

University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

1-2020

Aspects of causal inference.

John A. Craycroft
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>



Part of the [Applied Statistics Commons](#), and the [Biostatistics Commons](#)

Recommended Citation

Craycroft, John A., "Aspects of causal inference." (2020). *Electronic Theses and Dissertations*. Paper 3557.

<https://doi.org/10.18297/etd/3557>

This Doctoral Dissertation is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

ASPECTS OF CAUSAL INFERENCE

By

John Craycroft

B.A., B.S. The George Washington University, 1998

MBA, Emory University, 2004

M.S., University of Louisville, 2016

A Dissertation Submitted to the Faculty of the
School of Public Health and Information Sciences of the
University of Louisville
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, KY

December 2020

Copyright 2020 by John Anthony Craycroft

All rights reserved

ASPECTS OF CAUSAL INFERENCE

By

John Craycroft
B.A., B.S. The George Washington University, 1998
MBA, Emory University, 2004
M.S., University of Louisville, 2016

A Dissertation Approved on

October 30, 2020

by the following Dissertation Committee:

Dr. Maiying Kong, Committee Chair

Dr. Bertis Little

Dr. Douglas Lorenz

Dr. Subhadip Pal

Dr. Shesh Rai

DEDICATION

This dissertation is dedicated to my uncle, Fr. Leo Craycroft, who has always been a model of generosity, good humor, and genuine care for and interest in others, and who never fails to ask me about my progress.

ACKNOWLEDGMENTS

I would like to thank Dr. Maiying Kong, my dissertation advisor, for her guidance and patience. I would also like to thank the other committee members, Dr. Bertis Little, Dr. Douglas Lorenz, Dr. Subhadip Pal, and Dr. Shesh Rai for their time and input. In particular, I express thanks to Dr. Pal for his extra time with me on the Bayesian analysis.

I would also like to express thanks to my wife, Laurie, who has sacrificed as much as, if not more than I have during this time, even while providing me encouragement to push through. I am also grateful to my three sons, Ian, Leo, and Miles, and recognize this has been a challenging time for them too. Likewise, I thank my mother- and father-in-law, Sarah and Ed Hughes, for their support in so many ways, and my own parents and siblings, for their interest and support. Finally, I am so grateful to Joe Kufera, a true friend who has provided an open ear and insightful advice throughout this journey.

One last acknowledgment: in this (initial?) year of the COVID-19 pandemic (which, incidentally, provides a vivid example of the omnipresent need for rigorous, nuanced, and honest statistical analysis), I want to acknowledge those who have been directly affected – all those who have fallen ill, or lost a loved one – as well as those who have demonstrated empathy and acted, not primarily in self interest, but in recognition of our shared humanity.

ABSTRACT
ASPECTS OF CAUSAL INFERENCE

John A. Craycroft

December 12, 2020

Observational studies differ from experimental studies in that assignment of subjects to treatments is not randomized but rather occurs due to natural mechanisms, which are usually hidden from researchers. Yet objectives of the two studies are frequently the same: identify the causal – rather than merely associational – relationship between some treatment or exposure and an outcome. The statistical issues that arise in properly analyzing observational data for this goal are numerous and fascinating, and these issues are encompassed in the domain of causal inference. The research presented in this dissertation explores several distinct aspects of causal inference.

This dissertation is divided into four chapters. Chapter One gives an introduction to major concepts, underlying assumptions, and analytical frameworks encountered in the domain of causal inference. The next three chapters describe extensive research projects that are linked together by those threads.

Chapter Two deals with propensity score techniques and, more specifically, how to specify the propensity score model to achieve the best treatment effect estimates. This chapter not only provides a theoretical proof showing that one particular type of specification is best, but also demonstrates an original method for applying that result. The

method presented in Chapter Three has a similar purpose – obtaining precise and accurate estimates of causal effects – but views the challenge through a Bayesian, rather than a frequentist, lens. Here, a hierarchical Bayesian model is developed that is grounded in the framework of causal inference.

While Chapters Two and Three focus on scenarios involving causal inference from observational data, Chapter Four presents a method that has been designed to apply equally well to experimental data. The intent of the research here is to provide a method for identifying subgroups of the population in which the treatment effect differs from the overall population average treatment effect. Maintaining a central theme of causal inference, the research focuses on avoiding confounding bias while identifying effect modifiers that characterize the subgroups.

In all, this dissertation is intended to provide views of causal inference concepts from several distinct angles, demonstrating the complexity and richness of this domain.

TABLE OF CONTENTS

	PAGE
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	v
LIST OF TABLES	ix
LIST OF FIGURES	x
 CHAPTER 1: INTRODUCTION	 1
1.1 Potential outcomes and causal inference.....	1
1.2 Measures of causal effect.....	3
1.3 Assumptions required for valid causal inference.....	4
1.4 Structure of this dissertation.....	4
 CHAPTER 2: PROPENSITY SCORE SPECIFICATION FOR OPTIMAL ESTIMATION OF AVERAGE TREATMENT EFFECT WITH BINARY RESPONSE	 6
2.1 Introduction.....	6
2.2 Methods.....	11
2.3 Simulation.....	17
2.4 Case study.....	26
2.5 Discussion and conclusions.....	32
 CHAPTER 3: BAYESIAN CAUSAL INFERENCE USING MCMC	 36
3.1 Introduction.....	36
3.2 Methods.....	39
3.3 Simulation.....	43
3.4 Case study.....	50
3.5 Discussion and conclusions.....	54
 CHAPTER 4: INNOVATIVE APPROACH FOR SUBGROUP ANALYSIS	 56
4.1 Introduction.....	56
4.2 Methods.....	60
4.3 Simulation.....	65
4.4 Case study.....	72
4.5 Conclusions.....	78

REFERENCES	79
APPENDICES	87
A1: Proofs of theorems in Chapter 2.....	87
A2: Additional simulation results for Chapter 2.....	92
A3: Derivations of posterior conditional distributions in Chapter 3.....	95
A4: Alternative Bayesian model formulation using weighted likelihood.....	101
A5: Additional simulation results for Chapter 4.....	102
CURRICULUM VITA	104

LIST OF TABLES

TABLE		PAGE
2.1	Parameter values for simulation models varying covariate-treatment and covariate-outcome associations.....	19
2.2	Bias, standard error, and root MSE for selected simulation scenarios	25
2.3	Baseline distributions of variables by treatment and outcome	30
3.1	Comparison of Bayesian and IPW simulation study results	49
3.2	Baseline characteristics, stratified by treatment group	51
4.1	Settings for h -, g -, and k -functions in simulation scenarios	67
4.2	Simulation results from 1000 Monte Carlo repetitions for each of 15 scenarios, independent and dependent covariates.....	72
4.3	Baseline characteristics of patients, varenicline study	74
4.4	Subgroups identified in varenicline case study	77
A2.1	Bias, standard error, and root MSE for correlated simulation scenarios, moderate correlation (0.2).....	93
A2.2	Bias, standard error, and root MSE for correlated simulation scenarios, strong correlation (0.5).....	94
A5.1	Subgroup analysis simulation study results (N=400)	102
A5.2	Subgroup analysis simulation study results (N=200)	102
A5.3	Subgroup analysis simulation study results (N=100)	103

LIST OF FIGURES

FIGURE		PAGE
2.1	Causal diagrams	8
2.2	ATE estimates with 95% CIs	31
2.3	Radar chart of RMSE for tested methods	33
3.1	Boxplots of simulation study estimates of τ in 18 scenarios	48
3.2	Densities of posterior and bootstrap samples, case study	54
4.1	Structure of confounding in simulation scenarios	68
4.2	Case study outcome data, pre-study and 13-week study period	75
A2.1	Boxplots of ATE estimates, simulation study, independent covariates	92
A2.2	Boxplots of ATE estimates, simulation study, moderate correlation (.02)	93
A2.3	Boxplots of ATE estimates, simulation study, strong correlation (.05)	94

CHAPTER 1

INTRODUCTION

1.1 Potential outcomes and causal inference

Causal inference in statistics is the process of reasoning from specific samples of data to general population-level conclusions regarding causal relationships between some treatment or exposure and some outcome or result (Hernán & Robins, 2020). For any given study, “treatment/exposure” and “outcome/result” must be precisely defined, but conceptually, these terms can carry a broad variety of meanings. The former could refer to receiving an experimental drug; carrying a particular gene; participating in a nutritional diet; engaging in an educational program; or implementing a tax policy change. The latter could refer to some measurement of disease status (cured or not, degree of improvement); a specific phenotype; a particular health outcome; a future standardized test score, or future earnings; or future tax receipts, employment levels, business closings, or share of votes received in the next election. Any pair of concepts, or phenomena, or events that can be reasonably precisely defined could, in theory, be studied for their causal relationship with each other.

But what is meant by a “causal” relationship? To elucidate this idea from a statistical perspective, we introduce some notation. We will use the symbol T to refer to

the treatment/exposure concept, and the symbol Y to refer to the outcome/result concept. (Later, these symbols will be used to represent random variables.) What does it mean, then, for T to be a cause of Y , or equivalently, for Y to be an effect of T ? A helpful framework for precisely defining what is meant by a “causal” relationship is the potential outcomes framework (Rubin, 1974). In this system, each study unit is hypothesized to have a number of possible outcome values equal to the number of possible treatments that could be received. For example, with a binary treatment, say an experimental drug and a placebo (and so we might code a random variable as $T \in \{0,1\}$), each study unit would have two potential outcomes. We denote these two parts of Y as $Y^{(0)}$ and $Y^{(1)}$. Necessarily, only one of the potential outcomes is actually observed, specifically that one corresponding to the treatment level the study unit actually received. The other potential outcomes are unobserved, counterfactual (Lewis, 1973) values. While this example as well as the three methods described in this dissertation all focus on scenarios with binary treatments, the potential outcomes concept extends generally to non-binary treatment scenarios as well.

To say that there is a causal relationship between T and Y means that the individual potential outcome values of Y differ in some way. The causal effect of T on Y refers to the magnitude and nature of the variations among the potential outcomes. Subject-level causal effects are computed by comparing the potential outcomes for that subject under each treatment level. Since only one potential outcome is observed (the “fundamental problem of causal inference” (Holland, 1986)), subject-level causal effects are never known. However, population-level causal effects may be estimated if the treatment groups are, on average, comparable. For the groups to be comparable requires a variety of assumptions, and those assumptions are met in different ways in the context of randomized controlled

experiments and observational studies; these assumptions are described in more detail in Section 1.3. In the next section, we discuss measures of causal effect.

1.2 Measures of causal effect

There are two dimensions to consider when thinking about measuring causal effect. The first dimension is the level at which subject-level causal effects are aggregated. As mentioned above, we can never know subject-level causal effects, because we can't observe more than one potential outcome per subject. Hence, we are always considering average effects over some group. But what group? One intuitive answer is that we want to compare the average response under treatment to the average response under control for the entire target population; we term this the average treatment effect (ATE). A second answer, at times more appropriate than the ATE, is that we want to know the average response for those in the treatment group *only compared to those same subjects*. For this objective we use the average treatment effect among the treated (ATT). As an example, suppose we want to know the effect of a blood pressure medication in terms of magnitude of blood pressure reduction. Let us assume what is likely the case, that the medication has a higher effect on individuals who already have high blood pressure. Then if we average the treatment effect over all individuals, the apparent average effect will be lower than if we had averaged the treatment effect only over those individuals who already had high blood pressure. In this case, it would be important to focus on the ATT, rather than the ATE.

