STATISTICAL MACHINE LEARNING METHODOLOGY FOR INDIVIDUALIZED TREATMENT RULE ESTIMATION IN PRECISION MEDICINE

Kushal S. Shah

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2023

Approved by:

Michael R. Kosorok

Haoda Fu

Tanya P. Garcia

Anna R. Kahkoska

Feng-Chang Lin

ABSTRACT

Kushal S. Shah: Statistical Machine Learning Methodology for Individualized Treatment Rule Estimation in Precision Medicine
(Under the direction of Michael R. Kosorok and Haoda Fu)

Precision medicine aims to deliver optimal, individualized treatments for patients by accounting for their unique characteristics. With a foundation in reinforcement learning, decision theory, and causal inference, the field of precision medicine has seen many advancements in recent years. Significant focus has been placed on creating algorithms to estimate individualized treatment rules (ITRs), which map from patient covariates to the space of available treatments with the goal of maximizing patient outcome.

In Chapter 1, we extend ITR estimation methodology in the scenario where variance of the outcome is heterogeneous with respect to treatment and covariates. Accordingly, we propose Stabilized Direct Learning (SD-Learning), which utilizes heteroscedasticity in the error term through a residual reweighting framework that models residual variance via flexible machine learning algorithms such as XGBoost and random forests. We also develop an internal cross-validation scheme which determines the best residual model among competing models. Further, we extend this methodology to multi-arm treatment scenarios.

In Chapter 2, we develop ITR estimation methodology for situations where clinical decision-making involves balancing multiple outcomes of interest. Our proposed framework estimates an ITR which maximizes a combination of the multiple clinical outcomes, accounting for the fact that patients may ascribe importance to outcomes differently (utility heterogeneity). This approach employs inverse reinforcement learning (IRL) techniques through an expert-augmentation solution, whereby physicians provide input to guide the utility estimation and ITR learning processes.

In Chapter 3, we apply an end-to-end precision medicine workflow to novel data from older adults with Type 1 Diabetes in order to understand the heterogeneous treatment effects of continuous glucose monitoring (CGM) and develop an interpretable ITR to reveal patients for which CGM confers a major safety benefit. The results from this analysis elucidate the demographic and clinical markers which moderate CGM's success, provide the basis for using diagnostic CGM to inform therapeutic CGM decisions, and serve to augment clinical decision-making.

Finally, in Chapter 4, as a future research direction, we propose a deep autoencoder framework which simultaneously performs feature selection and ITR optimization, contributing to methodology built for direct consumption of unstructured, high-dimensional data in the precision medicine pipeline.

To Bapa.
May my every action be a step towards You.

ACKNOWLEDGEMENTS

Bapa - my Guru, my guide, and best friend. I have experienced Your blessings at every step. You have never even given me the chance to feel lonely, insecure, or helpless. Rather, You have made the journey so light, so joyous. What is contained in these pages is nothing - I was simply holding the pen, but You were the force behind every stroke. With immense gratitude, Your *bhakt* moves forward in his journey, with a purpose so much deeper than the decorations the world may give or take away. May this life be a celebration of You, and nothing else. Thank You endlessly. *Ho Vandan Aganit.*

I am fortunate to have two advisors who I can look up to first and foremost as people, and then as researchers. **Dr. Kosorok**, you carry yourself with such humility that no one who meets you would think that you are amongst the world leaders in statistical precision medicine. I am amazed by the way that you treat everyone with the same level of respect and courtesy, whether they are a student or a world-renowned researcher. I strive to emulate this in my career as well as life in general. Whenever I describe you to someone, though you are a brilliant researcher, I first talk about you as a person - which I think is such a reflection of your character. Over five years of weekly meetings, I have never once seen you frustrated, but rather, always supportive, positive, and ready to tackle challenges. I'm very grateful that on top of academic growth, you supported my emotional and spiritual growth. Thank you for everything you have taught me, both personally and professionally. **Haoda**, you came in at such a critical part of my PhD journey and instilled the confidence I needed to transition into the research phase. From the beginning, you have reiterated that your only aim was to support me and help me develop as a statistician and machine learning researcher; your belief in me has been instrumental in my progression to where I am now. A lot of my excitement about machine learning stems from the passion you have for the field, which exudes from you in every conversation that we have. Your desire to continue

learning and continue innovating is very inspiring to me, and it is so clear that you love what you do. I have benefitted tremendously from your mentoring and hope to find the same passion in my career as you have. Thank you for your kindness and inspiration.

There are three individuals who actually deserve the full credit for this publication: Mom, dad, and Neil. I have never had to go outside of you three to find someone who will listen to me, who will hear me, who will laugh with me. You have kept me grounded, aligned with my true purpose, and are the real reason behind any success I experience. **Mom**, your optimism and your passion has given me such a lift whenever I've needed it the most. You've taught me that long hours and high spirits are not mutually exclusive, success and kindness are not at odds with one another, and self-growth does not preclude serving others. I saw you work tirelessly for years while remaining 100% present for your family, and have also seen you retire early to follow your true purpose - both phases have been equally inspiring. Thank you for keeping me motivated, while being so patient and loving at the same time. **Dad**, it is when I speak to you that I don't feel far from home. There is a warmth to your presence which is difficult to describe, but so easy to experience. I don't think you even realize it when you do this, but you have the ability to instill such confidence in me through just a few words. Thank you for being my role model since the very beginning and embodying the modesty and selflessness that I wish to live by. You have given constantly and never asked for anything in return; I owe you the world. **Neil**, you have the ability to improve the people around you, which is a rare quality and one that I admire deeply. I have tried to model myself off of you in so many ways, and each time, I have come out of it a better person. I am beyond excited to see the heights that you reach in whatever pursuit you undertake. Thank you for repeatedly having my back, always reminding me of my goals, and consistently being a treasure of a friend to me.

Nani, in 2018, when I moved to North Carolina, I had to write "University of North Carolina" and "PhD - Biostatistics" on small piece of paper for you, because you were so intent on remembering the details of my school and program. What a force of innocence and kindness you were in my life and the lives of everyone around you. Your blessings run so deeply in my life that

it is difficult to find an area you didn't touch, directly or indirectly. May I one day be able to give the people around me the same unconditional love that you did. May I dance in my 80s with the same energy that you did in your 80s. May I make you proud, not by what I accomplish in the world, but by the values I choose to live by.

To the entire extended **Shah and Jhaveri families**, thank you for the love you have given me at all times. You have always wanted the best for me, and you have always been there, enthusiastically, to share every bit of good news. I look up to each of you in unique ways, and am thankful for everything you have taught me. You make me feel that I am moving forward with more blessings than I can count.

Aastha, what a light of energy and positivity you are in my life. You have felt like one of my own since the day I met you. There is a softness in the way you carry yourself, and a depth to your character, which is already influencing me to be a better person. I have felt your support every day and feel so lucky to have met you and your beautiful family. Thank you for loving me, cheering me on, and keeping me smiling. I am so excited for our journey together.

Harsha Auntie and **Umesh Uncle**, it was at your house, with your family, that I learned the importance of *satsang*. So much of my spiritual journey has been shared with you and your amazing family. You have been my Chapel Hill parents since day one, and I know that our connection goes long back - much more than just five years. Saying I am thankful for you would be an understatement; you made me feel that I have always had a home here, and because of you, I have not once felt lonely. I am so grateful that you have been a medium for my connection during such an important phase of my life. Thank you for taking me in so graciously, supporting and loving me unconditionally, and making me feel so comfortable in your presence.

I'm also fortunate to have been surrounded by friends who brought so much joy to my time in North Carolina. **Marissa**, I genuinely mean it when I say that I think you're capable of anything you put your mind to. You have such an impressive work ethic, self-awareness, mental strength, and reliability. Your friendship has helped me be a better person in so many ways, and I've learned an endless amount from you and am so amazed by you. **Kyle**, I remember going

through second year coursework thinking you were one of the smartest people I had ever met. My time since then has only confirmed this initial hypothesis. However, not only smartest, but also one of the kindest and most courteous people I've gotten to know. Thank you for everything. **Christina**, friendship with you felt so natural from the beginning. It's so easy to be open with you about everything. I've learned so much from the way you express genuine excitement in the smallest successes of others. Thank you for being there to unconditionally support and encourage me throughout my time at UNC. **Jay**, I cherish our friendship - our climbing sessions, hikes, dinners, Hurricanes games, etc. are some of my fondest memories from my time in NC. I'm so inspired by your resilience, positive attitude, and friendliness towards everyone you meet. Thank you for being an amazing friend and *kalyanmitra*, and for sharing the most important moments of the last couple years with me. I've found a friend for life in you, and am excited to see the success that you achieve in all spheres.

I would also like to thank **Anna Kahkoska** and **Marianne Muhlebach** for your support of my growth, for teaching me the applications of statistics and machine learning in human health, and for pushing me to be a better researcher, **Tanya Garcia** and **Feng-Chang Lin** for graciously serving on my dissertation committee, and all **PHAIR Lab members** for inspiring me to learn about precision medicine and pushing me to think critically. Finally, a special thank you to **Melissa Hobgood** - you are a special light in the Biostatistics department, and your unending support of all students does not go unnoticed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIDS	Acquired Immunodeficiency Syndrome
AIPWE	Augmented Inverse Probability Weighted Estimator
APE	Average Prediction Error
BDC	Brownian Distance Covariance
BGM	Blood Glucose Monitoring
BUEI	Bounded Uniform Entropy Integrable
CATE	Conditional Average Treatment Effect
CCA	Canonical Correlation Analysis
CGM	Continuous Glucose Monitoring
CITY	CGM Intervention in Teens and Young Adults with T1D
CV	Coefficient of Variation
DDROWL	Deep Doubly Robust Outcome Weighted Learning
DKA	Diabetic Ketoacidosis
DNN	Deep Neural Network
DRF	Distributional Random Forests
EARL	Efficient Augmentation and Relaxation Learning
FDR	False Discovery Rate
FGLS	Feasible Generalized Least Squares
FWLS	Feasible Weighted Least Squares
GLMM	Generalized Linear Mixed Model
HIV	Human Immunodeficiency Virus
IRL	Inverse Reinforcement Learning
IQR	Interquartile Range
ITR	Individualized Treatment Rule
LASSO	Least Absolute Shrinkage and Selection Operator

MCMC	Markov Chain Monte Carlo
MDG	Mean Decrease Gini
MH	Metropolis-Hastings
MLE	Maximum Likelihood Estimate
MOML	Multicategory Outcome Weighted Margin-based Learning
MSE	Mean Squared Error
OLS	Ordinary Least Squres
OWL	Outcome Weighted Learning
PCA	Principal Component Analysis
PM	Pointwise Measurable
RCT	Randomized Clinical Trial
RF	Random Forest
RKHS	Reproducing Kernel Hilbert Space
RWL	Residual Weighted Learning
SD	Standard Deviation
SEM	Standard Error of the Mean
SGD	Stochastic Gradient Descent
SL	SuperLearner
SUTVA	Stable Unit Treatment Value Assumption
SVM	Support Vector Machine
SVR	Support Vector Regression
T1D	Type 1 Diabetes
T-SNE	T-Distributed Stochastic Neighbor Embedding
VAE	Variational Autoencoder
WISDM	Wireless Innovation for Seniors with Diabetes Mellitus
WLS	Weighted Least Squares
XGB	XGBoost

CHAPTER 1: STABILIZED DIRECT LEARNING FOR EFFICIENT ESTIMATION OF INDIVIDUALIZED TREATMENT RULES

1.1 Introduction

Precision medicine is a framework at the intersection of statistics, machine learning, and causal inference, for leveraging patient heterogeneity to improve patient outcomes. Unlike the standard "one-size-fits-all" approach designed for the average patient, individualized treatment rules (ITRs) may recommend personalized actions based on patient demographic information, clinical biomarkers, or genetic data in order to maximize expected treatment benefit across a population. One of the primary goals of precision medicine has been formalized as decision support - the estimation of optimal and near-optimal regimes (Kosorok and Laber (2019)). In this vein, it is important to develop algorithms which work efficiently with the data at hand.

The performance of an ITR is commonly evaluated by its value function, a measure of the expected population mean outcome if all patients were to follow the decision rule. Various precision medicine algorithms have been developed to identify the optimal treatment rule which maximizes the value function (Imai and Li (2021)). The value function can be estimated through cross-validation techniques as a weighted combination of the individual patient outcomes (Qian and Murphy (2011); Jiang (2020)).

It is also clinically relevant to identify biomarkers which are influential in choosing an optimal treatment. Kosorok and Laber (2019) breaks down biomarkers which provide clinical information into three types: prognostic/predictive (useful in predicting patient mean outcome with respect to a clinical endpoint of interest), moderating (useful in predicting contrasts in the effects of candidate treatments on a mean outcome), and prescriptive (useful in selecting an optimal treatment to maximize the clinical outcome). Although all moderating biomarkers are prognos-

tic, and all prescriptive biomarkers are both moderating and prognostic, it may be the case that a prognostic biomarker is not prescriptive. The discovery of prescriptive biomarkers is relevant to precision medicine because it can inform the choice of an optimal action (treatment selection).

1.1.1 Common Methods for ITR Estimation

In recent years, an extensive literature has been developed in the area of estimating optimal ITRs to maximize a single outcome of interest. Traditionally, algorithms have fallen into one of two categories: *model-based* vs. *policy-search* approaches.

Model-based approaches may also be considered "regression-based" or "indirect" in that they first model the conditional response of interest and then invert the relationship between patient covariates, treatment, and outcome to estimate an optimal rule (Kosorok and Moodie (2016)). Primary examples of model-based approaches are Q-Learning (Qian and Murphy (2011)) and A-Learning (Murphy (2003); Robins (2004)). Q-Learning approaches model the outcome conditional on covariates, whereas A-Learning approaches model regret functions or contrast functions between treatments (Schulte *et al.* (2014)). The Qian and Murphy (2011) method is a two-step procedure: it first estimates the mean outcome, conditional on covariates and treatment, and then compares the conditional mean outcome across individual treatments in order to determine the optimal treatment.

Model-based methods are convenient because they allow for the use of well-known regression algorithms to model the conditional response of interest (e.g. linear regression, random forest, gradient boosting trees, etc.). In this sense, they are flexible and easy to implement, as well as logical: model the response as a function of each treatment individually, and pick the treatment that gives the optimal predicted response. However, this modeling approach requires the specification of a mean response model, which results in estimating many nuisance parameters (main effect parameters) which are not directly of interest for treatment selection. Performance, therefore, is highly dependent on correct specification of the mean response model, and can easily suffer from misspecification. Additionally, model-based methods often favor prediction accuracy

(good models for predicting outcome/response) over treatment decision accuracy (good ITRs for maximizing patient outcomes) (Murphy (2005)) due to a mismatch between the target of outcome regression modeling and the goal of learning the optimal ITR (Zhang and Zhang (2018)).

Policy-search approaches, on the other hand, maximize value functions directly instead of modeling the conditional mean. For this reason, they are also known as "value-search" or "direct-search" (Kosorok and Moodie (2016)). Many policy-search approaches have reframed the maximization of clinical outcomes as a weighted classification problem with the goal of minimizing weighted classification error, including the outcome weighted learning (OWL) family of methods (Zhao *et al.* (2012); Zhou *et al.* (2017); Zhang *et al.* (2020)), which utilizes weighted support vector machines (SVMs) (Vapnik (1999)) for ITR estimation, and various tree-based extensions (Cui, Zhu, and Kosorok (2017); Zhu *et al.* (2017); Kallus (2018)). By sidestepping the modeling step and directly searching for an optimal rule among a class of policies, policy-search algorithms may avoid model misspecification (Xiao *et al.* (2019)). However, policy-search approaches such as OWL involve maximizing a discontinuous objective function, which can become computationally burdensome.

A third category of algorithms to estimate ITRs also exists, consisting of hybrid methods which have attempted to combine the advantages of model-based and policy-search approaches while maintaining the classification framework. Often, these methods use the augmented inverse probability weighted estimator (AIPWE) within the classification framework, which requires a regression model for outcome to be posited. Such approaches are robust in the sense that they enjoy greater protection against model misspecification and increased efficiency when both models are correctly specified (Zhang and Zhang (2018); Liu *et al.* (2018); Zhao *et al.* (2019)). Further, residual weighted learning (RWL) of Zhou *et al.* (2017) can be considered as a modification of the AIPWE approach which enjoys efficiency gained from replacing the outcome in OWL by the residual (which requires an outcome model to be estimated). It is important to note that these approaches still maintain the classification-based framework and must handle the discontinuous objective function through a variety of possible methods (e.g. SVM-like estimation using a con-

cave relaxation like hinge loss (Hastie *et al.* (2009)), the GENOUD algorithm for discontinuous optimization (Mebane and Sekhon (2011)), etc.). Notably, however, estimation is supported by a mean outcome model in order to gain efficiency and robustness.

The following is a summary of the three categories of approaches, adapted from Zhang and Zhang (2018):

- Model-based: Response (outcome or regret) regression modeling followed by selection of a treatment which optimizes response. The form of the optimal ITR is completely determined by specification of mean response models (Q- or A-functions).
- Policy-search: Direct maximization of value function through classification framework, without incorporating information from outcome regression modeling. The form of the resulting ITR is dictated by optimization method.
- Classification-based Hybrid: Direct maximization of value function through classification framework, supplemented by information borrowed from outcome regression models (good outcome regression models can augment performance). However, outcome regression models to not dictate the form of the optimal ITR.

1.1.2 Direct Learning: A Unique Regression-Based Hybrid

Tian *et al.* (2014) developed a "modified-covariate" approach for ITR estimation, which is a hybrid method but is unique in that it maintains the regression-based framework to model the treatment-covariate interaction effect directly, without having to specify a main effect model or conditional mean outcome function. Later, Qi and Liu (2018) coined the term "Direct Learning" (D-Learning) for the Tian *et al.* (2014) method and extended it to nonlinear decision rules and multi-arm treatment settings. D-Learning is a simple method which allows for regression to be used, and is yet a one-step approach which sidesteps mean outcome modeling altogether. Therefore, it comes the benefits of many aforementioned methods because it maintains the flexi-

bility and simplicity of regression, models the treatment-covariate interaction effect directly, and bypasses the estimation of main effect nuisance parameters.

Key extensions to D-Learning are Angle-based Direct Learning (AD-Learning) (Qi *et al.* (2020)), which improves D-Learning in the multi-arm treatment case, and Robust Direct Learning (RD-Learning) (Meng and Qiao (2021)), which replaces the outcome by the residual similarly to RWL, thereby achieving a double robustness property.

Now consider, for example, the "AIDS Clinical Trial Group Study 175" (ACTG175), a randomized clinical trial (RCT) which compared the effectiveness of four treatments in increasing CD4 cell counts in HIV-1 patients (Hammer, S. M. et al. (1996)). Previous studies have suggested that the response of change in CD4 cell count from this data may have skewed, heteroscedastic errors (Xiao *et al.* (2019), Zhang *et al.* (2021)), which we confirm in Section 1.5. In such a situation, when the variance of the clinical outcome is a function of the covariates (or treatment), the D-Learning family of estimators remains consistent for the optimal ITR, but gives each observation equal weight by default in model training. A reweighting approach which utilizes this error structure to prioritize observations with smaller expected outcome variance is beneficial because it can attain greater efficiency when estimating an ITR. This example motivates the the ITR estimation approach of this paper.

1.1.3 Proposed Method

In this article, we propose Stabilized D-Learning (SD-Learning), a method to increase the efficiency of D-Learning estimates in situations where the variance of the error term is non-homogeneous and a function of the treatment and covariates. SD-Learning can be viewed as a special case of the framework of Liang and Yu (2020) with a single-index model. Liang and Yu (2020) find the efficient score for a semiparametric, and hence general, class of estimators of the decision function, but the estimation procedure does not lead to optimality. The SD-Learning methodology specializes on a smaller class of decision rules and achieves optimality within that class. These differences are highlighted in Section 1.2. From another perspective, SD-Learning

may be considered an adaptation of feasible weighted least squares (FWLS) (Olive (2017)) to the precision medicine setting where optimal ITR estimation with two or more treatments is of interest. Similarly to FWLS, SD-Learning is motivated by efficient estimation and controls on variance (Miller and Startz (2018)). We contribute to existing literature in the following ways:

1. We bring the work of Tian *et al.* (2014), Qi and Liu (2018), Qi *et al.* (2020), and Meng and Qiao (2022) into a single framework such that the estimated parameters from either of these methods can be improved by a single-iteration update.

2. Our method applies concepts from weighted least squares (WLS) theory to increase the precision of ITR parameter estimation under heteroscedasticity. This entails a residual reweighting framework where residual variance is modeled through flexible machine learning methods. We develop an internal cross-validation scheme allowing for selection of an optimal model amongst methods such as XGBoost (Chen and Guestrin (2016)) and random forests (Breiman (2001)).

3. We allow for even the multiple-treatment scenario ($K \geqslant 3$ treatments) to fit into the least squares framework through a vectorization approach. As a result of this, parameter estimation has a simple implementation leading to a closed-form solution; hence, the algorithm is efficient and does not require iterative optimization techniques to solve.

4. We show that SD-Learning parameter estimates are consistent, asymptotically normal in binary and multi-arm treatment scenarios under heterogeneous error, have greater efficiency than D-Learning estimates, and establish value function convergence bounds.

The rest of this paper is organized as follows: In Section 1.2, we introduce the methodology of SD-Learning. Specifically, Section 1.2.1 reviews recent developments in D-Learning and Section 1.2.2 introduces the mathematical motivation behind SD-Learning and outlines the reweighting solution. Section 1.2.3 extends the reweighting solution to scenarios with multi-arm treatments. In Section 1.2.4, the residual model fitting step of the method is described in greater

detail, and a stepwise implementation of the method is delineated. In Section 1.3, theoretical results for SD-Learning including consistency, asymptotic normality, asymptotic efficiency, and value bounds are established for binary and multi-arm settings. Head-to-head simulations comparing SD-Learning to D-Learning, AD-Learning, and RD-Learning based on average prediction error (APE), misclassification rate, and empirical value are provided in Section 1.4, and value comparisons from analysis of the ACTG175 RCT data are made in Section 1.5. Concluding discussions and areas for future work are presented in Section 1.6.

1.2 Stabilized Direct Learning (SD-Learning)

Although SD-Learning works with observational data, for simplicity, we first consider an RCT setting to demonstrate the methodology. For n patients, we observe independent realizations of the random triplet $(\boldsymbol{X}, A, R)$. Patient covariates are represented by the p-dimensional vector, $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$, which includes an intercept. We start with the binary treatment scenario, $A \in \mathcal{A} = \{-1, 1\}$. Clinical outcome is represented by $R \in \mathbb{R}$, and it is assumed, without loss of generality, that larger R corresponds to better outcome. The probability of receiving treatment a, given covariates $\boldsymbol{x}$, is represented by $\pi(a, \boldsymbol{x}) = \Pr(A = a | \boldsymbol{X} = \boldsymbol{x})$. An ITR, $d(\boldsymbol{X})\colon \mathcal{X} \mapsto \mathcal{A}$, is a mapping from covariates to treatments. Let $\mathbb{1}(\cdot)$ represent the indicator function, $\boldsymbol{Z}^\top$ denote the transpose of matrix $\boldsymbol{Z}$, and $\mathbb{P}_n(\cdot)$ represent empirical average (e.g. $\mathbb{P}_n(\boldsymbol{X}) = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i$, where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are realizations of the random variable, $\boldsymbol{X}$). Let Vec$(\boldsymbol{Z})$ represent vectorization (e.g. for matrix $\boldsymbol{Z} = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, Vec$(\boldsymbol{Z}) = [a\ c\ b\ d]^\top$).

Let $R^*(-1)$ and $R^*(1)$ represent potential outcomes that would have been observed had a patient received treatment -1 or 1, respectively. From the framework of Rubin (1974), we make the usual assumptions for the precision medicine context (Kosorok and Moodie (2016), Hernn and Robins (2019), Kosorok and Laber (2019)): (i) Stable unit treatment value assumption (SUTVA): $R = R^*(A)$, (ii) No unmeasured confounding (conditional exchangeability): $A \perp \{R^*(-1), R^*(1)\} \mid \boldsymbol{X}$, and (iii) Positivity: $\pi(A, \boldsymbol{X}) > c > 0, \forall\ A \in \mathcal{A}, \boldsymbol{X} \in \mathcal{X}$. Prior to outlining

the proposed SD-Learning methodology, we review key findings from the D-Learning family of methods.

1.2.1 D-Learning Background

It is known from Qian and Murphy (2011) that the expected response under an ITR, d, can be represented by the value function:

$$V(d) = E\{R \mid A = d(\boldsymbol{X})\} = E\left[\frac{R \cdot \mathbb{1}\{A = d(\boldsymbol{X})\}}{\pi(A, \boldsymbol{X})}\right],$$

and we define an optimal ITR, d^{opt}, as the decision rule that maximizes the expected average response: $d^{opt}(\cdot) = \underset{d \in \mathcal{D}}{\operatorname{argmax}}\, V(d)$, where $\mathcal{D}$ is a prespecified class of decision rules. Using the potential outcomes notation, $V(d) = E\{R^*(d)\} = \sum_{a \in \{-1,1\}} E\{R^*(a)\} P\{d(\boldsymbol{X}) = a\}$, representing the counterfactual population mean outcome under the ITR, d.

1.2.1.1 D-Learning

In the two-arm setting, assume that the outcome can be expressed by:

$$R = m(\boldsymbol{X}) + \delta(\boldsymbol{X})A + \eta, \tag{1.1}$$

where $m(\boldsymbol{X})$ and $\delta(\boldsymbol{X})$ are measurable functions representing the main and interaction effects, respectively, and η is a mean-zero random error term. (1.1) is a general multivariate regression setup for characterizing interactions between treatment and covariates. Note the following:

$$d^{opt}(\boldsymbol{X}) = \text{sign}\left\{E(R|\boldsymbol{X}, A = 1) - E(R|\boldsymbol{X}, A = -1)\right\} := \text{sign}\left\{f^{opt}(\boldsymbol{X})\right\}, \tag{1.2}$$

$$f^{opt}(\boldsymbol{X}) = E\left\{\frac{RA}{\pi(A, \boldsymbol{X})}\middle|\boldsymbol{X}\right\} = 2\delta(\boldsymbol{X}). \tag{1.3}$$

The no unmeasured confounders assumption ensures that a sufficient set of predictors has been included so that treatment assignment depends only on measured covariates. Due to SUTVA and the no unmeasured confounders assumptions, $f^{opt}(\boldsymbol{X})$ in (1.2) is causally interpreted as the conditional average treatment effect (CATE), as outlined in Pu and Zhang (2021) and in (2.3) of Jacob (2021):

$$\begin{aligned} f^{opt}(\boldsymbol{X}) &= E(R|\boldsymbol{X}, A=1) - E(R|\boldsymbol{X}, A=-1) \\ &= E\left\{R^*(1)|\boldsymbol{X}, A=1\right\} - E\left\{R^*(-1)|\boldsymbol{X}, A=-1\right\} \textbf{ (SUTVA)} \\ &= E\left\{R^*(1)|\boldsymbol{X}\right\} - E\left\{R^*(-1)|\boldsymbol{X}\right\} \textbf{ (NUC)} \\ &= E\left\{R^*(1) - R^*(-1)|\boldsymbol{X}\right\}. \end{aligned}$$

Positivity is necessary for d^{opt} to be optimal, as it ensures that every covariate-treatment combination has positive probability of being observed and that propensity score estimation does not produce extreme weights (Kosorok and Moodie (2016); Schulte *et al.* (2014)). Since the CATE is a contrast between the effects of two treatments (-1 and 1), the link between $f^{opt}(\boldsymbol{X})$ and $\delta(\boldsymbol{X})$ is intuitive because $\delta(\boldsymbol{X})(1) - \delta(\boldsymbol{X})(-1) = 2\delta(\boldsymbol{X})$, so estimating one is equivalent to estimating the other.

Tian *et al.* (2014) made the connection between the optimal ITR in (1.2) and formulation of the optimal decision function in (1.3), which forms the basis of D-Learning, as $f^{opt}(\boldsymbol{X})$ can now be directly learned through a regression method of choice. Lemma 1 of Qi and Liu (2018) shows that an estimation framework for $f^{opt}(\boldsymbol{X})$ in (1.3) is:

$$f^{opt}(\boldsymbol{X}) \in \underset{f}{\operatorname{argmin}}\, E\left[\frac{\{2RA - f(\boldsymbol{X})\}^2}{\pi(A, \boldsymbol{X})}\right]. \tag{1.4}$$

Considering the class $\mathcal{F} = \{f(\boldsymbol{X}) = \boldsymbol{X}^\top\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^p\}$ to approximate $f^{opt}(\boldsymbol{X})$, the estimation problem can be solved with ordinary least squares (OLS) with or without regularization.

1.2.1.2 AD-Learning

Qi and Liu (2018) proposed pairwise D-Learning for the case where $A \in \{1, 2, ..., K\}$. This was improved with AD-Learning (Qi *et al.* (2020)), which uses the angle-based approach of Zhang and Liu (2014) to project treatment A into K simplex vertices defined in $\mathbb{R}^{K-1}$. Let treatment A be represented by the vector $\boldsymbol{u}_A \in \mathbb{R}^{K-1}$:

$$\boldsymbol{u}_A = \begin{cases} \frac{1}{\sqrt{K-1}}\mathbf{1}_{K-1}, & A = 1 \\ \sqrt{\frac{K}{K-1}}\boldsymbol{e}_{A-1} - \frac{1+\sqrt{K}}{\sqrt{(K-1)^3}}\mathbf{1}_{K-1}, & 2 \leqslant A \leqslant K. \end{cases} \tag{1.5}$$

Here, $\boldsymbol{e}_i$ is a $(K-1)$-dimensional vector of zeroes with 1 in the i^{th} location. Let the random vector $\boldsymbol{U}$ be such that $\boldsymbol{U} \mid (\boldsymbol{X}, A) \overset{a.s.}{=} \boldsymbol{u}_A$. The working model is:

$$R = \mu(\boldsymbol{X}) + \sum_{k=1}^{K} \delta_k(\boldsymbol{X})\mathbb{1}(A = k) + \eta, \tag{1.6}$$

where $\mu(\boldsymbol{X})$ is the main effect, $\delta_k(\boldsymbol{X})$ is the interaction effect between the k^{th} treatment and covariates, and η is the mean-zero random error. The contrast $\delta_k(\boldsymbol{X}) - \delta_j(\boldsymbol{X})$ can be causally interpreted as the CATE between treatments k and j. The optimal ITR can then be expressed as:

$$\begin{aligned} d^{opt}(\boldsymbol{X}) &= \underset{k\in\{1,...,K\}}{\operatorname{argmax}} E(R|\boldsymbol{X} = \boldsymbol{x}, A = k) \\ &= \underset{k\in\{1,...,K\}}{\operatorname{argmax}} \boldsymbol{u}_k^\top E\left\{ \frac{R\boldsymbol{U}}{\pi(A, \boldsymbol{X})} \middle| \boldsymbol{X} \right\} \\ &= \underset{k\in\{1,...,K\}}{\operatorname{argmax}} \boldsymbol{u}_k^\top \sum_{k=1}^{K} \delta_k(\boldsymbol{X})\boldsymbol{u}_k \\ &:= \underset{k\in\{1,...,K\}}{\operatorname{argmax}} \boldsymbol{u}_k^\top f^{opt}(\boldsymbol{X}) \\ &= \underset{k\in\{1,...,K\}}{\operatorname{argmax}} \delta_k(\boldsymbol{X}), \end{aligned} \tag{1.7}$$

where $f^{opt}(\boldsymbol{X})\colon \mathbb{R}^{p+1} \mapsto \mathbb{R}^{K-1}$. As shown in Lemma 1 of Qi *et al.* (2020), for independent responses, this leads to an estimation problem for $f^{opt}(\boldsymbol{X})$ in (1.7) of the form:

$$f^{opt}(\boldsymbol{X}) \in \underset{f\in\mathbb{R}^{K-1}}{\operatorname{argmin}} E\left[\frac{\{KR\boldsymbol{U}-f(\boldsymbol{X})\}^{\top}\{KR\boldsymbol{U}-f(\boldsymbol{X})\}}{\pi(A,\boldsymbol{X})}\right],$$

which, in Lemma 2, is shown to be equivalent to the following estimation framework:

$$f^{opt}(\boldsymbol{X}) \in \underset{f\in\mathbb{R}^{K-1}}{\operatorname{argmin}}\ E\left[\frac{1}{\pi(A,\boldsymbol{X})}\left\{\frac{K}{K-1}R-\boldsymbol{U}^{\top}f(\boldsymbol{X})\right\}^{2}\right]. \tag{1.8}$$

1.2.1.3 RD-Learning

Meng and Qiao (2022) develop RD-Learning, which replaces the outcome r_i in D-Learning with the residual $r_i - \widehat{m}(\boldsymbol{x}_i)$, where $\widehat{m}(\boldsymbol{X})$ is an estimator for the main effect, $m(\boldsymbol{X})$ (similarly to Zhou *et al.* (2017)). This reduces the variance and leads to doubly robust estimation of the treatment effect in the sense that consistency is guaranteed if either the main effect model or propensity score model is correctly specified.

