stochastic Tree Search for Estimating Optimal Dynamic Treatment Regimes, *Submitted*.

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

Tao, Y., L. Wang, and D. Almirall (2018), Tree-based Reinforcement Learning for Estimating Optimal Dynamic Treatment Regimes, *The Annals of Applied Statistics*, *12*(3), 1914–1938, doi:10.1214/18-AOAS1137.

Thall, P., R. Millikan, and H. Sung (2000), Evaluating Multiple Treatment Courses in Clinical Trials, *Statistics in Medicine*, *19*(8), 1011–1028.

Thall, P., H. Sung, and G. Estey (2002), Selecting Therapeutic Strategies Based on Efficacy and Death in Multicourse Clinical Trials, *Journal of the American Statistical Association*, *97*(457), 29–39, doi:10.1198/016214502753479202.

Thall, P., C. Logothetis, L. Pagliara, S. Wen, M. Brown, D. Williams, and R. Millikan (2007), Adaptive Therapy for Androgen-Independent Prostate Cancer: A Randomized Selection Trial of Four Regimens, *Journal of the National Cancer Institute*, *99*(21), 1613–1622.

Therneau, T., E. Atkinson, and Mayo Foundation (2019), An Introduction to Recursive Partitioning Using the RPART Routines, *CRAN R Network*.

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

van Walraven, C., P. Austin, A. Jennings, and others (2009), A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data, *Medical Care*, *47*, 626–633.

Vilagut, G., et al. (2013), The mental component of the short-form 12 health survey (SF-12) as a measure of depressive disorders in the general population: results with three alternative scoring methods, *Value Health*, *16*(4), 564–573, doi:10.1016/j.val.2013.01.006.

Vincent, J., R. Moreno, J. Takala, S. Willatts, et al. (1996), The SOFA (Sequential Organ Failure Assessment) score to describe organ dysfunction/failure, *Intensive Care Medicine*, *22*(7), 707–710.

Wallace, M., E. Moodie, and D. Stephens (2016), SMART Thinking: A Review of Recent Developments in Sequential Multiple Assignment Randomized Trials, *Current Epidemiology Reports*, *3*(3), 225–232.

Wand, M. (2003), Smoothing and mixed models, *Computational Statistics*, *18*, 223–249.

Wang, L., A. Rotnitzky, X. Lin, R. Millikan, and P. 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.

Ware Jr, J., M. Kosinski, and S. Keller (1996), A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity, *Medical Care*, *34*, 220–233.

Watkins, C. (1989), *Doctoral Thesis: Learning from delayed rewards*, University of Cambridge, Cambridge, UK.

Wright, N. (2015), *Thesis: Choosing covariates in the analysis of cluster randomised trials*, Queen Mary University of London, London, UK.

Wright, N., N. Ivers, S. Eldridge, M. Taljaard, and S. Bremner (2015), A review of the use of covariates in cluster randomized trials uncovers marked discrepancies between guidance and practice, *Journal of Clinical Epidemiology*, *68*(6), 603–609, doi:10.1016/j.jclinepi.2014.12.006.

Xu, Y., P. Mueller, A. Wahed, and P. Thall (2016), Bayesian Nonparametric Estimation for Dynamic Treatment Regimes with Sequential Transition Times, *Journal of the American Statistical Association*, *111*(515), 921–950, doi:10.1080/1621459.2015.1086353.

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

Zhang, B., A. Tsiatis, E. Laber, and M. Davidian (2012b), A Robust Method for Estimating Optimal Treatment Regimes, *Biometrics*, *68*(4), 1010–1018.

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

Zhang, G., and R. Little (2009), A comparative study of doubly robust estimators of the mean with missing data, *Journal of Statistical Computation and Simulation*, *81*(12), 2039–2058, doi:10.1080/00949655.2010.516750.

Zhang, Y., E. Laber, A. Tsiatis, and M. Davidian (2015), Using Decision Lists to Construct Interpretable and Parsimonious Treatment Regimes, *Biometrics*, *71*(4), 895–904, doi:10.1111/biom.12354.

Zhang, Y., E. Laber, M. Davidian, and A. Tsiatis (2018), Interpretable Dynamic Treatment Regimes, *Journal of the American Statistical Association*, *113*(524), 1541–1549, doi:10.1080/01621459.2017.1345743.

Zhao, Y., M. Kosorok, and D. Zeng (2009), Reinforcement learning design for cancer clinical trials, *Statistics in Medicine*, *28*(26), 3294–3315, doi:10.1002/sim.3720.

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

Zhao, Y., D. Zeng, J. Rush, and M. Kosorok (2012), Estimating Individualized Treatment Rules Using Outcome Weighted Learning, *Journal of the American Statistical Association*, *107*(449), 1106–1118.

Zhao, Y., D. Zeng, E. Laber, and M. Kosorok (2015), New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes, *Journal of the American Statistical Association*, *110*(510), 583–598, doi:10.1080/01621459.2014.937488.

Zhou, G., M. Lee, H. Atieli, J. Githure, A. Githeko, J. Kazura, and G. Yan (2020), Adaptive interventions for optimizing malaria control: an implementation study protocol for a block-cluster randomized, sequential multiple assignment trial, *Trials*, *21*(665), doi:10.1086/s13063-020-04573-y.

Zhou, T., M. Elliott, and R. Little (2019), Penalized Spline of Propensity Methods for Treatment Comparison, *Journal of the American Statistical Association*, *114*(525), 1–19, doi:10.1080/01621459.2018.1518234.