The second dimension to consider when measuring causal effect is which type of measurement of association is most meaningful. The risk difference, risk ratio, and odds ratio are three different measures that, depending on circumstances, may be more or less appropriate for one's intended research purpose. In the work presented in this dissertation, the focus is always on the risk difference. Hence we define the average treatment effect as:

$$ATE = E[Y^{(1)}] - E[Y^{(0)}].$$

1.3 Assumptions required for valid causal inference

Several assumptions must hold true for causal inference to be valid (Hernán & Robins, 2020). First, *exchangeability* of the treated and untreated subjects means that the potential outcomes are independent of the treatment group, conditional on the covariates. That is, $Y^{(t)} \perp T | \mathbf{X}$. The second assumption that must hold is *positivity*, which means that every subject must have some chance of appearing in both the treatment group and the control group, i.e., $0 < p_i(\mathbf{X}_i) < 1$, for all i , where $p_i(\mathbf{X}_i) = \Pr(T_i = 1 | \mathbf{X}_i)$. The third assumption is *consistency*; this simply means that the observed outcome matches the potential outcome for each subject: $(Y_i | T_i = t) = Y_i^{(t)}$, for all i .

1.4 Structure of this dissertation

After this introductory chapter, we present three distinct methods relating to various aspects of causal inference. Chapter Two deals with propensity score techniques and, more specifically, how to specify the propensity score model to achieve the best treatment effect

estimates. This chapter not only provides a theoretical proof showing that one particular type of specification is best, but also demonstrates an original method for applying that result. The method presented in Chapter Three has a similar purpose – obtaining precise and accurate estimates of causal effects – but views the challenge through a Bayesian, rather than a frequentist, lens. Here, a hierarchical Bayesian model is developed that is grounded in the framework of causal inference. Chapter Four presents a method for identifying subgroups of the population in which the treatment effect differs from the overall population average treatment effect. Maintaining a central theme of causal inference, the research focuses on avoiding confounding bias while identifying effect modifiers that characterize the subgroups.

CHAPTER 2

PROPENSITY SCORE SPECIFICATION FOR OPTIMAL ESTIMATION

OF AVERAGE TREATMENT EFFECT WITH BINARY RESPONSE¹

2.1 Introduction

Observational studies are critical for scientific advancement. While randomized controlled trials (RCTs) may be considered the gold standard for establishing evidence supporting some causal relationship, there are many circumstances in which conducting an RCT is impractical, unfeasible, unethical, or impossible. Meanwhile, observational data abound, and it is quite natural to look to it for scientific insights. Yet countless examples demonstrate that a real danger lurks in the careless use of observational data, the fallacy of conflating association with causation.

One class of methods for the causal analysis of observational data is propensity score-based methods. Formalized in a landmark paper by Rosenbaum and Rubin (1983), these techniques begin with estimating for each subject a propensity score. The propensity score serves two roles simultaneously: it is the conditional probability of treatment group

¹ The material in this chapter has been published in *Statistical Methods in Medical Research*. The full citation is: “Craycroft, J. A., Huang, J., & Kong, M. (2020). Propensity score specification for optimal estimation of average treatment effect with binary response. *Statistical Methods in Medical Research*, 29(12), 3623-3640.”

given covariates, and it is also a *balancing score*, the adjustment for which results in similar distributions of covariates for treatment and control groups (Rosenbaum & Rubin, 1983).

Given the propensity scores, any of a variety of methods may be used to estimate the average treatment effect (ATE). Stratification, matching, weighting, and covariate adjustment are commonly used methods. Many accessible descriptions and illustrations of propensity score methods are available in the literature (Abdia, Kulasekera, Datta, Boakye, & Kong, 2017; Austin, 2011; Lunceford & Davidian, 2004; Yan et al., 2019). However, regardless of the ATE estimation method, and before the ATE estimates may be computed, the propensity scores themselves must be estimated. Typically, logistic regression or some nonparametric method, such as gradient boosting (McCaffrey et al., 2013), is used. An important question receiving much attention in past research is which of the available covariates should be included in the propensity score model to most effectively remove confounding bias from, and reduce the variance of, the ATE estimates. There are examples in the propensity score literature (Austin, 2011; McCaffrey et al., 2013) suggesting that all available covariates are used in estimating the propensity scores, an approach that treats all covariates as potential confounding variables to be adjusted for (Figure 2.1a). In reality, covariates may be of several types (Figure 2.1b): true confounders (\mathbf{X}_C), affecting both treatment and outcome; instrumental variables (\mathbf{X}_I), affecting the treatment but not the outcome; predictor variables (\mathbf{X}_P), affecting the outcome but not the treatment; and spurious, or noise, variables (\mathbf{X}_S), affecting neither the treatment nor the outcome. (We use bold font for the covariate type labels throughout this paper to indicate that they typically refer to vectors of covariates, rather than single covariates.) The identification of individual covariates with their particular type is sometimes informed by existing subject matter

knowledge, or by prior research. Other times, there is little *a priori* guidance, and it is the analyst's job to accurately categorize the available covariates.

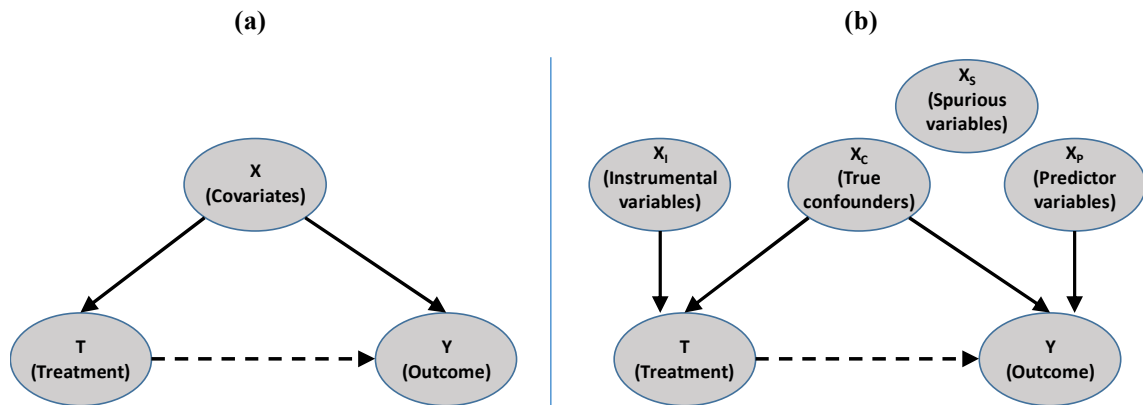


Figure 2.1. Overly simplistic (a) and more realistic (b) causal diagrams.

For a propensity score estimation method to yield unbiased estimates of ATE, all true confounders must be included in the propensity score model (Pearl, 2009) (unless the doubly robust ATE estimation method is used, which requires either the propensity score model or the outcome model be correctly specified) (Lunceford & Davidian, 2004). Also, multiple authors (Austin, 2007; Brookhart et al., 2006; Franklin, Eddings, Glynn, & Schneeweiss, 2015; Garrido et al., 2014; Patrick et al., 2011) have demonstrated that 1) including predictor variables in the propensity score model, while not needed for unbiased estimation of treatment effects, is in general beneficial for improving the precision of treatment effect estimates; and 2) including instrumental variables in the propensity score model increases the variance of the estimated treatment effect. Finally, including noise variables in the propensity score model should not affect bias, but will likely increase variability of treatment effect estimates; thus, the noise variables should be disregarded.

Consequently, in employing propensity score methods to estimate treatment effects, we are faced with a variable selection problem in which, given a potentially large set of covariates, we wish to identify only a particular subset to include in our propensity score model – specifically, the true confounders and the predictor variables. The analytical benefit of the propensity score derives from its role as a balancing score. When deciding which variables belong in the propensity score model, the question to answer is *not* how best to predict treatment group membership, but rather which covariates should be balanced in order to obtain the least biased and most precise estimates of treatment effect.

Two methods have been proposed recently for efficiently estimating propensity scores, but each method has its own drawback. Shortreed and Ertefaie (2017) introduced the outcome-adaptive lasso (OAL), which is a modification of the adaptive lasso (Zou, 2006). OAL aims to select true confounders and predictor variables into the propensity score estimation model via an adaptive lasso method where the weights in the lasso penalty term are the inverse of the estimated coefficients in the outcome regression model. In so doing, the OAL method assigns higher penalties to terms that are not related to the outcome (i.e., instrumental and spurious covariates) and lower penalties to terms that are related to the outcome (i.e., predictors and true confounders). Consequently, the minimization of the loss function results in forcing coefficients of instrumental and spurious covariates to zero, keeping only the predictors and confounders in the propensity score estimation model, as desired. As OAL is a variation of penalized regression, a tuning parameter, λ , controls the overall strength of the penalty term. In OAL, λ is selected so as to minimize the weighted absolute mean difference (wAMD) between the two treatment groups. The wAMD is a measure of covariate balance. Thus OAL cleverly incorporates both variable selection and

covariate balancing. After the propensity scores are thus estimated, any propensity score application method may be used to estimate ATE; the authors demonstrate their approach using the inverse probability weighting (IPW) method. The drawback to OAL is that, because of the initial outcome regression, the model breaks down when $p > n$.

Imai and Ratkovic (2014) recently introduced the covariate balancing propensity score (CBPS) with the recognition that the most important role of the propensity score is to balance the covariates between the treatment groups. The “just-identified” CBPS is obtained by using the covariate balance score only, while the “overidentified” CBPS is obtained by using both the covariate balance score and the score function. The drawback to CBPS is that it balances *all* covariates, rather than attempting to select only the \mathbf{X}_C and \mathbf{X}_P types.

The objective of the current study is two-fold: first, we provide a theoretical proof that the most efficient treatment effect estimates are obtained by specifying the propensity score as a function of all covariates related to outcome and excluding covariates related only to treatment. Second, we present an approach for estimating propensity scores that should mitigate the shortcomings of the OAL and CBPS methods mentioned above. The structure of the remainder of the paper is as follows: in Section 2.2, we first provide background on the potential outcomes framework, ATE, and causal inference from observational data, including assumptions that must hold for valid causal inference. We then present a theoretical proof showing why the subset of \mathbf{X} consisting only of \mathbf{X}_C and \mathbf{X}_P in the propensity score model results in the most efficient estimates of ATE. We also describe our proposed method for obtaining propensity score estimates. In Section 2.3, we describe the extensive simulation studies we conducted to compare several propensity

score specification methods. In Section 2.4, we describe the application of the methods to a case study dataset with the objective of determining if patients’ preoperative blood clotting factor is causally related to 30-day mortality following cardiac surgery. Finally, in Section 2.5 we provide discussion of the results and general conclusions.