1.2.2 SD-Learning

For the binary treatment RCT setting, $\pi(A,\boldsymbol{X})$ is known, and assuming (1.1), the D-Learning estimation problem in (1.4) induces the following working model:

$$2RA = f(\boldsymbol{X}) + \epsilon,$$

showing that the estimation of $f(\boldsymbol{X})$ can proceed without needing to model $m(\boldsymbol{X})$. Assume that $E(\epsilon|A,\boldsymbol{X}) = 0$ and $var(\epsilon|A,\boldsymbol{X}) = \sigma_0^2(A,\boldsymbol{X})$. Note that this error term is very general; it can be an arbitrary function of the treatment and covariates. In this case, the D-Learning estimator of the treatment effect is consistent, but due to the potential heteroscedasticity, it may lack efficiency as it gives each observation equal weight. Considering decision functions in $\mathcal{F}$, we propose a

modified D-Learning objective function based on reweighting to gain efficiency:

$$\widehat{\boldsymbol{\beta}}_n^S = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, \mathbb{P}_n \left\{ \frac{(2RA - \boldsymbol{X}^\top\boldsymbol{\beta})^2}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})} \right\}, \tag{1.9}$$

where $w(A, \boldsymbol{X})$ is an arbitrary set of weights which need to be specified and/or estimated. The following assumptions establish the basic conditions needed to find optimal weights. For all $A \in \mathcal{A}$ and $\boldsymbol{X} \in \mathcal{X}$ almost surely:

Assumption 1.1. *$E(\boldsymbol{X}\boldsymbol{X}^\top)$ is full rank and $E\,\|\boldsymbol{X}\|^2 < \infty$.*

Assumption 1.2. *$0 < c_1 \leqslant \sigma_0^2(A, \boldsymbol{X}) \leqslant c_2 < \infty$ almost surely.*

Assumption 1.1 imposes a finite second moment restriction and assumes nonsingularity of the covariates. Assumption 1.2 ensures that the true residual variance function is finite and nonzero (bounded above and below).

Proposition 1.1. *Under Assumptions 1.1 and 1.2, setting $w(A, \boldsymbol{X}) = \frac{\sigma_0^2(A,\boldsymbol{X})}{\pi(A,\boldsymbol{X})}$ minimizes the estimator of the asymptotic variance of (1.9).*

Proposition 1.1 offers a simple way to perform the reweighting. Let $\widehat{\boldsymbol{\beta}}_n^D$ be a consistent estimate of $\boldsymbol{\beta}_0$, which can be obtained by fitting a traditional D-Learning model (Qi and Liu (2018)). Since $\epsilon = 2AR - \boldsymbol{X}^\top\boldsymbol{\beta}$, $\sigma_0^2(A, \boldsymbol{X})$ can be estimated by regressing $\left(2AR - \boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_n^D\right)^2$ on $(A, \boldsymbol{X})$ through a parametric or nonparametric model. The resulting prediction function can be denoted as $\widehat{\sigma}_n^2(A, \boldsymbol{X})$. This procedure breaks down into the following implementation steps:

1. Obtain a D-Learning estimator:

$$\widehat{\boldsymbol{\beta}}_n^D = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n} \frac{\left(2r_i a_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2}{\pi(a_i, \boldsymbol{x}_i)}.$$

2. Regress the squared residuals from Step 1, $\left(2AR - \boldsymbol{X}^\top\widehat{\boldsymbol{\beta}}_n^D\right)^2$, on the treatment and covariates, $(A, \boldsymbol{X})$, to obtain prediction function $\widehat{\sigma}_n^2(A, \boldsymbol{X})$.

3. Find $\widehat{\boldsymbol{\beta}}_n^S$ using:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n^S &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \frac{\pi\left(a_i, \boldsymbol{x}_i\right)\left(2r_i a_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2}{\widehat{\sigma}_n^2\left(a_i, \boldsymbol{x}_i\right)\pi\left(a_i, \boldsymbol{x}_i\right)} \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \frac{\left(2r_i a_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2}{\widehat{\sigma}_n^2\left(a_i, \boldsymbol{x}_i\right)}.\end{aligned} \tag{1.10}$$

Thus, the SD-Learning estimator for the binary treatment case is formulated as a least squares problem, reweighted by the inverse of the estimated residual variance. A procedure for obtaining an improved estimate of the parameters has therefore been provided for binary D-Learning in the case of heteroscedasticity. The same reweighting framework can be used in the case of RD-Learning, where the only differences are that a model for the main effect, $m(\boldsymbol{X})$, must be estimated, and the augmented outcome becomes $R^* = R - \widehat{m}(\boldsymbol{X})$.

1.2.3 Extension of SD-Learning to Multiple Treatments

Now, we expand the treatment space to K treatments, indexed as $A \in \{1, 2, ..., K\}$. Let $\boldsymbol{u}_A \in \mathbb{R}^{K-1}$ be defined as per (1.5). Assuming (1.6), the AD-Learning estimation problem in (1.8) induces the following working model:

$$\frac{K}{K-1}R = \boldsymbol{U}^\top f(\boldsymbol{X}) + \epsilon.$$

We use this working model under the same scenario as Section 1.2.2: $E(\epsilon|A, \boldsymbol{X}) = 0$ and $var(\epsilon|A, \boldsymbol{X}) = \sigma_0^2(A, \boldsymbol{X})$. The class of linear decision functions is defined as $\mathcal{F} = \{f(\boldsymbol{X}) = \boldsymbol{B}^\top\boldsymbol{X} : \boldsymbol{B} \in \mathbb{R}^{p\times(K-1)}\}$.

Adding an arbitrary weight term, $w(A, \boldsymbol{X})$, in the denominator, similarly to the binary case, we propose the SD-Learning objective function as a modified version of AD-Learning:

$$\widehat{\boldsymbol{B}}_n^S = \underset{B\in\mathbb{R}^{p\times(K-1)}}{\operatorname{argmin}} \mathbb{P}_n\left\{\frac{1}{w(A, \boldsymbol{X})\pi(A, \boldsymbol{X})} \times \left(\frac{K}{K-1}R - \boldsymbol{U}^\top\boldsymbol{B}^\top\boldsymbol{X}\right)^2\right\}. \tag{1.11}$$

Again, $w(A, \boldsymbol{X})$ must be optimally chosen. We can reframe this objective function so that it is easier to optimize. Using the identity Vec $(\boldsymbol{ABC}) = \left(\boldsymbol{C}^\top \otimes \boldsymbol{A}\right)$ Vec$(\boldsymbol{B})$, where $\otimes$ denotes the Kronecker product, and the fact that $\boldsymbol{U}^\top \boldsymbol{B}^\top \boldsymbol{X}$ is a scalar:

$$\begin{aligned}\boldsymbol{U}^\top \boldsymbol{B}^\top \boldsymbol{X} &= \text{Vec}\left(\boldsymbol{U}^\top \boldsymbol{B}^\top \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}^\top \otimes \boldsymbol{U}^\top\right) \text{Vec}\left(\boldsymbol{B}^\top\right) \\ &= \boldsymbol{X}_*^\top \boldsymbol{B}_*,\end{aligned}$$

where $\boldsymbol{X}_* = \left(\boldsymbol{X}^\top \otimes \boldsymbol{U}^\top\right)^\top$ and $\boldsymbol{B}_* = \text{Vec}\left(\boldsymbol{B}^\top\right)$. This allows for an equivalent reformulation of the SD-Learning estimation problem:

$$\widehat{\boldsymbol{B}}_n^S = \underset{B \in \mathbb{R}^{p\times(K-1)}}{\text{argmin}} \, \mathbb{P}_n \left\{ \frac{1}{w(A, \boldsymbol{X})\pi(A, \boldsymbol{X})} \times \left(\frac{K}{K-1} R - \boldsymbol{X}_*^\top \boldsymbol{B}_* \right)^2 \right\}. \tag{1.12}$$

Note that $\boldsymbol{B}_*$ and $\boldsymbol{X}_*$ in (1.12) are vectors in $\mathbb{R}^{p(K-1)}$, unlike $\boldsymbol{B}$ and $\boldsymbol{X}$ in (1.11), which are a matrix in $\mathbb{R}^{p\times(K-1)}$ and vector in $\mathbb{R}^p$, respectively. $w(A, \boldsymbol{X})$ can now be optimized in a fashion akin to the binary SD-Learning case:

Proposition 1.2. *Under Assumption 1.1 for $\boldsymbol{X}_*$ instead of $\boldsymbol{X}$ and Assumption 1.2, setting $w(A, \boldsymbol{X}) = \frac{\sigma_0^2(A,\boldsymbol{X})}{\pi(A,\boldsymbol{X})}$ minimizes the estimator of the asymptotic variance of (1.12).*

Note that these are the same weights as found in Proposition 1.1 for the binary treatment case. Having found the optimal weights, $w(A, \boldsymbol{X})$, we switch back to non-vectorized notation (using $\boldsymbol{U}^\top \boldsymbol{B}^\top \boldsymbol{X}$ instead of the equivalent $\boldsymbol{X}_*^\top \boldsymbol{B}_*$). $\sigma_0^2(A, \boldsymbol{X})$ can be estimated by regressing $\left\{ \frac{K}{K-1} R - \boldsymbol{u}^\top \left(\widehat{\boldsymbol{B}}_n^{AD}\right)^\top \boldsymbol{X} \right\}^2$ on $(A, \boldsymbol{X})$ through a parametric or nonparametric model, with the estimate denoted by $\widehat{\sigma}_n^2(A, \boldsymbol{X})$. Let $\widehat{\boldsymbol{B}}_n^{AD}$ represent a consistent estimate of $\boldsymbol{B}_0$, obtained via an AD-Learning model (Qi *et al.* (2020)). The implementation of this procedure is as follows:

1. Obtain an AD-Learning estimator:

$$\widehat{\boldsymbol{B}}_n^{AD} = \underset{\boldsymbol{B}\in\mathbb{R}^{p\times(K-1)}}{\text{argmin}} \, \frac{1}{n} \sum_{i=1}^n \frac{1}{\pi(a_i, \boldsymbol{x}_i)} \left(\frac{K}{K-1} r_i - \boldsymbol{u}_{a_i}^\top \boldsymbol{B}^\top \boldsymbol{x}_i \right)^2.$$

2. Regress the squared residuals from Step 1, $\left\{ \frac{K}{K-1} R - \boldsymbol{u}^\top \left(\widehat{\boldsymbol{B}}_n^{AD} \right)^\top \boldsymbol{X} \right\}^2$, on $(A, \boldsymbol{X})$, to obtain prediction function $\widehat{\sigma}_n^2(A, \boldsymbol{X})$.

3. Find $\widehat{\boldsymbol{B}}_n^S$ using:

$$\widehat{\boldsymbol{B}}_n^S = \underset{\boldsymbol{B} \in \mathbb{R}^{p \times (K-1)}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\widehat{\sigma}_n^2(a_i, \boldsymbol{x}_i)} \left(\frac{K}{K-1} r_i - \boldsymbol{u}_{a_i}^\top \boldsymbol{B}^\top \boldsymbol{x}_i \right)^2. \tag{1.13}$$

Thus, SD-Learning in the multi-arm scenario remains a least squares problem, and under non-homogeneous error structures, provides an increased-efficiency estimation approach through the angle-based framework (refer to Theoretical Results in Section 1.3).

As per covariate dimensionality and sparsity assumptions, the OLS steps of the implementation (Steps 1 and 3) in the binary or multi-arm case can be replaced with LASSO, Ridge, Elastic Net, or other regularized least squares techniques. Detailed proofs of Propositions 1.1 and 1.2 can be found in Appendix A.1, and the extension of SD-Learning to observational data can be found in Appendix A.3.

We note that Liang and Yu (2020) use the same working model in developing a semiparametric efficiency framework for a large class of estimators ($f(\boldsymbol{X}) = g(\boldsymbol{\beta}^\top \boldsymbol{X})$, where g is an arbitrary function), leading to an efficient score function which also includes inverse variance estimates. However, due to the many conditional expectations involved in the score function, the inverse variance terms are difficult to estimate directly. As a result of this, their actual estimation procedure ignores the variance terms and reduces to D-Learning of Tian *et al.* (2014) if a single-index model is used and $E(\epsilon|\boldsymbol{X}) = 0$ is assumed. In this case, if a linear decision function is assumed, the leading variance term can be added back and the resulting efficient score function can be solved with the SD-Learning algorithm. Thus, SD-Learning enhances the estimation procedure of Liang and Yu (2020) in the linear decision function case. Additionally, SD-Learning allows the error term to depend on A as well as $\boldsymbol{X}$, which is often encountered in practice.

Similarly, Mo and Liu (2021) develop Efficient Learning (E-Learning), a semiparametric approach incorporating an inverse variance term, but this method necessitates the specification

of a main effect model. Moreover, multi-arm treatment estimation in E-Learning no longer maintains the least squares framework, hence requiring an accelerated proximal gradient method to solve (Nesterov (2013)). Compared to both methods, we specialize to linear decision rules, and within this class, achieve 1) the optimal estimator and 2) an estimation procedure which reaches a closed-form solution instead of requiring optimization techniques.

1.2.4 Residual Model Fitting

The residual modeling step in the implementation of SD-Learning is important, with various parametric and nonparametric options. We propose LASSO, random forest, and/or tree-based XGBoost for residual modeling in order to include a diversity of approaches through parametric assumptions, bagging, and/or boosting, respectively. Additionally, all three methods were chosen for their speed, LASSO and XGBoost for their ability to handle sparsity (Zhang and Huang (2008), Fauzan and Murfi (2018), Chen and Guestrin (2016)), and in the case of random forests and XGBoost, flexibility. They also have a relatively low number of hyperparameters to tune, making their implementation easier. A SuperLearner algorithm can also be used, which combines candidate parametric and nonparametric methods to find an optimal combination which minimizes cross-validated risk. The SuperLearner has been shown to perform asymptotically as well as or better than any of the constituent candidate learners (van der Laan *et al.* (2007)).

Instead of picking one of the above residual modeling algorithms directly, we propose to first compare them using an internal cross-validation step. After squared residuals from the initial D-Learning fit, $Z = \left(2AR - \boldsymbol{X}'\widehat{\boldsymbol{\beta}}_n^D\right)^2$, are obtained, each algorithm is fit by regressing Z against treatment and covariates, $(A, \boldsymbol{X})$. After the hyperparameters for each algorithm are tuned, they can be compared using K-fold cross-validation with mean squared error (MSE) as the evaluation metric. This results in finding the best residual modeling method. We describe the algorithm in detail in Table 1.1.

Table 1.1: Detailed overview of the SD- and SABD-Learning algorithms.

(1)	**Initial Estimate:** Obtain a consistent estimate, $\widehat{\boldsymbol{\beta}}_n^D$, of the parameters of the decision function through D-Learning in the binary treatment case or AD-Learning in the multi-arm treatment case (unweighted).
(2)	**Hyperparameter Tuning:** Obtain $Z = \left(2AR - \boldsymbol{X}'\widehat{\boldsymbol{\beta}}_n^D\right)^2$ as squared residuals from Step (1). Find optimal LASSO (L), random forest (RF), XGBoost (XG), and/or SuperLearner (SL) parameters for predicting Z from treatment and covariates, $(A, \boldsymbol{X})$, resulting in candidate prediction functions $\widehat{\sigma}_L^2(A, \boldsymbol{X})$, $\widehat{\sigma}_{RF}^2(A, \boldsymbol{X})$, $\widehat{\sigma}_{XG}^2(A, \boldsymbol{X})$, and $\widehat{\sigma}_{SL}^2(A, \boldsymbol{X})$.
(3)	**Internal K-Fold CV:** Randomly partition the original sample into K equal-sized (or nearly equal) training folds and let z_{ij}, a_{ij}, and r_{ij} represent the squared residual, treatment, and outcome, respectively, corresponding to the i^{th} observation in the j^{th} testing fold (of size n_j), where $i \in \{1, ..., n_j\}$ and $j \in \{1, ..., K\}$. For the method $m \in \{L, RF, XG, SL\}$, the average test set error can be determined as $MSE_m = \frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{n_j}\sum_{i=1}^{n_j}\left\{z_{ij} - \widehat{\sigma}_{mj}^2(a_{ij}, \boldsymbol{x}_{ij})\right\}^2\right]$, where $\widehat{\sigma}_{mj}^2(A, \boldsymbol{X})$ represents the prediction function resulting from method m when trained on training data from fold j. Pick the method with the lowest average error, $m_{opt} = \underset{m}{\operatorname{argmin}}\ \{MSE_m\}$.
(4)	**Obtaining Weights:** Using m_{opt} from Step (3), obtain the predicted squared residual for each of n observations, $\widehat{\sigma}_{m_{opt}}^2(A, \boldsymbol{X})$.
(5)	**Reweighted Estimate:** Using $\widehat{\sigma}_{m_{opt}}^2(A, \boldsymbol{X})$ from Step (4) as weights, obtain stabilized parameter estimates through SD-Learning as in (1.10) or SABD-Learning as in (1.13).

1.3 Theoretical Results

In this section, we establish the theoretical properties of SD-Learning for settings with fixed dimension, p. We establish consistency and asymptotic normality of the SD-Learning estimator, along with bounds for the empirical value function. Detailed proofs for all theorems and remarks can be found in Appendix A.2. We state two additional assumptions below, and will delineate which assumptions are needed for each theorem.

Assumption 1.3. *$\widehat{\sigma}_n^2(A, \boldsymbol{X})$ is uniformly consistent for $\sigma_0^2(A, \boldsymbol{X})$. In other words, $||\widehat{\sigma}_n^2(A, \boldsymbol{X}) - \sigma_0^2(A, \boldsymbol{X})||_{\infty,(\mathcal{A},\mathcal{X})} \xrightarrow{P} 0$, where $|| \cdot ||_{\infty,(\mathcal{A},\mathcal{X})}$ represents the uniform norm over $(\mathcal{A}, \mathcal{X})$, and $(\mathcal{A}, \mathcal{X})$ is in a bounded set. In the case of observational data, we also require $||\widehat{\pi}_n(A, \boldsymbol{X}) - \pi_0(A, \boldsymbol{X})||_{\infty,(\mathcal{A},\mathcal{X})} \xrightarrow{P} 0$.*

Assumption 1.4. *$\forall\, \gamma > 0, \exists\, \mathcal{G}$ which is P-Donsker such that $\Pr\left\{\frac{\boldsymbol{X}\epsilon}{\widehat{\sigma}_n^2(A,\boldsymbol{X})} \in \mathcal{G}\right\} > 1 - \gamma,\ \forall\, n$ large enough.*

In Remark 1.1, below, we propose two estimation methods for $\widehat{\sigma}_n^2(A, \boldsymbol{X})$ and provide justification that they satisfy Assumptions 1.3 and 1.4. Then, in Theorem 1.1, we establish consistency of the SD-Learning estimator in the binary treatment setting, and with consistency established, asymptotic normality of the estimator is shown in Theorem 1.2.

Remark 1.1. *Estimating $\sigma_0^2(A, \boldsymbol{X})$ by regressing squared residuals, Z, against treatment and covariates, $(A, \boldsymbol{X})$, with (1) Linear regression with arbitrary features and (2) Random forests satisfies Assumptions 1.3 and 1.4.*

Theorem 1.1. *If Assumptions 1.1-1.3 are met, $\widehat{\boldsymbol{\beta}}_n^S \xrightarrow{P} \boldsymbol{\beta}_0$.*

Theorem 1.2. *Let $Y^* = 2AY$ and $E(Y^*|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_0$. Denote $\boldsymbol{U}_0 = E\left\{\dfrac{\boldsymbol{X}\boldsymbol{X}^\top}{\sigma_0^2(A, \boldsymbol{X})}\right\}$. If Assumption 1.4 is additionally met, $\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_n^S - \boldsymbol{\beta}_0\big)$ is asymptotically normal with variance $\boldsymbol{U}_0^{-1}$.*

$\widehat{\boldsymbol{\beta}}_n^S$ achieves the lower bound of the asymptotic variance shown in Proposition 1.1, and is thus the optimal estimator for $\boldsymbol{\beta}_0$ among the weighted choices of (1.9). Thus we have achieved consistency, convergence, and asymptotic efficiency.

Since Section 1.2.3 unifies the framework between SD-Learning in binary vs. multi-arm treatment scenarios, the extension of Theorems 1.1 and 1.2 to multi-arm treatment are natural. This development is outlined in Theorem 1.3:

Theorem 1.3. *Let* $\boldsymbol{X}_* = \left(\boldsymbol{X}^\top \otimes \boldsymbol{U}^\top\right)^\top$, $\boldsymbol{B}_* = Vec\left(\boldsymbol{B}^\top\right)$, $Y^* = \frac{K}{K-1}R$ *where* $A \in \{1, 2, ..., K\}$, *and* $E\left(Y^* | \boldsymbol{X}_*\right) = \boldsymbol{X}_*^\top \boldsymbol{B}_*$. *Denote* $\boldsymbol{U}_0 = E\left\{\dfrac{\boldsymbol{X}_* \boldsymbol{X}_*^\top}{\sigma_0^2\left(A, \boldsymbol{X}\right)}\right\}$. *Under Assumption 1.1 for* $\boldsymbol{X}_*$ *instead of* $\boldsymbol{X}$ *and Assumptions 1.2-1.3,* $\widehat{\boldsymbol{B}}_*^S \xrightarrow{P} \boldsymbol{B}_*$. *Moreover, if Assumption 1.4 is additionally met,* $\sqrt{n}\left(\widehat{\boldsymbol{B}}_*^S - \boldsymbol{B}_*\right)$ *is asymptotically normally distributed with variance* $\boldsymbol{U}_0^{-1}$.

Thus, all results established for the binary treatment case extend to the multi-arm setting. Combining the asymptotic normality result of Theorem 1.3 with Theorem 1 of Qi *et al.* (2020) which shows that:

$$V(d^{opt}) - V(\widehat{d}_n) \leqslant \frac{2K(K-1)}{1 - C(K)} \left(E \left\| f^{opt} - \widehat{f}_n \right\|_2^2\right)^{1/2},$$

where constant $C(K)$ only depends on K, $\sqrt{n}$-convergence of $V(\widehat{d}_n)$ to $V(d^{opt})$ is established.

1.4 Numerical Results: Simulation Studies

We perform head-to-head comparisons between SD-Learning and the D-Learning family of methods (D-Learning, AD-Learning, and RD-Learning), while noting that D-Learning has been extensively compared to other precision medicine methods in existing literature (Qi and Liu (2018); Wang *et al.* (2020); Qi *et al.* (2020); Meng and Qiao (2021); Mo and Liu (2021)). In all simulations, LASSO is used to obtain estimates of the decision function parameters (Steps 1 and 5 of Table 1.1), and we use internal cross-validation to pick between LASSO, random forest, and XGBoost at the intermediate residual modeling steps (Steps 2 and 3 of Table 1.1). Methods are evaluated based on three criteria:

1. Average Prediction Error (APE): Coefficient accuracy based on MSE of true vs. predicted decision functions. For binary simulation settings, $APE = n^{-1} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_0 - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_n\right)^2$; for

multi-arm settings, $APE = n^{-1}\sum_{i=1}^{n}\left\{f^{opt}(\boldsymbol{x}_i) - \widehat{f}(\boldsymbol{x}_i)\right\}^2 = n^{-1}\sum_{i=1}^{n}\left\{\sum_{k=1}^{K}\delta_k(\boldsymbol{x}_i)\boldsymbol{u}_k - \sum_{k=1}^{K}\widehat{\delta}_k(\boldsymbol{x}_i)\boldsymbol{u}_k\right\}^2$.

2. Misclassification Rate: % incorrect treatment assignment.

3. Empirical Value: $\widehat{V}(d) = \dfrac{\mathbb{P}_n\left[R\cdot\mathbb{1}\{A = d(\boldsymbol{X})\}/\pi(A,\boldsymbol{X})\right]}{\mathbb{P}_n\left[\mathbb{1}\{A = d(\boldsymbol{X})\}/\pi(A,\boldsymbol{X})\right]}$.

Better performance corresponds to lower APE, lower misclassification rate, and higher empirical value. In all simulations, performance based on these criteria is determined on a test data set with 10000 observations. 100 replications of each simulation setting are performed.

We compare all binary methods with $p = \{30, 60, 120\}$ and multi-arm methods with $p = \{20, 40, 60\}$. Simulated observations are independent with continuous covariates generated according to a U[-1,1] distribution. To allow for heteroscedasticity, the outcome is generated according to the working model in (1.1) for binary treatments and (1.6) for multi-arm treatments, but with η generated according to $\sigma_0(\boldsymbol{X}) * Z$, where $Z \sim N(0,1)$ and $\sigma_0(\boldsymbol{X}) > 0$. Here, non-constant $\sigma_0(\boldsymbol{X})$ introduces heteroscedasticity.

1.4.1 Binary Treatment Simulations

We compare SD-Learning and D-Learning with four simulation settings where $n = 200$:

1. $m(\boldsymbol{X}) = 1 + 2X_1 + X_2 + 0.5X_3$; $\delta(\boldsymbol{X}) = 0.5(0.9 - X_1)$;

   $\sigma_0^2(\boldsymbol{X}) = 1$; $\pi(A = 1, \boldsymbol{X}) = 0.5$ (RCT).

2. $m(\boldsymbol{X}) = 1 + 12X_1 + 6X_2 + 3X_3$; $\delta(\boldsymbol{X}) = 4X_1$;

   $\sigma_0^2(\boldsymbol{X}) = 0.25 + (X_2 + 1)^2$; $\pi(A = 1, \boldsymbol{X}) = 0.25 + 0.5 * \mathbb{1}(X_2 > 0)$.

3. $m(\boldsymbol{X}) = 1 + 10X_1 + 10X_2 + 20X_3 + 5X_4$; $\delta(\boldsymbol{X}) = 4(0.3 - X_1 - X_2)$;

   $\sigma_0^2(\boldsymbol{X}) = 1 + 4X_3^2$; $\pi(A = 1, \boldsymbol{X}) = 0.5$ (RCT).

4. $m(\boldsymbol{X}) = 1 + 10X_1 + 6X_1^2 - 6X_2^2 + 10X_3$; $\delta(\boldsymbol{X}) = 2X_2^2 + 1.5X_3 + 3X_4$;

   $\sigma_0^2(\boldsymbol{X}) = .25 + (.3 - X_1 - X_2)^2$; $\pi(A = 1, \boldsymbol{X}) = 0.5$ (RCT).

Scenario (1) is very similar to the third linear decision boundary scenario in Qi and Liu (2018), and has homoscedastic error. In this case, an intermediate residual reweighting step is not necessary since the optimal weights are $w(\boldsymbol{X}) = 1$ by design, which is the default for D-Learning. (2)-(4) introduce heteroscedasticity through the error term. (2) and (3) meet linear decision function assumptions, and in both, the interaction effect has variables in common with the main effect but not with the error function. Specifically, (2) is an observational data setting, where we model the propensity score through random forest for binary classification. Finally, (4) is a nonlinear decision boundary scenario where SD-Learning is misspecified. This scenario will help test the robustness of SD-Learning in situations where the true parameters of the decision function cannot be consistently estimated.

Figure 1.1 shows the APE for all four scenarios, and Table 1.2 contains misclassification rates and the estimated empirical value function on the test dataset. Since Scenario (1) has homoscedastic error, D-Learning performs "optimally" in the sense of correctly specified weights; hence, stabilizing the D-Learning estimates is technically not needed. However, SD-Learning still performs similarly to D-Learning, with comparable APE and only slightly lower misclassification rate and empirical value. For scenarios (2) and (3), SD-Learning outperforms D-Learning because of the heterogeneous error structure. SD-Learning prioritizes observations with smaller expected outcome variance, and therefore estimates parameters more efficiently, as shown by the lower APEs in Figure 1.1. This results in better classification and therefore, higher empirical value. In Scenario (4), although the true decision boundary is nonlinear, SD-Learning's flexible modeling of heteroscedasticity gives it an advantage over D-Learning.

We also perform simulations to show the advantage of stabilizing the estimates of RD-Learning. For these simulations, $n = 100$. For the main effect modeling step, LASSO was used, but nonparametric methods may also be used as per Meng and Qiao (2022). The simulation settings for SRD-Learning vs. RD-Learning are as follows:

5. $m(\boldsymbol{X}) = 1 + 10X_1 + 10X_2 + 20X_3 + 20X_5 + 10X_1X_2$; $\delta(\boldsymbol{X}) = 1.25X_3 + 2.5X_4$; $\sigma_0^2(\boldsymbol{X}) = 1 + 0.1(1 + X_1^2)$; $\pi(A = 1, \boldsymbol{X}) = 0.5$ (RCT).

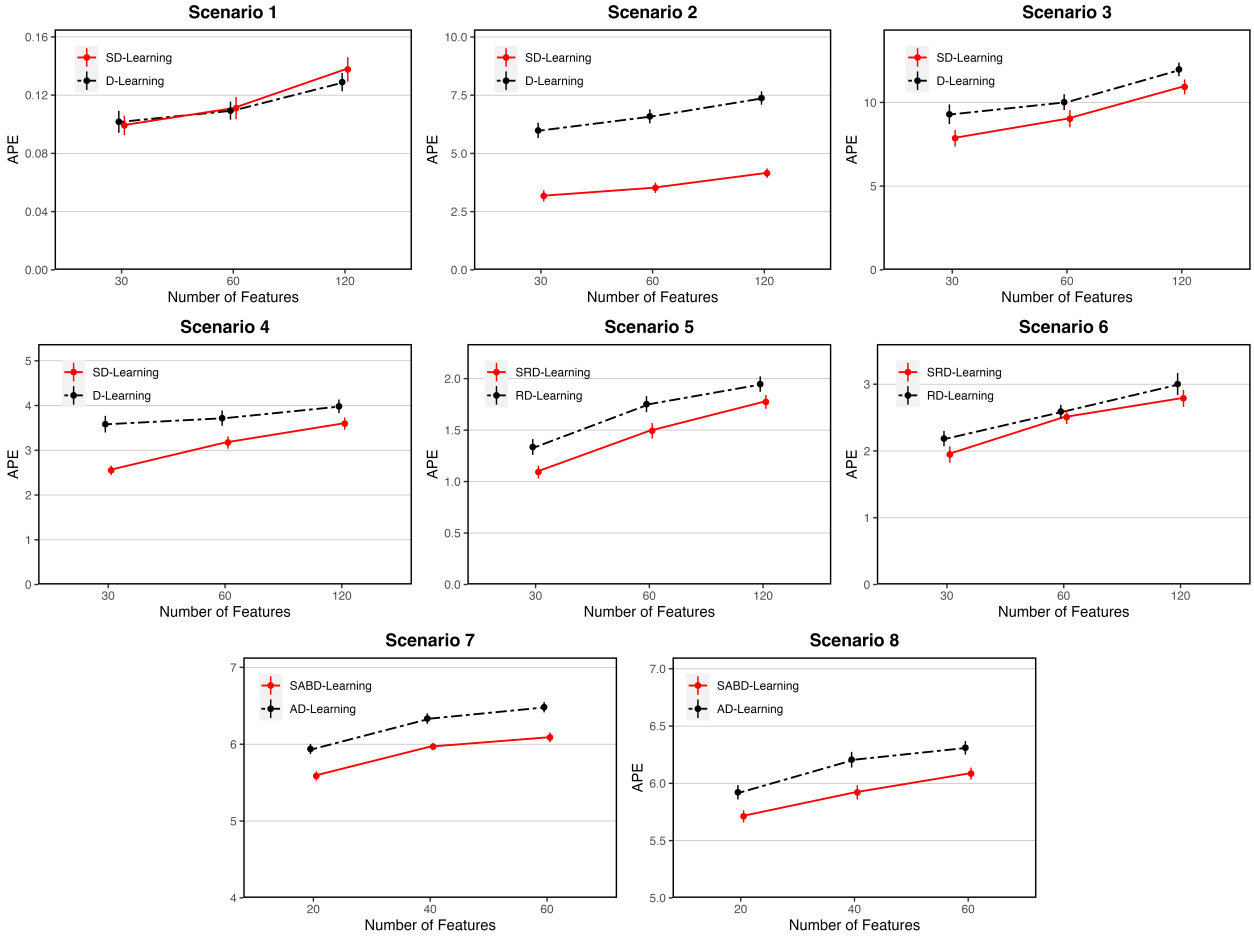

Figure 1.1: Average Prediction Error (APE) results, along with standard error of the mean (SEM) bars, of four binary simulation scenarios comparing D- to SD-Learning ($n = 200$), two binary simulation scenarios comparing RD- to SRD-Learning ($n = 100$), and two multi-arm simulation scenarios comparing AD- to SABD-Learning ($n = 200$). In binary scenarios, p varies from 30 to 120, and in multi-arm scenarios, p varies from 20 to 60. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Table 1.2: Mean empirical value and misclassification rate, along with standard error of the mean (SEM), for four binary D- vs. SD-Learning simulations and two binary RD- vs. SRD-Learning simulations for 30, 60, and 120 covariates. The best-performing method for each category is bolded.