2.2 Methods

2.2.1 Potential outcomes and ATE

A great deal of the causal inference literature works from the potential outcomes framework (Little & Rubin, 2000). In this framework, we suppose that every subject in the data set has a number of potential outcomes equal to the number of distinct treatment groups. Only one of these potential outcomes for each subject is ever actually observed; the other is a “counterfactual” outcome. However, we assume that subjects who are very similar in as many aspects as possible will have very similar potential outcomes. Hence, the ATE may be computed by comparing subjects who are very similar in all aspects except their treatment group membership (Holland, 1986).

This paper focuses on the binary treatment and binary outcome scenario. We use T to indicate treatment group and Y to indicate outcome, with $T, Y \in \{0,1\}$. We denote the two potential outcomes for subject $i, i=1, \dots, n$, as $Y_i^{(0)}$ and $Y_i^{(1)}$, corresponding respectively to the outcomes for control and treatment. As only one of these is actually observed, the observed outcome is $Y_i = T_i Y_i^{(1)} + (1 - T_i) Y_i^{(0)}$. The objective is to estimate the ATE, which is denoted by τ and defined as

$$\tau = E[Y^{(1)} - Y^{(0)}] = E[Y^{(1)}] - E[Y^{(0)}].$$

The propensity score is defined as the conditional probability of treatment group membership, given the subject's covariates. In notation, $p_i(\mathbf{X}) = P(T_i = 1|\mathbf{X}_i)$, where p (or $p(\mathbf{X})$) represents the propensity score, and \mathbf{X} represents the set of covariates. In this chapter, we use the IPW method to estimate ATE, whereby each unit is weighted by the inverse of the probability of the treatment status which that unit actually received. The IPW estimator of ATE based on n observations is

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i}{p_i} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - T_i) Y_i}{1 - p_i} \quad (1)$$

with $T_i, Y_i \in \{0,1\}$. Under the assumptions of exchangeability and positivity, described below, this estimator has been shown to be unbiased (Lunceford & Davidian, 2004).

2.2.2 Assumptions required for causal inference

Several assumptions must hold true for causal inference, whether stemming from observational or experimental data, to be justified. First, exchangeability of the treated and untreated subjects means that the potential outcomes are independent of the treatment group, conditional on the covariates. That is, $Y^{(t)} \perp T|\mathbf{X}$. Pearl (2009) proved that if the set \mathbf{X} includes all true confounders, \mathbf{X}_C , then exchangeability holds: $Y^{(t)} \perp T|\mathbf{X}_C$. Rosenbaum and Rubin (1983) showed that if exchangeability holds given \mathbf{X} , then it also holds given the propensity score: $Y^{(t)} \perp T|p(\mathbf{X})$. The second assumption that must hold is positivity, which means that every subject must have some chance of appearing in both the treatment group and the control group, i.e., $0 < p_i(\mathbf{X}_i) < 1$, for all i . The third assumption is consistency; this simply means that the observed outcome matches the potential outcome for each subject: $(Y_i|T_i = t) = Y_i^{(t)}$, for all i .

2.2.3 *Theoretical study for variance reduction of ATE estimator by including \mathbf{X}_C and \mathbf{X}_P in propensity score estimation*

Hahn (1998) showed that under exchangeability, the asymptotic variance bound for an estimator of τ is

$$E \left[\frac{\sigma_1^2(\mathbf{X})}{p(\mathbf{X})} + \frac{\sigma_0^2(\mathbf{X})}{1-p(\mathbf{X})} + (\tau(\mathbf{X}) - \tau)^2 \right], \quad (2)$$

where $\sigma_0^2(\mathbf{X}) = \text{var}(Y^{(0)}|\mathbf{X})$, $\sigma_1^2(\mathbf{X}) = \text{var}(Y^{(1)}|\mathbf{X})$, and $\tau(\mathbf{X}) = E(Y^{(1)}|\mathbf{X}) - E(Y^{(0)}|\mathbf{X})$. Hirano et al.(2003), showed that the ATE estimator (1) achieves the lower bound (2) under some regularity conditions. In the following, we show that ATE estimates with different adjustment sets of pre-treatment covariates are all unbiased; however, the ATE estimate with the adjustment set $\{\mathbf{X}_C, \mathbf{X}_P\}$ is the most efficient, i.e., it has the smallest variance.

Proposition 1 (conditional independence) (Hernán & Robins, 2020; Pearl, Glymour, & Jewell, 2016): Assume that the variables all follow the directed acyclic graph shown in Figure 1b. The set of variables \mathbf{X}_C blocks the backdoor path from T to Y, $T \leftarrow \mathbf{X}_C \rightarrow Y$. This means that if we ignore the direct path from $T \rightarrow Y$, then the following conditional independences given \mathbf{X}_C all hold: (i) $T \perp (Y^{(0)}, Y^{(1)})|\mathbf{X}_C$; (ii) $(\mathbf{X}_I, T) \perp (Y^{(0)}, Y^{(1)})|\mathbf{X}_C$; (iii) $T \perp (\mathbf{X}_P, Y^{(0)}, Y^{(1)})|\mathbf{X}_C$; and (iv) $(\mathbf{X}_I, T) \perp (\mathbf{X}_P, Y^{(0)}, Y^{(1)})|\mathbf{X}_C$.

Theorem 1: Under Proposition 1, exchangeability holds given any of the following sets of covariates: (i) \mathbf{X}_C ; (ii) $\mathbf{X}_C, \mathbf{X}_I$; (iii) $\mathbf{X}_C, \mathbf{X}_P$; or (iv) $\mathbf{X}_C, \mathbf{X}_I, \mathbf{X}_P$. That is, $T \perp (Y^{(0)}, Y^{(1)}) | \mathbf{X}^{(*)}$, where $\mathbf{X}^{(*)}$ is any set of covariates (henceforth referred to as “adjustment sets”) specified in (i), (ii), (iii), or (iv). Furthermore, assuming positivity holds under each adjustment set, the four IPW estimators resulting from applying formula (1) with each adjustment set are all unbiased. That is,

$$E(\hat{\tau}^{(*)}) = \tau \quad (3)$$

where * indicates one of the four adjustment sets, and $\hat{\tau}^{(*)}$ has the same expression as (1) except that in the denominator, p_i is defined as $Pr(T_i = 1 | \mathbf{X}^{(*)})$. It should be highlighted that $\tau(\mathbf{X}^{(*)}) = E(Y^{(1)} | \mathbf{X}^{(*)}) - E(Y^{(0)} | \mathbf{X}^{(*)})$, the conditional treatment effect given one of the adjustment sets. Meanwhile, $\hat{\tau}^{(*)}$ indicates the IPW estimator of ATE in (1) where the propensity score model for p includes only $\mathbf{X}^{(*)}$.

The proof of the exchangeability assumptions and the unbiasedness of (3) stated in Theorem 1 is provided in Appendix 1, as is the proof for Theorem 2, below.

Theorem 2: The IPW estimator for the ATE is most efficient when the propensity score model includes true confounders \mathbf{X}_C and predictors \mathbf{X}_P only.

$$\begin{aligned} & \text{i. } E \left[\frac{\sigma_1^2(X_C, X_P)}{p(X_C, X_P)} + \frac{\sigma_0^2(X_C, X_P)}{1 - p(X_C, X_P)} + (\tau(X_C, X_P) - \tau)^2 \right] \\ & \leq E \left[\frac{\sigma_1^2(X_I, X_C, X_P)}{p(X_I, X_C, X_P)} + \frac{\sigma_0^2(X_I, X_C, X_P)}{1 - p(X_I, X_C, X_P)} + (\tau(X_I, X_C, X_P) - \tau)^2 \right] \quad (4) \end{aligned}$$

$$\leq E \left[\frac{\sigma_1^2(X_I, X_C)}{p(X_I, X_C)} + \frac{\sigma_0^2(X_I, X_C)}{1 - p(X_I, X_C)} + (\tau(X_I, X_C) - \tau)^2 \right]$$

Because the ATE estimate $\hat{\tau}$ in Equation (1) achieves the asymptotic variance bound specified in Equation (2), the inequalities in (4) imply that $\text{Var}(\hat{\tau}^{(CP)}) \leq \text{Var}(\hat{\tau}^{(ICP)}) \leq \text{Var}(\hat{\tau}^{(IC)})$.

$$\begin{aligned} \text{ii. } & E \left[\frac{\sigma_1^2(X_C, X_P)}{p(X_C, X_P)} + \frac{\sigma_0^2(X_C, X_P)}{1 - p(X_C, X_P)} + (\tau(X_C, X_P) - \tau)^2 \right] \\ & \leq E \left[\frac{\sigma_1^2(X_C)}{p(X_C)} + \frac{\sigma_0^2(X_C)}{1 - p(X_C)} + (\tau(X_C) - \tau)^2 \right] \quad (5) \\ & \leq E \left[\frac{\sigma_1^2(X_I, X_C)}{p(X_I, X_C)} + \frac{\sigma_0^2(X_I, X_C)}{1 - p(X_I, X_C)} + (\tau(X_I, X_C) - \tau)^2 \right] \end{aligned}$$

Again, the ATE estimate $\hat{\tau}$ in Equation (1) achieves the asymptotic variance bound specified in Equation (2), so the inequalities in (5) imply that $\text{Var}(\hat{\tau}^{(CP)}) \leq \text{Var}(\hat{\tau}^{(C)}) \leq \text{Var}(\hat{\tau}^{(IC)})$.

2.2.4 Proposed method: variable selection via elastic net

Theorem 2, along with the literature cited in Section 2.1, demonstrates that the propensity score IPW estimator of the ATE is most efficient if the propensity score model includes true confounders and predictors only. The proposed method selects these covariates by leveraging the elastic net penalized regression procedure. The elastic net, a hybrid between ridge regression and lasso regression, was developed to help in situations with large p , and for situations in which groups of covariates are correlated with each other (Zou & Hastie, 2005). The proposed propensity score estimation method includes two

steps: we first build an elastic net regression model using *all* available covariates *and* the treatment group indicator as the regressors and the outcome Y as the dependent variable. All covariates in this elastic net model that receive a non-zero coefficient are indicated as relating to the outcome; hence, all of these variables are considered to be either the true confounder or the predictor type, and they are therefore the variables to be included in our propensity score model. Next, only those variables selected by the elastic net method in the outcome model are sent to either a regular logistic regression (with treatment, T , as the dependent variable) or to the CBPS estimating procedure to estimate propensity scores. The resulting estimated propensity scores are then used in whichever propensity score-based method is desired. (In this chapter we use the IPW method, where the propensity scores in (1) are replaced by their estimates; common alternative methods include matching, doubly robust regression, and stratification.) Because it leverages the advantage of the elastic net, this approach is expected to work effectively even when $p > n$, thus addressing a shortcoming of OAL; this approach should also yield the most precise ATE estimates, because covariates related only to treatment and spurious covariates are excluded from the propensity score specification.

The objective function solved in elastic net is:

$$\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^N l(y_i, \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) + \lambda \left[(1 - \alpha_{EN}) \frac{\|\boldsymbol{\beta}\|_2^2}{2} + \alpha_{EN} \|\boldsymbol{\beta}\|_1 \right],$$

where $l(y_i, \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)$ is the negative log-likelihood contribution for observation i (Hastie & Qian, 2014). The overall strength of the penalty is controlled by the tuning parameter λ , which is selected via cross-validation. The elastic net penalty is a convex combination of the L2-norm ridge penalty ($\|\boldsymbol{\beta}\|_2^2$) and the L1-norm lasso penalty ($\|\boldsymbol{\beta}\|_1$), with the relative

weighting between the two controlled by the parameter α_{EN} . When $\alpha_{EN} = 0$, elastic net reduces to ridge regression; when $\alpha_{EN} = 1$, elastic net reduces to lasso. To effectively use the elastic net method, we consider both α_{EN} and λ as tuning parameters. In our proposed method, the best value for α_{EN} for a particular dataset is determined by performing the elastic net outcome regression under a variety of α_{EN} settings ($\alpha_{EN} \in [0, 1]$), and then selecting that value that results in the model with the lowest cross-validation error. It is the covariates selected by this particular elastic net model, then, that are sent to a logistic regression or to the CBPS algorithm for estimation of the propensity scores. We refer to this approach as the “EN-optimal” method in our simulation study and case study.

Note that the drawback of OAL is that it will break down in situations with large p relative to n , because the regression step for the outcome model will not converge. CBPS makes no variable selection, but rather balances *all* the available covariates. The proposed method addresses both of these drawbacks.

2.3 Simulation

We conducted simulation studies to explore the properties of the proposed method and to compare operating characteristics among alternative methods. Our simulations examined the effects of sample size, number of measured covariates, correlation between covariates of the same type, strengths of associations between covariates and treatment, and strengths of associations between covariates and outcome, following the general principles provided by Morris, White, and Crowther (2019).

2.3.1 Simulation procedure

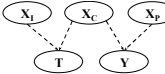
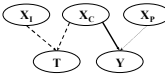
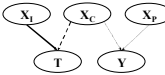
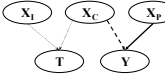
The data generating process consisted of the following steps:

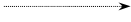


1. Generate an $n \times p$ design matrix of Gaussian-distributed covariates, with two columns each of confounders, instrumental variables, and predictor variables, and $p - 6$ columns of spurious covariates. We represent a vector of covariates for one observation as $\mathbf{X} = (X_{C1}, X_{C2}, X_{I1}, X_{I2}, X_{P1}, X_{P2}, X_{S1}, X_{S2}, \dots, X_{S(p-6)})$. The spurious covariates were always mutually independent. For the other three covariate types, we tested different correlation coefficients, and the correlations were imposed within each type. Simulation settings at this step included sample size ($n \in \{500, 1000\}$), correlation coefficients $((\rho_C, \rho_I, \rho_P) \in \{(0, 0, 0); (0.2, 0.2, 0.2); (0.5, 0.5, 0.5)\})$, and number of covariates ($p \in \{20, 100, 600\}$).
2. Generate n observations of the treatment, $T \in \{0,1\}$, as a binomial random variable with $P(T = 1|\mathbf{X})$ obtained from the underlying model $\text{logit}[P(T = 1|\mathbf{X})] = \beta_0 + \beta_1 X_{C1} + \beta_1 X_{C2} + \beta_2 X_{I1} + \beta_2 X_{I2}$. Simulation settings at this step included the vector of β values, controlling the strength of the associations between confounders and treatment (β_1), and between instrumental variables and treatment (β_2), where $\beta_1, \beta_2 \in \{0.6, 1.0, 1.6\}$ for low, medium, or high association strengths.
3. Generate n observations of the outcome, $Y \in \{0,1\}$, as a binomial random variable with $P(Y = 1|\mathbf{X}, T)$ obtained from the underlying model $\text{logit}[P(Y = 1|\mathbf{X}, T)] = \alpha_0 + \alpha_1 X_{C1} + \alpha_1 X_{C2} + \alpha_2 X_{P1} + \alpha_2 X_{P2} + \tau T$. Simulation settings at this step included the vector of α values, controlling the strength of the associations between confounders and outcome (α_1), and between predictor variables and outcome (α_2), where $\alpha_1, \alpha_2 \in \{0.6, 1.0, 1.6\}$ for low, medium, or high association strengths. τ , the

coefficient for T in the underlying model, controls the true ATE (risk difference) and is set to 0 for comparing operating characteristics of the methodologies.

The intercept for each model (α_0 and β_0) was set to $\log(0.25)$ to obtain a prevalence of treatment and outcome of approximately $\text{expit}(\log(0.25)) = 0.2$. This represents a situation where both treatment and outcome are not rare in the population. The strengths of associations between the various covariates and the outcome and the treatment comprise four individual factors ($\alpha_1, \alpha_2, \beta_1, \beta_2$, corresponding to the 4 solid arrows in Figure 2.1b), each with three possible settings (low, medium, or high), yielding $3^4 = 81$ distinct settings for covariate association strengths. In our results presented below, we focus on four of these, which are defined in Table 2.1 and labeled as Models A through D. The first column of Table 2.1 shows the causal structure for each of the four models, and the arrow weights indicate the strengths of associations among covariates.

Table 2.1: Parameter values for simulation models varying covariate-treatment & covariate-outcome associations.

MODEL	Relationship with Treatment (T) $\text{logit}(T = 1 X_I, X_C) = X_C\beta_C + X_I\beta_I$		Relationship with Outcome (Y) $\text{logit}(Y = 1 X_C, X_P) = X_C\alpha_C + X_P\alpha_P$	
	Instrumental variables	True confounders	True confounders	Predictors
A: 	$\beta_I = 1$ Medium	$\beta_C = 1$ Medium	$\alpha_C = 1$ Medium	$\alpha_P = 1$ Medium
B: 	$\beta_I = 1$ Medium	$\beta_C = 1$ Medium	$\alpha_C = 1.6$ High	$\alpha_P = 0.6$ Low
C: 	$\beta_I = 1.6$ High	$\beta_C = 1$ Medium	$\alpha_C = 0.6$ Low	$\alpha_P = 0.6$ Low
D: 	$\beta_I = 0.6$ Low	$\beta_C = 0.6$ Low	$\alpha_C = 1$ Medium	$\alpha_P = 1.6$ High

Key: Arrow weights indicate strength of association Low:  Medium:  High: 

Note: Low/Medium/High settings correspond to parameter values 0.6/1.0/1.6, respectively.

Combining all four sets of simulation settings (sample size, number of covariates, correlation among covariates, and association strengths) yields $2 \times 3 \times 3 \times 81 = 1458$ possible combinations of settings. We chose a selection of settings to explore the general patterns for how the methodologies behaved. For each set of parameter settings, we generated 1000 Monte Carlo datasets, each with 100 bootstrap repetitions for estimating the standard error of the ATE estimators. For each dataset, we estimated the ATE using the IPW method. These estimations were made under 20 propensity score specification methods, as follows:

- i. The first 10 methods employed elastic net to select covariates related to the outcome. We tested nine different settings for α_{EN} : 0.0, 0.1, 0.3, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. Four of these are shown in Table 2.2 and are labeled as EN0, EN0.5, EN0.7, and EN1. We also included an “EN-optimal” method (EN.opt in Table 2.2), which, for each Monte Carlo dataset, chose the α_{EN} setting (out of the nine just listed) yielding the model with the lowest cross-validation error. For each of these ten methods, then, the covariates receiving non-zero coefficients were sent to a logistic regression model with treatment status as the dependent variable, and the estimated propensity scores were the predicted values obtained from the logistic regression models.
- ii. The outcome-adaptive lasso method (OAL).
- iii. “Overidentified” and “just-identified” CBPS using all available covariates (Table 2.2, columns CB.over and CB.just); and “overidentified” and “just-identified” CBPS using only covariates selected by the EN.optimal method (Table 2.2, columns EN.CB.o and EN.CB.j).

- iv. Five reference models using the known sets of covariates: true confounders only (Table 2.2, column \mathbf{X}_C); confounders and predictors, i.e., the target set (column $\mathbf{X}_C\mathbf{X}_P$); confounders and instrumental variables (column $\mathbf{X}_C\mathbf{X}_I$); confounders, predictors, and instrumental variables (column $\mathbf{X}_C\mathbf{X}_P\mathbf{X}_I$); and all available covariates (column $\mathbf{X}_C\mathbf{X}_P\mathbf{X}_I\mathbf{X}_S$).

Note that EN0 (ridge regression) is exactly the same as the $\mathbf{X}_C\mathbf{X}_P\mathbf{X}_I\mathbf{X}_S$ reference group. The elastic net method is being used here solely for choosing covariates to include in the propensity score model, not directly for predicted values; since ridge regression (EN0) does no subset selection, all available covariates are always included in the propensity score logistic regression model. Table 2.2 presents the bias, standard error, and root mean squared error (RMSE) of the ATE estimates for the different scenarios and the various propensity score specification methods.

In this simulation study we were interested in examining how the methods would perform in cases with high and low ratios of sample size to total number of parameters. Also, for parameter values in the treatment and outcome models, we roughly followed settings used in some similar prior simulation studies (Shortreed & Ertefaie, 2017), mainly focusing on having enough difference between “high,” “medium,” and “low” strength settings such that any performance differences between the tested methods due to this factor would be apparent. All simulations and analyses were performed in R 3.2.1 (OAL) or R 3.5.2 (all others) (R Core Team, 2018). The CBPS package (Ratkovic, Imai, & Fong, 2012) was used for all CBPS estimates, and the glmnet package (Hastie & Qian, 2014) was used for all elastic net estimates. For OAL, the authors’ provided R code was used

(supplementary materials included with the online version of (Shortreed & Ertefaie, 2017)), with adaptation for the binary response scenario considered here.

2.3.2 Simulation results

We highlight below a selection of results to illustrate the major takeaways from the study. Table 2.2 summarizes seven different simulation scenarios, with each scenario being defined according to the strengths of the associations between the covariates and the treatment, the strengths of the associations between the covariates and outcome, the correlations among the simulated covariates, and the n/p ratio. For each scenario, Table 2.2 shows the bias, standard error, and RMSE for the estimated ATE ($\hat{\tau}$) under 15 different specifications of the propensity score. Five of those propensity score specifications are reference models using known sets of covariates, while the other eleven are tested models. The results in Table 2.2 are for the fully independent set of covariates (i.e., $\rho_I = \rho_C = \rho_P = 0$), and the boxplots of the corresponding ATE estimates are presented in Figure A2.1 in Appendix 2. Results for scenarios with correlated covariates with $\rho_I = \rho_C = \rho_P = 0.2$ and 0.5 are presented, respectively, in Tables A2.1 and A2.2, as well as in Figures A2.2 and A2.3 in Appendix 2; in those scenarios, the relative performances of the various methods are consistent with the results shown below.