	$p = 30$		$p = 60$		$p = 120$	
	Value	Misclass.	Value	Misclass.	Value	Misclass.
			Scenario 1			
D-Learning	**1.43** (0.01)	**0.07** (0.01)	**1.43** (0.01)	**0.07** (0.01)	**1.44** (0.01)	**0.08** (0.01)
SD-Learning	1.42 (0.01)	0.08 (0.01)	1.42 (0.01)	0.09 (0.01)	1.43 (0.01)	0.09 (0.01)
			Scenario 2			
D-Learning	2.46 (0.01)	0.14 (0.01)	2.27 (0.01)	0.14 (0.01)	2.17 (0.01)	0.15 (0.01)
SD-Learning	**2.48** (0.01)	**0.13** (0.01)	**2.30** (0.01)	**0.13** (0.01)	**2.21** (0.01)	**0.14** (0.01)
			Scenario 3			
D-Learning	2.64 (0.09)	0.29 (0.01)	2.40 (0.09)	0.32 (0.01)	2.16 (0.08)	0.36 (0.01)
SD-Learning	**2.82** (0.08)	**0.26** (0.01)	**2.56** (0.09)	**0.30** (0.01)	**2.21** (0.08)	**0.35** (0.01)
			Scenario 4			
D-Learning	1.81 (0.05)	0.31 (0.01)	1.91 (0.05)	0.32 (0.01)	1.90 (0.04)	0.33 (0.01)
SD-Learning	**2.08** (0.03)	**0.24** (0.01)	**2.02** (0.05)	**0.29** (0.01)	**1.99** (0.04)	**0.31** (0.01)
			Scenario 5			
RD-Learning	1.94 (0.03)	0.20 (0.01)	1.91 (0.05)	0.27 (0.01)	1.43 (0.05)	0.30 (0.01)
SRD-Learning	**1.98** (0.03)	**0.18** (0.01)	**2.02** (0.04)	**0.24** (0.01)	**1.55** (0.05)	**0.27** (0.01)
			Scenario 6			
RD-Learning	6.16 (0.04)	0.21 (0.01)	5.94 (0.04)	0.23 (0.01)	5.65 (0.05)	0.25 (0.02)
SRD-Learning	**6.24** (0.04)	**0.19** (0.01)	**5.97** (0.04)	0.23 (0.01)	**5.70** (0.05)	**0.24** (0.01)

6. $m(\boldsymbol{X}) = 1 + 5\cos^2(X_1) + 10X_1X_2 + 20X_2 + 30X_5$; $\delta(\boldsymbol{X}) = 3X_1 + 2X_2 + 2X_5^2$;

   $\sigma_0^2(\boldsymbol{X}) = 0.5 + 0.5\left(1 - 0.25X_6\right)^3$; $\pi(A = 1, \boldsymbol{X}) = 0.5$ (RCT).

Both scenarios have heterogeneous error. Scenario (5) has a true linear decision function, whereas (6) has a nonlinear decision function along with a nonlinear *cosine* term in the main effect. Figure 1.1 shows the APE for both scenarios, and Table 1.2 contains misclassification rates and the estimated empirical value function on the test dataset. Although RD-Learning is already robust in the sense that the main effect is removed before model fitting, the reweighing step of SRD-Learning adds efficiency in situations with heteroscedasticity - even in the presence of a misspecified decision function and nonlinearity in the main effect.

1.4.2 Multi-Arm Treatment Simulations

We compare SABD-Learning and AD-Learning with two multi-arm treatment scenarios under heteroscedasticity, setting $K = 4$ and $n = 200$ in both:

7. $m(\boldsymbol{X}) = 1 + 2X_1 + 2X_2$;

   $\sigma_0^2(\boldsymbol{X}) = .25 + 0.2(1.5 - X_2)^2$;

$$\begin{cases} \delta(\boldsymbol{X}) = .75 + 1.5X_1 + 1.5X_2 + 1.5X_3 + 1.5X_4;\ \pi(A, \boldsymbol{X}) = 0.25, & A = 1 \\ \delta(\boldsymbol{X}) = .75 + 1.5X_1 - 1.5X_2 - 1.5X_3 + 1.5X_4;\ \pi(A, \boldsymbol{X}) = 0.25, & A = 2 \\ \delta(\boldsymbol{X}) = .75 + 1.5X_1 - 1.5X_2 + 1.5X_3 - 1.5X_4;\ \pi(A, \boldsymbol{X}) = 0.25, & A = 3 \\ \delta(\boldsymbol{X}) = .75 - 1.5X_1 + 1.5X_2 - 1.5X_3 + 1.5X_4;\ \pi(A, \boldsymbol{X}) = 0.25, & A = 4 \end{cases}$$

8. $m(\boldsymbol{X}) = 1 + X_5 + 3X_6 + 2X_1X_2$;

   $\sigma_0^2(\boldsymbol{X}) = .25 + 2X_2 * \mathbb{1}(X_2 > 0) + X_3 * \mathbb{1}(X_3 > 0, A = 1) + X_4 * \mathbb{1}(X_4 > 0, A = 2)$;

$$\begin{cases} \delta(\boldsymbol{X}) = 0.5 + 2X_1 + X_2 + X_3;\ \pi(A, \boldsymbol{X}) = 0.25{*}\mathbb{1}(X_1 < 0) + 0.4{*}\mathbb{1}(X_1 > 0), & A = 1 \\ \delta(\boldsymbol{X}) = 1 + X_1 - X_2 - X_3;\ \pi(A, \boldsymbol{X}) = 0.25{*}\mathbb{1}(X_1 < 0) + 0.2{*}\mathbb{1}(X_1 > 0), & A = 2 \\ \delta(\boldsymbol{X}) = 1.5 + 3X_1 - X_2 + X_3;\ \pi(A, \boldsymbol{X}) = 0.25{*}\mathbb{1}(X_1 < 0) + 0.2{*}\mathbb{1}(X_1 > 0), & A = 3 \\ \delta(\boldsymbol{X}) = 1 - X_1 - X_2 + X_3;\ \pi(A, \boldsymbol{X}) = 0.25{*}\mathbb{1}(X_1 < 0) + 0.2{*}\mathbb{1}(X_1 > 0), & A = 4. \end{cases}$$

Table 1.3: Mean empirical value and misclassification rate, along with standard error of the mean (SEM), for two multi-arm scenarios comparing AD-Learning to SABD-Learning. All simulations are repeated with 20, 40, and 60 covariates. The best-performing method for each category is bolded.

	$p = 20$		$p = 40$		$p = 60$	
	Value	Misclass.	Value	Misclass.	Value	Misclass.
Scenario 7						
AD-Learning	3.39 (0.02)	0.34 (0.01)	3.23 (0.03)	0.39 (0.01)	3.19 (0.03)	0.42 (0.01)
SABD-Learning	**3.49** (0.01)	**0.30** (0.01)	**3.39** (0.02)	**0.33** (0.01)	**3.39** (0.02)	**0.35** (0.01)
Scenario 8						
AD-Learning	2.71 (0.03)	0.52 (0.01)	2.68 (0.03)	0.56 (0.01)	2.67 (0.03)	0.58 (0.01)
SABD-Learning	**2.86** (0.02)	**0.47** (0.01)	**2.86** (0.02)	**0.50** (0.01)	**2.88** (0.02)	**0.51** (0.01)

Both scenarios meet linear assumptions for the treatment-covariate interaction effects. In Scenario (7), heterogeneous error is a quadratic function of the covariate X_2, but in Scenario (8), it is a spline function of X_2, X_3, and X_4 dependent on treatments $A = 1$ and $A = 2$. Scenario (8) is additionally an observational data setting, with the propensity score modeled through random forest for multiclass classification. Figure 1.1 reports APE for both scenarios, and Table 1.3 displays misclassification and value results. Reweighting in the multi-arm scenario under heteroscedasticity also improves efficiency, resulting in lower APEs and misclassification rates, and higher empirical values. SABD-Learning improves the performance of AD-Learning when an error structure can be learned and utilized to obtain decision rules built by favoring observations which are more likely to represent signal than noise.

1.5 Data Analysis: AIDS Clinical Trial

We observe that in the ACTG175 data of 2139 patients, heteroscedasticity is present through covariates and treatments; this analysis is detailed in Appendix A.4. Hence, we apply SD-Learning to the data, which may benefit from a reweighting approach. This double-blinded study evaluated monotherapy vs. combination approaches to increasing CD4 cell counts in HIV-1-infected patients with initial cell counts of 200-500 cells/mm^3 (Hammer, S. M. et al. (1996)). AIDS-defining

illnesses have been shown to decrease as CD4 cell count increases (Mocroft, A. et al. (2013)), so larger increases in CD4 cell count are preferable.

Patients were randomly assigned to one of four daily regimens with equal probability:

1. 600 mg zidovudine (Z)
2. 600 mg zidovudine + 400 mg didanosine (ZD)
3. 600 mg zidovudine + 2.25 mg zalcitabine (ZZ)
4. 400 mg didanosine (D)

The outcome we use for this analysis is the change in CD4 cell count from baseline to 20 weeks, as done in Qi and Liu (2018) and Qi *et al.* (2020). 12 covariates are selected as per Fan *et al.* (2017); five continuous: weight (kg), age (years), Karnofsky score (0-100), baseline CD4 count (cells/mm^3), baseline CD8 count (cells/mm^3); and seven binary: hemophilia (1=yes, 2=no), homosexual activity (1=yes, 0=no), history of intravenous drug use (1=yes, 0=no), race (1=non-white, 0=white), gender (1=male, 0=female), antiretroviral history (1=experienced, 0=naive), and symptomatic status (1=symptomatic, 0=asymptomatic). For all comparisons, LASSO is used to obtain estimates of the decision function parameters. LASSO, random forest, and XGBoost are tuned and used for the residual modeling step of SD- and SABD-Learning, with the optimal method picked by internal cross-validation.

1.5.1 Binary Scenario

We compare the performance of D-Learning and its corresponding stabilized version, SD-Learning, for each pairwise set of treatments from the four choices. We randomly split the data into a training set of n observations, using the rest of the observations as testing data. n is selected to be 100, 200, 400, and 800. For generating empirical value estimates, $\widehat{V}(d)$, we perform Monte Carlo Cross-Validation (repeated random subsampling) with 1000 iterations at each n. The corresponding binary treatment results are shown in Table 1.4.

Table 1.4: Empirical value estimates for binary AIDS data scenarios comparing performance of D- and SD-Learning on each pairwise set of treatments (Z, ZD, ZZ, D) at varying sample sizes of training data ($n = 100, 200, 400, 800$). At each sample size, results are averaged from 1000 replications, and corresponding standard error of the mean (SEM) is shown. The best-performing method at each level of sample size is bolded. When both methods converge upon recommending a single treatment (in over 99% of patients across all replications), the treatment is specified instead of the (nearly identical) value estimates.

		D-Learning	SD-Learning			D-Learning	SD-Learning
ZD vs. ZZ	$n = 100$	49.12 (0.31)	**49.24** (0.30)	Z vs. ZD	$n = 100$	51.75 (0.21)	**52.18** (0.20)
	$n = 200$	51.76 (0.21)	**52.12** (0.20)		$n = 200$	52.92 (0.15)	**53.43** (0.14)
	$n = 400$	52.74 (0.18)	**53.15** (0.17)		$n = 400$	53.48 (0.16)	**53.90** (0.16)
	$n = 800$	53.16 (0.35)	**53.55** (0.36)		$n = 800$	53.84 (0.35)	**54.30** (0.35)
ZD vs. D	$n = 100$	48.67 (0.29)	**48.92** (0.28)	Z vs. ZZ	$n = 100$	15.62 (0.27)	15.62 (0.27)
	$n = 200$	50.59 (0.23)	**51.02** (0.23)		$n = 200$	17.96 (0.14)	**18.10** (0.15)
	$n = 400$	52.88 (0.19)	**53.28** (0.19)		$n = 400$	*ZZ for over 99% of patients.*	
	$n = 800$	55.99 (0.33)	**56.42** (0.33)		$n = 800$	*ZZ for over 99% of patients.*	
ZZ vs. D	$n = 100$	23.57 (0.13)	**23.64** (0.13)	Z vs. D	$n = 100$	**24.57** (0.19)	24.55 (0.19)
	$n = 200$	**23.89** (0.13)	23.88 (0.14)		$n = 200$	*D for over 99% of patients.*	
	$n = 400$	24.52 (0.15)	**24.57** (0.15)		$n = 400$	*D for over 99% of patients.*	
	$n = 800$	**25.55** (0.25)	25.48 (0.25)		$n = 800$	*D for over 99% of patients.*	

In terms of empirical value, SD-Learning improves upon the performance of D-Learning in most scenarios, especially for comparisons involving treatment ZD. SD- and D-Learning perform approximately equally well for ZZ vs. D, with empirical values differing by less than 0.10 in all cases. For pairwise comparisons of Z vs. ZZ and Z vs. D, both methods eventually converge upon recommending a single treatment to over 99% of patients and therefore have very similar value estimates. Overall, SD-Learning either outperforms D-Learning (ZD vs. ZZ, ZD vs. D, and Z vs. ZD) or performs equally as well (ZZ vs. D, Z vs. ZZ, Z vs. D).

1.5.2 Multi-Arm Treatment Scenario

We now compare the performance of AD- and SABD-Learning at varying sample sizes while considering all four treatments simultaneously. We randomly split the data into training sets of $n = 100, 200, 400, 800$, and 1200, using the rest of the observations as testing data. The procedure is repeated 1000 times for each value of n. All results are shown in Table 1.5.

SABD-Learning has a distinct advantage over AD-Learning at all observed training data sample sizes. The stabilization step weighs patients differentially based on predicted squared

Table 1.5: Empirical value estimates for multi-arm AIDS data scenarios comparing the performance of AD- and SABD-Learning in selecting amongst four treatments simultaneously. Varying sample sizes of the training data were chosen to be $n = 100$, 200, 400, 800, and 1200. At each sample size, results are averaged from 1000 replications, and the corresponding standard error of the mean (SEM) is shown. The best-performing method at each level of sample size is bolded.

	AD-Learning	SABD-Learning
$n = 100$	43.27 (0.43)	**43.98** (0.43)
$n = 200$	47.19 (0.37)	**48.29** (0.35)
$n = 400$	50.57 (0.25)	**51.74** (0.22)
$n = 800$	53.35 (0.18)	**54.25** (0.17)
$n = 1200$	54.27 (0.23)	**55.24** (0.23)

residual from the initial AD-Learning step. This results in more efficiently estimated treatment rules and therefore higher values, even in scenarios with very low sample sizes.

1.6 Discussion

In this article, we propose SD-Learning, which boosts the efficiency of D-Learning in a wide range of scenarios with more general error functions, thus enhancing the utility of D-Learning on datasets which may be encountered in practice. The performance of SD-Learning relies on sufficiently modeling residuals from an initial D-Learning fit, which can be achieved through a variety of parametric or nonparametric methods. When the true residual variance is homogeneous, SD-Learning reduces to D-Learning. Our results suggest that under this homogeneous error, SD-Learning pays a minor efficiency price, but under heterogeneous error, it can offer substantial efficiency gains. Additionally, SD-Learning parameter estimates are asymptotically normal when OLS is used for the estimation steps, allowing post-modeling inference even in the multi-arm treatment scenario.

The implementation benefits of SD-, SRD-, and SABD-Learning lie in the fact that they are straightforward to use and can simply be stacked on top of D-, RD-, and AD-Learning which have been shown to perform well in a multitude of settings. Additionally, our methodology even in the multi-arm treatment setting (SABD-Learning) remains estimable with a least squares framework and closed-form solution, not requiring the use of optimization algorithms.

For practical use, we recommend SRD-Learning, since Meng and Qiao (2022) showed that incorporation of the mean outcome model results in protection against incorrect specification of the propensity score model. This would be especially helpful in the case of observational data. In general, unless there is prior evidence of homoscedasticity, we recommend SD-Learning methods because they involve reweighting, which is a simple add-on that increases the efficiency of already-proven methods of Qi and Liu (2018) and Tian *et al.* (2014). We also suggest considering a variety of nonparametric prediction algorithms for the residual modeling step in order to gain an understanding of the heteroscedasticity structure of the dataset at hand. Random forests appeared to work well under a wide variety of scenarios, with the number of trees as the most important tuning parameter.

Several future extensions of this work are possible. Theoretical results in Section 1.3, including Remark 1.1 for random forests, can be extended to the case where covariate dimension, p, increases with n. Although our method can be used with higher-order polynomial terms and interactions by replacing $\boldsymbol{X}$ with a collection of bases of $\boldsymbol{X}$, a more natural methodological extension would be to allow for SD-Learning to estimate nonlinear decision rules using Reproducing Kernel Hilbert Space (RKHS) techniques (Fan *et al.* (2019)). Additionally, SD-Learning may be broadened to work for binary and survival outcomes under heteroscedasticity. As proposed in this paper, SD-Learning assumes independence (but not identically distributed errors) between observations. It would also be valuable to use SD-Learning in scenarios with correlation between observations. This would be akin to feasible generalized least squares (FGLS) (Olive (2017)) for ITR estimation.

CHAPTER 2: INVERSE REINFORCEMENT LEARNING FOR PHYSICIAN-ASSISTED ESTIMATION OF INDIVIDUALIZED TREATMENT RULES WITH MULTIPLE OUTCOMES OF INTEREST

2.1 Introduction

The majority of precision medicine methods, including those outlined in Section 1.1.1, have been developed with a goal of optimizing one outcome of interest. However, it is often the case that treatment selection involves balancing trade-offs between multiple outcomes of interest. Take, for example, *H. pylori* infection, a common cause of stomach ulcers. Two main treatment strategies include 14-day triple therapy (14T) and dual therapy of omeprazole plus amoxycillin or clarithromycin (OAC). De Boer and Tytgat (1995) present the question, "Should efficacy or side-effect profile determine our choice?", and then outline that 14T provides the best eradication results but comes with severe side effects, whereas OAC causes less patient burden but also lacks the efficacy of 14T. Similar examples are seen in schizophrenia, where antipsychotic medications are more efficacious than other treatments but come with greater side effects (Swartz *et al.* (2008)), and bipolar disorder, where treatment must balance trade-offs between depression and mania (Wu *et al.* (2015)).

2.1.1 Precision Medicine with Multiple Outcomes of Interest

The need for precision medicine is often described to result from treatment effect heterogeneity (Wager and Athey (2018); Kent *et al.* (2020); Fang *et al.* (2022)), where patients exhibit diversity in their responses to treatment. However, in the above examples, when multiple clinical endpoints are of interest, patients may display utility heterogeneity and perceive the importance of individual endpoints differently (Irony (2017)). In such scenarios, individual composite util-

ity (as a weighted combination of the multiple outcomes) varies on the basis of patient-specific factors (e.g. a subgroup that displays increased vulnerability to side effects). Such utility heterogeneity also begets a need for personalization of decision making.

Beginning to formalize the setting with heterogeneous utilities, we note that patient-specific utilities will depend on:

(1) Characteristics of the outcomes (i.e. certain outcomes, overall, are more important than other outcomes)

(2) Characteristics of patients (i.e. some patients may prescribe greater importance to certain outcomes than other patients)

The latter drives the patient-specific tailoring, but both must be accounted for.

There are a number of approaches for constructing composite utilities. Classical approaches include canonical correlation analysis (CCA), which maximizes correlation between a linear combination of outcomes and covariates (Thompson (1984)), and nonlinear extensions (Nandy and Cordes (2003)). A more recent approach related to CCA finds a univariate outcome "most-easily predicted" by the covariates and treats it as an optimal summary outcome (Benkeser *et al.* (2021)). Latent variable analysis has strived to turn multiple outcomes into a lower-dimensional representation through unsupervised grouping, with a-posteriori analysis to attribute meaning to the dimensions (Sobel (1994)). A number of methods can be used to estimate ITRs in multiple outcome situations: Murray *et al.* (2016) converts the outcomes to a pre-specified composite for all patients, Laber *et al.* (2014) recommends sets of treatments at each decision point if different treatments optimize different outcomes, and Chen *et al.* (2021) learn ITRs through finding discrete latent constructs that underlie observed outcomes. Notably, however, none of the aforementioned methods estimate individual patient-specific utility functions. Rather, they assume that a single composite outcome can be estimated to reflect utility across all patients.

2.1.2 Strategies for Observational Data

Most precision medicine strategies are predicated on the assumption that the data at hand reflect mixed treatment assignment (i.e. some patients treated optimally and others treated non-optimally), and that a treatment rule can be derived from the data which is more effective than the rule used to generate the data. Of course, this is most often the case. However, Wallace *et al.* (2018) argue that there are certain observational data situations in which clinicians assign treatments well; that is, they assign treatment optimally. For this situation, they propose to ignore patient outcomes entirely and focus on predicting observed treatment from covariates. This method is called Reward Ignorant Modeling of ITRs. If it is assumed that the data already reflect the optimal treatment rule, it is natural that directly predicting observed treatment is more efficient than using common ITR estimation strategies which are built to determine patients who were treated well and patients who were treated poorly. This regression is independent of the number of observed outcomes, so this strategy works for multiple outcomes in a roundabout way. However, the assumption that motivates this approach is strong, nothing is learned about the patient-specific outcome preferences (heterogeneous utility), and this method is not transferable to the randomized clinical trial (RCT) setting.

Luckett *et al.* (2021) extends the Reward Ignorant Modeling framework by assuming that physicians follow an approximately optimal policy, not an optimal one. Instead of assuming that physicians treat all patients optimally, they estimate the probability that each patient is treated optimally. This basis is used to develop estimators of individual patient utilities, and further, decision rules which optimize patient-specific composite outcomes.

It should be noted that these papers are inherently related to the field of inverse reinforcement learning (IRL), which Ng and Russell (2000) describe as the problem of attempting to extract the reward function (i.e., utility function) given observed, optimal behavior of an agent. IRL can also be considered a branch of imitation learning as it works to learn a policy through examples or reproduce demonstrated behavior (Zhifei and Meng Joo (2012)). Examples of application areas of IRL are self-driving cars, which attempt to learn reward structures from the behavior of drivers

in various situations (Arora and Doshi (2021)), teaching a computer a control policy (Hussein *et al.* (2017)), or even teaching a robot to play table tennis from demonstrations (Muelling *et al.* (2014)).

IRL however, assumes that there is a single environment in which the agent makes decisions. In the context of utility heterogeneity, the utility function varies across patients, and hence the decisions of the expert are not only outcome-specific, but also patient-specific (as defined in Section 2.1.1). Thus, whereas the work of Wallace *et al.* (2018) is pure IRL, the work of Luckett *et al.* (2021), which attempts to understand and recreate expert behavior with multiple patients, should be considered a variant of IRL with multiple environments.

That physicians make optimal or near-optimal decisions with respect to the patient is a strong assumption, but may apply based on prior knowledge of certain clinical settings. However, these methods do not readily extend to RCT settings, in which no judgement was made to inform treatment.

2.1.3 Patient Preference Elicitation

In RCT settings, some form of additional information must be collected in order to learn about patient-specific reward structures amongst multiple outcomes. Butler *et al.* (2018) incorporates patient preference information to augment the ITR estimation framework for two-outcome settings. A questionnaire is first composed with binary-choice questions for patients to answer, assuming that each patient has an underlying preference that parameterizes a weighting between the outcomes. The methodology then utilizes concepts from the field of item response theory (Embretson and Reise (2000)) to implement a latent trait model (Rasch model; Kean *et al.* (2018)) with Q-learning to incorporate these preferences into a treatment plan. Butler (2016) incorporates a nonparametric model to determine the weights of the linear utility function.

Notably, the work of Luckett *et al.* (2021) for observational data and Butler (2016) for patient preferences in RCT settings works for two outcomes of interest. Because both methods are framed with Q-learning for ITR estimation, they can work with > 2 treatments, although these

cases are not covered as examples in the papers. Butler (2016) requires that preference information be collected for each incoming patient. It would be worthwhile to develop methodology that links preference information with covariates such that preferences for new patients can be known directly from the covariates they present with.

2.1.4 Proposed Method

In this article, we propose an ITR estimation framework, augmented by physician input, for settings where multiple outcomes are of interest. Although extensive research has attempted to develop composite outcomes where weights are assigned based on expert (physician) opinion, these methods take a "one utility fits all" approach and do not account for utility function variation across patients. Our methodology involves a physician questionnaire and attempts to find an individualized utility function as a weighted combination of the observed outcomes. Each weighted combination serves as a composite outcome that can be used in the precision medicine pipeline for ITR estimation. A covariate-dependent set of weights is interpreted as a preference describing the relative importance of the different outcomes for a specific patient.

2.2 Physician-assisted ITR Estimation with Heterogeneous Utilities

2.2.1 Setup and Notation

We assume that observed patient data, for patients $i \in \{1, ..., n\}$ with outcomes $l \in \mathcal{L} = \{1, ..., L\}$, constitute independent realizations of the random triplet $\{\boldsymbol{X}, A, \boldsymbol{R}\}_i$, where patient covariates are represented by the $(d+1)$-dimensional vector, $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^{d+1}$, which includes an intercept, A represents one of K possible treatments in $\mathcal{A} = \{1, ..., K\}$, and $\boldsymbol{R} = \{R_1, ..., R_L\}$ is a vector of the L unique outcomes. The probability of receiving treatment a, given covariates $\boldsymbol{x}$, is represented by $\pi(a, \boldsymbol{x}) = \Pr(A = a | \boldsymbol{X} = \boldsymbol{x})$. Outcomes where smaller is better (e.g. side effect burden) can be negated without loss of generality, so it is assumed in the methodology that larger R_l corresponds to better outcome.

Patient Data		
Gender	Female	
Age	74	
Diagnosis Age	55	
PHQ-9 Depression Questionnaire	17 (Moderately Severe Depression)	
Baseline HbA1c (%)	7.2	
Time in Hypoglycemia (%)	4.4	
Diabetic Ketoacidosis Events	2	
C-Peptide Detected?	No	
Preferred Outcome (Please select a box below):		
	Outcome A	**Outcome B**
HbA1c Reduction (%)	1.4	0.8
Hypoglycemia Reduction (% Time)	2.2	1.9
PROMIS Physical Function	54.1	64.7
PROMIS Mental Scale	45.8	56.0
Check one →	Prefer *Outcome A* ☐	Prefer *Outcome B* ☐

Figure 2.1: Example choice scenario from a physician questionnaire. Here, the physician is presented with a diabetic patient's baseline demographic and clinical information, along with two potential outcomes containing built-in trade-offs (Outcome A with greater reduction of HbA1c and hypoglycemia, and Outcome B with better physical and mental functioning). Based on the patient, the physician must decide on a preferred outcome amongst the two.

For physicians $p \in \{1, ..., m\}$, assume that each physician is presented with c_q choice scenarios, indexed by $q \in \{1, ..., c_p\}$. Let realizations of the individual choice scenarios be represented by the random vector $\{\boldsymbol{X}^*, \boldsymbol{R}_A, \boldsymbol{R}_B, S\}_{pq}$, where $\boldsymbol{R}_A = \{R_{A1}, ..., R_{AL}\}$ and $\boldsymbol{R}_B = \{R_{B1}, ..., R_{BL}\}$ are two competing outcome profiles for a hypothetical patient presenting with covariates $\boldsymbol{X}^*$, and $S \in \{0, 1\}$ reflects the "better" outcome profile selected by the physician for that patient, with $S = 1$ corresponding to a selection of $\boldsymbol{R}_A$ and $S = 0$ to $\boldsymbol{R}_B$. Note that choice scenarios between different physicians are independent from one another, but choice scenarios within a physician are not. An example choice scenario is shown in Figure 2.1. Practical considerations for developing a physician questionnaire are discussed in Section 2.2.5.

An ITR, $d(\boldsymbol{X}) \colon \mathcal{X} \mapsto \mathcal{A}$, is a mapping from covariates to treatments. Let $\mathbb{1}(\cdot)$ represent the indicator function.

2.2.2 Utility Function Characterization

We assume that individual utility can be represented by the function $U(\boldsymbol{R}|\boldsymbol{X}) = w_1(\boldsymbol{X})R_1 + ... + w_L(\boldsymbol{X})R_L$, a convex combination of the individual outcomes, with covariate-dependent weights, $\{w_1(\boldsymbol{X}), ..., w_L(\boldsymbol{X})\}$, constrained such that $\sum_{l=1}^{L} w_l(\boldsymbol{X}) = 1$.

Parameterizing weights with $\boldsymbol{\beta} = (\boldsymbol{0}, \boldsymbol{\beta_2}, ..., \boldsymbol{\beta_L})$ such that $w_l(\boldsymbol{X}; \boldsymbol{\beta}) = e^{\boldsymbol{\beta_l}^\top \boldsymbol{x}} / \left(1 + \sum_{l=2}^{L} e^{\boldsymbol{\beta_l}^\top \boldsymbol{x}}\right)$, where $\boldsymbol{\beta_1}^\top = \boldsymbol{0}$ and $\boldsymbol{\beta_l}^\top = \{\beta_{l0}, ..., \beta_{ld}\}$ for $l \geqslant 2$, the utility function can be represented by the following multinomial logit-like model:

$$U_{\boldsymbol{\beta}}(\boldsymbol{R}|\boldsymbol{X} = \boldsymbol{x}) = \frac{R_1 + \sum_{l=2}^{L} e^{\boldsymbol{\beta_l}^\top \boldsymbol{x}} R_l}{1 + \sum_{l=2}^{L} e^{\boldsymbol{\beta_l}^\top \boldsymbol{x}}}. \tag{2.1}$$

This representation of the utility function maintains convexity (by enforcing that $w_l(\boldsymbol{X}) > 0 \; \forall \; l \in \mathcal{L}$ and $\sum_{l=1}^{L} w_l(\boldsymbol{X}) = 1$) and allows for the outcome-specific weights to be modeled as functions of covariates. Such a convex utility is also rational (i.e. if one outcome profile is better than another profile at every individual outcome, its utility will always be greater) (Butler *et al.* (2018)). Note that $\boldsymbol{\beta_l}^\top \boldsymbol{x} = \beta_{l0} + \beta_{l1}x_1 + ... + \beta_{ld}x_d$, so the intercept parameters, $\{\beta_{l0} \in \boldsymbol{\beta} : 2 \leqslant l \leqslant L\}$, establish population-average importance weights of the outcomes, and all other parameters, $\{\beta_{l1}, ..., \beta_{ld} \in \boldsymbol{\beta} : 2 \leqslant l \leqslant L\}$, establish individual perturbations to the weights through interaction with $\boldsymbol{X}$. Thus, estimating $\boldsymbol{\beta}$ amounts to achieving patient-specific weighting of the outcomes.

2.2.3 Utility Function Parameter Estimation

In this section, we demonstrate the following: (1) A design of physician questionnaires such that the parameters of (2.1) can be estimated from response data, and (2) The ensuing estimation procedure.

The difference in utility of the two outcome profiles, $\boldsymbol{R}_A$ and $\boldsymbol{R}_B$, can be represented by:

$$U_{\boldsymbol{\beta}}(\boldsymbol{R}_A - \boldsymbol{R}_B|\boldsymbol{X} = \boldsymbol{x}) = \frac{(R_{A1} - R_{B1}) + \sum_{l=2}^{L} e^{\boldsymbol{\beta_l}^\top \boldsymbol{x}} (R_{Al} - R_{Bl})}{1 + \sum_{l=2}^{L} e^{\boldsymbol{\beta_l}^\top \boldsymbol{x}}}. \tag{2.2}$$

Physician preference information is collected with a goal of estimating $\boldsymbol{\beta}$, thus establishing a link between $\mathcal{X}$ and the utility function. We design the questionnaire by simulating various choice scenarios with a hypothetical patient (randomly-generated $\boldsymbol{X}^*$) and two potential outcome profiles (randomly generated $\boldsymbol{R}_A$ and $\boldsymbol{R}_B$). For the given patient presenting with covariates $\boldsymbol{X}^*$, the physician compares the utility of $\boldsymbol{R}_A$ and $\boldsymbol{R}_B$ and makes a selection, S, of a preferred outcome profile. Let p physicians each be presented with c_p such choice scenarios, amongst which they must make selections. Due to possible dependence within each physician's answers, we incorporate physician-specific random intercepts, $\boldsymbol{b}_p = \{b_{p2}, ..., b_{pL}\}$, for all $p \in \{1, ..., m\}$. This allows for dependence amongst the random effects pertaining to each outcome. Thus, the utility difference for two outcome profiles, specific to arbitrary physician p, is:

$$U_{\boldsymbol{\beta},p}(\boldsymbol{R}_A - \boldsymbol{R}_B|\boldsymbol{X} = \boldsymbol{x}) = \frac{(R_{A1} - R_{B1}) + \sum_{l=2}^{L} e^{\boldsymbol{\beta_l}^\top \boldsymbol{x} + b_{pl}}(R_{Al} - R_{Bl})}{1 + \sum_{l=2}^{L} e^{\boldsymbol{\beta_l}^\top \boldsymbol{x} + b_{pl}}}. \tag{2.3}$$

Let $\Pr_{pq}(\boldsymbol{X}^*, \boldsymbol{R}_A, \boldsymbol{R}_B|\boldsymbol{\beta}, \boldsymbol{b}_p)$ represent the probability that outcome $\boldsymbol{R}_A$ is selected over $\boldsymbol{R}_B$ in choice scenario q for physician p, where the hypothetical patient has covariates $\boldsymbol{X}^*$. We establish the following logit framework to link probability of selection with difference in utility:

$$\text{logit}\left\{\Pr_{pq}(\boldsymbol{X}^*, \boldsymbol{R}_A, \boldsymbol{R}_B|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} = U_{\boldsymbol{\beta},p}(\boldsymbol{R}_A - \boldsymbol{R}_B|\boldsymbol{X}^*), \tag{2.4}$$

Finally, let the prior distribution for $\boldsymbol{b}_p$ be $N(\boldsymbol{0}_{L-1}, \boldsymbol{\Sigma}_{(L-1)*(L-1)})$, where $\boldsymbol{\Sigma}$ is an arbitrary exchangeable covariance matrix, indexed by ρ, a scalar representing correlation, and $\boldsymbol{\sigma}^2$, a representation of the element-wise variances of $\boldsymbol{\beta}$. The full data likelihood for fixed parameters can

now be written as:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \rho) &= p(\boldsymbol{S}|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \rho) \\
&= \int p(\boldsymbol{S}, \boldsymbol{b}|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \rho) d\boldsymbol{b} \\
&= \int p(\boldsymbol{S}|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \rho, \boldsymbol{b}) \cdot p(\boldsymbol{b}|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \rho) d\boldsymbol{b} \\
&= \int p(\boldsymbol{S}|\boldsymbol{\beta}, \boldsymbol{b}) \cdot p(\boldsymbol{b}|\boldsymbol{\sigma}^2, \rho) d\boldsymbol{b} \\
&= \int \prod_{p=1}^{m} \left\{ \prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p) \right\} \cdot p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) d\boldsymbol{b}_p \\
&= \prod_{p=1}^{m} \int \left\{ \prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p) \right\} \cdot p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) d\boldsymbol{b}_p,
\end{aligned} \tag{2.5}$$

where, letting $\Pr_{pq}(\boldsymbol{\beta}, \boldsymbol{b}_p)$ represent $\Pr_{pq}(\boldsymbol{X}^*, \boldsymbol{R}_A, \boldsymbol{R}_B|\boldsymbol{\beta}, \boldsymbol{b}_p)$ for notational convenience, $p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p) = \{Pr_{pq}(\boldsymbol{\beta}, \boldsymbol{b}_p)\}^{s_{pq}} * \{1 - Pr_{pq}(\boldsymbol{\beta}, \boldsymbol{b}_p)\}^{1-s_{pq}}$.