Examining only the five reference models (last five columns in Table 2.2), the $\mathbf{X}_C\mathbf{X}_P$ specification has the lowest RMSE for each scenario. This demonstrates the statements made in Section 2.1, and for which we showed a proof in Section 2.2.3. In some scenarios, the improvement over the \mathbf{X}_C model from including \mathbf{X}_P is substantial, while in other scenarios it is only slight. Moreover, in every scenario, by comparing columns \mathbf{X}_C

and $\mathbf{X}_C\mathbf{X}_I$, or columns $\mathbf{X}_C\mathbf{X}_p$ and $\mathbf{X}_C\mathbf{X}_p\mathbf{X}_I$ in Table 2.2, it is clear that including \mathbf{X}_I increases the variance of the ATE estimates. In general, we conclude the following: (i) all ATE estimates are unbiased when all true confounders (\mathbf{X}_C) are included in the propensity score model; (ii) the standard error and RMSE of ATE estimates are lowest when the propensity score model uses confounders and predictors ($\mathbf{X}_C\mathbf{X}_p$) only; (iii) the standard error and RMSE of ATE estimates increase when the propensity score model includes the instrumental variables (\mathbf{X}_I); and (iv) including noise variables in the propensity score model further increases the standard error and RMSE of the ATE estimates (column $\mathbf{X}_C\mathbf{X}_p\mathbf{X}_I\mathbf{X}_S$). The results in column $\mathbf{X}_C\mathbf{X}_p$ provide benchmarks for comparison with the tested models.

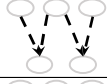
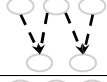
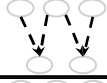
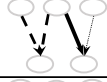
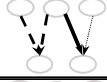
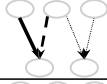
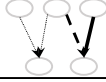
We pointed out that CBPS balances *all* available covariates. An examination of columns CB.over and CB.just relative to other columns in Table 2.2 demonstrates that, although they are improvements over the logistic regression model with all covariates (columns EN.0 and $\mathbf{X}_C\mathbf{X}_p\mathbf{X}_I\mathbf{X}_S$, which are equivalent), CB.over and CB.just are the worst, or very nearly worst, of all the tested specifications. However, when the CBPS approach uses only the variables selected via the EN.optimal method (columns EN.CB.o and EN.CB.j), the CBPS performance is improved, and the just-identified CBPS is competitive with OAL and elastic net specifications (columns EN0.5, EN0.7, EN1, EN.opt and OAL).

The first five columns in Table 2.2 show that, as the tuning parameter α_{EN} grows larger, bias typically stays unchanged, but the variance of the ATE estimates decreases (though not necessarily monotonically). This demonstrates that the regularized regression penalty is resulting in regression coefficients for covariates unrelated to Y being shrunk to zero as the weighting between the lasso penalty and the ridge penalty increasingly favors

lasso. The performance under lasso (EN1) and EN.optimal (EN.opt) is competitive with the target model ($\mathbf{X}_C\mathbf{X}_P$).

Comparing Scenario 3 ($n/p = 5$) to Scenario 1 ($n/p = 50$), or Scenario 5 ($n/p = 10$) to Scenario 4 ($n/p = 50$), we see that the ratio $\frac{RMSE(\mathbf{X}_C\mathbf{X}_P\mathbf{X}_I\mathbf{X}_S)}{RMSE(\mathbf{X}_C\mathbf{X}_P)}$ is higher for the cases with the lower n/p ratio (scenarios 3 and 5). In other words, the detriment to the ATE estimates of failing to do variable selection is higher in models with higher proportions of spurious and instrumental variables. Moreover, as seen in Scenario 2, where $n/p < 1$ (i.e., $p > n$), we cannot obtain results for OAL or for CB.over or CB.just. In this scenario, performing variable selection ahead of time via elastic net successfully excludes spurious and instrumental variables and results in ATE estimates with bias, standard error, and RMSE very close to those of the target ($\mathbf{X}_C\mathbf{X}_P$) specification.

Table 2.2: Bias, Standard Error, and Root MSE for selected simulation scenarios.

#	SCENARIO		TESTED MODELS									REFERENCE MODELS					
			ENO	ENO.5	ENO.7	EN1	EN.opt	OAL	CB.over	CB.just	EN.CB.o	EN.CB.j	X_c	$X_c X_p$	$X_c X_i$	$X_c X_p X_i$	$X_c X_p X_i X_s$
1	Model A: 	Bias	0.001	0.002	0.002	0.002	0.002	0.004	0.027	-0.007	0.015	-0.004	0.002	<i>0.001</i>	0.002	0.002	0.001
		SE	0.060	0.032	0.031	0.030	0.035	0.033	0.042	0.048	0.032	0.034	0.032	<i>0.029</i>	0.059	0.057	0.060
		RMSE	0.060	0.032	0.031	0.030	0.035	0.033	0.050	0.049	0.036	0.034	0.032	<i>0.029</i>	0.059	0.057	0.060
2	Model A: 	Bias	*	0.006	0.004	0.003	0.005	*	*	*	0.035	-0.003	0.006	<i>0.004</i>	0.012	0.010	*
		SE	*	0.045	0.044	0.045	0.044	*	*	*	0.043	0.042	0.051	<i>0.044</i>	0.082	0.079	*
		RMSE	*	0.045	0.044	0.045	0.045	*	*	*	0.055	0.042	0.051	<i>0.044</i>	0.083	0.079	*
3	Model A: 	Bias	-0.007	0.005	0.004	0.004	0.004	-0.005	0.075	0.042	0.026	-0.004	0.003	<i>0.002</i>	0.010	0.011	-0.007
		SE	0.179	0.045	0.042	0.041	0.043	0.039	0.060	0.063	0.041	0.043	0.045	<i>0.041</i>	0.081	0.079	0.179
		RMSE	0.179	0.045	0.043	0.042	0.043	0.040	0.095	0.076	0.049	0.043	0.045	<i>0.041</i>	0.082	0.080	0.179
4	Model B: 	Bias	0.004	0.006	0.004	0.003	0.004	0.006	0.043	-0.010	0.026	-0.006	0.003	<i>0.003</i>	0.004	0.004	0.004
		SE	0.059	0.032	0.029	0.026	0.031	0.030	0.040	0.045	0.029	0.032	0.026	<i>0.025</i>	0.055	0.054	0.059
		RMSE	0.059	0.032	0.029	0.026	0.031	0.031	0.059	0.046	0.039	0.032	0.026	<i>0.025</i>	0.055	0.055	0.059
5	Model B: 	Bias	-0.008	0.005	0.003	0.003	0.003	0.007	0.087	-0.009	0.026	-0.008	0.003	<i>0.003</i>	0.006	0.006	-0.008
		SE	0.083	0.030	0.028	0.027	0.028	0.031	0.038	0.046	0.030	0.031	0.028	<i>0.027</i>	0.055	0.054	0.083
		RMSE	0.084	0.030	0.028	0.027	0.028	0.031	0.095	0.047	0.040	0.032	0.028	<i>0.027</i>	0.055	0.055	0.084
6	Model C: 	Bias	0.013	0.003	0.003	0.002	0.004	0.008	0.026	-0.002	0.015	0.001	0.001	<i>0.001</i>	0.013	0.013	0.013
		SE	0.123	0.041	0.040	0.041	0.046	0.051	0.078	0.090	0.043	0.045	0.042	<i>0.040</i>	0.114	0.112	0.123
		RMSE	0.124	0.041	0.040	0.041	0.046	0.052	0.082	0.090	0.046	0.045	0.042	<i>0.040</i>	0.114	0.113	0.124
7	Model D: 	Bias	0.004	0.001	0.000	0.000	0.001	0.003	0.024	-0.001	0.015	-0.002	-0.001	<i>0.000</i>	0.002	0.004	0.004
		SE	0.064	0.043	0.042	0.041	0.044	0.044	0.047	0.051	0.041	0.041	0.048	<i>0.041</i>	0.062	0.056	0.064
		RMSE	0.064	0.043	0.042	0.041	0.044	0.045	0.053	0.051	0.044	0.041	0.048	<i>0.041</i>	0.062	0.056	0.064

Notes: EN=Elastic Net (displayed settings include $\alpha_{EN}=0, 0.5, 0.7, 1,$ and EN.optimal); OAL=Outcome-adaptive Lasso; CB=Covariate Balancing Propensity Score; EN.CB=CBPS after Elastic Net (either over- or just-identified).

SE=Standard Error; RMSE=Root Mean Squared Error.

* indicates estimates not available, unable to use propensity score specification since $p > n$.

Models are defined by strengths of associations between covariates and treatment/outcome. See Figure 1b and Table 1.

Scenarios are defined by Model and n/p ratio.

ENO is equivalent to $X_c X_p X_i X_s$; these both use all available covariates in estimating the propensity scores.

$X_c X_p$ is "Target" propensity score specification and is emphasized in italics.

Bold font is used to emphasize tested propensity score estimation method with the lowest RMSE of each scenario.

The importance of conducting variable selection for the propensity score specification also varies according to the strength of the associations between covariates and treatment, and between covariates and outcome. When the associations between covariates and the outcome are weak, as in Scenario 6, including *all* covariates in the propensity score specification results in ATE estimates with relatively high bias. This is consistent with Brookhart et al. (2006); Austin, Grootendorst, and Anderson (2007); Patrick et al. (2011); and Zhu, Schonbach, Coffman, and Williams (2015), among others. Meanwhile, when the associations between covariates and treatment are weak, as in Scenario 7, there is not much cost in terms of bias for using all covariates in the propensity score specification. Nevertheless, there remains a cost in terms of precision; the standard error of the estimated ATE in Scenario 7 decreases by about one-third for the target $\mathbf{X}_C\mathbf{X}_P$ model relative to the $\mathbf{X}_C\mathbf{X}_P\mathbf{X}_T\mathbf{X}_S$ model. The lasso (EN1) and the just-identified CBPS after elastic net (EN.CB.just) are the most effective here in attaining minimum RMSE. Note that the results for Scenarios 6 and 7 are very similar as long as variable selection is done via *some* method. Indeed, elastic net shows huge improvement over no variable selection even with α_{EN} set as low as 0.1 (not shown).

2.4 Case Study

2.4.1 Case study background

We demonstrate the proposed process on an interesting data set involving the association of preoperative conditions with 30-day mortality following cardiac surgery. A major challenge of perioperative management for cardiac surgery arises from intraoperative and postoperative bleeding. Bleeding can occur due to the trauma of surgery,

physiologic changes associated with cardiopulmonary bypass, thrombocytopenia, platelet defects, fibrinolysis or coagulation factor deficiency. In these surgeries, blood count, coagulation studies and blood group determination are regarded as routine preoperative investigations in virtually all patients (Cornelissen & Arrowsmith, 2006). Preoperative coagulation studies, including baseline international normalized ratio (INR), platelet counts, and platelet function tests, are performed in order to identify risk factors and optimize hemostasis. Coagulation studies are of special interest because excessive bleeding occurs in between 7% and 53% of patients after cardiac surgery, and bleeding related re-exploration is associated with high in-hospital mortality and morbidity (Biancari, Mikkola, Heikkinen, & al., 2012). Studies on whether these preoperative coagulation tests can predict clinical outcomes, or whether correcting any negative conditions found would improve outcomes, are sparse and controversial.