We optimize the likelihood above through a Bayesian approach: setting prior distributions for the components of $\boldsymbol{\beta}$ and $\boldsymbol{b}$ and using Markov chain Monte Carlo (MCMC) sampling with Metropolis-Hastings (MH) (Chib and Greenberg (1995)) implementation to obtain iterative draws from the posterior. The elements of $\boldsymbol{\beta}$ are set to initial values of 0, with priors according to $N(\mu, \sigma^2)$, where μ and σ^2 are pre-specified to $\mu = 0$ and σ^2-large. Given that the prior distribution for $\boldsymbol{b}_p$ is $N(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is an exchangeable covariance matrix, let $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{A}(\rho)\boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix with entries $\boldsymbol{d} = (\sigma_2, ..., \sigma_L)^\top$ and $\boldsymbol{A}(\rho) = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{j}\boldsymbol{j}^\top$, where $\boldsymbol{I}$ is a $(L-1) \times (L-1)$ identity matrix, and $\boldsymbol{j}$ is an $(L-1)$-vector of ones. σ_l^2 is pre-specified with a flat prior distribution of $\Gamma(0.01, 0.01)$ for all $l \in \{2, ..., L\}$.

Constant priors would ensure that the posterior is effectively proportional to the likelihood function, resulting in the posterior mode being numerically identical to the maximum likelihood estimate (Song (2007)), thus the preference for flat priors (unless prior information is available for certain elements of $\boldsymbol{\beta}$). MCMC automatically integrates over the random effect (Pollock (2002)), thus the posterior distribution of $\boldsymbol{\beta}$ reflects the marginal, which is of interest. All parameters are estimated as the average of the MCMC draws following a specified burn-in period.

2.2.4 ITR Estimation

In this section, we describe two approaches for ITR estimation. Having already obtained $\widehat{\boldsymbol{\beta}}_n$, let the estimator for $U_{\boldsymbol{\beta}}(\boldsymbol{R}|\boldsymbol{X} = \boldsymbol{x})$ be $U_{\widehat{\boldsymbol{\beta}}_n}(\boldsymbol{R}|\boldsymbol{X} = \boldsymbol{x}) = \sum_{l=1}^{L} w_l(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}_n) R_l$. We begin with a general method, and follow it up with a regression-based approach.

Method #1 (General): For all patients, calculate $U_{\widehat{\boldsymbol{\beta}}_n}(\boldsymbol{R}|\boldsymbol{X})$, which represents the estimated utility based on observed outcomes. $U_{\widehat{\boldsymbol{\beta}}_n}(\boldsymbol{R}|\boldsymbol{X})$ can directly be treated as the "new" observed outcome which can be used with any preexisting ITR estimation algorithm with desired properties, such as efficient augmentation and relaxation learning (EARL) for double robustness (Zhao *et al.* (2019)), Stabilized Direct Learning (SD-Learning) for heteroscedastic data (Shah *et al.* (2022)), Multicategory Outcome Weighted Margin-based Learning (MOML) for $\geqslant 2$ treatments (Zhang *et al.* (2020)), etc.

Method #2 (Q-Learning analog): For every outcome $l \in \mathcal{L}$, let $Q_l(\boldsymbol{X}, A) = E(R_l|\boldsymbol{X}, A)$, which can be estimated through regression. With the subsequent Q-function for overall utility, the optimal ITR, d, can be estimated according to the following:

$$Q_{U_{\widehat{\boldsymbol{\beta}}_n}}(\boldsymbol{X} = \boldsymbol{x}, A = a) = \sum_{l=1}^{L} w_l(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}_n) \widehat{Q}_l(\boldsymbol{x}, a).$$

$$\widehat{d}(\boldsymbol{x}) = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q_{U_{\widehat{\boldsymbol{\beta}}_n}}(\boldsymbol{x}, a).$$

This estimation approach amounts to a Q-Learning analog that finds an ITR to maximize individual patient utility.

While Method #2 utilizes more information because it predicts each outcome separately, Method #1 gives the flexibility of treating estimated utility as the single observed outcome and then using any precision medicine algorithm to estimate the ITR.

2.2.5 Practical Considerations for Developing Physician Questionnaire

Utility function estimation is an inverse reinforcement learning task; at this step, instead of creating the optimal behavior/actions, we strive to find the reward function (utility function) from expert demonstrations that display informed behavior (physician questionnaire). This will be followed by the reinforcement learning step: treatment selection in order to optimize patient-specific utility.

In the questionnaire, each physician is presented with a series of choice tasks, each of which presents a hypothetical patient, $\boldsymbol{X}^*$, along with two possible sets of outcomes, $\boldsymbol{R}_A = \{R_{A1}, ..., R_{AL}\}$ and $\boldsymbol{R}_B = \{R_{B1}, ..., R_{BL}\}$. Keeping that patient in mind, the physician must choose the more desirable set of outcomes. In the example choice scenario shown in Figure 2.1, a physician must examine a diabetic patient's baseline data and choose amongst two outcome vectors with a built-in trade-off (one favoring HbA1c and hypoglycemia reduction, the other favoring physical and mental functioning). Note that the patient covariates shown to physicians can be (and are likely to be) a subset of the collected baseline data. This decision can be made based on prior knowledge of the variables that affect patient utility.

Designing these choice scenarios poses two key challenges: (1) The question of whether the distribution of the hypothetical covariates and corresponding outcomes matches what could be observed in reality, and (2) Designing pairs of outcomes that have trade-offs built in. For (1), if the distributions are not close, the resultant data drift (Shandhi and Dunn (2022)) could cause significant inefficiency in attempting to estimate the utility function. This issue is further aggravated with curse of dimensionality. For (2), choice scenarios with outcomes lacking trade-offs would be redundant in the sense that no information is gained by a physician picking the outcome vector which is clearly optimal at each individual outcome.

To alleviate these challenges, we first propose that the hypothetical patient data ($\boldsymbol{X}^*$) presented to physicians are actually chosen as random samples from $\boldsymbol{X}$, the observed patient data. This ensures that the distribution of patients presented to physicians matches covariates that are observed in reality. In order to generate realistic hypothetical outcomes for the given covariates,

we propose a solution through Distributional Random Forests (DRF) (Cevid *et al.* (2022)). DRF is a regression framework that finds the estimated conditional outcome distribution, rather than simply the conditional expectation of the outcome (a point prediction). Further, the methodology allows for prediction of distributions of multivariate response vectors. Therefore, we train a DRF model to estimate the conditional distribution of $\boldsymbol{R}$ given $\boldsymbol{X}$. Then for choice scenario q for physician p, we obtain samples from $p(\boldsymbol{R}|\boldsymbol{X}^*_{pq})$ until two that have a trade-off are found. These two samples are considered as $\boldsymbol{R}_A$ and $\boldsymbol{R}_B$ for that scenario, and this is repeated to obtain hypothetical outcomes for all choice scenarios for each physician.

2.3 Theory

In this section, we delineate the theoretical properties of the utility function parameter estimation technique in Section 2.2.3. We establish consistency and asymptotic normality of the parameters of the estimator in the scenario that the number of physicians in the questionnaire, m, diverges, and covariate dimensionality is fixed. Detailed proofs for all theorems are shown in the Supporting Information in Appendix B. In order to show consistency of the utility function parameter estimates, we begin by stating a standard assumption of boundedness and rank of the second moment, along with a constraint on ρ:

Assumption 2.1. *$E(\boldsymbol{X}\boldsymbol{X}^\top)$ is full rank, $E\left\|\boldsymbol{X}\right\|^2 < c_x < \infty$, $\max_{2\leqslant l\leqslant L}(E|R_{Al}|^2 + E|R_{Bl}|^2) < \infty$, and $0 < \sigma_l^2 < \infty$ for all $l \in \mathcal{L}$.*

Assumption 2.2. *Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \rho\} \in \boldsymbol{\Theta}$. The true values of the fixed parameters, $\boldsymbol{\theta}_0 = \{\boldsymbol{\beta}_0, \boldsymbol{\sigma}_0^2, \rho_0\}$, lie in the interior of $\boldsymbol{\Theta}_0$, which is a known, compact subset of $\boldsymbol{\Theta}$, the parameter space. Additionally, the exchangeable correlation, ρ, of the exchangeable correlation matrix, $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{A}(\rho)\boldsymbol{D}$, is bounded such that $\frac{-1}{L-2} < \rho < 1$.*

Theorem 2.1. *Consistency: Let $\widehat{\boldsymbol{\theta}} = \left\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\rho}\right\}$. If Assumptions 2.1 and 2.2 are met, then the utility function parameter estimation is consistent. That is, $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$.*

With consistency proven, we can move towards additionally showing asymptotic normality of the utility function parameter estimates. This will require one additional assumption:

Assumption 2.3. *Log-concavity: The Fisher Information, $\boldsymbol{I}(\boldsymbol{\theta}_0)$, is positive-definite for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.*

Theorem 2.2. *Asymptotic normality: Let $\widehat{\boldsymbol{\theta}} = \left\{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\sigma}}^2, \widehat{\rho}\right\}$. With Assumptions 2.1-2.3 met, and consistency established, $\sqrt{m}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}$. Thus, the maximum likelihood estimator achieves the asymptotic efficiency bound.*

In Theorem 2.3, we also show that the inverse Fisher Information, found in Theorem 2.2 to be the asymptotic covariance, can be consistently estimated.

Theorem 2.3. *Let the observed information matrix be represented by $\boldsymbol{I}(\widehat{\boldsymbol{\theta}})$:*

$$\boldsymbol{I}(\widehat{\boldsymbol{\theta}}) = \frac{1}{m}\left(\sum_{p=1}^{m} \frac{dl_p}{d\boldsymbol{\theta}}\right)\left(\sum_{p=1}^{m} \frac{dl_p}{d\boldsymbol{\theta}}\right)^{\top}\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \tag{2.6}$$

Then, $\boldsymbol{I}(\widehat{\boldsymbol{\theta}})^{-1}$ is a consistent estimator of the inverse Fisher Information matrix, $\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}$, which is the asymptotic covariance of $\sqrt{m}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$.

Having shown properties of the asymptotic distribution of the estimated utility function, we turn our attention to the resulting ITR estimation and value function. Let Y represent an arbitrary, scalar outcome of interest. Under an ITR, d, the expected population outcome can be represented by the value function:

$$V(d) = E\left\{Y \mid A = d(\boldsymbol{X})\right\}, \tag{2.7}$$

and the optimal ITR with respect to that outcome, d^{opt}, is that which maximizes the expected outcome: $d^{opt}(\cdot) = \underset{d \in \mathcal{D}}{\operatorname{argmax}}\, V(d)$, where $\mathcal{D}$ is a prespecified class of decision rules. Accurate evaluation of the performance any decision rule is critical in determining its usability in practice, thus an unbiased and consistent estimator for $V(d)$ has been established by Qian and Murphy

(2011) as:

$$\widehat{V}(d) = \frac{\sum_{i=1}^{n} Y_i \cdot \mathbb{1}\left\{A_i = d(\boldsymbol{X}_i)\right\} / \pi(A_i|\boldsymbol{X}_i)}{\sum_{i=1}^{n} \mathbb{1}\left\{A_i = d(\boldsymbol{X}_i)\right\} / \pi(A_i|\boldsymbol{X}_i)}. \tag{2.8}$$

Now, the methodology in this paper strives to optimize patient utility, thus setting $Y = U$ (see Method #1 in Section 2.2.4). However, U is unobserved and estimated as $\widehat{U}$ by plugging $\widehat{\boldsymbol{\beta}}$ into Equation 2.1. As a result, instead of plugging $Y_i = U_i$ into $\widehat{V}(d)$, our version of the value function estimate, denoted as $\widehat{V}_{\widehat{U}}(d)$, will set $Y_i = \widehat{U}_i$ (the estimated outcome). Since $\widehat{V}_{\widehat{U}}(d)$ uses an estimated instead of observed outcome of interest, its consistency for $V(d)$ cannot be assumed from previous results and must be established. We demonstrate this result in Theorem 2.4:

Theorem 2.4. *Under Assumptions 2.1 and 2.2, along with consistency of the utility function parameters (established in Theorem 2.1) and $\pi(A|\boldsymbol{X})$ being known (or estimated such that $\widehat{\pi}(A|\boldsymbol{X}) \xrightarrow{P} \pi(A|\boldsymbol{X})$), $\widehat{V}_{\widehat{U}}(d)$ is a consistent estimator of $V(d)$.*

The proof of 2.4 is dependent on the consistency of the parameters involved in the calculation of $\widehat{U}$ as shown in Theorem 2.1. Consistency of $\widehat{V}_{\widehat{U}}(d)$ for the truth establishes that $\widehat{U}$ can be used in value calculations. As a result, standard precision medicine results hold, such as the convergence of $\widehat{V}_{\widehat{U}}(d)$ to the value function for the optimal decision rule, $V(d^{opt})$.

2.4 Numerical Results: Simulation Study

We outline a simulation to demonstrate the value of this methodology for scenarios with multiple outcomes and utility heterogeneity. In this scenario, there are $L = 2$ outcomes of interest, and 2 covariates are shown to physicians. The two competing outcomes, R_1 and R_2, are efficacy and side effect reduction, with both outcomes coded so that larger values are preferred. Assume binary treatment, $A \in \mathcal{A} = \{-1, 1\}$, and covariates $\boldsymbol{X} = (X_0, X_1, X_2)$, where $X_0 = 1$ is the intercept, and $X_1 \in [0, 2]$ represents a depression assessment scale where 0 reflects no depression, 1 reflects mild depression, and 2 reflects severe depression, and X_2 is an arbitrary covariate that is measured but does not factor into the utility function. Let $\boldsymbol{\beta} = (\boldsymbol{\beta_1}, \boldsymbol{\beta_2})$, where $\boldsymbol{\beta_1} = (0, 0, 0)$ by

definition and $\boldsymbol{\beta_2} = (-1, 1, 0)$. Thus, the true utility function is:

$$U_{\boldsymbol{\beta}}(\boldsymbol{R}|\boldsymbol{X} = \boldsymbol{x}) = \frac{R_1 + e^{x_1 - 1} R_2}{1 + e^{x_1 - 1}}.$$

The resulting utility functions for patients presenting with different levels of depression are shown below:

$$U_{\boldsymbol{\beta}}(\boldsymbol{R}|X_1 = 0) = 0.73R_1 + 0.27R_2,$$

$$U_{\boldsymbol{\beta}}(\boldsymbol{R}|X_1 = 1) = 0.5R_1 + 0.5R_2,$$

$$U_{\boldsymbol{\beta}}(\boldsymbol{R}|X_1 = 2) = 0.27R_1 + 0.73R_2.$$

This shows that a patient with no depression favors efficacy (R_1), one with mild depression places equal importance on each outcome, and one with severe depression favors side effect reduction (R_2).

We simulate data for 200 patients with $X_1, X_2 \sim U[0, 2]$, and A based on an RCT setting equal probability either treatment being assigned. Letting $\epsilon \sim N(0, 1)$, patient outcomes are simulated according to the following:

$$R_1 = 5 - X_1 + 2A + \epsilon, \tag{2.9}$$

$$R_2 = 5 - X_1 - 2A + 0.25X_1A + \epsilon. \tag{2.10}$$

Note that this simulation was designed such that treatment $A = 1$ generally leads to greater efficacy and $A = -1$ leads to greater side effect reduction. Thus, patients with no depression are more likely to favor $A = 1$, and those with severe depression are likely to favor $A = -1$. Patients with mild depression ($X_1 = 1$) favor $A = 1$ slightly because they value both outcomes equally, and $A = 1$ boosts R_1 more than it decreases R_2.

Finally, we simulate physician responses at 3 sample sizes: (1) $m = 10$ physicians each answering $c_p = 10$ questions, (2) $m = 20$ physicians each answering $c_p = 20$ questions, and

(3) $m = 30$ physicians each answering $c_p = 30$ questions. We randomly sample patient covariates and link them with hypothetical outcome vectors according to the DRF method outlined in Section 2.2.5, which ensures that there is always a trade-off between $\boldsymbol{R}_A$ and $\boldsymbol{R}_B$ (i.e., there are no obvious choice scenarios where one outcome profile beats another at both outcomes). We also generate $\boldsymbol{b} = \{b_1, ..., b_m\}$, which are scalar physician-specific random effects when $L = 2$, according to $N(0, \sigma_2^2)$ where $\sigma_2^2 = 1$. Recall that Pr_{pq}, the probability that physician p selects $\boldsymbol{R}_A$ as the better outcome in choice scenario q, is equal to $\frac{e^{U_{\boldsymbol{\beta},p}(\boldsymbol{R}_A - \boldsymbol{R}_B|\boldsymbol{X}^*)}}{1+e^{U_{\boldsymbol{\beta},b}(\boldsymbol{R}_A - \boldsymbol{R}_B|\boldsymbol{X}^*)}}$ based on Equation (2.4), and thus the resulting selection, S_{pq}, is simulated as $S_{pq} \sim \text{Bernoulli}(\text{Pr}_{pq})$, where $S_{pq} = 1$ corresponds to selecting $\boldsymbol{R}_A$ and $S_{pq} = 0$ to $\boldsymbol{R}_B$.

From the physician data, we use a MH algorithm to estimate $\{\boldsymbol{\beta}_2, \sigma_2^2\}$, initiating each element of $\boldsymbol{\beta_2}$ with a prior of $N(0, 1000)$, σ_2^2 with a prior of $\Gamma(0.001, 0.001)$, and treating the elements of $\boldsymbol{b}$ as fixed with priors of $N(0, \sigma_2^2)$. After performing 10,000 iterations with a burn-in period of 1,000, we obtain $\widehat{\boldsymbol{\beta}}_2 = (\widehat{\beta}_{20}, \widehat{\beta}_{21}, \widehat{\beta}_{22})$ and $\widehat{\sigma}_2^2$ by taking the mean of samples 1,000 through 10,000. This full simulation (including patient data generation, physician data generation, and MCMC estimation) is repeated 100 times, with parameter estimation results displayed in Figure 2.2. It is clearly seen that as the number of physicians and choice scenarios increases, parameter estimates fall closer to the truth.

We then use $\widehat{\boldsymbol{\beta}}_2$ to calculate each patient's estimated observed utility, $U_{\widehat{\boldsymbol{\beta}}_n}(\boldsymbol{R}|\boldsymbol{X})$. Using the estimated utilities as the new, scalar outcome, we perform Q-Learning (using random forest for the regression method) to estimate an ITR, $\widehat{d}^{\text{IRL}}$. We use the known utilities, U, to calculate the empirical value, $\widehat{V}(\widehat{d}^{\text{IRL}}) = \frac{\mathbb{P}_n\left[U \cdot \mathbb{1}\{A = \widehat{d}^{\text{IRL}}(\boldsymbol{X})\}/0.5\right]}{\mathbb{P}_n\left[\mathbb{1}\{A = \widehat{d}^{\text{IRL}}(\boldsymbol{X})\}/0.5\right]}$ (i.e., the expected outcome if treatment were to be assigned according to $\widehat{d}^{\text{IRL}}$). For comparison purposes, we also calculated the estimated utility derived from three other treatment assignment strategies: (1) Q-Learning optimizing only R_1 (efficacy), (2) Q-Learning optimizing only R_2 (side effect reduction), and (3) Coinflip (random) treatment assignment. The results are displayed in Figure 2.3. Expectedly, IRL outperforms the other methods and improves quickly with greater sample size of the physician data (thus more accurate estimation of the utility function).

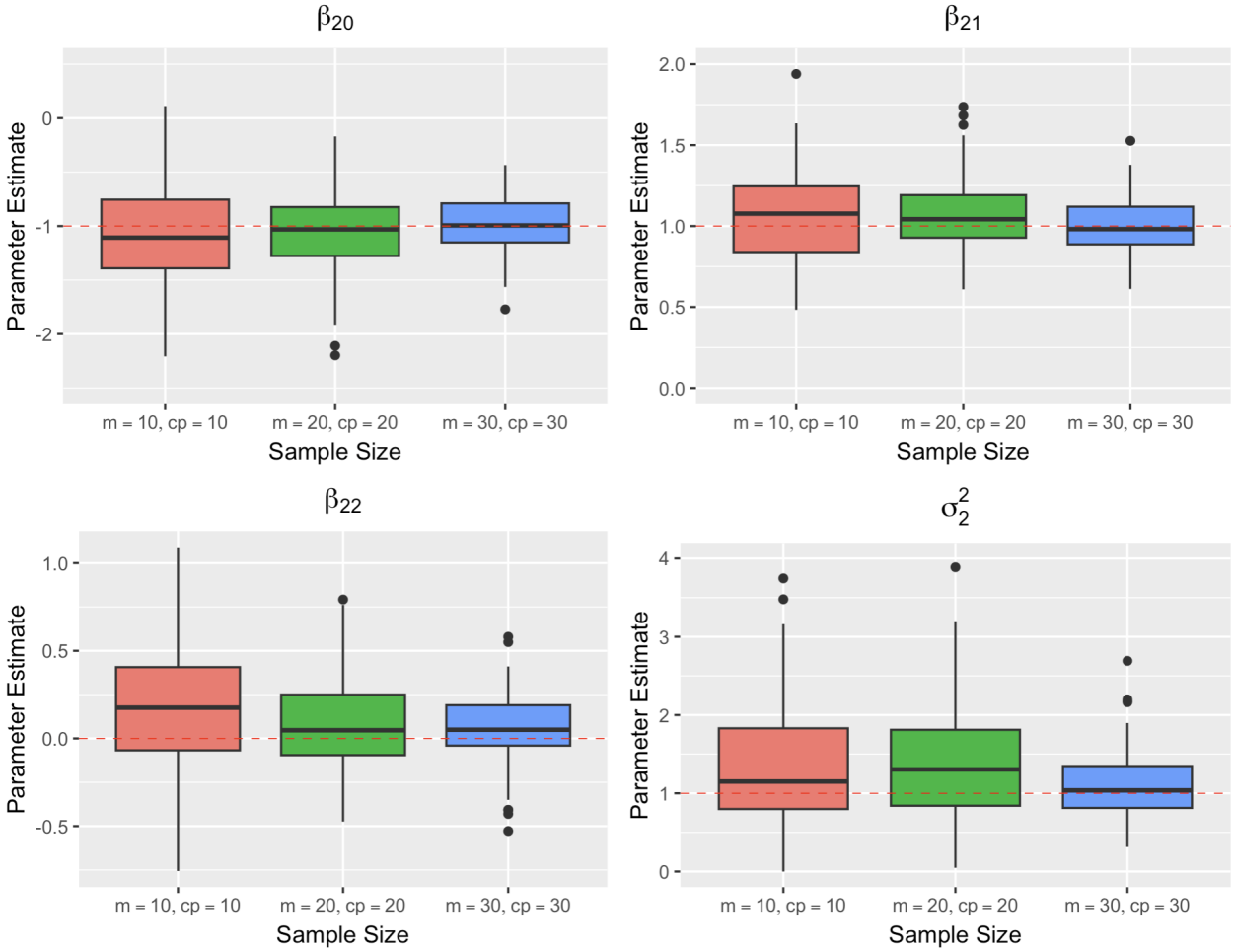

Figure 2.2: Boxplots displaying parameter estimation results of $\beta_{20}, \beta_{21}, \beta_{22}$ and σ_2^2 resulting from 100 replications of the simulation at 3 sample sizes of physician questionnaire data: (1) $m = 10, c_p = 10$; (2) $m = 20, c_p = 20$; (3) $m = 30, c_p = 30$. The true parameter values are reflected by the dashed red line. The sample size of the patient data for this simulation is $n = 200$.

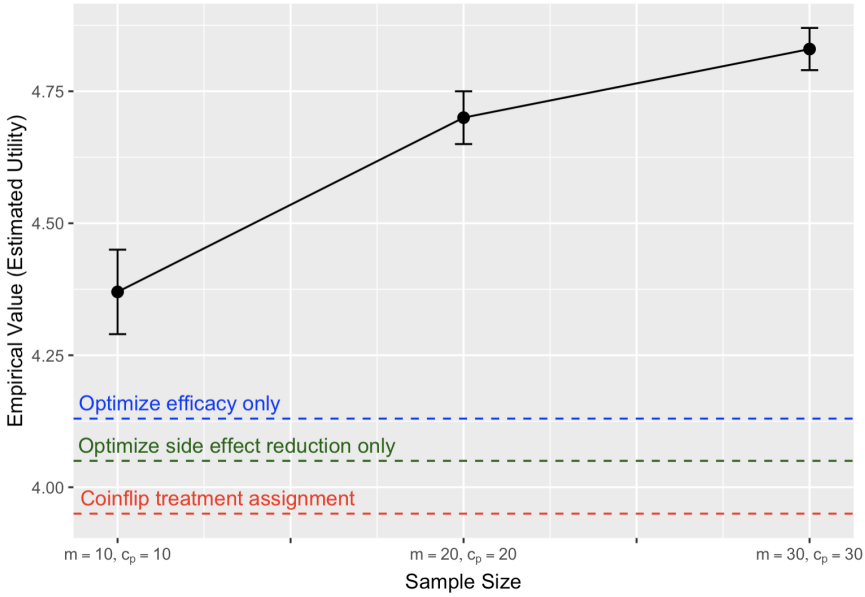

Figure 2.3: Expected utility, averaged across the 100 replications, along with standard error of the mean (SEM) bars, for the IRL methodology at 3 sample sizes of physician questionnaire data: (1) $m = 10, c_p = 10$; (2) $m = 20, c_p = 20$; (3) $m = 30, c_p = 30$. For comparison, expected utility results are displayed for 3 other methods: (1) Optimization of efficacy only, (2) Optimization of side effect reduction only, and (3) Coinflip (random) treatment assignment. For all methods, the expected utility estimates are based on the patient data with a sample size of $n = 200$.

2.5 Discussion

This work develops a method for linking physician input and patient covariates in such a way knowing $\boldsymbol{X}$ provides the entire utility information for a patient. After estimating a utility function, we demonstrate how ensuing ITR estimation can provide a decision rule, which, amidst utility heterogeneity across patients, optimizes individual patient utility.

The methodology contributed by this paper is an Inverse Reinforcement Learning (IRL) step that precedes the usual reinforcement learning (RL) framework of ITR estimation. In the IRL step, we strive to find the reward/utility function (i.e., uncover patient-specific importance weights for the outcomes based on demonstrations of expert behavior). In the RL step, we strive to optimize action (i.e., optimal treatment selection in relation to the individual-specific estimated utilities). By unifying the IRL and RL steps, this paper contributes a novel precision medicine

approach that allows for multiple outcomes to be considered in the treatment decision-making framework.

Unlike the patient preferences setting, where preference information has to be collected for every incoming patient, we use physician recommendations as a proxy for patient preference. Our methodology requires physician data to be collected only once, in order to draw a connection between covariates and utility function. Then, for all patients, this link can be used to estimate the patient-specific utilities.

Our method works for any number of outcomes of interest, although dimensionality considerations should be made when selecting the number of outcomes that will be used to construct the utility function. This method also works in scenarios with any number of treatments, because it allows for any ITR estimation method to be used, including approaches built for > 2 treatments.

An area in which the methodology presented in this paper could be fruitful is Type 1 Diabetes (T1D). The Wireless Innovation for Seniors with Diabetes Mellitus (WISDM) clinical trial (Pratley *et al.* (2020); Weinstein *et al.* (2023)) measured the effectiveness of the interventions of continuous glucose monitoring (CGM) and traditional blood glucose monitoring (BGM), while also measuring a variety of outcomes, both clinical (hypoglycemia, hyperglycemia, etc.) and psychosocial (mental/physical functioning, etc.). A precision medicine analysis has been conducted on this data in order to find an ITR that minimizes patient time spent in hypoglycemia (Kahkoska *et al.* (2023)). However, if balancing multiple outcomes is of interest (such as a patient's clinical variables along with mental and physical health), the methodology of this paper may lead to decision support algorithms that optimize overall patient utility rather than only one individual outcome.

As a future extension, the constraint of $w_l(\boldsymbol{X}) > 0$ could be relaxed to $w_l(\boldsymbol{X}) \geqslant 0$ with a selection strategy capable of excluding some outcomes altogether for certain subsets of patients. Additionally, the physician questionnaire could be designed with some form of internal validation in mind; a subset of the choice scenarios could be repeated between physicians, with the results statistically analyzed to determine the level of general agreement between physicians.

Finite-sample theory could also be developed to assist with strategical considerations of resource constraints (e.g. determining whether it is more beneficial to increase the number of choices in a questionnaire vs. increasing the number of physicians that respond to the survey). Finally, it is worth exploring construction of the questionnaire more deeply in order to determine whether more efficient designs are possible (better estimation of the parameters with lower required sample size).

CHAPTER 3: PRECISION MEDICINE IN DIABETES: ESTIMATION OF A DECISION RULE TO UNCOVER HETEROGENEOUS EFFECTS OF CONTINUOUS GLUCOSE MONITORING ON HYPOGLYCEMIA

3.1 Introduction

It is often quoted that the primary goal of precision medicine is to deliver the "right treatment to the right patient at the right time" (Zhang (2015); Kosorok and Laber (2019); Freeman *et al.* (2022)). We denote this as *decision making*. However, there may be scenarios in which this goal is not of primary importance.

Say, for example, that clinical contexts exists in which: (1) Clinicians are not looking for algorithms to replace human decision making, or (2) A treatment has already been well-established as the standard of care. What can precision medicine methodology contribute in these cases? In scenario (1), precision medicine can still offer *decision support* by helping to augment decisions that are already being made based on what is known scientifically. In scenario (2), precision medicine can contribute *scientific depth* by helping understand why a treatment works, and who it works well on.

Thus, although the precision medicine framework is built to offer the potential of *decision making* capabilities, it also contributes to the secondary goals of *decision support* and *scientific depth*. Many such clinical contexts exist in which these secondary goals are of prime interest. One such context is Type 1 Diabetes (T1D).

3.1.1 The Utility of Precision Medicine in Type I Diabetes

The Wireless Innovation for Seniors with Diabetes Mellitus (WISDM) study evaluated the effect of continuous glucose monitoring (CGM) on hypoglycemia (glucose $<$70 mg/dL) over

6 months, compared with standard BGM, among older adults ages 60 years with T1D (Pratley *et al.* (2020)). On average, use of CGM reduced hypoglycemia from approximately 73 minutes per day at the trial baseline to 39 minutes per day over the 6-month trial period, with a concurrent reduction in hemoglobin A1c (HbA1c). There were no significant changes in the BGM group, who experienced hypoglycemia for 68 minutes per day at baseline and 70 minutes per day over the trial period (Pratley *et al.* (2020)). Based on the WISDM trial and other studies (Ruedy *et al.* (2017); Toschi *et al.* (2020); Munshi *et al.* (2022)), contemporary clinical guidelines state that CGM should be considered for glucose monitoring for older adults with T1D (ADA Professional Practice Committee (2022a)), and specifically to reduce hypoglycemia.