The primary objective of the study was to determine if the level of the preoperative international normalized ratio (INR), which is a measure of how long it takes blood to clot, is causally associated with post-surgery outcomes, particularly 30-day mortality. The dataset was obtained from a retrospective review of 1390 patient medical records at Jewish Hospital, Louisville, KY. The patients all had cardiac surgeries performed from January 2008 to December 2013. The hospital data were linked with the Society of Thoracic Surgeons database for additional covariate data. Baseline covariates included age, gender, diabetes status, creatinine level, functional platelet number (i.e., the product of platelet count and platelet aggregation percentage), patients' other health conditions such as chronic lung disease, and operative characteristics such as use of an intra-aortic balloon pump. This was observational data, as the preoperative level of INR was considered the

treatment and was not randomly assigned to units but rather observed as a preoperative characteristic of each subject. The data indicated each subject as being in a high INR group or a low INR group. It was expected that high INR should be causally associated with a higher 30-day mortality, as a higher INR value indicates thinner blood that clots less easily.

2.4.2 Case study results

Table 2.3 presents the baseline distributions of all covariates in the case study data, stratified by treatment and outcome. The 1390 observations were divided nearly equally (51% to 49%) between low INR and high INR groups. 31 (2.2%) of the patients died within 30 days post-surgery; of these, 26 (84%) were in the high INR group. Therefore, the crude (unadjusted) estimate of the risk difference for 30-day mortality between the high and low INR groups was

$$\begin{aligned} Pr(Died | High\ INR\ grp) - Pr(Died | Low\ INR\ grp) = \\ \frac{26}{684} - \frac{5}{706} = 0.038 - 0.007 = 0.031, \end{aligned}$$

with a 95% confidence interval for the estimated risk difference of (0.015, 0.047). A chi-square test of independence between treatment and outcome was statistically significant ($X^2=15.2$, p-value < 0.001).

However, this was observational data, and consequently, there was risk of the high and low INR groups differing in systematic ways, resulting in confounding bias in the estimated risk difference. Indeed, as may be seen in Table 2.3, several covariates appeared unequally distributed in the two INR groups (chronic lung disease, diabetes, incidence, intra-aortic balloon pump, hypertension, congestive heart failure, and type of surgery). Of these, chronic lung disease, incidence, intra-aortic balloon pump, hypertension, congestive

heart failure, and type of surgery also appeared to be related to outcome, making these six variables candidates for true confounders (\mathbf{X}_C). Three other variables (gender, peripheral vascular disease, and myocardial infarction) also appeared related to 30-day mortality (outcome), but not to INR group (treatment), making these three variables potential predictors (\mathbf{X}_P).

We used the proposed method to estimate the 30-day mortality ATE between high INR and low INR groups. Applying the elastic net method on the outcome model (i.e., regressing mortality on all baseline covariates) identified three covariates that were associated with the outcome (gender, intra-aortic balloon pump, and type of surgery); hence, these three variables were used in a logistic regression with the INR group as the dependent variable to compute the estimated propensity score for each patient. The propensity score IPW estimator was then used to compute an adjusted ATE estimate, which was 0.021 (0.004, 0.039). While still positive, the magnitude of the estimated ATE was about one-third lower than the unadjusted estimate. From the prior discussion and analysis in this paper, we believe that the unadjusted estimate is biased.

Table 2.3: Baseline distributions of variables by treatment and outcome.

	Stratified by Treatment				Stratified by 30-Day Mortality				% Mortality
	Low INR		High INR		Survived		Died		
	Mean/No.	(SD/%)	Mean/No.	(SD/%)	Mean/No.	(SD/%)	Mean/No.	(SD/%)	
N	706	(50.8)	684	(49.2)	1359	(97.8)	31	(2.2)	2.2
CONTINUOUS VARIABLES									
Age	62.31	(11.62)	64.29	(12.47)	63.14	(12.01)	69.29	(13.69)	
Creatinine Level	1.11	(0.83)	1.23	(0.96)	1.16	(0.89)	1.45	(1.21)	
FPN	181.59	(72.41)	185.36	(87.44)	183.97	(80.28)	160.47	(71.69)	
Ejection Fraction	50.29	(15.22)	44.15	(17.50)	47.28	(16.64)	46.81	(17.69)	
CATEGORICAL VARIABLES									
Gender									
Female	246	(34.8)	216	(31.6)	442	(32.5)	20	(64.5)	4.3
Male	460	(65.2)	468	(68.4)	917	(67.5)	11	(35.5)	1.2
Chronic Lung Disease									
No	494	(70.0)	449	(65.6)	928	(68.3)	15	(48.4)	1.6
Mild	118	(16.7)	122	(17.8)	231	(17.0)	9	(29.0)	3.8
Moderate	63	(8.9)	69	(10.1)	128	(9.4)	4	(12.9)	3.0
Severe	31	(4.4)	44	(6.4)	72	(5.3)	3	(9.7)	4.0
Diabetes									
No	449	(63.6)	390	(57.0)	820	(60.3)	19	(61.3)	2.3
Yes	257	(36.4)	294	(43.0)	539	(39.7)	12	(38.7)	2.2
Status									
Elective	311	(44.1)	300	(43.9)	598	(44.0)	13	(41.9)	2.1
Urgent	392	(55.5)	376	(55.0)	751	(55.3)	17	(54.8)	2.2
Emergent	3	(0.4)	8	(1.2)	10	(0.7)	1	(3.2)	9.1
Incidence									
No	650	(92.1)	580	(84.8)	1206	(88.7)	24	(77.4)	2.0
Yes	56	(7.9)	104	(15.2)	153	(11.3)	7	(22.6)	4.4
Intra-Aortic Balloon Pump									
No	670	(94.9)	588	(86.0)	1239	(91.2)	19	(61.3)	1.5
Yes	36	(5.1)	96	(14.0)	120	(8.8)	12	(38.7)	9.1
Peripheral Vascular Disease									
No	604	(85.6)	583	(85.2)	1163	(85.6)	24	(77.4)	2.0
Yes	102	(14.4)	101	(14.8)	196	(14.4)	7	(22.6)	3.4
Hypertension									
No	76	(10.8)	103	(15.1)	174	(12.8)	5	(16.1)	2.8
Yes	630	(89.2)	581	(84.9)	1185	(87.2)	26	(83.9)	2.1
Myocardial Infarction									
No	360	(51.0)	375	(54.8)	716	(52.7)	19	(61.3)	2.6
Yes	346	(49.0)	309	(45.2)	643	(47.3)	12	(38.7)	1.8
Congestive Heart Failure									
No	337	(47.7)	237	(34.6)	569	(41.9)	5	(16.1)	0.9
Yes	369	(52.3)	447	(65.4)	790	(58.1)	26	(83.9)	3.2
Type of Surgery									
CABG	552	(78.2)	383	(56.0)	927	(68.2)	8	(25.8)	0.9
Others	19	(2.7)	50	(7.3)	65	(4.8)	4	(12.9)	5.8
Valve	64	(9.1)	136	(19.9)	191	(14.1)	9	(29.0)	4.5
Both	71	(10.1)	115	(16.8)	176	(13.0)	10	(32.3)	5.4
Treatment									
High INR									
No	706	(100.0)	0	(0.0)	701	(51.6)	5	(16.1)	0.7
Yes	0	(0.0)	684	(100.0)	658	(48.4)	26	(83.9)	3.8

FPN: Functional Platelet Number; CABG: Coronary Artery Bypass Grafting; INR: Int'l Normalized Ratio

For further comparison, the other methods explored in the simulation study above were also used for estimating the propensity scores, with the estimated ATEs subsequently computed in each case using IPW. The results of these various approaches at specifying the propensity score model are summarized in Figure 2.2.

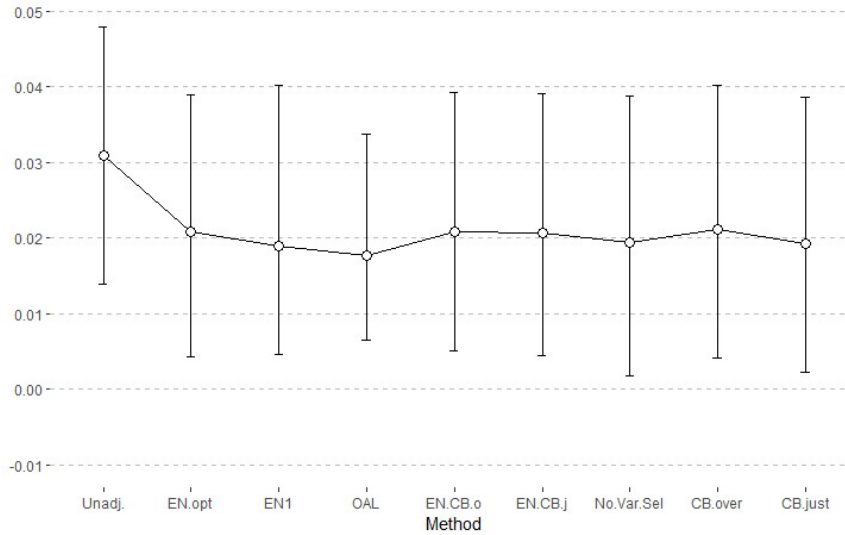


Figure 2.2: ATE estimates with 95% CIs for different methods for the case study on 30-day mortality. 95% CI is the bootstrap percentile CI, except for the unadjusted estimate.

From Figure 2.2 it is evident that, after adjusting for important covariates, the ATE for being in the high INR group was about 2%, indicating that there was about a 2 percentage point higher risk of 30-day mortality purely due to higher preoperative INR, as compared to lower preoperative INR. We also see in Figure 2.2 that the ATE estimate from the propensity score model using all covariates (no variable selection) is less precise than the other adjusted estimates, although the difference is not very pronounced for this particular data set. The similar performance is most likely due to the fact that the variables related to treatment are also related to outcomes, that is, there was no instrumental variables in this data set.

2.5 Discussion and conclusion

We have demonstrated from theory as well as via simulation studies that eliminating instrumental variables and including true confounders and predictor variables in the propensity score specification is beneficial for reducing variance of the ATE estimates. We propose using elastic net regression to select the covariates related to the outcome variable Y and then using only the selected variables to estimate propensity scores. The simulation study endorses the theoretical results developed in this paper, although the simulation study serves more as a proof of concept for the elastic net approach rather than a definitive evaluation. Based on this study, even a modest effort at removing unassociated covariates from the propensity score model can pay large dividends in terms of reducing the variance of the ATE estimates. Using lasso (EN1) to select variables for the propensity score model is consistently the best, or among the best, approaches. This is due to the lasso's shrinkage of many/most of the extraneous covariates (\mathbf{X}_{IS} and \mathbf{X}_{SS}).

The OAL method is usually very good, but OAL is not an option when $p > n$. The same problem holds with CBPS, whether overidentified or just-identified. The proposed method is expected to work when $p > n$ unless the number of true confounding variables and predictors is larger than n (i.e., $(p_C + p_P) > n$). Usually in an observational study, the number of observations is in thousands, and we expect the number of confounding variables and predictor variables to be smaller than n , even while the total number of variables could be large (say, $(p_C + p_P + p_I + p_S) > n$). In this latter scenario, the outcome regression in OAL at stage 1 would suffer, while the proposed method may not. When the number of true confounding variables and predictors is larger than n , the

convergence problems that hamper OAL at stage 1 will also pose a problem for this method at stage 2, and CBPS would also fail. We also found that for both the overidentified and just-identified CBPS methods, the variance of the ATE estimator is relatively high. However, in some circumstances, the just-identified CBPS *after* variable selection via elastic net is a very competitive method.

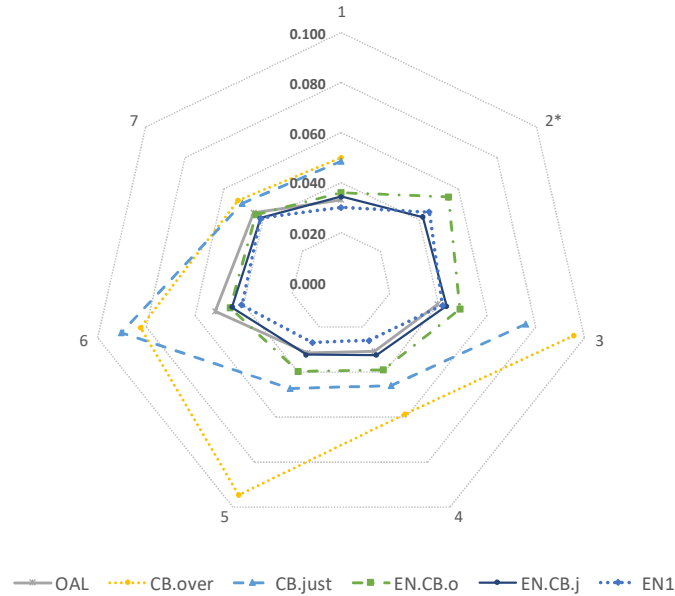


Figure 2.3: Radar chart of RMSE for tested methods, 7 scenarios. Vertex numbers indicate simulation scenarios (see Table 2.2). *ATE unestimable in Scenario 2 for OAL, CB.over, and CB.just.

The radar chart in Figure 2.3 plots the RMSE for six of the tested propensity score specification methods for each of the seven simulation scenarios summarized in Table 2.2. From the chart, we can see that CB.over and CB.just always have a RMSE greater than the other four methods (except for Scenario 2, for which these two methods and OAL produced no estimates, and hence have no RMSE); sometimes, as in scenarios 3, 5, and 6, the excess RMSE of these two methods compared to the other methods is substantial. Scenarios 3 and 5 are scenarios with n/p ratios that, while greater than one, are yet fairly low. The

importance of the variable selection step here is even higher than cases in which the n/p ratio is high. In Scenario 6, we have strong covariate associations with the treatment, so again, the importance of doing variable selection to exclude the instrumental variables is very high. In Scenario 7, there is strong covariate associations with the outcome, but not the treatment; in this case, there is not such a high cost in terms of RMSE (driven mostly by the variance of the ATE estimates) for not doing variable selection. Meanwhile, there is not a large difference in RMSE among OAL, EN.CB.over, EN.CB.just, and EN1; however, EN1 (lasso) does have, in most cases, the minimum RMSE, albeit sometimes by a very slight margin.

One issue occasionally mentioned with respect to propensity score methods is the ability to compute the propensity score without any reference to outcome data. For example, Rubin (2007) emphasizes that in planning an observational study, the choices for variables to be measured should be made without looking at the impact of those variables on the outcome. Rubin is concerned with what is sometimes termed “p-value fishing”: “rather than the outcome data Y_{obs} being ‘not in sight,’ they are used over and over again to fit various models, try different transformations, look at results discarding influential outliers, etc.” (p. 25). Rubin recommends that the decisions on propensity score variables “...should be done without ever looking at any outcome data, and thus without looking at any answers about causal effects” (p. 33). We totally agree that “p-value fishing” must be avoided. In our proposed approach, we are not looking at all at the impact on the estimated causal effect when determining which variables are included in the propensity score model. We are simply recognizing that an observational data set is likely to contain variables that are unrelated to the outcome and therefore irrelevant for propensity score methods. In

addition, the prognostic score, the score summarizing the covariates' association with outcomes, has been used to estimate ATE (Hansen, 2008); (Leacy & Stuart, 2014). Even in RCTs, variables known unrelated to outcome are likely to be – and should be – ignored; stratification preceding randomization is usually done based on covariates associated with outcome³⁴. Thus, we maintain that the procedure emphasized throughout this article, i.e., to specify the propensity score model by using only covariates associated with the outcome and excluding covariates only associated with the treatment, is not only acceptable, but highly recommended. Empirical results (such as those from the simulation study described in Section 2.3, and those in the literature referenced in Section 2.1) and the theoretical results (Section 2.2) clearly demonstrate the increased precision available with this approach.

Our study, although involving extensive simulations and a high number of parameter settings, does have some limitations. These limitations consequently speak to areas where further investigation is warranted. First, both the treatment variable and the outcome variable in the simulation had a fairly high prevalence, particularly compared to, say, the prevalence of most diseases. A future study should examine the performance of the various propensity score specification methods under rare prevalence of outcome, and perhaps under rare prevalence of treatment. Second, our simulation setup included relatively few \mathbf{X}_C , \mathbf{X}_I , and \mathbf{X}_P covariates (exactly two of each). Perhaps more of these covariate types would affect the relative performance of the methods. Nevertheless, this study did take into account a large number of parameters and settings, and it could be fairly easily extended to include some of these additional factors.

CHAPTER 3

BAYESIAN CAUSAL INFERENCE USING MCMC

3.1 Introduction

The application of Bayesian methods in the field of causal inference has a long history, as exemplified by Donald Rubin’s 1978 article in *The Annals of Statistics*, “Bayesian Inference for Causal Effects: The Role of Randomization.” Rubin has a career-long involvement in the field of causal inference, including on the frequentist side, being (along with Paul Rosenbaum) one of the two authors of the seminal 1983 *Biometrika* article “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” A great contribution made by Rubin to the field of causal inference was his extension of the potential outcomes framework, which was originated in 1923 by Jerzy Neyman, from randomized experiments to observational studies. The potential outcomes framework is today a common structure in which to study causal inference. In this framework, each study unit is hypothesized to have a number of possible outcome values corresponding to the number of possible treatments that could be received. For example, with a binary treatment, say an experimental drug and a placebo, each study unit would have two potential outcomes. Necessarily, only one of the potential outcomes is actually observed, specifically that one corresponding to the treatment level received. The other potential outcomes are unobserved, counterfactual values. Causal inference proceeds in this potential outcomes

framework by viewing these unobserved potential outcomes as missing data. Subject-level causal effects would be computed by comparing the outcomes under each treatment level. Since only one potential outcome is observed (the “fundamental problem of causal inference” (Holland, 1986)), subject-level causal effects are never known. However, population-level causal effects may be estimated if the treatment groups are, on average, comparable. For the groups to be comparable requires a variety of assumptions, and those assumptions are met in different ways in the context of randomized controlled experiments and observational studies.

Bayesian strategies enter the analysis typically in two different ways. First, as in the articles by Rubin and Keil, et al. (Keil, Daza, Engel, Buckley, & Edwards, 2018), due to the formulation of the problem in a way that implies missing data, Bayesian methods may be used for calculating the predictive distributions of the unobserved data (the non-realized potential outcomes) given the observed covariates and observed outcomes. Then, the estimation of causal effects takes place by directly comparing the individual level causal effects computed via the predictive distributions. A second common application of Bayesian methodologies in causal inference is in conducting sensitivity analysis on certain underlying assumptions. For example, an important assumption in the analysis of observational data for causal effects is that there are no unmeasured confounders. Viewing any unmeasured confounders as missing data, McCandless, Gustafson, and Levy (2007) use Bayesian inference on a hypothesized latent bias term representing all unmeasured confounders and assesses how large such a bias term need be to affect one’s directional conclusion regarding the causal estimand. In a separate paper, McCandless, Gustafson, and Austin (2009) use Bayesian methods to investigate how much accounting for uncertainty

in the estimation of propensity scores can affect the precision of estimated treatment effects derived from the use of those scores.

In the work presented in this chapter, we take a somewhat different tack. Indeed, this approach is aimed very directly at the underlying research question: what is the average causal effect of a treatment on an outcome. We do not model the posterior predictive distributions of unobserved outcomes, as per Rubin, nor do we conduct sensitivity analysis. Rather, we directly model the posterior conditional distribution of the causal effect itself. In the process, we incorporate uncertainties inherent in propensity score modeling, in order to achieve causal effect estimates with accurate precision.

To achieve these goals, we use a hierarchical Bayesian structure to describe the entire information set, which is understood as the full complement of data, including treatment, outcome, and covariates. We build the structure in a way that carries intuitive causal interpretation and that links with the common frequentist causal constructs such as the propensity score and the average treatment effect. We then parameterize the model with compellingly reasonable prior distributions; “reasonable” here is applied with respect to choices about the distribution families and structures, while the level of informativeness of those priors is then controlled by choices about hyperparameters. Bayesian Markov chain Monte Carlo sampling is then conducted directly on the posterior distribution of the causal estimand. The method shows promise in comparing admirably with the nonparametric inverse probability weighting (IPW) estimator that is commonly used in causal inference.

The structure of the remainder of this chapter is as follows: in Section 3.2, we provide background on causal inference for observational data and specify the causal risk difference estimand for the IPW method. We also describe a Bayesian hierarchical

structure for conceptualizing an observational data set, and we explain the intuition underlying our parameterization of the structure. In Section 3.3, we describe the simulation study we executed to test the proposed Bayesian methodology and evaluate it against the typical frequentist IPW approach. In Section 3.4, we illustrate application of the method to a case study data set. Finally, in Section 3.5, we provide further discussion of the results and general conclusions.

3.2 Methods

We are concerned with the inverse probability weighting (IPW) estimator of average treatment effect (ATE). The basic estimator, developed on techniques elucidated by Horvitz and Thompson (1952), is an unbiased estimate of ATE and is specified as

$$\hat{t}_{IPW} = \frac{1}{n} \sum_{i=1}^n \frac{y_i T_i}{p_i} - \frac{1}{n} \sum_{i=1}^n \frac{y_i (1 - T_i)}{(1 - p_i)}$$

where y_1, y_2, \dots, y_n are observed outcome values, T_1, T_2, \dots, T_n are observed treatment indicators, and $p_i = \Pr(T_i = 1 | \mathbf{X}_i)$ is the propensity score for each subject. Each \mathbf{X}_i is a vector of pre-treatment covariates. However, this estimator lacks robustness, a situation that can be improved upon by using “stabilized weights”:

$$\hat{t}_{IPW.SW} = \frac{\sum_{i=1}^n \frac{y_i T_i}{p_i}}{\sum_{i=1}^n \frac{T_i}{p_i}} - \frac{\sum_{i=1}^n \frac{y_i (1 - T_i)}{(1 - p_i)}}{\sum_{i=1}^n \frac{(1 - T_i)}{(1 - p_i)}}$$

as described by Lunceford and Davidian (2004).