Though there are likely benefits of CGM for all older adults with T1D (Toschi *et al.* (2020); ADA Professional Practice Committee (2022a)), it is known that estimates of average treatment effects of the primary outcome in a trial setting can mask the effects that individual participants may experience. Pre-specified subgroup analyses in the WISDM study showed the treatment effect of CGM, in terms of reduction in hypoglycemia, was greater among participants with increased baseline hypoglycemia and glycemic variability as measured by the coefficient of variation (%CV), and lower detectable C-peptide levels (Pratley *et al.* (2020)).

Yet, conventional subgroup analyses are limited in discovering heterogeneous treatment effects as they require moderating markers to be decided a priori (Kent *et al.* (2018)). Given that older adults are extremely diverse in terms of biopsychosocial profiles, clinical needs, historical diabetes self-management experiences, and preferences for care (Munshi *et al.* (2020); Kirkman *et al.* (2012)), an entirely a priori approach may not capture patient markers that were not pre-specified as previously recognized as potential moderating markers of a treatment response.

An alternative approach is to use rigorous machine learning methods from the field of precision medicine that estimate a decision rule, a mathematical function that maps patient-level information (demographic characteristics, clinical biomarkers, other measures) to a recommended treatment or intervention to optimize (i.e., maximize or minimize) an outcome of interest (Kosorok and Laber (2019)). Decision rules offer an entirely data-driven approach to discovery

of moderating markers (i.e., the patient-level variables that are predictive of the within-individual difference in effect of 2 or more treatments, as per Kosorok and Laber (2019)), and link therapies to subgroups of patients likely to show favorable responses (Kosorok and Laber (2019); Trusheim *et al.* (2007)).

3.1.2 Proposed Approach

In this post hoc analysis, our objective was to use a data-driven approach to explore the moderating markers (such as baseline CGM glucose management indices and participant characteristics) that are associated with heterogeneous effects of CGM on hypoglycemia in older adults with T1D. The significance of this work is to inform future hypotheses by supporting shared decision-making surrounding the utility of CGM for individual older adults living with T1D (*decision support*) and uncovering relationships between moderating markers and differential effects of CGM on hypoglycemia (*scientific depth*). In relation to the former objective, we strive to answer the question, "Are there edge-cases where a physician's decision-making can be augmented?" Towards the latter objective, we answer the question, "Can we uncover patterns in the heterogeneity of treatment success of CGM?"

3.2 Methodology

3.2.1 Design, Setting, and Participants

The WISDM study (ClinicalTrials.gov Identifier: NCT03240432) enrolled 203 older adults with T1D at 22 sites in the United States. Older adults (age 60 years) were eligible if they used an insulin pump or multiple daily injections, but they could not have used unblinded CGM as part of T1D management in the past 3 months. Participants were randomized to CGM (n = 103) or BGM (n = 100) for 6 months to study whether CGM could reduce hypoglycemia (% time with glucose $<$70 mg/dL) Pratley *et al.* (2020). The cohort has been extensively characterized, with both treatment groups having balanced baseline characteristics as per Table 1 and eTable 16 of

Pratley *et al.* (2020). Seven patients were lost to follow-up, requested withdrawal from the study, or discontinued the intervention, and 2 patients were missing education data, resulting in 194 older adults for this analysis (CGM [n = 100]; BGM [n = 94]).

3.2.2 Measures

The primary outcome was CGM-measured percentage of time spent in hypoglycemic range (<70 mg/dL) at follow-up, which used pooled data from approximately 2 weeks prior to randomization, and 1 week prior to the 8-, 16-, and 26-week visits, during which time participants in the BGM arm wore blinded CGM (Pratley *et al.* (2020)). For the WISDM participants in this analysis, the average baseline time with CGM readings was 16.3 days (fifth percentile of 12.8 days and 95th percentile of 23.8 days) with an average of 21.9 hours of CGM data per day (fifth percentile of 17.0 h and 95th percentile of 23.7 h). The average follow-up time with CGM readings was 6.9 days (fifth percentile of 5.8 days and 95th percentile of 7.0 days) with an average of 22.5 hours of CGM readings per day (fifth percentile of 17.8 h and 95th percentile of 24.0 h). For our decision rule, we specified a minimization of percentage of time in hypoglycemia as the optimal outcome for each participant. As a secondary analysis, we repeated the analysis using change in CGM-measured percentage of time spent in hypoglycemic range (<70 mg/dL) between baseline and follow-up (Pratley *et al.* (2020)). Here, a decrease in the percentage of time in hypoglycemia was defined as the optimal outcome.

In estimating the optimal decision rule, we considered the following baseline variables, which were collected from medical records and confirmed by participants (Pratley *et al.* (2020)): age, diabetes duration, age at diagnosis, sex, highest education, health insurance status, insulin pump use, screening HbA1c, detectable C-peptide levels, history of severe hypoglycemia events, and diabetic ketoacidosis events in the past 12 months. We used 3 variables measured during the baseline period of blinded CGM wear: percentage of time in hypoglycemia (glucose under 70 mg/dL), percentage of time in range (TIR; glucose in range of 70-180 mg/dL), and glycemic variability (defined as the %CV).

Highest education was categorized by "Bachelors Degree", "Less than a Bachelors Degree", "More than a Bachelors degree". Health insurance was split into "Private", "Private and Medicare", and "Medicare/other". C-peptide values were dichotomized into non-detectable ($<$0.003 nmol/L) and detectable ($\geqslant$0.003 nmol/L). Severe hypoglycemia and diabetic ketoacidosis were dichotomized to reflect whether an event occurred or not. For characterizing the resulting subgroups produced by the decision rule, we additionally compared the subgroups based on race and ethnicity, total daily insulin dose per kg, and body mass index.

3.2.3 Statistical Analysis

We estimated and evaluated a series of potential decision rules that estimate an optimal glucose monitoring modality for each WISDM study participant (CGM vs BGM) to minimize total time in hypoglycemia. As above, we repeated our approach as part of a secondary analysis to maximize the reduction (i.e., negative change) in time spent in hypoglycemia from baseline.

The decision rules were based on the 14 demographic, clinical, and laboratory measures specified above. We selected statistical approaches for precision medicine that estimate simple, interpretable treatment decision rules which can be applied clinically (Rudin (2019)). We estimated optimal decision rules by implementing policy trees (Zhou *et al.* (2022)) and decision lists (Zhang *et al.* (2015)).

The decision list algorithm results in decision rules in the form of a sequence of "if-then" statements, while the policy tree algorithm results in a decision tree. For both algorithms, the "depth" parameter is synonymous with the number of steps involved in using a decision rule to identify an optimal intervention: For a decision lists, depth is the number of "if-then" statements, and for a policy tree, depth is the number of layers of the tree. While the decision rule provides estimation of optimal therapies, it is not meant to provide deterministic treatment "assignments" or clinical recommendations for actual patients, but rather, to explore subgroups and the markers that define the subgroups, as treatment effects may vary based on their values. We refer to the subgroups as the CGM versus BGM treatment rule subgroups.

3.2.4 Decision Rule Evaluation

We evaluated the clinical benefit of potential decision rules using the value function, a scalar performance measure which estimates the expected outcome in the study population if every participant was to receive their optimal intervention as estimated by the rule. For the "total time in hypoglycemia" outcome, a lower value suggests better performance of the rule, as it is indicative of less time spent in hypoglycemia. For the "reduction in time spent in hypoglycemia" outcome, a higher value suggests better performance, as it indicates a larger decrease in percentage of time in hypoglycemia over the duration of the trial.

Estimation of the value of the resultant decision rules was carried out over a nested 5×5 cross-validation scheme. An initial outer loop was used to split the data into 5 training and test folds, and within the training set, an inner 5-fold cross-validation loop was used for algorithm selection and tuning of the depth parameter. Once the optimal model (algorithm and depth) was selected using the inner loop, for evaluation, its value was estimated on the held-out data using 5-fold cross-validation in the outer loop. This resulted in an honest cross-validation process in which the data by which the final model was selected and evaluated was kept separate from the training data.

For comparison, we estimated the value function for a "CGM-only" rule (i.e., all participants receive CGM) through 5-fold cross-validation in the outer loop. As context for the value of the optimal decision rule and the CGM-only rule, we also estimated the value of a "BGM-only" rule.

3.2.5 Subgroup Characterization

Using descriptive statistics, we compared the characteristics of the participants in the CGM versus BGM treatment rule subgroups, as per the optimal decision rule for the primary outcome. P-values for differences in means were calculated with a 2-sample t-test and differences in proportions with a 2-proportion Z-test at the 0.05 and 0.10 significance levels. All analyses were conducted in R, version 3.6.

3.2.6 Data Availability and Resource Sharing

The WISDM data are publicly available through the Jaeb Center for Health Research (JCHR). All codes to replicate the secondary analyses described herein are available on GitHub at the following link: `https://github.com/kushshah1/WISDM-Precision-Medicine`. This repository includes all steps of the precision medicine pipeline, including data processing (compilation of variables, treatment infromation, and clinical outcome, along with omission of patients lost to follow-up and selection of final features for analysis), model fitting (implementation of interpretable precision medicine algorithms - decision list and policy tree), model evaluation (empirical value function approximation for resulting ITRs, along with nested K-fold cross validation for parameter tuning, optimal model selection, and optimal model evaluation on held-out test set), and compilation of results (optimal decision rule, study participant characteristics, visualization of differential treatment effects, training/validation set estimates of potential decision rules, and held-out test set evaluation of optimal rule). Step-by-step instructions to run the workflow can be found in the online documentation for the GitHub repository.

3.3 Results

The optimal decision rule was found to be a decision list with a depth of 3 (Table 3.1). The decision rule is depicted in Figure 3.1. The first step of the decision rule moved WISDM participants with baseline time-below range $>$1.35% and no detectable C-Peptide to the CGM subgroup (n = 139), and the second step moved WISDM participants with baseline time-below range of $>$6.45% to the CGM subgroup (n = 18). The remaining participants (n = 37) were left in the BGM subgroup.

The characteristics of the 157 (81%) WISDM participants in the CGM versus the 37 (19%) WISDM participants in the BGM subgroup are shown in Table 3.2. The median baseline time spent in hypoglycemia in the CGM subgroup was 6.7% (interquartile range [IQR]: 6.5%) compared with 1.2% (IQR: 2.5%) in the BGM subgroup. The mean proportion of patients with de-

Table 3.1: Training and validation set value estimates of potential decision rules, along with test set evaluation of final rule, for the primary outcome (% time in hypoglycemia). The "optimal method" was decided as the method with optimal (lowest) inner validation set value; only that method was evaluated on the held-out test set in order to ensure honest cross validation.

Policy Tree - Parameter Tuning			Final Evaluation (of optimal method)
Depth	*Training Set Value*	*Inner Validation Set Value*	on Held-Out Test Set
1	2.87	3.07	–
2	2.66	3.09	–
3	2.46	3.29	–
Decision List - Parameter Tuning			
Depth	*Training Set Value*	*Inner Validation Set Value*	
1	2.87	3.13	–
2	2.87	3.13	–
3	2.87	**3.01**	**2.98***

*Note, for comparison, that the estimated value of "CGM-only" rule on the held-out test set was 3.09%.

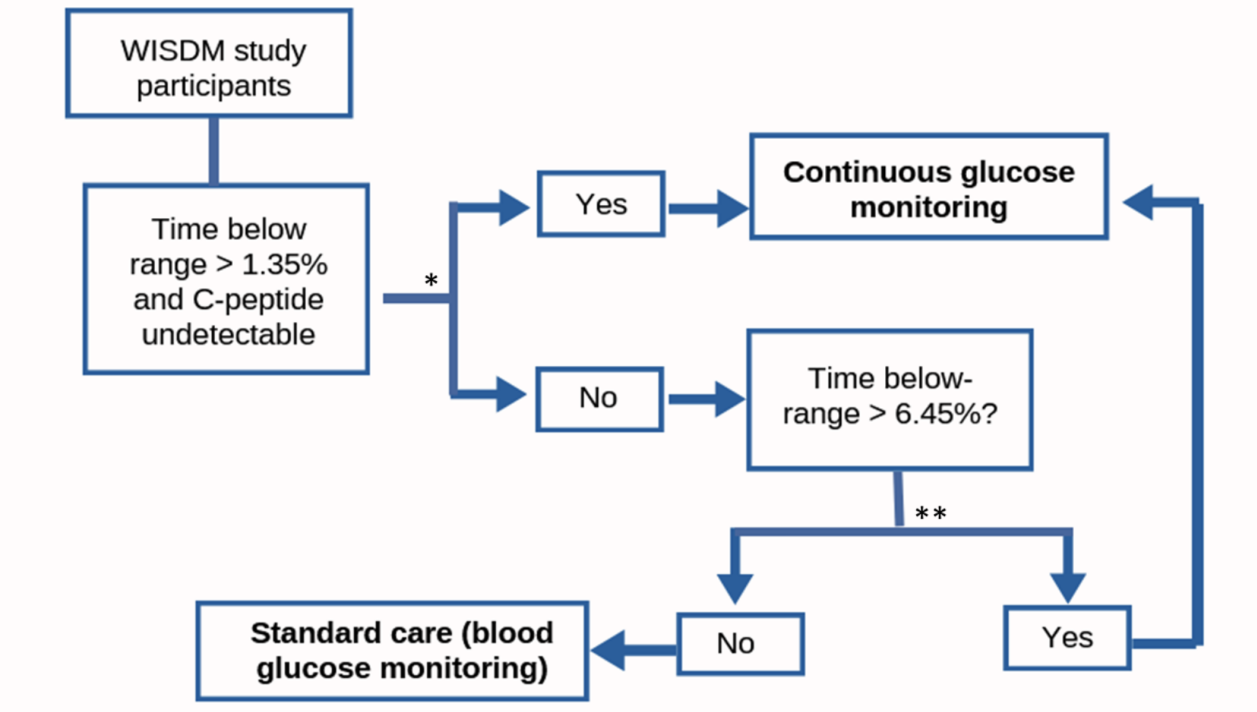

Figure 3.1: Visualization of the decision rule. *Denotes the first split, at which point 139 participants are assigned to the CGM group. **Denotes the second split, at which point 18 participants were assigned to the CGM group. The remaining 37 participants were assigned to the BGM group.

Table 3.2: Characteristics of study participants, stratified by decision rule subgroup. P values for differences in means were calculated with a 2-sample t-test and differences in proportions with a 2-proportion Z-test. Abbreviations: SD, standard deviation; IQR, interquartile range.

Characteristic, n (%) or mean (SD)	**Decision Rule Subgroup**		*P-value*
	CGM (n=157)	*BGM (n=37)*	
Moderating Markers			
Hypoglycemia (time <70 mg/dL), %; median (IQR)	6.7 (6.5)	1.2 (2.5)	<.0001[a]
Hypoglycemia (time <70 mg/dL), %	7.8 (5.0)	2.0 (1.7)	<.0001[a]
Detectable C-peptide	18 (11.5%)	28 (75.7%)	<.0001[a]
Demographic Characteristics			
Age, years	67.9 (5.6)	69.4 (6.0)	.17
Diabetes duration, years	38.3 (14.4)	25.3 (17.4)	.0001**
Age at diagnosis, years	29.6 (15.3)	44.1 (18.8)	<.0001**
Male sex	72 (45.9%)	21 (56.8%)	.31
Non-Hispanic ethnicity	152 (96.8%)	36 (97.3%)	1
White race	149 (94.9%)	34 (91.9%)	.75
Highest education			
Less than a bachelor's degree	66 (42.0%)	9 (24.3%)	.07*
Bachelor's degree	46 (29.3%)	16 (43.2%)	.15
Graduate or professional degree	45 (28.7%)	12 (32.4%)	.80
Health insurance			
Private	43 (27.4%)	9 (24.3%)	.86
Private and Medicare	53 (33.8%)	14 (37.8%)	.78
Medicare/other	61 (38.9%)	14 (37.8%)	1
Clinical Characteristics			
Insulin pump use	89 (56.7%)	13 (35.1%)	.03**
Coefficient of variation, %	43.1 (5.7)	35.4 (5.2)	<.0001**
Total daily insulin dose, units/kg	0.56 (0.20)	0.51 (0.283)	.32
Screening HbA1c, %	7.5 (0.9)	7.8 (0.9)	.13
Body mass index, kg/m^2	27.3 (4.5)	26.3 (5.0)	.30
$\geqslant$ 1 severe hypoglycemia event in past 12 months	25 (15.9%)	3 (8.1%)	.34
$\geqslant$ 1 diabetic ketoacidosis event in past 12 months	8 (5.0%)	0 (0%)	.35
Time with glucose in range of 70-180 mg/dL, %	56.6 (13.1)	54.0 (17.3)	.39

[a]Significance expected since decision rule is based on this marker.
*P $<$ .1, **P $<$ .05

tectable C-peptide levels in the CGM subgroup was 11.5% compared with 75.7% in the BGM subgroup. Compared with the BGM subgroup, the CGM subgroup also had, on average, a longer diabetes duration (38.3 years vs 25.3 years; P = .0001) and younger age at diagnosis (29.6 years vs 44.1 years; P $<$ .0001), in addition to a higher proportion with insulin pump use (56.7% vs 35.1%; P = .03) and larger mean %CV (43.1% vs 35.4%; P $<$ .0001). There were no significant differences in age, sex, or other demographic factors, nor total daily insulin dose, HbA1c, history of hypoglycemia in the preceding 6 months, or history of diabetic ketoacidosis (DKA) in the preceding 6 months.

The optimal decision rule was estimated to result in a total time in hypoglycemia of 2.98% (standard error of the mean [SEM]: 0.32%) across the full study population (Table 3.1), compared with 3.09% (SEM: 0.24%) with the "CGM-only" rule. For context, use of BGM-only was estimated to correspond with a total time in hypoglycemia of 6.23% (SEM: 0.44%).

The results of the secondary analyses, which focused on the secondary outcome, change in hypoglycemia, are shown in Appendix C. For this outcome, the optimal decision rule was found to be a policy tree algorithm with a depth of one, and the rule suggests that study participants with baseline %CV $>$34% would experience a greater reduction in hypoglycemia using CGM compared with BGM.

3.4 Discussion

We used data from the WISDM study and machine learning methods to discover how patient characteristics are associated with the estimated treatment effects of CGM on hypoglycemia among older adults with T1D. Our results suggest that there are 2 distinct subgroups of the WISDM participants - those with baseline time-below range $>$1.35% and no detectable C-peptide levels, and those with baseline time-below range of $>$6.45% - for whom CGM was estimated to result in lower hypoglycemia at the follow-up period compared with BGM. Together, these subgroups represented the majority of the study sample, with distinguishing features including a lower proportion of detectable C-peptide, higher glycemic variability, longer disease duration,

and higher proportion of insulin pump use. As such, we elucidate the moderating factors to identify for whom treatment effect is expected to be greater or smaller, and importantly, markers for the subgroups of older adults for whom CGM may be a particularly effective intervention for reducing hypoglycemia.

The benefits of CGM for older adults are both significant and well-established, and CGM now represents standard-of-care for T1D (Pratley *et al.* (2020); Ruedy *et al.* (2017); ADA Professional Practice Committee (2022a); Kirkman *et al.* (2012); ADA Professional Practice Committee (2022c); Holt *et al.* (2021); Toschi and Munshi (2020); Forlenza *et al.* (2017)). This is reflected in our results as well, where the majority of WISDM participants were estimated to benefit from CGM over BGM to minimize time in hypoglycemia. It is also important to note the small difference in mean expected hypoglycemia with the optimal treatment rule versus CGM-only rule (2.98% vs 3.09%), which underscores the value of CGM across the population. The decision rule herein also only focused on hypoglycemia, and thus does not focus on other glycemic or patient-oriented benefits associated with CGM. Above all, estimating a decision rule does not provide rationale for restricting an evidence-based therapy; rather, it offers insight into how the same therapy may results in different outcomes for individuals, as well as the subgroups for whom benefits in terms of one outcome may be more versus less pronounced, as well as the physiology and clinical histories that may underlie those differences.

Furthermore, despite the benefits and value of CGM for older adults, there remains a proportion of patients who do not use therapeutic CGM as part of their T1D self-management (Munshi *et al.* (2022); Divan *et al.* (2022); Gubitosi-Klug *et al.* (2022)). The decision to initiate CGM may be shaped by baseline glycemic management and priorities (Munshi *et al.* (2022)), perceived benefits and burdens of the technology (Divan *et al.* (2022)), or the impact of age-specific challenges to integrating technology into care (Krishnaswami *et al.* (2020)). As adults get older, managing the device and data can be perceived as challenging, and preferences surrounding technology use may evolve (Toschi and Munshi (2020)). Understanding heterogenous treatment effects and markers thereof may aid providers in identifying patients for whom CGM technology is criti-

cal to mitigate the risk of hypoglycemia. Providing more granular, individualized estimation of the value of CGM may help providers to find ways to balance clinical outcomes and individual preferences through shared decision-making.

The decision rule suggests that 2 groups of study participants would experience a lower proportion of time spent in hypoglycemia by using CGM compared with BGM: (1) those with baseline hypoglycemia $>1.35\%$ and with no detectable C-peptide, and (2) those with baseline hypoglycemia $>6.45\%$. The decision rule thus underscores an important link between baseline hypoglycemia and risk for future events, which may be mitigated by CGM use. It is worth noting that in a clinical setting, any duration of hypoglycemia greater than 1% may indicate a need for CGM for the reduction in non-severe hypoglycemia and prevention of severe episodes in the future (ADA Professional Practice Committee (2022a)), and 2023 Standards of Care in Diabetes recommend CGM to reduce hypoglycemia among older adults with T1D. An inverse association between age at diagnosis and pump use has been described previously (Casu *et al.* (2020)), and one possible explanation between the higher proportion of insulin pump use in the subgroup may reflect existing efforts to mitigate the burden of diabetes management, though this association cannot be considered causal.

Together, this subgroup of patients was also distinguished by a longer disease duration and higher glycemic variability at baseline, which may reflect, together, a phenotype of very long-standing T1D. An association between glycemic variability and severe hypoglycemia has been reported in older adults previously (DuBose *et al.* (2016); Weinstock *et al.* (2016)). A high CV ($>36\%$) has also been associated with duration of time spent in non-severe hypoglycemia among older adults with T1D (Toschi *et al.* (2020)). Interestingly, the a priori subgroup analysis of the WISDM study showed that among participants in the CGM arm, a higher baseline %CV was associated with a greater treatment effect (Pratley *et al.* (2020)). The results of the decision rule analysis are related, but not redundant, with that finding, as the subgroups reported herein are generated from the algorithm comparing predicted effects between CGM and BGM, for each individual within the data set, to assign them to an optimal therapy group. Our finding that de-

mographic features, markers of socioeconomic position, and HbA1c levels were not significantly different across subgroups is also consistent with the main trial analyses, which did not find other baseline characteristics, including age ($<$70 vs $\geqslant$70 years), socioeconomic status, presence of cognitive impairment, or HbA1c value to interact with the treatment effect (on hypoglycemia) (Pratley *et al.* (2020)).

Our secondary analysis showed that the moderating factors may shift based on the outcome the decision rule aims to optimize, where baseline glycemic variability greater than 34% was a marker in determining patients for whom CGM was estimated to produce the greatest reduction in hypoglycemia. A point of interest is that the %CV threshold from the entirely data-driven decision rule aligns closely with existing cutoffs for glycemic variability to minimize risk of hypoglycemia (Danne *et al.* (2017)), where a %CV of $<$34% has been specified as an optimal threshold to reduce such episodes (Danne *et al.* (2017); Gomez *et al.* (2019); Monnier *et al.* (2020)) and a %CV target of $<$33% for "additional protection against hypoglycemia" in insulin-treated diabetes (ADA Professional Practice Committee (2022b)). An association between glycemic variability and severe hypoglycemia has been reported in older adults previously (DuBose *et al.* (2016); Weinstock *et al.* (2016)).

Future studies can confirm the utility of hypoglycemia and C-peptide as a marker for heterogeneous effects of CGM on hypoglycemia in older adults. A potential implication of the decision rule is a role for laboratory values (i.e., C-peptide) and diagnostic CGM to inform decision-making surrounding therapeutic CGM, particularly for older adults who do not wish to permanently integrate new technology into their T1D self-management routines. Future studies may also investigate decision rules for CGM to optimize other glycemic outcomes, such as TIR. To understand the degree to which moderating markers generalize across age, it may be important to evaluate markers of CGM effects on hypoglycemia in different age groups, including published clinical trials in adolescents and young adults, such as the CGM Intervention in Teens and Young Adults with T1D (CITY) trial.

One of the most significant limitations of the study relates to the fact that the WISDM study participants do not reflect the same level of heterogeneity of the larger population of older adults with T1D. Not only is the sample a "young" older adult population, including individuals between 60 and 65 years of age, but this study population is also reflective of robust, highly-educated, and majority non-Hispanic White older adult patients. The value of the precision medicine-based statistical methods applied herein may be greater in the setting of a wider population, including variability in cognitive and physical function, living status, and diabetes self-management.

Other limitations of the study include its exploratory nature, small sample size, and the focus on hypoglycemia only. The WISDM study demonstrated treatment effects for hyperglycemia (glucose levels $>$180 mg/dL, $>$250 mg/dL, and $>$300 mg/dL), mean glucose concentration, and glycemic variability (Pratley *et al.* (2020)); each of these outcomes comprise important clinical benefits for patients. Several potential important variables were lacking for the analysis; data on cognitive status were not included, in addition to physical function, other behavioral or psychosocial measures such as impaired awareness of hypoglycemia, or living situation. Their inclusion may change the decision rule (Flatt *et al.* (2022)).

The strengths of this analysis include the application of novel machine learning methods from the field of statistical precision medicine. These methods are entirely data-driven and thus permit the discovery of subgroups of interest with characteristics/features or a combination thereof that were not defined a priori. Though benefits of CGM for older adults are both significant and well-established, our results may serve as the basis for future studies to explore heterogeneous treatment effects and their markers across a range of outcomes, in more diverse patient populations, and considering variables with known relevant to the clinical care of older adults, such as frailty status, cognitive impairment, and living situation. In the meantime, the results may also aid providers in identifying patients for whom technology is critical to mitigate the risk of hypoglycemia and support enhanced shared decision-making surrounding the individualized benefits of CGM for individual older T1D adults.

3.5 Conclusion

A precision medicine approach was utilized in this analysis in order to contribute towards two goals that were specified at the outset: *decision support* and *scientific depth*. In this section, we summarize the Discussion section by briefly restating the direct contributions towards these goals.

Towards the goal of *decision support*, the results from this analysis may be used to augment decision-making where resource allocation/prioritization is needed (patients for whom CGM is critical and immediately necessary vs. patients for whom the decision may be more relaxed). Additionally, this analysis opens the door to other factors being considered in treatment selection (e.g. the burden of change for certain patients who may struggle with the complexity of new technology), rather than only hypoglycemia. To this end, the methodology from Chapter 2 can be utilized in this context to simultaneously optimize hypoglycemia as well as other clinical outcomes (e.g. hyperglycemia, time in range, etc.) and psychosocial measures (e.g. fear of hypoglycemia, physical/mental functioning, etc.) which were also documented in this clinical trial.

Towards the goal of *scientific depth*, we have elucidated demographic and clinical markers which moderate CGM's success. This provides the basis for using diagnostic CGM (CGM use over a baseline trial period) to inform therapeutic CGM decisions. Additionally, the fact that the optimal decision rule was unable to improve significantly over the "CGM-only" rule also underscores the value of CGM at an individual level (whereas previous studies also showed the value of CGM, but at the average population level).

CHAPTER 4: FUTURE DIRECTIONS: AUTOENCODER-BASED REPRESENTATION LEARNING FOR HIGH DIMENSIONAL PRECISION MEDICINE

4.1 Introduction

In precision medicine settings, it is a challenge to tailor treatment based on high-dimensional patient data. Another layer of difficulty is added with data which is unstructured or correlated. One example of this is X-ray images, where pixels of an image have spatial dependence, and thus groups of pixels are needed to provide relevant information. Another example is continuous glucose monitoring (CGM) counts in diabetes. A patient's baseline CGM data, for example, may consist of weeks worth of continuous reporting of blood glucose levels at intervals of 1 to 15 minutes (Bergenstal (2018)). CGM data is time-correlated, and a single time point provides little value; trends and patterns in a subject's blood glucose levels must be understood in order to derive useful results.

4.1.1 General Approaches for Working with High-Dimensional Data

Many statistical and machine learning tools provide a framework for working with a large set of features. Variable relevance ranking approaches like random forest Mean Decrease Gini (MDG) (Han *et al.* (2016)) or Brownian distance covariance (BDC) (Szkely and Rizzo (2009)) can be used to order features according to various measures of importance, from which a pre-specified top p variables can be selected. BDC, for example, can be used to evaluate the nonlinear dependence between vectors of individual predictors and vectors of a response (or groups of responses), resulting in p-values for independence tests which reflect the predictive power of each feature. MDG measures how each covariate contributes to the homogeneity of nodes and leaves in a random forest model, resulting in a predictive importance ranking. However, variable

ranking approaches used directly in relation to an outcome of interest provide information on predictive relevance, but not prescriptive relevance, the latter of which is needed to tailor treatment. Moreover, these approaches work with variables that can be independently ranked, but not with unstructured data like CGM or image inputs.

For high dimensional and possibly unstructured data, there exist approaches which find low-dimensional representations of the inputs, such as Principal Component Analysis (PCA) (Ringnr (2008)), t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton (2008)), or autoencoders (Wang *et al.* (2016)). Autoencoders are neural networks which have an input layer, hidden middle layer, and output layer, and attempt to copy input to output with minimal reconstruction error. Undercomplete autoencoders are built with a middle layer with smaller dimension than the input, with a goal of learning a representation of the input which captures its most salient features Goodfellow *et al.* (2016). Examples of autoencoders in practice include Homayouni *et al.* (2020), who use an LSTM-autoencoder to detect anomalies in time-series data, and Gomez-Bombarelli *et al.* (2018), who use a variational autoencoder (VAE) to generatively find points in a continuous, low-dimensional latent space (hidden layer) that correspond to molecules in a discrete input space with desirable properties. Rashid *et al.* (2021a) use autoencoders to perform unsupervised feature extraction of X-ray images, which are then fed into a supervised convolutional neural network architecture for COVID-19 prediction.

4.1.2 Traditional Precision Medicine Techniques for High-Dimensional Data

It is difficult to directly input high-dimensional data as covariates in nonparametric mean response models, as this would lead to overfitting and poor performance on held-out test data. Traditional precision medicine approaches for handling a large feature set focus on two steps: (1) Derive a smaller feature set through variable ranking or representation learning approaches, as described in Section 4.1.1. (2) Input these features as well-defined covariates into the traditional ITR estimaton framework.

Examples of such an approach include Lou *et al.* (2018), who discuss derivation of a smaller feature set through correlation-based screening (Li *et al.* (2012)) before fitting an ITR, and Zhu *et al.* (2017), who perform pre-screening of genes based on marginal variance prior to implementing a subgroup identification approach. These approaches suffer from the fact that derived variables are not necessarily prescriptive; they were selected with for their usefulness in outcome prediction (prognostic variables) or input reconstruction (representative features), but not for their ability to help with personalized treatment decision making. Rashid *et al.* (2021b) extend BDC to prescriptive selection in a precision medicine pipeline, but the method is still limited to structured input list of covariates. Additional methods that perform prescriptive screening are Zhou and Kosorok (2017), which proposes an adaptive causal k-nearest neighbor regime that simultaneously performs metric selection and variable selection, and Athey and Imbens (2016), which discusses various ways for splitting causal trees on prescriptive covariates rather than predictive ones.

Nezhad *et al.* (2016) propose a patient risk stratification approach with feature space reduction through autoencoders and supervised risk classification occur sequentially. Such frameworks, with autoencoders being trained independently of the ensuing modeling approach, produce a feature set which is biased towards accurate reconstruction of the input only. However, a framework where autoencoder input reconstruction is optimized simultaneously with ITR value maximization using backpropagation and stochastic gradient descent (SGD) could produce a feature set which is both representative and prescriptive. The combination of these two properties could be powerful in the treatment recommendation framework.

4.1.3 Deep Learning in Precision Medicine

More recently, deep learning has begun to be utilized in the precision medicine pipeline. Liang *et al.* (2018) use a neural network classifier within an OWL-like ITR estimation framework for mutiple combination therapies in order to avoid local minimizers and achieve scalable computation amidst a nonsmooth/nonconvex loss function. The proposed solution to the optimization

problem involves backpropagation and SGD as described in Goodfellow *et al.* (2016). Here, the deep learning framework is used mainly as a method to minimize objective functions for ITR estimation.

Jiang (2020) builds on the work of Liang *et al.* (2018) and proposes deep doubly robust outcome weighted learning (DDROWL), a variant of RWL that can be solved by deep neural networks, thereby achieving double robustness in the deep learning optimization framework. This method exhibits flexibility and accommodates large covariates spaces which cannot be modeled parametrically due to dimensionality, unknown model specification, etc. DDROWL is one of the first methods to directly consume unstructured data with a large number of dimensions into the precision medicine workflow. Specifically, it analyzes input MRI images of large dimension (e.g. 162x30x4096), which many existing methods are not able to handle. Mi *et al.* (2019) also develops a deep neural network (DNN) ensemble based on the original OWL optimization problem and analyzes sequencing data from around 1000 human cancer cell lines for optimal treatment allocation.

4.1.4 Proposed Method

Current ITR estimation approaches lack the ability to create learned representations of prescriptive variables from unstructured, high-dimensional data. The question that motivates this paper is: Can we develop a one-step approach which directly uses high-dimensional, unstructured features in the precision medicine pipeline, without needing to make an initial conversion to low dimensions (which runs the risk of losing prescriptive information)?

In learned representations of high-dimensional inputs, the representation is evaluated by reconstruction error. That is, how well can the low-dimensional representation be used to reconstruct the input data? This goal is the motivating factor behind autoencoders, an unsupervised technique which utilizes a neural network architecture with a built-in bottleneck, thus forcing a compressed representation of the original input (Sainath *et al.* (2012)).

In precision medicine, three broad classes of covariates are specified: prognostic, moderating, and prescriptive variables. The goal of identifying prescriptive variables is fundamental because only prescriptive variables are useful in selecting treatments which maximize value (Kosorok and Laber (2019)). The value of a decision rule is the expected population mean outcome if all patients were to follow the rule; this is the key metric by which optimality of a rule is gauged. The identification of prescriptive variables is what gives an ITR the potential to have higher value than a "one-size-fits-all" approach.

We propose methodology for a novel approach for high-dimensional data that simultaneously (1) uses an autoencoder to find a low-dimensional representation of the input which minimizes reconstruction error, and (2) fits an ITR to maximize treatment benefit. We combine two objective functions (reconstruction loss minimization and value function maximization) into a single optimization procedure that includes a parameter to control the trade-off between these two goals. Thus, by balancing reconstruction loss minimization and value function maximization, this approach finds a *prescriptive* low-dimensional representation of the input data (features which are both representative of the original input and prescriptive for ITR estimation).

4.2 Methodology

For the autoencoder framework, let $\boldsymbol{X}$ represent patient covariates (input layer), $g_{\boldsymbol{\theta}_1}(\boldsymbol{X})$ represent the encoded data (hidden layer), and $r_{\boldsymbol{\theta}_2}\left\{g_{\boldsymbol{\theta}_1}(\boldsymbol{X})\right\}$ represent the decoded data (output layer). Then, let $L_1\left[\boldsymbol{X}, r_{\boldsymbol{\theta}_2}\left\{g_{\boldsymbol{\theta}_1}(\boldsymbol{X})\right\}\right]$ be the mean squared-error autoencoder reconstruction loss. For value maximization, we use the outcome weighted learning (Zhao *et al.* (2012)) loss function with the hidden layer, $g_{\boldsymbol{\theta}_1}(\boldsymbol{X})$, as covariate information:

$$L_2\left\{g_{\boldsymbol{\theta}_1}(\boldsymbol{X}), A, R\right\} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\widehat{\pi}(A_i, \boldsymbol{X}_i)}\left(1 - A_i\left[f_{\boldsymbol{\theta}_3}\left\{g_{\boldsymbol{\theta}_1}(\boldsymbol{X}_i)\right\}\right]\right)^{+} + \lambda_n||f_{\boldsymbol{\theta}_3}||^2$$

We can then combine L_1 and L_2 into one optimization procedure:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, \alpha_n L_1\left[\boldsymbol{X}, r_{\boldsymbol{\theta}_2}\left\{g_{\boldsymbol{\theta}_1}(\boldsymbol{X})\right\}\right] + L_2\left\{g_{\boldsymbol{\theta}_1}(\boldsymbol{X}), A, R\right\}, \tag{4.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \boldsymbol{\theta_3})$ and Θ reflects the joint parameter space for g, r, and f.

In (4.1), the reconstruction loss serves as a regularization term because of the bottleneck forced by the middle layers of the neural network. This helps focus on finding representative features which are useful in recovery of the input. (4.1) can be solved via a backpropagation algorithm as per the architecture in Figure 4.1, which uses CGM input data as an example.

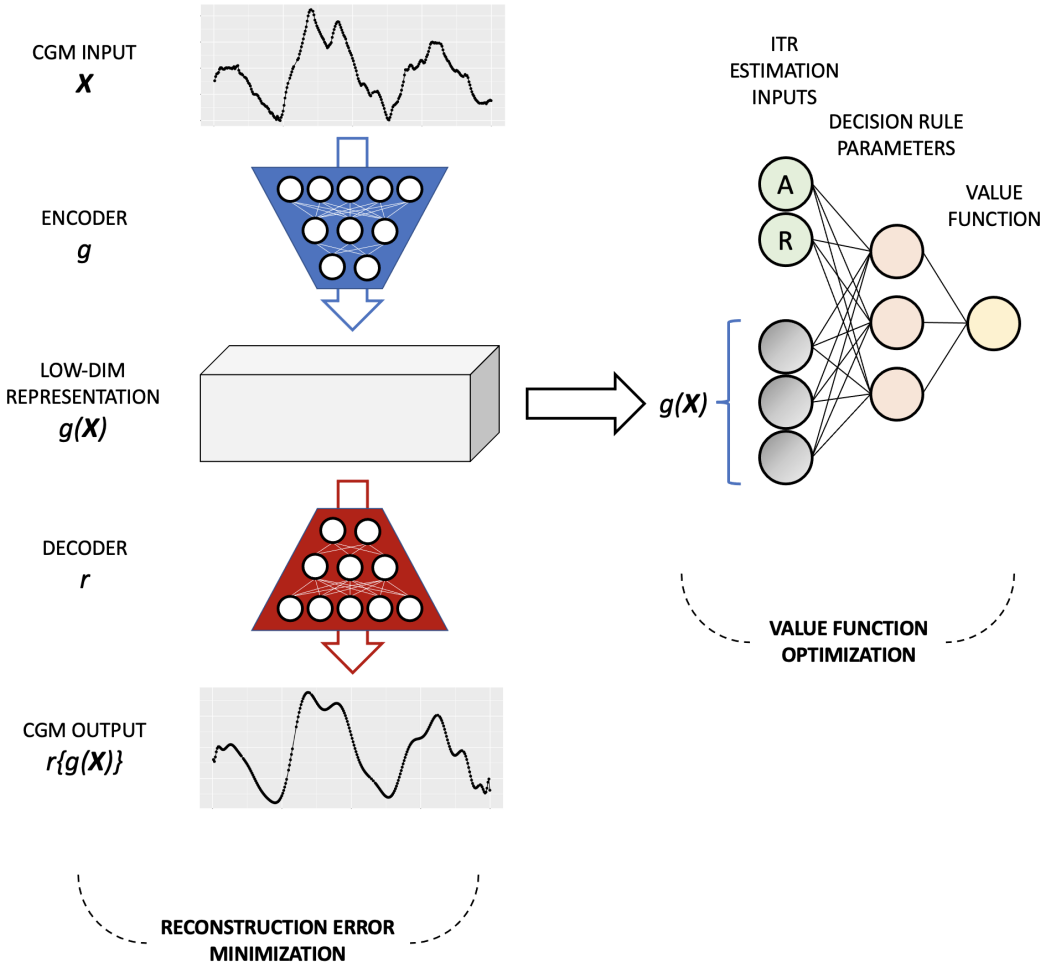

Figure 4.1: Neural network architecture for combined reconstruction loss minimization and value function maximization. The ITR estimation portion of the optimization function takes the low dimensional representation from the autoencoder, $g(\boldsymbol{X})$, as input, along with patient treatment and observed outcome information. This figure is inspired by Gomez-Bombarelli *et al.* (2018).

4.3 Parameter Tuning and Evaluation

A simple method for parameter tuning is to perform a grid search and select the value of α_n that results in an ITR which maximizes the value function on a held-out test set. However, it is likely that many estimated ITRs will perform similarly. In that case, we can pick the largest value of α_n (out of all values of α_n) for which the resulting value function is within some specified percentage of the optimal value function. Thus, we select an ITR with near-optimal performance and an emphasis on reconstruction error minimization, resulting in a more robust decision rule.

4.4 Discussion

In general, traditional dimension reduction approaches which precede ITR estimation do not consider value maximization in the feature derivation step. The approach presented in this paper varies considerably from the traditional approach, because we present a single loss function framework to simultaneously handle feature derivation and ITR value maximization. By including the latter in the dimensionality reduction framework, our methodology biases the latent space representation of the original input to be more helpful in treatment prescription than outcome prediction. This framework contributes to methodology built to consume image data, CGM data, sound waves, etc. directly in the precision medicine pipeline.

APPENDIX A: SUPPORTING MATERIALS FOR CHAPTER 1

A.1 Proofs of Propositions

This section includes proofs of Proposition 1.1 and Proposition 1.2.

Proof of Proposition 1.1. Let $l_n(\boldsymbol{\beta}) = \mathbb{P}_n \left\{ \dfrac{(2RA - \boldsymbol{X}^\top \boldsymbol{\beta})^2}{w(A, \boldsymbol{X})\pi(A, \boldsymbol{X})} \right\}$, and define the first and second derivatives of l_n:

$$\frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbb{P}_n \left\{ \frac{(2RA - \boldsymbol{X}^\top \boldsymbol{\beta})\boldsymbol{X}}{w(A, \boldsymbol{X})\pi(A, \boldsymbol{X})} \right\}, \tag{A.1}$$

$$\frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = 2\mathbb{P}_n \left\{ \frac{\boldsymbol{X}\boldsymbol{X}^\top}{w(A, \boldsymbol{X})\pi(A, \boldsymbol{X})} \right\}. \tag{A.2}$$

Assume that $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$. Using Taylor Series approximations:

$$0 = \frac{\partial l_n(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} + o_p(1) \right\} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right). \tag{A.3}$$

From this, we have:

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) &= -\sqrt{n} \left\{ \frac{\partial^2 l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \right\}^{-1} \frac{\partial l_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + o_p(1) \\
&= \sqrt{n} \left\{ \mathbb{P}_n \frac{\boldsymbol{X}\boldsymbol{X}^\top}{w(A, \boldsymbol{X})\pi(A, \boldsymbol{X})} \right\}^{-1} \left\{ \mathbb{P}_n \frac{(2AR - \boldsymbol{X}^\top \boldsymbol{\beta})\boldsymbol{X}}{w(A, \boldsymbol{X})\pi(A, \boldsymbol{X})} \right\} + o_p(1) \\
&= E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top}{w(A, \boldsymbol{X})\pi(A, \boldsymbol{X})} \right]^{-1} \left\{ \sqrt{n}\mathbb{P}_n \frac{\epsilon \boldsymbol{X}}{w(A, \boldsymbol{X})\pi(A, \boldsymbol{X})} \right\} + o_p(1), \text{ and}
\end{aligned} \tag{A.4}$$

$$
\begin{aligned}
&var\left\{\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\} \\
&\quad= E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})}\right]^{-1} var\left[\frac{\epsilon\boldsymbol{X}}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})}\right] E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})}\right]^{-1} + o(1) \\
&\quad= E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})}\right]^{-1} E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top E\left(\epsilon^2 \mid A,\boldsymbol{X}\right)}{w(A,\boldsymbol{X})^2\pi(A,\boldsymbol{X})^2}\right] E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})}\right]^{-1} + o(1) \\
&\quad= E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})}\right]^{-1} E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top\sigma_0^2(A,\boldsymbol{X})}{w(A,\boldsymbol{X})^2\pi(A,\boldsymbol{X})^2}\right] E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})}\right]^{-1} + o(1) \\
&\quad= \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} + o(1),
\end{aligned}
\tag{A.5}
$$

since $\sigma_0^2(A,\boldsymbol{X}) = var\left(\epsilon \mid A,\boldsymbol{X}\right) = E\left(\epsilon^2 \mid A,\boldsymbol{X}\right)$. For ease of notation, we will let $\boldsymbol{A} = E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})}\right]$ and $\boldsymbol{B} = E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top\sigma_0^2(A,\boldsymbol{X})}{w(A,\boldsymbol{X})^2\pi(A,\boldsymbol{X})^2}\right]$. Letting $w_t(A,\boldsymbol{X}) = w(A,\boldsymbol{X}) + ts(A,\boldsymbol{X})$ be a perturbation of the weights, we take a functional derivative and show that no matter the choice of $s(A,\boldsymbol{X})$, the first derivative of the variance expression is zero when $w(A,\boldsymbol{X}) = \frac{\sigma_0^2(A,\boldsymbol{X})}{\pi(A,\boldsymbol{X})}$ (e.g. the variance is minimized). The variance term can be written as:

$$
\begin{aligned}
&var\left\{\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right\} \\
&\quad= E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top}{w_t(A,\boldsymbol{X})\pi(A,\boldsymbol{X})}\right]^{-1} E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top\sigma_0^2(A,\boldsymbol{X})}{w_t(A,\boldsymbol{X})^2\pi(A,\boldsymbol{X})^2}\right] E\left[\frac{\boldsymbol{X}\boldsymbol{X}^\top}{w_t(A,\boldsymbol{X})\pi(A,\boldsymbol{X})}\right]^{-1} + o(1) \\
&\quad= \boldsymbol{A}_t^{-1}\boldsymbol{B}_t\boldsymbol{A}_t^{-1} + o(1).
\end{aligned}
\tag{A.6}
$$

The derivatives of $\boldsymbol{A}_t^{-1}$ and $\boldsymbol{B}_t$ are:

$$
\frac{\partial\boldsymbol{A}_t^{-1}}{\partial t} = -\boldsymbol{A}_t^{-1}\frac{\partial\boldsymbol{A}_t}{\partial t}\boldsymbol{A}_t^{-1} = -\boldsymbol{A}_t^{-1}E\left\{\frac{-\boldsymbol{X}\boldsymbol{X}^\top s(A,\boldsymbol{X})}{w_t(A,\boldsymbol{X})^2\pi(A,\boldsymbol{X})}\right\}\boldsymbol{A}_t^{-1} = \boldsymbol{A}_t^{-1}\boldsymbol{C}_t\boldsymbol{A}_t^{-1}, \tag{A.7}
$$

$$
\frac{\partial\boldsymbol{B}_t}{\partial t} = E\left\{\frac{-2\boldsymbol{X}\boldsymbol{X}^\top\sigma_0^2(A,\boldsymbol{X})s(A,\boldsymbol{X})}{w_t(A,\boldsymbol{X})^3\pi(A,\boldsymbol{X})^2}\right\}, \tag{A.8}
$$

where $\boldsymbol{C}_t = E\left\{\dfrac{\boldsymbol{X}\boldsymbol{X}^\top s(A,\boldsymbol{X})}{w_t(\boldsymbol{X})^2\pi(A,\boldsymbol{X})}\right\}$. Having noted this, we can take the first derivative of the variance expression:

$$\begin{aligned}
&\frac{\partial}{\partial t}\left[var\left\{\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\}\right] \\
&\quad= \frac{\partial \boldsymbol{A}_t^{-1}}{\partial t}\boldsymbol{B}_t\boldsymbol{A}_t^{-1} + \boldsymbol{A}_t^{-1}\frac{\partial \boldsymbol{B}_t}{\partial t}\boldsymbol{A}_t^{-1} + \boldsymbol{A}_t^{-1}\boldsymbol{B}_t\frac{\partial \boldsymbol{A}_t^{-1}}{\partial t} \\
&\quad= \boldsymbol{A}_t^{-1}\boldsymbol{C}_t\boldsymbol{A}_t^{-1}\boldsymbol{B}_t\boldsymbol{A}_t^{-1} + \boldsymbol{A}_t^{-1}E\left\{\frac{-2\boldsymbol{X}\boldsymbol{X}^\top\sigma_0^2(A,\boldsymbol{X})s(A,\boldsymbol{X})}{w_t(A,\boldsymbol{X})^3\pi(A,\boldsymbol{X})^2}\right\}\boldsymbol{A}_t^{-1} + \boldsymbol{A}_t^{-1}\boldsymbol{B}_t\boldsymbol{A}_t^{-1}\boldsymbol{C}_t\boldsymbol{A}_t^{-1}.
\end{aligned} \tag{A.9}$$

Therefore:

$$\begin{aligned}
&\frac{\partial}{\partial t}\left[var\left\{\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\}\right]\Bigg|_{t=0} \\
&\quad= \boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}E\left\{\frac{-2\boldsymbol{X}\boldsymbol{X}^\top\sigma_0^2(A,\boldsymbol{X})s(A,\boldsymbol{X})}{w(A,\boldsymbol{X})^3\pi(A,\boldsymbol{X})^2}\right\}\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1}.
\end{aligned} \tag{A.10}$$

Now, letting $w(A,\boldsymbol{X}) = \frac{\sigma_0^2(A,\boldsymbol{X})}{\pi(A,\boldsymbol{X})}$:

$$E\left\{\frac{-2\boldsymbol{X}\boldsymbol{X}^\top\sigma_0^2(A,\boldsymbol{X})s(A,\boldsymbol{X})}{w(A,\boldsymbol{X})^3\pi(A,\boldsymbol{X})^2}\right\} = -2\boldsymbol{C}, \tag{A.11}$$

$$\boldsymbol{A}^{-1}\boldsymbol{B} = E\left\{\frac{\boldsymbol{X}\boldsymbol{X}^\top}{\sigma_0^2(A,\boldsymbol{X})}\right\}^{-1}E\left\{\frac{\boldsymbol{X}\boldsymbol{X}^\top}{\sigma_0^2(A,\boldsymbol{X})}\right\} = 1. \tag{A.12}$$

Therefore:

$$\begin{aligned}
\frac{\partial var\left\{\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right\}}{\partial t}\Bigg|_{t=0} &= \boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} - 2\boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} \\
&= 0.
\end{aligned} \tag{A.13}$$

We have shown that the sandwich variance is minimized at $w(A,\boldsymbol{X}) = \dfrac{\sigma_0^2(A,\boldsymbol{X})}{\pi(A,\boldsymbol{X})}$. Thus, the proof is complete. □

Proof of Proposition 1.2. Let $l_n(\boldsymbol{B}_*) = \mathbb{P}_n \left\{ \frac{1}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})} \left(\frac{K}{K-1} R - \boldsymbol{X}_*^\top \boldsymbol{B}_* \right)^2 \right\}$, and define the first and second derivatives of l_n:

$$\frac{\partial l_n(\boldsymbol{B}_*)}{\partial \boldsymbol{B}_*} = -2\mathbb{P}_n \left\{ \frac{1}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})} \left(\frac{K}{K-1} R - \boldsymbol{X}_*^\top \boldsymbol{B}_* \right) \boldsymbol{X}_* \right\}, \tag{A.14}$$

$$\frac{\partial^2 l_n(\boldsymbol{B}_*)}{\partial \boldsymbol{B}_*^2} = 2\mathbb{P}_n \left\{ \frac{1}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})} \boldsymbol{X}_* \boldsymbol{X}_*^\top \right\}. \tag{A.15}$$

Assuming that $\widehat{\boldsymbol{B}_*} \xrightarrow{P} \boldsymbol{B}_*$ and using Taylor Series approximations similar to Proposition 1.1,

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\boldsymbol{B}_*} - \boldsymbol{B}_*\right) \\
&= \sqrt{n} E \left\{ \frac{\boldsymbol{X}_* \boldsymbol{X}_*^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})} \right\}^{-1} \left\{ \mathbb{P}_n \frac{\boldsymbol{X}_*}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})} \left(\frac{K}{K-1} R - \boldsymbol{X}_*^\top \boldsymbol{B}_* \right) \right\} + o_p(1) \\
&= E \left\{ \frac{\boldsymbol{X}_* \boldsymbol{X}_*^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})} \right\}^{-1} \left\{ \sqrt{n}\mathbb{P}_n \frac{\epsilon \boldsymbol{X}_*}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})} \right\} + o_p(1), \text{ and}
\end{aligned} \tag{A.16}$$

$$\begin{aligned}
&var\left\{ \sqrt{n}\left(\widehat{\boldsymbol{B}_*} - \boldsymbol{B}_*\right) \right\} \\
&= E \left\{ \frac{\boldsymbol{X}_* \boldsymbol{X}_*^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})} \right\}^{-1} E \left\{ \frac{\boldsymbol{X}_* \boldsymbol{X}_*^\top \sigma_0^2(A,\boldsymbol{X})}{w(A,\boldsymbol{X})^2 \pi(A,\boldsymbol{X})^2} \right\} E \left\{ \frac{\boldsymbol{X}_* \boldsymbol{X}_*^\top}{w(A,\boldsymbol{X})\pi(A,\boldsymbol{X})} \right\}^{-1} + o(1).
\end{aligned} \tag{A.17}$$

The format of (A.17) is now very similar to that of Proposition 1.1. Taking a functional derivative, as in the proof for Proposition 1.1, shows that the sandwich variance is minimized at $w(A,\boldsymbol{X}) = \frac{\sigma_0^2(A,\boldsymbol{X})}{\pi(A,\boldsymbol{X})}$. Thus, the proof is complete.

□

A.2 Proofs of Remarks and Theorems

This section includes proofs of Remark 1.1, Theorem 1.1, Theorem 1.2, and Theorem 1.3.

Proof of Remark 1.1. For the first case, linear regression, assume that $\Phi(A, \boldsymbol{X})$ is a finite-dimensional collection of features in $(A, \boldsymbol{X})$ and functions of the features. $\Phi(A, \boldsymbol{X})$ includes main effects, and can additionally include squared and intercept terms, cubic terms, etc. Assume that $E\left\{\Phi(A, \boldsymbol{X})\Phi(A, \boldsymbol{X})^\top\right\}$ is bounded and positive definite and let $\widehat{\sigma}_n^2(A, \boldsymbol{X})$ be a linear function predicting the squared residuals, $Z = \left(2AR - \boldsymbol{X}'\widehat{\boldsymbol{\beta}}_n^D\right)^2$, from the features in the set $\Phi(A, \boldsymbol{X})$. Standard methods from Tian *et al.* (2014) show that $\widehat{\boldsymbol{\beta}}_n^D$ is consistent for $\boldsymbol{\beta}_0$. Then, the prediction function, $\widehat{\sigma}_n^2(A, \boldsymbol{X})$, is uniformly consistent for $\sigma_0^2(A, \boldsymbol{X})$, and moreover is a vector process which belongs to a VC-class via Lemma 9.6 of Kosorok (2008). This implies that it is bounded uniform entropy integrable (belongs to a BUEI class) and therefore in a Donsker class because of the measurability conferred by Lemma 8.12 of Kosorok (2008).

For the second case, Assumptions 2-4 of Cho *et al.* (2020) specify basic conditions related to the covariate distribution (weakly dependent features) and the random forest (terminal node size growth rate and α-regular random split trees) under which it is established that the tree kernels of the random forest are Donsker. Theorem 1 of Cho *et al.* (2020) demonstrates uniform consistency of the resulting random forest.

Further, it is known that random forest predictions can be characterized by $\widehat{\sigma}_n^2(A, \boldsymbol{X}) = \sum_{j=1}^{\infty} \alpha_j c_j \mathbb{I}\{(a, \boldsymbol{x}) \in R_j\}$, where each R_j is a hyperrectangle, c_j is the expected value of Y over the region R_j, and $\sum_{j=1}^{\infty} \alpha_j = 1$. The random forest prediction function, $\widehat{\sigma}_n^2(A, \boldsymbol{X})$, is therefore a convex combination of expected values over hyperrectangular regions. Since the class of hyperrectangles of fixed dimensions is a VC-class and $\widehat{\sigma}_n^2(A, \boldsymbol{X})$ is a convex combination of them, $\widehat{\sigma}_n^2(A, \boldsymbol{X})$ belongs to a convex hull class, $\mathcal{C}$.

For a convex hull class, Corollary 9.5 of Kosorok (2008) applies and indicates that the covering number of $\mathcal{C}$ is bounded above by $K\left(\frac{1}{\epsilon}\right)^{2-2/V}$, where V is the VC-index of $\mathcal{C}$. Taking the square root results in $\epsilon^{-\alpha}$ where $\alpha < 1$, resulting in a bounded integral when integrated. Therefore, $\widehat{\sigma}_n^2(A, \boldsymbol{X})$ belongs to a BUEI class and can also be shown to be Pointwise Measurable (PM) (see, e.g., Section 8.2 of Kosorok (2008)).

We wish first to show that $\frac{1}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}$ also belongs to a class that is BUEI and PM. For $g(x) = \frac{1}{x}$, $\frac{\partial}{\partial x} g(x) = \frac{-1}{x^2} \leqslant \frac{-1}{c_1^2}$ since x represents the quantity $\widehat{\sigma}_n^2(A,\boldsymbol{X})$. Therefore, $g(x)$ is a Lipschitz continuous function, and Lemma 9.17 (vi) of Kosorok (2008) shows that $\frac{1}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}$ is BUEI and PM using $c_1 < \infty$ as an envelope. Finally, using Lemma 9.17 (v), $\frac{\boldsymbol{X}\epsilon}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}$ also belongs to a class which is BUEI and PM, using $E(|\boldsymbol{X}\epsilon|) \leqslant \|\boldsymbol{X}\| \sigma_0$ as an envelope. Since $\frac{\boldsymbol{X}\epsilon}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}$ belongs to a class which is BUEI and PM, by Theorem 8.19, the class is also Donsker. □

Proof of Theorem 1.1. Considering $f^{opt}(\boldsymbol{X}) \in \mathcal{F}$, where $f^{opt}(\boldsymbol{X}) = E\left\{ \frac{RA}{\pi(A,\boldsymbol{X})} \middle| \boldsymbol{X} \right\} = 2\delta(\boldsymbol{X})$, let $Y^* = 2RA$, and characterize the estimated variance function, $\widehat{\sigma}_n^2(A,\boldsymbol{X})$, as the matrix $\widehat{\boldsymbol{\Sigma}}_n \in \mathbb{R}^{n\times n}$ where $\widehat{\boldsymbol{\Sigma}}_n = \text{diag}\left\{\widehat{\sigma}_n^2(a_1,\boldsymbol{x}_1), ..., \widehat{\sigma}_n^2(a_n,\boldsymbol{x}_n)\right\}$. The least squares SD-Learning estimate for $\boldsymbol{\beta}_0$ can be written as:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_n^S &= (\boldsymbol{X}^\top \widehat{\boldsymbol{\Sigma}}_n^{-1} \boldsymbol{X})^{-1} (\boldsymbol{X}^\top \widehat{\boldsymbol{\Sigma}}_n^{-1} Y^*) \\ &= \boldsymbol{\beta}_0 + \left(\frac{1}{n} \boldsymbol{X}^\top \widehat{\boldsymbol{\Sigma}}_n^{-1} \boldsymbol{X} \right)^{-1} \left(\frac{1}{n} \boldsymbol{X}^\top \widehat{\boldsymbol{\Sigma}}_n^{-1} \epsilon \right). \end{aligned} \tag{A.18}$$

We show that the second term in this equation converges to zero. Note that since $\sigma_0^2(A,\boldsymbol{X}) \geqslant c_1$ by Assumption 1.2 and $\widehat{\sigma}_n^2(A,\boldsymbol{X})$ is uniformly consistent for $\sigma_0^2(A,\boldsymbol{X})$ by Assumption 1.3, there exists N for which $\widehat{\sigma}_n^2(A,\boldsymbol{X}) \geqslant \frac{c_1}{2}$ for all $n \geqslant N$, with probability going to 1. Letting $\boldsymbol{t} \in \mathbb{R}^p$ be an arbitrary vector on the unit sphere,

$$\begin{aligned} &\left| \frac{1}{n} \sum_{i=1}^n \frac{\boldsymbol{t}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{t}}{\widehat{\sigma}_n^2(a_i,\boldsymbol{x}_i)} - \frac{1}{n} \sum_{i=1}^n \frac{\boldsymbol{t}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{t}}{\sigma_0^2(a_i,\boldsymbol{x}_i)} \right| \\ &\qquad \leqslant \frac{1}{n} \sum_{i=1}^n \left(\boldsymbol{t}^\top \boldsymbol{x}_i\right)^2 \left| \frac{1}{\widehat{\sigma}_n^2(a_i,\boldsymbol{x}_i)} - \frac{1}{\sigma_0^2(a_i,\boldsymbol{x}_i)} \right| \\ &\qquad \leqslant \frac{1}{n} \sum_{i=1}^n \|\boldsymbol{x}_i\|^2 \cdot \frac{2}{c_1^2} \left\| \widehat{\sigma}_n^2(A,\boldsymbol{X}) - \sigma_0^2(A,\boldsymbol{X}) \right\|_{\mathcal{A},\mathcal{X}} + o_p(1) \\ &\qquad = \left\{ \frac{2}{c_1^2} \left\| \widehat{\sigma}_n^2(A,\boldsymbol{X}) - \sigma_0^2(A,\boldsymbol{X}) \right\|_{\mathcal{A},\mathcal{X}} \right\} \cdot \frac{1}{n} \sum_{i=1}^n \|\boldsymbol{x}_i\|^2 + o_p(1) \\ &\qquad = o_p(1) \cdot O_p(1) + o_p(1), \text{and} \end{aligned} \tag{A.19}$$

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i\epsilon_i}{\widehat{\sigma}_n^2(a_i,\boldsymbol{x}_i)}-\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{X}_i\epsilon_i}{\sigma_0^2(a_i,\boldsymbol{x}_i)}\right| \\
&\qquad=\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\epsilon_i\left\{\frac{1}{\widehat{\sigma}_n^2(a_i,\boldsymbol{x}_i)}-\frac{1}{\sigma_0^2(a_i,\boldsymbol{x}_i)}\right\}\right| \\
&\qquad\leqslant\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\epsilon_i\left\{\frac{2\left\|\widehat{\sigma}_n^2(A,\boldsymbol{X})-\sigma_0^2(\boldsymbol{X})\right\|_{\mathcal{A},\mathcal{X}}}{c_1^2}\right\}\right|+o_p(1) \\
&\qquad=\frac{2\left\|\widehat{\sigma}_n^2(A,\boldsymbol{X})-\sigma_0^2(\boldsymbol{X})\right\|_{\mathcal{A},\mathcal{X}}}{c_1^2}\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\epsilon_i\right|+o_p(1),
\end{aligned}
\tag{A.20}
$$

where $n^{-1}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|^2$ is $O_p(1)$ because it converges to $E\left(\boldsymbol{X}\boldsymbol{X}^\top\right)$ and $\left|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\epsilon_i\right|$ converges to $|E(\boldsymbol{X}_i\epsilon_i)|=0$. (A.19) shows that $n^{-1}\boldsymbol{X}^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}\xrightarrow{P}n^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X}$, since it is true for arbitrary $\boldsymbol{t}$ on the unit sphere, and (A.20) shows that $n^{-1}\boldsymbol{X}^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\epsilon\xrightarrow{P}n^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}_0^{-1}\epsilon$. Since inversion is a continuous operation when the limit is a positive definite matrix, $\left(n^{-1}\boldsymbol{X}^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}\right)^{-1}\xrightarrow{P}\left(n^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X}\right)^{-1}$ by the Continuous Mapping Theorem. Since $\left(n^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X}\right)^{-1}$ is finite and positive definite while $n^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}_0^{-1}\epsilon$ converges to zero using the Kolmogorov Law of Large Numbers, $\left(n^{-1}\boldsymbol{X}^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}\right)^{-1}\left(n^{-1}\boldsymbol{X}^\top\widehat{\boldsymbol{\Sigma}}_n^{-1}\epsilon\right)$ converges to zero, and $\widehat{\boldsymbol{\beta}}_n^S\xrightarrow{P}\boldsymbol{\beta}_0$. □

Proof of Theorem 1.2. We show that the preconditions are met for Lemma 4.4 of Kosorok (2008). Assumptions 1.1 and 1.2 guarantee that U_0 is positive definite. Additionally, $\widehat{\boldsymbol{\beta}}_n^S\xrightarrow{P}\boldsymbol{\beta}_0$ based on Theorem 1.1. $\mathbb{P}_n\left\{\frac{\boldsymbol{X}\boldsymbol{X}'}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}-\frac{\boldsymbol{X}\boldsymbol{X}'}{\sigma_0^2(A,\boldsymbol{X})}\right\}\to 0$ was shown by (A.19). Now, consider

the following:

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{t}^{\top}\boldsymbol{x}_i\epsilon_i\left\{\frac{1}{\widehat{\sigma}_n^2(a_i,\boldsymbol{x}_i)}-\frac{1}{\sigma_0^2(a_i,\boldsymbol{x}_i)}\right\}\right]^2\right| \\
&\qquad=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{t}^{\top}\boldsymbol{x}_i)^2\sigma_0^2(a_i,\boldsymbol{x}_i)\left\{\frac{1}{\widehat{\sigma}_n^2(a_i,\boldsymbol{x}_i)}-\frac{1}{\sigma_0^2(a_i,\boldsymbol{x}_i)}\right\}^2+o_p(1) \\
&\qquad\leqslant\frac{c_2}{n}\sum_{i=1}^{n}(\boldsymbol{t}^{\top}\boldsymbol{x}_i)^2\left\{\frac{1}{\widehat{\sigma}_n^2(a_i,\boldsymbol{x}_i)}-\frac{1}{\sigma_0^2(a_i,\boldsymbol{x}_i)}\right\}^2 \\
&\qquad\leqslant\frac{4c_2}{c_1^4}\cdot\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{t}^{\top}\boldsymbol{x}_i)^2\left\|\widehat{\sigma}_n^2(a_i,\boldsymbol{x}_i)-\sigma_0^2(a_i,\boldsymbol{x}_i)\right\|_{\mathcal{A},\mathcal{X}}^2+o_p(1) \\
&\qquad\leqslant\frac{4c_2}{c_1^4}\left\|\widehat{\sigma}_n^2(A,\boldsymbol{X})-\sigma_0^2(A,\boldsymbol{X})\right\|_{\mathcal{A},\mathcal{X}}^2\cdot\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^2+o_p(1) \\
&\qquad=o_p(1)\cdot O_p(1)+o_p(1).
\end{aligned}
\tag{A.21}
$$

Therefore, $\mathbb{P}_n\left\{\dfrac{\boldsymbol{X}\epsilon}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}-\dfrac{\boldsymbol{X}\epsilon}{\sigma_0^2(A,\boldsymbol{X})}\right\}^2 \xrightarrow{P} 0$, and the conditions for the Lemma 4.4 of Kosorok (2008) are met. Note the following two facts:

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_n^S&=\mathbb{P}_n\left\{\frac{\boldsymbol{X}\boldsymbol{X}'}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}\right\}^{-1}\mathbb{P}_n\left\{\frac{XY^*}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}\right\} \\
&\implies\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^S-\boldsymbol{\beta}_0\right)=\left[\mathbb{P}_n\left\{\frac{\boldsymbol{X}\boldsymbol{X}'}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}\right\}^{-1}\right]\cdot\sqrt{n}\mathbb{P}_n\left\{\frac{\boldsymbol{X}\epsilon}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}\right\},\ \text{and}
\end{aligned}
\tag{A.22}
$$

$$
\begin{aligned}
&\sqrt{n}\left[\mathbb{P}_n\left\{\frac{\boldsymbol{X}\epsilon}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}\right\}-E\left\{\frac{\boldsymbol{X}\epsilon}{\sigma_0^2(A,\boldsymbol{X})}\right\}\right]\to N\left[0,E\left\{\frac{\boldsymbol{X}\boldsymbol{X}'}{\sigma_0^2(A,\boldsymbol{X})}\right\}\right] \\
&\implies\sqrt{n}\mathbb{P}_n\left\{\frac{\boldsymbol{X}\epsilon}{\widehat{\sigma}_n^2(A,\boldsymbol{X})}\right\}\to N\left[0,E\left\{\frac{\boldsymbol{X}\boldsymbol{X}'}{\sigma_0^2(A,\boldsymbol{X})}\right\}\right],
\end{aligned}
\tag{A.23}
$$

where $E\left\{\dfrac{\boldsymbol{X}\epsilon}{\sigma_0^2(A,\boldsymbol{X})}\right\}=0$ in (A.23) stems from Assumption 1.4. With the combination of (A.22) and (A.23), we can now establish:

$$
\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n^S-\boldsymbol{\beta}_0\right)&\to E\left\{\frac{\boldsymbol{X}\boldsymbol{X}'}{\sigma_0^2(A,\boldsymbol{X})}\right\}^{-1}\cdot N\left[0,E\left\{\frac{\boldsymbol{X}\boldsymbol{X}'}{\sigma_0^2(A,\boldsymbol{X})}\right\}\right] \\
&=N\left[0,E\left\{\frac{\boldsymbol{X}\boldsymbol{X}'}{\sigma_0^2(A,\boldsymbol{X})}\right\}^{-1}\right].
\end{aligned}
\tag{A.24}
$$

□

Proof of Theorem 1.3. Since the SD-Learning framework is unified between the binary and multi-arm case, this proof is identical to the proofs of Theorems 1.1 and 1.2, with $\boldsymbol{X}_* \in \mathbb{R}^{p(K-1)}$ instead of $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{B}_* \in \mathbb{R}^{p(K-1)}$ instead of $\boldsymbol{\beta}_0 \in \mathbb{R}^p$. □

A.3 Extension to Observational Data

It was assumed in the development of the methodology that the data stem from an RCT setting. In practice, however, it is often the case that drug efficacy is evaluated retrospectively through observational data. In this case, treatment assignment probabilities are not known. If the conditional exchangeability assumption holds, the SD-Learning methodology remains intact, but $\pi(A, \boldsymbol{X})$ must be estimated given the observed covariates. As discussed in Chen, Zeng, and Kosorok (2016), the estimate $\widehat{\pi}(A, \boldsymbol{X})$ could be obtained by a parametric model such as logistic regression (multinomial regression in the multi-arm case), or a nonparametric method such as boosting, random forests, or support vector regression (SVR).

A.4 Heteroscedasticity Analysis

Here, we analyze the heteroscedasticity present in the ACTG175 dataset after fitting a linear model to predict change in CD4 cell count from covariates and treatments (using treatment Z as the reference). We first implement the Studentized Score χ^2 Test, proposed by Koenker (1981), which tests whether the variance of errors from a linear regression model is dependent on the values of the independent variables. Similarly to the SD-Learning methodology, the Studentized Score Test does not rely on an assumption of normally distributed errors. Testing the null hypothesis that errors are independent and identically distributed (homoscedasticity) resulted in $\chi^2_{15} = 69.2$ ($p < 0.0001$), indicating the presence of heteroscedasticity.

Let W represent absolute residuals from the aforementioned linear model. Having confirmed heteroscedasticity through the score test, we fit a new regression model predicting W from co-

variates and treatments, in order to determine the variables responsible for heteroscedasticity. To account for multiple variables being tested, we perform false discovery rate-based (FDR) adjustment of the resulting p-values (Benjamini and Hochberg (1995)) and find that three variables are significantly associated with absolute residuals at the $\alpha = 0.05$ level: baseline CD4 ($p < 0.0001$), treatment ZD ($p < 0.0001$), and treatment D ($p = 0.0148$).

We thus find enough evidence of heteroscedasticity in the ACTG175 data to justify incorporating SD-Learning as an ITR estimation approach.

APPENDIX B: SUPPORTING MATERIALS FOR CHAPTER 2

B.1 Proofs of Theorems

This section includes proofs to Theorems 2.1 and 2.2.

Proof of Theorem 2.1. Since we assume compactness of the parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \rho\}$, it remains to show that the first derivative of the log-likelihood is bounded, which would ensure continuity. Note the following:

$$log\left\{\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \rho)\right\} = \sum_{p=1}^{m} log\left[\int \left\{\prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} \cdot p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) d\boldsymbol{b}_p\right]. \tag{B.1}$$

Denoting the inside of the summation by l_p, we establish that:

$$\frac{\partial l_p}{\partial \boldsymbol{\theta}} = \frac{\int \left\{\prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} \cdot p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) \cdot \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{S}_p) d\boldsymbol{b}_p}{\int \left\{\prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} \cdot p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) d\boldsymbol{b}_p}, \tag{B.2}$$

where $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{S}_p) = [s_{\boldsymbol{\beta}}(\boldsymbol{S}_p) \;\; s_{\boldsymbol{\sigma}^2}(\boldsymbol{S}_p) \;\; s_{\rho}(\boldsymbol{S}_p)]^\top = \frac{\partial}{\partial \boldsymbol{\theta}} log\left[\left\{\prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} \cdot p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho)\right]$, the score of the inside of the integral. Say $|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{S}_p)| < \boldsymbol{c}_{\boldsymbol{\theta}}$. Then:

$$\begin{aligned}
\frac{\partial l_p}{\partial \boldsymbol{\theta}} &\leqslant \frac{\int \left\{\prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) \cdot |\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{S}_p)| \cdot d\boldsymbol{b}_p}{\int \left\{\prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) \cdot d\boldsymbol{b}_p} \\
&\leqslant \frac{\int \left\{\prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) \cdot \boldsymbol{c}_{\boldsymbol{\theta}} \cdot d\boldsymbol{b}_p}{\int \left\{\prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) \cdot d\boldsymbol{b}_p} \\
&= \boldsymbol{c}_{\boldsymbol{\theta}} \frac{\int \left\{\prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) d\boldsymbol{b}_p}{\int \left\{\prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p)\right\} p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) d\boldsymbol{b}_p} \\
&= \boldsymbol{c}_{\boldsymbol{\theta}},
\end{aligned} \tag{B.3}$$

and $\frac{\partial l}{\partial \boldsymbol{\theta}} = \sum_{p=1}^{m} \frac{\partial l_p}{\partial \boldsymbol{\theta}} \leqslant \sum_{p=1}^{m} \boldsymbol{c_\theta} = m\boldsymbol{c_\theta}$. Therefore, showing that the elements of $\boldsymbol{s_\theta}(\boldsymbol{S}_p)$ are bounded completes the proof and establishes boundedness.

We now derive $\boldsymbol{s_\beta}(\boldsymbol{S}_p)$, $\boldsymbol{s_{\sigma^2}}(\boldsymbol{S}_p)$, and $s_\rho(\boldsymbol{S}_p)$, and show the boundedness of each. We begin with $\boldsymbol{s_\beta}(\boldsymbol{S}_p)$:

$$
\begin{aligned}
\boldsymbol{s_\beta}(\boldsymbol{S}_p) &= \frac{\partial}{\partial \boldsymbol{\beta}} \log \left[\left\{ \prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p) \right\} \cdot p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) \right] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \sum_{q=1}^{c_p} \log p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p) \right\} + \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \log p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) \right\} \\
&= \sum_{q=1}^{c_p} \frac{\partial}{\partial \boldsymbol{\beta}} \log p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p) \\
&= \sum_{q=1}^{c_p} \frac{\partial}{\partial \boldsymbol{\beta}} \log \left\{ \left(\frac{e^{U_{\boldsymbol{\beta},p}}}{1+e^{U_{\boldsymbol{\beta},p}}} \right)^{s_{pq}} \left(1 - \frac{e^{U_{\boldsymbol{\beta},p}}}{1+e^{U_{\boldsymbol{\beta},p}}} \right)^{1-s_{pq}} \right\} \\
&= \sum_{q=1}^{c_p} \frac{\partial}{\partial \boldsymbol{\beta}} \log \left\{ \frac{e^{s_{pq}U_{\boldsymbol{\beta},p}}}{1+e^{U_{\boldsymbol{\beta},p}}} \right\} \\
&= \sum_{q=1}^{c_p} \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ s_{pq}U_{\boldsymbol{\beta},p} - \log\left(1+e^{U_{\boldsymbol{\beta},p}}\right) \right\} \\
&= \sum_{q=1}^{c_p} \left\{ s_{pq}\dot{U}_{\boldsymbol{\beta},p} - \frac{\dot{U}_{\boldsymbol{\beta},p}e^{U_{\boldsymbol{\beta},p}}}{1+e^{U_{\boldsymbol{\beta},p}}} \right\} \\
&= \sum_{q=1}^{c_p} \dot{U}_{\boldsymbol{\beta},p} \left\{ s_{pq} - \frac{e^{U_{\boldsymbol{\beta},p}}}{1+e^{U_{\boldsymbol{\beta},p}}} \right\},
\end{aligned}
\tag{B.4}
$$

In (B.4), $\left(S_{pq} - \frac{e^{U_{\boldsymbol{\beta},b}}}{1+e^{U_{\boldsymbol{\beta},b}}} \right)$ is bounded by 1, thus we work on showing boundedness of $\dot{U}_{\boldsymbol{\beta},p}$:

$$
\begin{aligned}
\frac{\partial U_{\boldsymbol{\beta},p}}{\partial \boldsymbol{\beta}_l} = &\frac{\left(1+\sum_{l=2}^{L} e^{\boldsymbol{\beta}_l^\top x + b_{pl}}\right) \left\{ x e^{\boldsymbol{\beta}_l^\top x + b_{pl}} \left(R_{Al} - R_{Bl}\right) \right\}}{\left(1+\sum_{l=2}^{L} e^{\boldsymbol{\beta}_l^\top x + b_{pl}}\right)^2} \\
&- \frac{\left\{ \left(R_{Al} - R_{Bl}\right) + \sum_{l=2}^{L} e^{\boldsymbol{\beta}_l^\top x + b_{pl}} \left(R_{Al} - R_{Bl}\right) \right\} \left(x e^{\boldsymbol{\beta}_l^\top x + b_{pl}} \right)}{\left(1+\sum_{l=2}^{L} e^{\boldsymbol{\beta}_l^\top x + b_{pl}}\right)^2}.
\end{aligned}
\tag{B.5}
$$

The denominator is $\geqslant 1$, thus it suffices to look at the numerator terms:

$$\begin{aligned}\frac{\partial U_{\boldsymbol{\beta},b}}{\partial \boldsymbol{\beta}_l} &\leqslant ||X|| \max_{2\leqslant l\leqslant L} \left(|R_{Al}| + |R_{Bl}|\right) + ||X|| \max_{2\leqslant l\leqslant L} \left(|R_{Al}| + |R_{Bl}|\right) \\ &= 2||X|| \max_{2\leqslant l\leqslant L} \left(|R_{Al}| + |R_{Bl}|\right).\end{aligned} \tag{B.6}$$

Thus, with Assumption 2.1, $\dot{U}_{\boldsymbol{\beta},b}$ is bounded with respect to $\boldsymbol{\beta}_l$ for all $l \in \{2, ..., L\}$. Since $\dot{U}_{\boldsymbol{\beta},p}$ and $\left\{s_{pq} - \frac{e^{U_{\boldsymbol{\beta},p}}}{1+e^{U_{\boldsymbol{\beta},p}}}\right\}$ are bounded, $\sum_{q=1}^{c_p} \dot{U}_{\boldsymbol{\beta},p} \left\{s_{pq} - \frac{e^{U_{\boldsymbol{\beta},p}}}{1+e^{U_{\boldsymbol{\beta},p}}}\right\}$ is also bounded. Denote the bound as $|\boldsymbol{s}_{\boldsymbol{\beta}}(\boldsymbol{S}_p)| < c_{\boldsymbol{\beta}}$.

We now move to $\boldsymbol{s}_{\boldsymbol{\sigma}^2}(\boldsymbol{S}_p)$:

$$\begin{aligned}\boldsymbol{s}_{\sigma_l^2}(\boldsymbol{S}_p) &= \frac{\partial}{\partial \sigma_l^2} \log \left[\left\{ \prod_{q=1}^{c_p} p(s_{pq}|\boldsymbol{\beta}, \boldsymbol{b}_p) \right\} \cdot p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) \right] \\ &= \frac{\partial}{\partial \sigma_l^2} \log p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2, \rho) \\ &= \frac{\partial}{\partial \sigma_l^2} \log \left\{ (2\pi)^{-(L-1)/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(\frac{1}{2}\boldsymbol{b}_p^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{b}_p\right) \right\} \\ &= \frac{\partial}{\partial \sigma_l^2} \left\{ -\frac{1}{2} \log\left(|\boldsymbol{D}|^2 |\boldsymbol{A}(\rho)|\right) - \frac{1}{2}\boldsymbol{b}_p^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{b}_p \right\} \\ &= \frac{\partial}{\partial \sigma_l^2} \left\{ -\frac{1}{2} \log |\boldsymbol{D}|^2 - \frac{1}{2}\boldsymbol{b}_p^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{b}_p \right\} \\ &= -\frac{1}{2\sigma_l^2} - \frac{1}{2} \frac{\partial}{\partial \sigma_l^2} \left\{ \boldsymbol{b}_p^\top \boldsymbol{D}^{-1} \boldsymbol{A}^{-1}(\rho) \boldsymbol{D}^{-1} \boldsymbol{b}_p \right\},\end{aligned} \tag{B.7}$$

where the fourth equality stems from the fact that the determinant of a matrix product is equal to the product of the determinants. In order to determine $\boldsymbol{A}^{-1}(\rho)$, note that by the Sherman-Morrison formula (Press and Teukolsky (2007)), $\boldsymbol{A}^{-1}(\rho) = \frac{1}{1-\rho} \cdot \boldsymbol{I} - \frac{\{\rho/(1-\rho)^2\}}{1 + \{\rho/(1-\rho)\}\boldsymbol{j}^\top \boldsymbol{j}} \cdot \boldsymbol{j}\boldsymbol{j}^\top = \frac{1}{1-\rho}\left\{\boldsymbol{I} - \frac{\rho}{1+\rho(L-2)}\boldsymbol{j}\boldsymbol{j}^\top\right\}$. Let the diagonal and non-diagonal entries of $\boldsymbol{A}^{-1}(\rho)$ be represented by c_d and c_n, respectively, where:

$$c_d = \frac{1 + \rho(L-3)}{(1-\rho)\{1+\rho(L-2)\}}, \tag{B.8}$$

$$c_n = -\frac{\rho}{(1-\rho)\{1+\rho(L-2)\}}. \tag{B.9}$$

Thus, with some linear algebra:

$$\boldsymbol{b}_p^\top \boldsymbol{D}^{-1}\boldsymbol{A}^{-1}(\rho)\boldsymbol{D}^{-1}\boldsymbol{b}_p = \sum_{j=2}^{L}\sum_{i=2}^{L}\frac{b_{pi}b_{pj}}{\sigma_i\sigma_j}\left\{c_d\mathbb{1}(i=j)+c_n\mathbb{1}(i\neq j)\right\}, \tag{B.10}$$

and it can be seen that:

$$\frac{\partial}{\partial\sigma_l^2}\left\{\boldsymbol{b}_p^\top \boldsymbol{D}^{-1}\boldsymbol{A}^{-1}(\rho)\boldsymbol{D}^{-1}\boldsymbol{b}_p\right\} = \frac{1}{2}\sum_{i=2}^{L}\frac{b_{pi}b_{pl}}{\sigma_i\sigma_l^3}\left\{c_d\mathbb{1}(i=l)+c_n\mathbb{1}(i\neq l)\right\}. \tag{B.11}$$

Therefore:

$$\boldsymbol{s}_{\sigma_l^2}(\boldsymbol{S}_p) = -\frac{1}{2\sigma_l^2}-\frac{1}{4}\sum_{i=2}^{L}\frac{b_{pi}b_{pl}}{\sigma_i\sigma_l^3}\left\{c_d\mathbb{1}(i=l)+c_n\mathbb{1}(i\neq l)\right\}. \tag{B.12}$$

Note that c_d and $c_n < \infty$ if $\rho < 1$ and $\rho > -\frac{1}{L-2}$, which is guaranteed by Assumption 2.2. In combination with Assumption 2.1, which guarantees σ_l^2 to be nonzero for all l, $\boldsymbol{s}_{\sigma_l^2}(\boldsymbol{S}_p)$ is bounded, which can be denoted by $\left|\boldsymbol{s}_{\sigma_l^2}(\boldsymbol{S}_p)\right| < c_{\sigma_l^2}$, and thus $|\boldsymbol{s}_{\boldsymbol{\sigma^2}}(\boldsymbol{S}_p)| < \boldsymbol{c}_{\boldsymbol{\sigma}^2}$.

Finally, we move to $s_\rho(\boldsymbol{S}_p)$:

$$\begin{aligned}
s_\rho(\boldsymbol{S}_p) &= \frac{\partial}{\partial\rho}\log\left[\left\{\prod_{q=1}^{c_p}p(s_{pq}|\boldsymbol{\beta},\boldsymbol{b}_p)\right\}\cdot p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2,\rho)\right]\\
&= \frac{\partial}{\partial\rho}\left\{\log p(\boldsymbol{b}_p|\boldsymbol{\sigma}^2,\rho)\right\}\\
&= \frac{\partial}{\partial\rho}\log\left\{(2\pi)^{-(L-1)/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(\frac{1}{2}\boldsymbol{b}_p^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{b}_p\right)\right\}\\
&= \frac{\partial}{\partial\rho}\left\{-\frac{1}{2}\log\left(|\boldsymbol{D}|^2|\boldsymbol{A}(\rho)|\right)-\frac{1}{2}\boldsymbol{b}_p^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{b}_p\right\}\\
&= \frac{\partial}{\partial\rho}\left\{-\frac{1}{2}\log|\boldsymbol{A}(\rho)|-\frac{1}{2}\boldsymbol{b}_p^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{b}_p\right\}.
\end{aligned} \tag{B.13}$$

To find $|\boldsymbol{A}(\rho)|$, we perform an eigenvalue decomposition. Let $\boldsymbol{v}_1 = \boldsymbol{j}/\sqrt{L-1}$. Then:

$$\begin{aligned}
\boldsymbol{A}(\rho)\boldsymbol{v}_1 &= \left\{(1-\rho)\boldsymbol{I}+\rho\boldsymbol{j}\boldsymbol{j}^\top\right\}\boldsymbol{v}_1\\
&= (1-\rho)\boldsymbol{v}_1+\rho(L-1)\boldsymbol{v}_1,
\end{aligned} \tag{B.14}$$

and setting $\boldsymbol{A}(\rho)\boldsymbol{v}_1 = \lambda_1 \boldsymbol{v}_1$, we find that $\lambda_1 = 1 + (L-2)\rho$. Now, let $\boldsymbol{v}_2$ be such that $\boldsymbol{v}_1^\top \boldsymbol{v}_2 = 0$ and $\boldsymbol{v}_2^\top \boldsymbol{v}_2 = 1$. Then:

$$\begin{aligned}\boldsymbol{A}(\rho)\boldsymbol{v}_2 &= (1-\rho)\boldsymbol{v}_2 + \rho \boldsymbol{j}\boldsymbol{j}^\top \boldsymbol{v}_2 \\ &= (1-\rho)\boldsymbol{v}_2 + \rho\sqrt{L-1}\boldsymbol{j}\boldsymbol{v}_1^\top \boldsymbol{v}_2 \\ &= (1-\rho)\boldsymbol{v}_2,\end{aligned} \tag{B.15}$$

and setting $\boldsymbol{A}(\rho)\boldsymbol{v}_2 = \lambda_2 \boldsymbol{v}_2$, we find that $\lambda_2 = 1-\rho$, with multiplicity $L-2$ since the space orthogonal to $\boldsymbol{v}_1$ has rank $L-2$. Therefore, since the determinant of a matrix is equal to the product of its eigenvalues, $|\boldsymbol{A}(\rho)| = \{1+(L-2)\,\rho\}\,(1-\rho)^{L-2}$, and we can continue deriving $s_\rho(\boldsymbol{S}_p)$:

$$\begin{aligned}s_\rho(\boldsymbol{S}_p) &= \frac{\partial}{\partial \rho}\left(-\frac{1}{2}\log\left[\{1+(L-2)\,\rho\}\,(1-\rho)^{L-2}\right] - \frac{1}{2}\sum_{j=2}^{L}\sum_{i=2}^{L}\frac{b_{pi}b_{pj}}{\sigma_i\sigma_j}\left\{c_d \mathbb{1}(i=j) + c_n \mathbb{1}(i\neq j)\right\}\right) \\ &= -\frac{1}{2}\left[\frac{L-2}{1+(L-2)\,\rho} - \frac{L-2}{1-\rho} + \sum_{j=2}^{L}\sum_{i=2}^{L}\frac{b_{pi}b_{pj}}{\sigma_i\sigma_j}\left\{\frac{\partial c_d}{\partial \rho}\mathbb{1}(i=j) + \frac{\partial c_n}{\partial \rho}\mathbb{1}(i\neq j)\right\}\right].\end{aligned} \tag{B.16}$$

Taking the necessary derivatives:

$$\frac{\partial c_d}{\partial \rho} = \frac{(1-\rho)\,\{1+\rho(L-2)\}\,(L-3) - \{1+\rho(L-3)\}\,\{(L-3)-2\rho(L-2)\}}{\left[(1-\rho)\,\{1+\rho(L-2)\}\right]^2}, \tag{B.17}$$

$$\frac{\partial c_n}{\partial \rho} = -\frac{(1-\rho)\,\{1+\rho(L-2)\} - \rho\,\{(L-3)-2\rho(L-2)\}}{\left[(1-\rho)\,\{1+\rho(L-2)\}\right]^2}. \tag{B.18}$$

It can be seen that every term in $s_\rho(\boldsymbol{S}_p)$ has nonzero denominator for ρ corresponding to Assumption 2.2 and σ_l^2 for all $l \in \mathcal{L}$ corresponding to Assumption 2.1. Therefore, $s_\rho(\boldsymbol{S}_p)$ is bounded, which we denote by $|s_\rho(\boldsymbol{S}_p)| < c_\rho$.

We have thus shown that all elements of $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{S}_p)$ are bounded, and therefore, the first derivative of the log-likelihood is bounded. This proves consistency of the estimator. □

Proof of Theorem 2.2. The consistency of the maximum likelihood estimator (MLE), along with differentiability of the log-likelihood, was proven in Theorem 2.1. By inspection, the components of the score vector are smooth and all second derivatives exist. Thus, under mild regularity conditions, standard MLE theory applies, and $\sqrt{m}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ is asymptotically normal with mean $\mathbf{0}$ and covariance equal to the joint Fisher Information, $\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)^{-1}$. Given the calculations of the score in the proof of Theorem 2.1, $\boldsymbol{I}\left(\boldsymbol{\theta}_0\right)^{-1}$ can be calculated as the inverse of the expected outer product of the score vector, $E\left\{\left(\frac{\partial l}{\partial \boldsymbol{\theta}_0}\right)\left(\frac{\partial l}{\partial \boldsymbol{\theta}_0}\right)^{\top}\right\}^{-1}$. The details for this calculation follow standard arguments. □

Proof of Theorem 2.3. Here, we sketch a proof that $\boldsymbol{I}(\widehat{\boldsymbol{\theta}})^{-1} \xrightarrow{P} \boldsymbol{I}\left(\boldsymbol{\theta}_0\right)^{-1}$.

Let the score, $\frac{\partial l}{\partial \boldsymbol{\theta}}$, be reflected by $\boldsymbol{s}(\boldsymbol{\theta})$. Additionally, let $\boldsymbol{\Theta}_\epsilon = \{\theta : \|\theta - \theta_0\| \leqslant \epsilon\}$, where $\epsilon > 0$ is small enough that $\boldsymbol{\Theta}_\epsilon \in \boldsymbol{\Theta}_0$, and $\mathcal{F} = \{\boldsymbol{s}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}_\epsilon\}$. Let $\mathbb{P}_n(\boldsymbol{X}) = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i$ denote the empirical average (e.g. $\mathbb{P}_n(\boldsymbol{X}) = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i$, where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are realizations of the random variable, $\boldsymbol{X}$).

It is not difficult to verify that $\mathcal{F}$ is Glivenko-Cantelli with square integrable envelope, denoted by F. Now, Glivenko-Cantelli preservation results verify that $\mathcal{F} \cdot F$ is also Glivenko-Cantelli with integrable envelope F^2. Therefore:

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_\epsilon} \left| \mathbb{P}_n \left\{ \boldsymbol{s}(\boldsymbol{\theta}) \boldsymbol{s}(\boldsymbol{\theta})^{\top} \right\} - E_{\boldsymbol{\theta}_0} \left\{ \boldsymbol{s}(\boldsymbol{\theta}) \boldsymbol{s}(\boldsymbol{\theta})^{\top} \right\} \right| \xrightarrow{a.s.} 0. \tag{B.19}$$

By continuity of the map $\boldsymbol{\theta} \mapsto E_{\boldsymbol{\theta}_0}\left\{\boldsymbol{s}(\boldsymbol{\theta})\boldsymbol{s}(\boldsymbol{\theta})^{\top}\right\}$ along with consistency of $\widehat{\boldsymbol{\theta}}$, $\boldsymbol{I}(\widehat{\boldsymbol{\theta}}) \xrightarrow{P} \boldsymbol{I}(\boldsymbol{\theta}_0)$. By continuity of matrix inversion at a positive definite matrix, positive-definiteness of $\boldsymbol{I}(\boldsymbol{\theta}_0)$, and the Continuous Mapping Theorem, it holds that $\boldsymbol{I}(\widehat{\boldsymbol{\theta}})^{-1} \xrightarrow{P} \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}$.

□

Proof of Theorem 2.4. The goal of this proof is to show consistency of $\widehat{V}_{\widehat{U}}(d)$ for the true value function, $V(d)$ (i.e., $\widehat{V}_{\widehat{U}}(d) \xrightarrow{P} V(d)$). It is already known that $\widehat{V}(d) \xrightarrow{P} V(d)$ (Qian and Murphy

(2011)). Thus, proving that $\widehat{V}_{\widehat{U}}(d) \xrightarrow{P} \widehat{V}(d)$ would be sufficient. First,

$$\begin{aligned}\left|\widehat{V}_{\widehat{U}}(d) - \widehat{V}(d)\right| &= \left|\frac{\sum_{i=1}^{n} \frac{\widehat{U}_i \cdot \mathbb{1}\{A_i = d(\boldsymbol{X}_i)\}}{\pi(A_i|\boldsymbol{X}_i)}}{\sum_{i=1}^{n} \frac{\mathbb{1}\{A_i = d(\boldsymbol{X}_i)\}}{\pi(A_i|\boldsymbol{X}_i)}} - \frac{\sum_{i=1}^{n} \frac{U_i \cdot \mathbb{1}\{A_i = d(\boldsymbol{X}_i)\}}{\pi(A_i|\boldsymbol{X}_i)}}{\sum_{i=1}^{n} \frac{\mathbb{1}\{A_i = d(\boldsymbol{X}_i)\}}{\pi(A_i|\boldsymbol{X}_i)}}\right| \\ &\leqslant \sum_{i=1}^{n} w_i \left|\widehat{U}_i - U_i\right| \\ &\leqslant \max_i \left|\widehat{U}_i - U_i\right|,\end{aligned} \tag{B.20}$$

where $w_i = \dfrac{\mathbb{1}\left\{A_i = d(\boldsymbol{X}_i)\right\}/\pi(A_i|\boldsymbol{X}_i)}{\sum_{i=1}^{n} \mathbb{1}\left\{A_i = d(\boldsymbol{X}_i)\right\}/\pi(A_i|\boldsymbol{X}_i)}$ and the last inequality is because $\sum_{i=1}^{n} w_i = 1$. So it remains to show that $\left|\widehat{U}_i - U_i\right| \xrightarrow{P} 0$, which would complete the proof. Observe that:

$$\left|\widehat{U}_i - U_i\right| = \left|\frac{R_{i1} + \sum_{l=2}^{L} exp\left(\widehat{\boldsymbol{\beta}}_l^\top \boldsymbol{x}_i\right) R_{il}}{1 + \sum_{l=2}^{L} exp\left(\widehat{\boldsymbol{\beta}}_l^\top \boldsymbol{x}_i\right)} - \frac{R_{i1} + \sum_{l=2}^{L} exp\left(\boldsymbol{\beta}_l^\top \boldsymbol{x}_i\right) R_{il}}{1 + \sum_{l=2}^{L} exp\left(\boldsymbol{\beta}_l^\top \boldsymbol{x}_i\right)}\right|, \tag{B.21}$$

and let $\left|\widehat{U}_i - U_i\right|$ be denoted by $\left|\dfrac{a_1}{b_1} - \dfrac{a_2}{b_2}\right|$, where:

$$a_1 = R_{i1} + \sum_{l=2}^{L} exp\left(\widehat{\boldsymbol{\beta}}_l^\top \boldsymbol{x}_i\right) R_{il}, \tag{B.22}$$

$$a_2 = R_{i1} + \sum_{l=2}^{L} exp\left(\boldsymbol{\beta}_l^\top \boldsymbol{x}_i\right) R_{il}, \tag{B.23}$$

$$b_1 = 1 + \sum_{l=2}^{L} exp\left(\widehat{\boldsymbol{\beta}}_l^\top \boldsymbol{x}_i\right) \tag{B.24}$$

$$b_2 = 1 + \sum_{l=2}^{L} exp\left(\boldsymbol{\beta}_l^\top \boldsymbol{x}_i\right). \tag{B.25}$$

Using the equality, $\dfrac{a_1}{b_1} - \dfrac{a_2}{b_2} = \dfrac{a_1 - a_2}{b_1} - \dfrac{a_2(b_1 - b_2)}{b_1 b_2}$, the fact that $\widehat{\boldsymbol{\beta}}_l$ converges to a bounded quantity, and that the covariates are bounded, we inspect individual parts:

$$\frac{1}{b_1} = \frac{1}{\sum_{i=1}^{n}\left\{exp\left(-\widehat{\boldsymbol{\beta}}_l^\top \boldsymbol{x}_i\right)\right\}} \leqslant \frac{1}{\sum_{i=1}^{n}\left\{exp\left(-\left\|\widehat{\boldsymbol{\beta}}_l\right\| \|\boldsymbol{x}_i\|\right)\right\}} = O_p(1), \tag{B.26}$$

$$\frac{1}{b_1 b_2} = O_p(1)O_p(1) = O_p(1) \tag{B.27}$$

$$a_2 = O_p(1), \tag{B.28}$$

where $O_p(1)$ is universal over all i. This leads to:

$$\left|\widehat{U}_i - U_i\right| \leqslant \left|O_p(1)(a_1 - a_2) + O_p(1)(b_1 - b_2)\right|. \tag{B.29}$$

Inspecting $a_1 - a_2$, we see that:

$$\begin{aligned} exp\left(\widehat{\boldsymbol{\beta}}_l^\top \boldsymbol{x}_i - \boldsymbol{\beta}_l^\top \boldsymbol{x}_i\right) &\leqslant \left|exp\left(\widehat{\boldsymbol{\beta}}_l^\top \boldsymbol{x}_i - \boldsymbol{\beta}_l^\top \boldsymbol{x}_i\right)\right| \\ &\leqslant exp\left(\boldsymbol{\beta}_l^\top \boldsymbol{x}_i\right)\left|exp\left\{(\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)^\top \boldsymbol{x}_i\right\} - 1\right|. \end{aligned} \tag{B.30}$$

Here, $exp\left(\boldsymbol{\beta}_l^\top \boldsymbol{x}_i\right)$ is bound by universal constant $exp\left(\|\boldsymbol{\beta}_l\| \|\boldsymbol{x}_i\|\right)$, and $\left|\left(\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l\right)^\top \boldsymbol{x}_i\right| \leqslant o_p(1) \|\boldsymbol{x}_i\| = o_p(1)$. Given that $|exp\{o_p(1)\} - 1| = o_p(1)$, $a_1 - a_2 = o_p(1)$, and showing $b_1 - b_2 = o_p(1)$ follows similarly. This leads to:

$$\begin{aligned} \left|\widehat{U}_i - U_i\right| &\leqslant |O_p(1)o_p(1) + O_p(1)o_p(1)| \\ &\leqslant |o_p(1) + o_p(1)| = o_p(1). \end{aligned} \tag{B.31}$$

With $\left|\widehat{U}_i - U_i\right| \xrightarrow{P} 0$, the proof is complete, and $\widehat{V}_{\widehat{U}}(d) \xrightarrow{P} V(d)$.

For $\pi(A_i|\boldsymbol{X}_i)$ estimated such that $\widehat{\pi}(A_i|\boldsymbol{X}_i) \xrightarrow{P} \pi(A_i|\boldsymbol{X}_i)$, it is also known that $\widehat{V}_{\widehat{\pi}}(d) \xrightarrow{P} V(d)$ (Jiang (2020)). In this case, the proof for $\widehat{V}_{\widehat{U},\widehat{\pi}}(d) \xrightarrow{P} \widehat{V}_{\widehat{\pi}}(d)$ would follow the exact same steps as above, with $\widehat{\pi}(A_i|\boldsymbol{X}_i)$ replacing $\pi(A_i|\boldsymbol{X}_i)$. □

APPENDIX C: SECONDARY ANALYSES FOR CHAPTER 3

C.1 Optimal Decision Rule for Secondary Outcome

For the outcome of *change in hypoglycemia*, interestingly, the optimal decision rule was found to be a policy tree algorithm with a depth of just 1 (Table C.1). The rule suggests that study participants with baseline %CV $>34\%$ would experience a greater reduction in hypoglycemia using CGM compared to BGM.

Table C.1: Training and validation set value estimates of potential decision rules, along with test set evaluation of final rule, for the secondary outcome (% reduction of time in hypoglycemia). The "optimal method" was decided as the method with optimal (highest) inner validation set value; only that method was evaluated on the held-out test set in order to ensure honest cross validation.

Policy Tree - Parameter Tuning			Final Evaluation (of optimal method)
Depth	*Training Set Value*	*Inner Validation Set Value*	on Held-Out Test Set
1	3.64	**3.51**	**3.68**
2	3.98	3.21	–
3	4.30	3.14	–
Decision List - Parameter Tuning			
Depth	*Training Set Value*	*Inner Validation Set Value*	
1	3.48	3.49	–
2	3.49	3.44	–
3	3.53	3.49	–

*Note, for comparison, that the estimated value of "CGM-only" rule on the held-out test set was 3.50%.

C.2 Evaluation of Decision Rule

Figure C.1 visualizes the differential treatment effects of CGM and BGM in terms of estimated decrease in hypoglycemia across varying levels of the moderating marker, %CV. The decision rule suggests greater benefit of CGM for 173 (89%) WISDM participants and greater benefit with BGM for 21 (11%) WISDM participants (Table C.2). The mean %CV in the CGM group was 42% compared with 30.9% in the BGM group. The participants for whom CGM was of greater benefit than BGM had higher baseline time spent in hypoglycemia (hypoglycemia (7.3% versus 1.7%; $p<0.0001$), in addition to higher baseline daily insulin doses (0.056 units/kg versus 0.44 units/kg; $p<0.05$), baseline HbA1c (7.59% versus 7.26%; $p<0.10$), earlier age at diagnosis (31.4 years versus 40.1 years; $p<0.01$), and higher percent with undetectable C-peptide levels (79% versus 57%); $p<0.10$) compared to BGM (n=21). The optimal decision rule was estimated to reduce hypoglycemia by an average of 3.68% (SEM 0.27%) across the full study population, compared to 3.50% (SEM 0.26%) with the "CGM-only" rule. For context, use of "BGM-only" was estimated to reduce hypoglycemia 0.43% (SEM 0.37%).

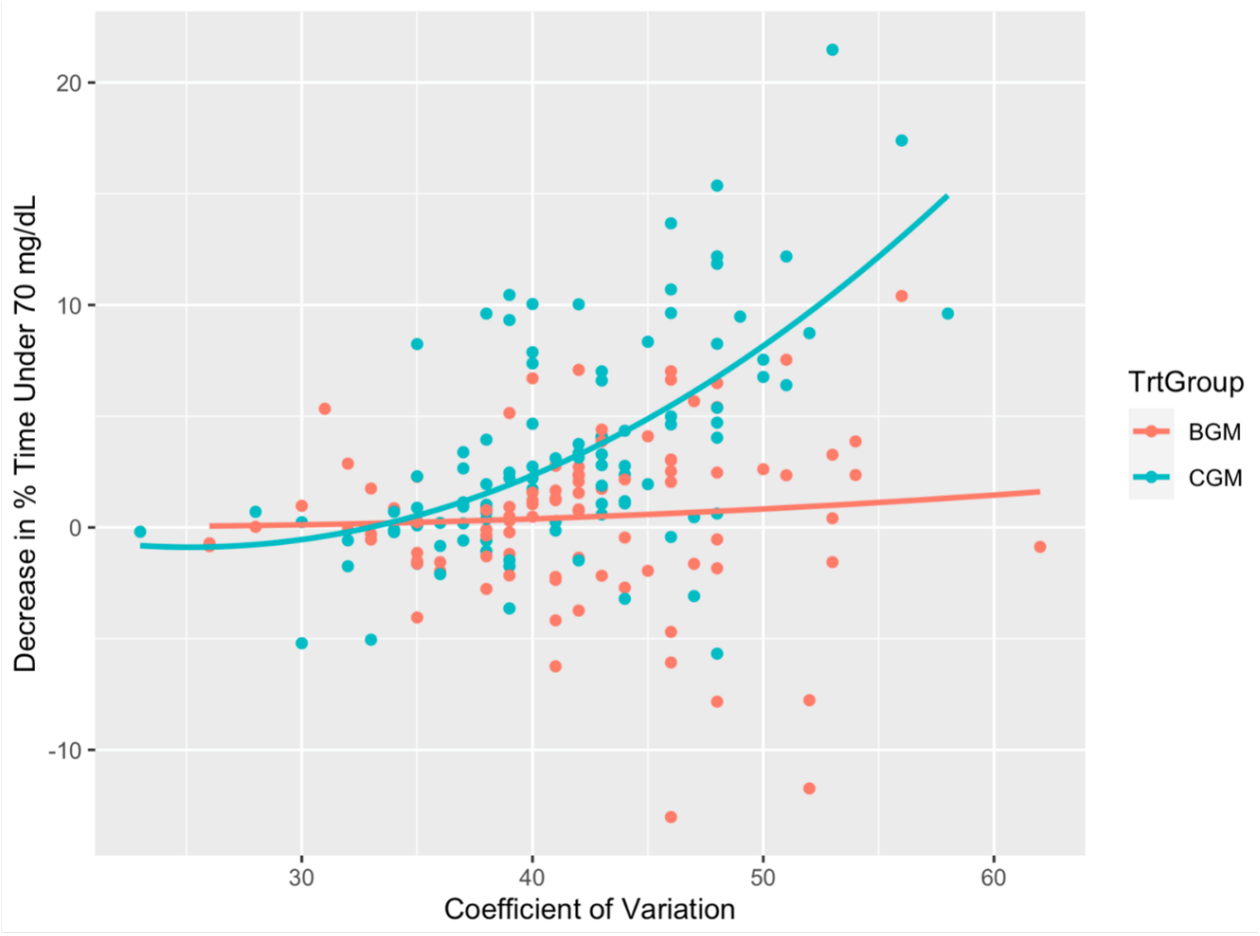

Figure C.1: Differences in treatment effect of CGM vs BGM by %CV, among WISDM study participants. The outcome depicted is reduction in hypoglycemia. Curves reflect polynomial fits (degree 2): $Y = \beta_0 + \beta_1 * \%CV + \beta_2 * \%CV^2$. For the CGM treatment group, estimated parameters for β_0, β_1, and β_2, respectively, are 0.035, 0.285, and 0.081; for the BGM group, estimated parameters are 0.005, 0.026, 0.006.

Table C.2: Characteristics of study participants, stratified by decision rule subgroup. P-values for differences in means were calculated with a 2-sample t-test and differences in proportions with a 2-proportion Z-test. Abbreviations: SD, standard deviation.

Characteristic, n (%) or mean (SD)	**Decision Rule Subgroup**		*P-value*
	CGM (n=173)	*BGM (n=21)*	
Moderating Marker			
Coefficient of variation, %	42.9 (5.3)	30.9 (3.1)	<.0001[a]
Demographic Characteristics			
Age, years	67.9 (5.7)	69.8 (6.0)	.19
Diabetes duration, years	36.5 (15.4)	29.8 (18.7)	.12
Age at diagnosis, years	31.4 (16.5)	40.0 (19.3)	.06*
Male sex	83 (48.0%)	10 (47.6%)	1
Non-Hispanic ethnicity	157 (90.8%)	21 (100%)	.30
White race	162 (93.6%)	21 (100%)	.49
Highest education			
Less than a bachelor's degree	70 (40.5%)	5 (23.8%)	.21
Bachelor's degree	55 (31.8%)	7 (33.3%)	1
Graduate or professional degree	48 (27.7%)	9 (42.9%)	.24
Health insurance			
Private	49 (28.3%)	3 (14.3%)	.27
Private and Medicare	58 (33.5%)	9 (42.9%)	.54
Medicare/other	66 (38.2%)	9 (42.9%)	.86
Clinical Characteristics			
Insulin pump use	93 (53.8%)	9 (42.9%)	.48
Screening HbA1c, %	7.59 (0.91)	7.26 (0.79)	.09*
Detectable C-peptide	37 (21.4%)	9 (42.9%)	.06*
Total daily insulin dose, units/kg	0.561 (0.211)	0.441 (0.217)	.03**
Body mass index, kg/m^2	26.51 (4.30)	26.89 (4.34)	.84
$\geqslant$ 1 severe hypoglycemia event in the past 12 months	27 (15.6%)	1 (4.8%)	.31
$\geqslant$ 1 diabetic ketoacidosis event in the past 12 months	8 (4.6%)	0 (0%)	.67
Hypoglycemia (time w glucose <70 mg/dL), %	7.3 (5.0)	1.7 (1.6)	<.0001[a]
Time with glucose in range of 70-180 mg/dL, %	55.6 (12.7)	60.9 (21.5)	.28

[a]Significance expected since decision rule is based on this marker.
*P $<$ .1, **P $<$ .05

REFERENCES

ADA Professional Practice Committee (2022a). 13. Older adults: standards of medical care in diabetes2022. *Diabetes Care*, **45**(suppl 1), S195–S207.

ADA Professional Practice Committee (2022b). 6. Glycemic targets: standards of medical care in diabetes2022. *Diabetes Care*, **45**, S83–S96.

ADA Professional Practice Committee (2022c). 7. Diabetes technology: standards of medical care in diabetes2022. *Diabetes Care*, **45**(suppl 1), S97–S112.

Arora, S. and Doshi, P. (2021). A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, **297**, 103500.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, **113**(27), 7353–7360.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**(1), 289–300.

Benkeser, D., Mertens, A., Colford, J. M., Hubbard, A., Arnold, B. F., Stein, A., and Laan, M. J. v. d. (2021). A machine learning-based approach for estimating and testing associations with multivariate outcomes. *The International Journal of Biostatistics*, **17**(1), 7–21.

Bergenstal, R. M. (2018). Understanding Continuous Glucose Monitoring Data. In *Role of Continuous Glucose Monitoring in Diabetes Treatment*. American Diabetes Association, Arlington (VA).

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

Butler, E. L. (2016). *Using patient preferences to estimate optimal treatment strategies for competing outcomes*. Ph.D., The University of North Carolina at Chapel Hill, United States – North Carolina.

Butler, E. L., Laber, E. B., Davis, S. M., and Kosorok, M. R. (2018). Incorporating Patient Preferences into Estimation of Optimal Individualized Treatment Rules. *Biometrics*, **74**(1), 18–26.

Casu, A., Kanapka, L., Foster, N., and al, e. (2020). Characteristics of adult-compared to childhood-onset type 1 diabetes. *Diabet Med*, **37**(12), 2109–2115.

Cevid, D., Michel, L., Nf, J., Bhlmann, P., and Meinshausen, N. (2022). Distributional Random Forests: Heterogeneity Adjustment and Multivariate Distributional Regression. *Journal of Machine Learning Research*, **23**(333), 1–79.

Chen, G., Zeng, D., and Kosorok, M. R. (2016). Personalized Dose Finding Using Outcome Weighted Learning. *Journal of the American Statistical Association*, **111**(516), 1509–1521.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA. Association for Computing Machinery.

Chen, Y., Zeng, D., and Wang, Y. (2021). Learning Individualized Treatment Rules for Multiple-Domain Latent Outcomes. *Journal of the American Statistical Association*, **116**(533), 269–282.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, **49**(4), 327–335.

Cho, H., Holloway, S. T., and Kosorok, M. R. (2020). Multi-stage optimal dynamic treatment regimes for survival outcomes with dependent censoring. *arXiv:2012.03294 [stat]*.

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

Danne, T., Nimri, R., Battelino, T., and al, e. (2017). International consensus on use of continuous glucose monitoring. *Diabetes Care*, **40**(12), 1631–1640.

De Boer, W. A. and Tytgat, G. N. J. (1995). Review: The Best Therapy for Helicobacter pylori Infection: Should Efficacy or Side-Effect Profile Determine Our Choice? *Scandinavian Journal of Gastroenterology*, **30**(5), 401–407.

Divan, V., Greenfield, M., Morley, C., and Weinstock, R. (2022). Perceived burdens and benefits associated with continuous glucose monitor use in type 1 diabetes across the lifespan. *J Diabetes Sci Technol*, **16**(1), 88–96.

DuBose, S., Weinstock, R., Beck, R., and al, e. (2016). Hypoglycemia in older adults with type 1 diabetes. *Diabetes Technol Ther*, **18**(12), 765–771.

Embretson, S. E. and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Psychology Press, New York.

Fan, C., Lu, W., Song, R., and Zhou, Y. (2017). Concordance-Assisted Learning for Estimating Optimal Individualized Treatment Regimes. *JRSSB*, **79**(5), 1565–1582.

Fan, J., Lv, F., Shi, L., ,Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong, China, and ,School of Mathematical Sciences, Shanghai Key Laboratory for Contemporary Applied Mathematics, Fudan University, Shanghai, 200433, China (2019). An RKHS approach to estimate individualized treatment rules based on functional predictors. *Mathematical Foundations of Computing*, **2**(2), 169–181.

Fang, E. X., Wang, Z., and Wang, L. (2022). Fairness-Oriented Learning for Optimal Individualized Treatment Rules. *Journal of the American Statistical Association*, pages 1–14.

Fauzan, M. A. and Murfi, H. (2018). The Accuracy of XGBoost for Insurance Claim Prediction. *Int. J. Advance Soft Compu*, **10**(2), 13.

Flatt, A., Chen, E., Peleckis, A., and al, e. (2022). Evaluation of clinical metrics for identifying defective physiologic responses to hypoglycemia in long-standing type 1 diabetes. *Diabetes Technol Ther*, **24**(10), 737–748.

Forlenza, G., Argento, N., and Laffel, L. (2017). Practical considerations on the use of continuous glucose monitoring in pediatrics and older adults and nonadjunctive use. *Diabetes Technol Ther*, **19**(S3), S13–S20.

Freeman, N. L. B., Browder, S. E., McGinigle, K. L., and Kosorok, M. R. (2022). Dynamic treatment regime characterization via value function surrogate with an application to partial compliance. arXiv:2212.00650 [stat].

Gomez, A., Henao, D., Imitola Madero, A., and al, e. (2019). Defining high glycemic variability in type 1 diabetes: comparison of multiple indexes to identify patients at risk of hypoglycemia. *Diabetes Technol Ther*, **21**(8), 430–439.

Gomez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernndez-Lobato, J. M., Snchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, **4**(2), 268–276.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gubitosi-Klug, R., Braffett, B., Bebu, I., and al, e. (2022). Continuous glucose monitoring in adults with type 1 diabetes with 35 years duration from the DCCT/EDIC study. *Diabetes Care*, **45**, 659–665.

Hammer, S. M. et al. (1996). A Trial Comparing Nucleoside Monotherapy with Combination Therapy in HIV-Infected Adults with CD4 Cell Counts from 200 to 500 per Cubic Millimeter. *New England Journal of Medicine*, **335**(15), 1081–1090.

Han, H., Guo, X., and Yu, H. (2016). Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 219–224. ISSN: 2327-0594.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Support Vector Machines and Flexible Discriminants. In T. Hastie, R. Tibshirani, and J. Friedman, editors, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, pages 417–458. Springer, New York, NY.

Hernn, M. A. and Robins, J. M. (2019). *Causal Inference: What If*. CRC Press.

Holt, R., DeVries, J., Hess-Fischl, A., and al, e. (2021). The management of type 1 diabetes in adults. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care*, **44**(11), 2589–2625.

Homayouni, H., Ghosh, S., Ray, I., Gondalia, S., Duggan, J., and Kahn, M. G. (2020). An Autocorrelation-based LSTM-Autoencoder for Anomaly Detection on Time-Series Data. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 5068–5077, Atlanta, GA, USA. IEEE.

Hussein, A., Gaber, M. M., Elyan, E., and Jayne, C. (2017). Imitation Learning: A Survey of Learning Methods. *ACM Computing Surveys*, **50**(2), 21:1–21:35.

Imai, K. and Li, M. L. (2021). Experimental Evaluation of Individualized Treatment Rules. *Journal of the American Statistical Association*, pages 1–15.

Irony, T. Z. (2017). The Utility in Composite Outcome Measures: Measuring What Is Important to Patients. *JAMA*, **318**(18), 1820–1821.

Jacob, D. (2021). CATE meets ML – The Conditional Average Treatment Effect and Machine Learning. *arXiv:2104.09935 [econ]*. arXiv: 2104.09935.

Jiang, X. (2020). *Developing Machine Learning Methodology for Precision Health*. Ph.D., The University of North Carolina at Chapel Hill, United States – North Carolina. ISBN: 9798635265857.

Kahkoska, A. R., Shah, K. S., Kosorok, M. R., Miller, K. M., Rickels, M., Weinstock, R. S., Young, L. A., and Pratley, R. E. (2023). Estimation of a Machine Learning-Based Decision Rule to Reduce Hypoglycemia Among Older Adults With Type 1 Diabetes: A Post Hoc Analysis of Continuous Glucose Monitoring in the WISDM Study. *Journal of Diabetes Science and Technology*, page 19322968221149040. Publisher: SAGE Publications Inc.

Kallus, N. (2018). Balanced Policy Evaluation and Learning. *Advances in neural information processing systems*, **31**.

Kean, J., Brodke, D. S., Biber, J., and Gross, P. (2018). An introduction to Item Response Theory and Rasch Analysis of the Eating Assessment Tool (EAT-10). *Brain impairment : a multidisciplinary journal of the Australian Society for the Study of Brain Impairment*, **19**(Spec Iss 1), 91–102.

Kent, D., Steyerberg, E., and van Klaveren, D. (2018). Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*, **363**, k4245.

Kent, D. M., Paulus, J. K., van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P., Patrick-Lake, B., Morton, S., Pencina, M., Raman, G., Ross, J. S., Selker, H. P., Varadhan, R., Vickers, A., Wong, J. B., and Steyerberg, E. W. (2020). The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. *Annals of Internal Medicine*, **172**(1), 35–45.

Kirkman, M., Briscoe, V., Clark, N., and al, e. (2012). Diabetes in older adults: a consensus report. *J Am Geriatr Soc*, **60**(12), 2342–2356.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, **17**(1), 107–112.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York, New York, NY.

Kosorok, M. R. and Laber, E. B. (2019). Precision Medicine. *Annual review of statistics and its application*, **6**, 263–286.

Kosorok, M. R. and Moodie, E. E. M. (2016). *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. ASA-SIAM statistics and applied probability. Society for Industrial and Applied Mathematics, Philadelphia.

Krishnaswami, A., Beavers, C., Dorsch, M., and al, e. (2020). Gerotechnology for older adults with cardiovascular diseases: JACC state-of-the-art review. *J Am Coll Cardiol*, **76**(22), 2650–2670.

Laber, E. B., Lizotte, D. J., and Ferguson, B. (2014). Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, **70**(1), 53–61.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, **40**(3), 1846–1877.

Liang, M. and Yu, M. (2020). A Semiparametric Approach to Model Effect Modification. *Journal of the American Statistical Association*, **0**(0), 1–13.

Liang, M., Ye, T., and Fu, H. (2018). Estimating individualized optimal combination therapies through outcome weighted deep learning algorithms. *Statistics in Medicine*, **37**(27), 3869–3886.

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., and Zeng, D. (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. *Statistics in Medicine*, **37**(26), 3776–3788.

Lou, Z., Shao, J., and Yu, M. (2018). Optimal treatment assignment to maximize expected outcome with multiple treatments. *Biometrics*, **74**(2), 506–516.

Luckett, D. J., Laber, E. B., Kim, S., and Kosorok, M. R. (2021). Estimation and Optimization of Composite Outcomes. *Journal of Machine Learning Research*, **22**(167), 1–40.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**(86), 2579–2605.

Mebane, W. R. and Sekhon, J. S. (2011). Genetic Optimization Using Derivatives: The rgenoud Package for R. *Journal of Statistical Software*, **42**, 1–26.

Meng, H. and Qiao, X. (2021). Doubly Robust Direct Learning for Estimating Conditional Average Treatment Effect. *arXiv:2004.10108 [stat]*.

Meng, H. and Qiao, X. (2022). Augmented direct learning for conditional average treatment effect estimation with double robustness. *EJS*, **16**(1), 3523–3560.

Mi, X., Zou, F., and Zhu, R. (2019). Bagging and deep learning in optimal individualized treatment rules. *Biometrics*, **75**(2), 674–684.

Miller, S. and Startz, R. (2018). Feasible Generalized Least Squares Using Machine Learning. SSRN Scholarly Paper ID 2966194, Social Science Research Network, Rochester, NY.

Mo, W. and Liu, Y. (2021). Efficient learning of optimal individualized treatment rules for heteroscedastic or misspecified treatment-free effect models. *JRSSB*, **84**(2), 440–472.

Mocroft, A. et al. (2013). The Incidence of AIDS-Defining Illnesses at a Current CD4 Count greater than 200 Cells/microliter in the Post-Combination Antiretroviral Therapy Era. *Clinical Infectious Diseases*, **57**(7), 1038–1047.

Monnier, L., Wojtusciszyn, A., Molinari, N., and al, e. (2020). Respective contributions of glycemic variability and mean daily glucose as predictors of hypoglycemia in type 1 diabetes: are they equivalent? *Diabetes Care*, **43**(4), 821–827.

Muelling, K., Boularias, A., Mohler, B., Schlkopf, B., and Peters, J. (2014). Learning strategies in table tennis using inverse reinforcement learning. *Biological Cybernetics*, **108**(5), 603–619.

Munshi, M., Meneilly, G., Rodrguez-Maas, L., and al, e. (2020). Diabetes in ageing: pathways for developing the evidence base for clinical guidance. *Lancet Diabetes Endocrinol*, **8**(10), 855–867.

Munshi, M., Slyne, C., Davis, D., and al, e. (2022). Use of technology in older adults with type 1 diabetes: clinical characteristics and glycemic metrics. *Diabetes Technol Ther*, **24**(1), 1–9.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *JRSSB*, **65**(2), 331–355.

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

Murray, T. A., Thall, P. F., and Yuan, Y. (2016). Utility-based designs for randomized comparative trials with categorical outcomes. *Statistics in Medicine*, **35**(24), 4285–4305.

Nandy, R. R. and Cordes, D. (2003). Novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data. *Magnetic Resonance in Medicine*, **50**(2), 354–365.

Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming*, **140**(1), 125–161.

Nezhad, M. Z., Zhu, D., Li, X., Yang, K., and Levy, P. (2016). SAFS: A Deep Feature Selection Approach for Precision Medicine. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 501–506. arXiv:1704.05960 [cs, stat].

Ng, A. Y. and Russell, S. J. (2000). Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 663–670, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Olive, D. J. (2017). WLS and Generalized Least Squares. In *Linear Regression*, pages 163–173. Springer International Publishing, Cham.

Pollock, K. H. (2002). The use of auxiliary variables in capture-recapture modelling: An overview. *Journal of Applied Statistics*, **29**(1-4), 85–102.

Pratley, R., Kanapka, L., Rickels, M., and al, e. (2020). Effect of continuous glucose monitoring on hypoglycemia in older adults with type 1 diabetes: a randomized clinical trial. *JAMA*, **323**(23), 2397–2406.

Press, W. H. and Teukolsky, S. A. (2007). *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press.

Pu, H. and Zhang, B. (2021). Estimating Optimal Treatment Rules with an Instrumental Variable: A Partial Identification Learning Approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **83**(2), 318–345. arXiv: 2002.02579.

Qi, Z. and Liu, Y. (2018). D-learning to estimate optimal individual treatment rules. *EJS*, **12**(2), 3601–3638.

Qi, Z., Liu, D., Fu, H., and Liu, Y. (2020). Multi-Armed Angle-Based Direct Learning for Estimating Optimal Individualized Treatment Rules With Various Outcomes. *Journal of the American Statistical Association*, **115**(530), 678–691.

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

Rashid, N., Hossain, M. A. F., Ali, M., Islam Sukanya, M., Mahmud, T., and Fattah, S. A. (2021a). AutoCovNet: Unsupervised feature learning using autoencoder and feature merging for detection of COVID-19 from chest X-ray images. *Biocybernetics and Biomedical Engineering*, **41**(4), 1685–1701.

Rashid, N. U., Luckett, D. J., Chen, J., Lawson, M. T., Wang, L., Zhang, Y., Laber, E. B., Liu, Y., Yeh, J. J., Zeng, D., and Kosorok, M. R. (2021b). High-Dimensional Precision Medicine From Patient-Derived Xenografts. *Journal of the American Statistical Association*, **116**(535), 1140–1154.

Ringnr, M. (2008). What is principal component analysis? *Nature Biotechnology*, **26**(3), 303–304.

Robins, J. M. (2004). Optimal Structural Nested Models for Optimal Sequential Decisions. In D. Y. Lin and P. J. Heagerty, editors, *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, Lecture Notes in Statistics, pages 189–326. Springer, New York, NY.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**(5), 688–701.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*, **1**(5), 206–215.

Ruedy, K., Parkin, C., Riddlesworth, T., Graham, C., and Group, D. S. (2017). Continuous glucose monitoring in older adults with type 1 and type 2 diabetes using multiple daily injections of insulin: results from the DIAMOND trial. *J Diabetes Sci Technol*, **11**(6), 1138–1146.

Sainath, T. N., Kingsbury, B., and Ramabhadran, B. (2012). Auto-encoder bottleneck features using deep belief networks. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4153–4156.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Stat Sci*, **29**(4), 640–661.

Shah, K. S., Fu, H., and Kosorok, M. R. (2022). Stabilized direct learning for efficient estimation of individualized treatment rules. *Biometrics*.

Shandhi, M. M. H. and Dunn, J. P. (2022). AI in medicine: Where are we now and where are we going? *Cell Reports Medicine*, **3**(12), 100861.

Sobel, M. E. (1994). Causal inference in latent variable models. In *Latent variables analysis: Applications for developmental research*, pages 3–35. Sage Publications, Inc, Thousand Oaks, CA, US.

Song, P. X.-K. (2007). Mixed-Effects Models: Bayesian Inference. In P. X.-K. Song, editor, *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer Series in Statistics, pages 195–215. Springer, New York, NY.

Swartz, M. S., Stroup, T. S., McEvoy, J. P., Davis, S. M., Rosenheck, R. A., Keefe, R. S. E., Hsiao, J. K., and Lieberman, J. A. (2008). What CATIE Found: Results From the Schizophrenia Trial. *Psychiatric services (Washington, D.C.)*, **59**(5), 500–506.

Szkely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, **3**(4), 1236–1265.

Thompson, B. (1984). *Canonical Correlation Analysis: Uses and Interpretation*. SAGE.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. *Journal of the American Statistical Association*, **109**(508), 1517–1532.

Toschi, E. and Munshi, M. (2020). Benefits and challenges of diabetes technology use in older adults. *Endocrinol Metab Clin North Am*, **49**(1), 57–67.

Toschi, E., Slyne, C., Sifre, K., and al, e. (2020). The relationship between CGM-derived metrics, A1C, and risk of hypoglycemia in older adults with type 1 diabetes. *Diabetes Care*, **43**(10), 2349–2354.

Trusheim, M., Berndt, E., and Douglas, F. (2007). Stratified medicine: strategic and economic implications of combining drugs and clinical biomarkers. *Nat Rev Drug Discov*, **6**(4), 287–293.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, **6**(1).

Vapnik, V. (1999). *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, **113**(523), 1228–1242.

Wallace, M. P., Moodie, E. E. M., and Stephens, D. A. (2018). Reward ignorant modeling of dynamic treatment regimes. *Biometrical Journal*, **60**(5), 991–1002.

Wang, D., Fu, H., and Loh, P.-L. (2020). Boosting Algorithms for Estimating Optimal Individualized Treatment Rules. Technical Report arXiv:2002.00079, arXiv.

Wang, Y., Yao, H., and Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing*, **184**, 232–242.

Weinstein, J. M., Berkowitz, S. A., Pratley, R. E., Shah, K. S., and Kahkoska, A. R. (2023). Statistically Adjusting for Wear Time in Randomized Trials of Continuous Glucose Monitors as a Complement to Intent-to-Treat and As-Treated Analyses: Application and Evaluation in Two Trials. *Diabetes Technology & Therapeutics*, **25**(7), 457–466.

Weinstock, R., DuBose, S., Bergenstal, R., and al, e. (2016). Risk factors associated with severe hypoglycemia in older adults with type 1 diabetes. *Diabetes Care*, **39**(4), 603–610.

Wu, F., Laber, E. B., Lipkovich, I. A., and Severus, E. (2015). Who will benefit from antidepressants in the acute treatment of bipolar depression? A reanalysis of the STEP-BD study by Sachs et al. 2007, using Q-learning. *International Journal of Bipolar Disorders*, **3**(1), 7.

Xiao, W., Zhang, H. H., and Lu, W. (2019). Robust Regression for Optimal Individualized Treatment Rules. *Statistics in Medicine*, **38**(11), 2059–2073.

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

Zhang, C. and Liu, Y. (2014). Multicategory angle-based large-margin classification. *Biometrika*, **101**(3), 625–640.

Zhang, C., Chen, J., Fu, H., He, X., Zhao, Y.-Q., and Liu, Y. (2020). Multicategory Outcome Weighted Margin-based Learning for Estimating Individualized Treatment Rules. *Statistica Sinica*, **30**, 1857–1879.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *The Annals of Statistics*, **36**(4), 1567–1594.

Zhang, J., Troxel, A. B., and Petkova, E. (2021). Robust index of confidence weighted learning for optimal individualized treatment rule estimation. *Stat*, **10**(1), e374.

Zhang, X. D. (2015). Precision Medicine, Personalized Medicine, Omics and Big Data: Concepts and Relationships. *Journal of Pharmacogenomics & Pharmacoproteomics*, **06**(02).

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

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, **107**(499), 1106–1118.

Zhao, Y., Laber, E. B., Ning, Y., Saha, S., and Sands, B. E. (2019). Efficient augmentation and relaxation learning for individualized treatment rules using observational data. *Journal of Machine Learning Research*, **20**, 48.

Zhifei, S. and Meng Joo, E. (2012). A survey of inverse reinforcement learning techniques. *International Journal of Intelligent Computing and Cybernetics*, **5**(3), 293–311.

Zhou, X. and Kosorok, M. R. (2017). Causal nearest neighbor rules for optimal treatment regimes. arXiv:1711.08451 [stat].

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017). Residual Weighted Learning for Estimating Individualized Treatment Rules. *Journal of the American Statistical Association*, **112**(517), 169–187.

Zhou, Z., Athey, S., and Wager, S. (2022). Offline multi-action policy learning: Generalization and optimization. *Operations Research*, **70**(4), 832–849.

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S., and Zhao, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, **73**(2), 391–400.