In this article, we deal with the setting with continuous outcome and binary treatment, i.e., $Y_i \in \mathbb{R}$ and $T_i \in \{0,1\}$. The covariates \mathbf{X}_i are considered fixed, but may be binary, nominal or continuous. We assume the following conditional distributions for the data Y_i and T_i :

$$Y_i|T_i \sim \begin{cases} \text{Normal}(\theta + \tau, p_i \sigma^2), & \text{if } T_i = 1 \\ \text{Normal}(\theta, (1 - p_i) \sigma^2), & \text{if } T_i = 0 \end{cases}$$

and

$$T_i|p_i \sim \text{Bernoulli}(p_i).$$

In the above, then, θ is the expected response for a subject in the control group, and $\theta + \tau$ is the expected response for a subject in the treatment group. τ captures the treatment effect (it is clear that $E[Y_i|T_i = 1] - E[Y_i|T_i = 0] = (\theta + \tau) - \theta = \tau$). Next, $p_i = \Pr(T_i = 1|\mathbf{X}_i)$ is the propensity score. The scale factor in the variance (p_i if $T_i = 1$ and $(1 - p_i)$ if $T_i = 0$) is closely related to inverse probability weighting, where the weight is $\frac{1}{p_i}$ if $T_i = 1$ and $\frac{1}{(1-p_i)}$ if $T_i = 0$. Unlike in the frequentist approach, the weights here are considered to have variability and are updated over the posterior sampling.

Next, we need to specify priors for the parameters. We hypothesize the following hierarchical scheme that appears fairly complex at first, yet contains logical motivations for each element:

$$\tau \sim \text{Normal}(0, \sigma_\tau^2),$$

$$\theta \sim \text{Normal}(0, \sigma_\theta^2),$$

and

$$\sigma^2 \sim \text{Inv. Gamma}(a, b).$$

These three elements of the prior structure are quite straightforward in Bayesian analysis of normally distributed data (Gelman et al., 2013). The hyperparameters σ_τ^2 and σ_θ^2 would usually be provided large values, resulting in flat, non-informative priors, thus giving majority of influence to the data rather than the priors.

Next,

$$p_i | \alpha_i, \lambda \sim \text{Beta}(\lambda \alpha_i, \lambda).$$

Because p_i is a probability it must be between 0 and 1, and thus the Beta distribution is a natural choice for its prior. The parameterization here results in an expected value of

$$E[p_i | \alpha_i, \lambda] = \frac{\lambda \alpha_i}{\lambda \alpha_i + \lambda} = \frac{\alpha_i}{\alpha_i + 1}. \text{ The expected value is independent of } \lambda.$$

Next,

$$\alpha_i | \boldsymbol{\beta}, v^2 \sim \text{Lognormal}(\mathbf{x}_i^T \boldsymbol{\beta}, v^2),$$

$$\boldsymbol{\beta} \sim \text{Multivar. Normal}(\mathbf{0}, \Sigma_\beta),$$

$$v^2 \sim \text{Inv. Gamma}(c, d),$$

and

$$\lambda \sim \text{Gamma}(f, g).$$

Thus, α_i provides the connection between the covariates \mathbf{x}_i and the probability of treatment p_i , while λ (along with α_i) controls the variance of the Beta distribution: if λ gets very large, $\text{Var}(p_i)$ gets very small, indicating strong knowledge about the $\text{Pr}(T_i = 1 | \mathbf{X}_i)$. The hyperparameters for v^2 and λ are typically given values to result in non-informative priors.

The likelihood function that results from the data $y_i, T_i, i = 1, \dots, n$ may be expressed as:

$$L(\tau, \theta) = \frac{\exp \left\{ - \sum_{i=1}^n \left[\frac{T_i(y_i - \tau - \theta)^2}{2p_i\sigma^2} + \frac{(1 - T_i)(y_i - \theta)^2}{2(1 - p_i)\sigma^2} \right] \right\}}{\sigma^n (\sqrt{2\pi})^n \prod_{i=1}^n (1 - p_i)^{\frac{1-T_i}{2}} p_i^{\frac{T_i}{2}}}$$

It may be shown that the value of $\hat{\tau}$ that optimizes this likelihood function is the same as the equation for $\hat{\tau}_{IPW.SW}$ given above. This demonstrates that we have formulated our Bayesian hierarchical model such that the parameter τ has a very natural interpretation as the average treatment effect, with that estimand understood to be the difference in expected values of the outcome under the two treatment levels.

The benefit achieved from the complexity of this hierarchical structure is, first, that we incorporate into the posterior estimation of the treatment effect all of the variability that exists in estimating the propensity scores for each subject; and second, that we have an extremely flexible model that is capable of providing posterior estimates of the treatment effect even under a variety of true underlying data structures and which captures the variation inherent in the process of estimating the propensity scores by employing a subset of the available measured covariates.

The hierarchical model described above is by no means the only approach at a Bayesian strategy for directly sampling the causal estimand, but an additional advantage it carries is flexibility in allowing for the inclusion of constraints on the parameter space that be required to analyze real data sets arising from various applications. Such constraints may be incorporated by an appropriate prior specification, or modification of the particular

parameters affected. An example of this flexibility is provided in Section 3.4 in the analysis of the case study data.

We would like to demonstrate that this is a viable construction for estimating average treatment effects in observational data. We follow typical Bayesian methods and compute the full joint posterior distribution, as well as the conditional posterior distributions for all of the parameters (see Appendix 3 for full details on the derivations of the posterior conditional distributions). We note that the parameters τ , θ , σ^2 , ν^2 , and $\boldsymbol{\beta}$ all have conjugate constructions and hence have specifiable conditional posterior distributions. These five parameters may be sampled using a Gibbs sampling approach. The remaining three parameters, \boldsymbol{p} , $\boldsymbol{\alpha}$, and λ , have conditional posterior distributions that are unknown; for these, we will need to employ some Markov chain Monte Carlo (MCMC) approach for the posterior sampling (Gelman et al., 2013).

3.3 Simulation

Although we have a number of parameters for which we plan to take posterior samples, our focus is primarily on τ , the average treatment effect. In order to test and demonstrate that the proposed model may be used to gain accurate information about τ , we have designed a simulation study to test various aspects of the approach.

3.3.1 Simulation procedure

To describe the simulation study, we use the ADEMP framework advised by Morris et al. (2019). ADEMP stands for Aims, Data generating mechanisms, Estimands, Methods, and Performance metrics.

Aims: What specifically do we want to learn from the simulation study?

This study is more to demonstrate proof-of-concept, rather than to delineate conditions under which the proposed method is superior to other methods (see Morris, p. 2077). We wish to evaluate the Bayesian sampler created to implement the Bayesian causal inference method devised. In particular, we intend to evaluate the small-sample bias of the proposed method; evaluate the variance of the estimator; and evaluate the method's robustness under varying magnitudes and sources of uncertainty in the data being studied.

Data-generating mechanisms: How will simulated data sets be generated?

We begin by generating a design matrix of dimension $n \times 5$ ($n \in \{100, 500, 1000\}$), with one column for the intercept and 4 columns for mutually independent and normally distributed covariates. We use this design matrix to generate n observations of the binary treatment, $T_i \in \{0,1\}, i = 1,2, \dots, n$, each distributed as *Bernoulli*(p_i) where $p_i = \Pr(T_i = 1|\mathbf{X}_i)$. The p_i s are obtained from the underlying model $\text{logit}[\Pr(T = 1|\mathbf{X})] = \mathbf{X}\boldsymbol{\beta} + \epsilon_p$. We use $\boldsymbol{\beta} = \{0, 4, -2, 0, 0\}$, and $\epsilon_p \sim N(0, \sigma_p^2)$. σ_p^2 is a parameter varied in different simulation settings; we refer to this parameter as “treatment-level noise,” and it reflects the fact that the treatment probabilities are viewed as random variables, not fixed probabilities. Finally, having generated the design matrix \mathbf{X} and the treatment status T , we

generate an outcome value $Y \sim N(\theta + T\tau, \sigma_y^2)$. σ_y^2 is another parameter varied in different simulation settings. It controls the random noise associated with the outcome.

It is worth emphasizing that we are not generating the simulated data according to the Bayesian hierarchical structure described in Section 3.2. That structure is used to facilitate a Bayesian approach at estimating the treatment effect, but it would be far-fetched to imagine that any particular data set would have exactly that structure as its underlying data generating mechanism. Furthermore, we want an analysis method that is robust and broadly applicable, in other words, that is effective under a variety of underlying data generating mechanisms. Our construction of the simulated data sets provides us the flexibility of testing many different potential data generating mechanisms, all sharing a general structure.

Estimands: What is the target of the study?

The proposed method is intended to produce estimates for τ , the ATE. Consequently, the simulation study targets that estimand. We compare to the nonparametric IPW estimate, $\hat{\tau}_{IPW.SW}$. Note that the proposed method involves several additional parameters. These weakly identified parameters are not part of the target of the study. We monitor them to understand aspects of how the method is working, but we do not require them to converge to “actual” values.

Methods: What methods are to be tested or compared?

Here, “method” refers to an approach for estimating causal treatment effects, as understood within the causal inference framework. We are primarily comparing the

proposed method described in Section 3.2 with the nonparametric IPW method using stabilized weights (Lunceford & Davidian, 2004).

Specific details about the implementation of the Bayesian method are partially general to the model, and partially specific to the simulation setup. For example, regardless of the simulation setup, certain parameters have specifiable posterior conditional distributions, and hence we use a Gibbs sampling approach for them. Other parameters have posterior conditional distributions of unknown form. For those, in this simulation, we use a Metropolis-Hastings approach with a random walk technique for drawing new values, i.e., we start with the previous value, add a small random perturbation to it, and then assess the acceptance probability of the new value. Other approaches could be used, such as drawing new values (rather than just incremental change magnitudes from existing value) from an actual known probability distribution. Details of the derivations of the conditional posterior distributions are provided in Appendix 3.

Performance measures: By what criteria will the methods be measured and compared?

Evaluation of the proposed method is conducted first by assessing convergence of the posterior samples for the measures of primary interest. For comparison of the proposed method against the IPW method, we use $Bias(\hat{t}_{IPW})$, $Var(\hat{t}_{IPW})$, $MSE(\hat{t}_{IPW})$; $Bias(\hat{t}_{BYSN})$, $Var(\hat{t}_{BYSN})$, $MSE(\hat{t}_{BYSN})$; and the coverage rates of the different estimators. Note that Bayesian sampling diagnostics, such as convergence and absence of autocorrelation, are not performance measures; rather, these are prerequisites for establishing that the proposed Bayesian approach behaves acceptably in producing output appropriate for analysis.

3.3.2 *Simulation results*

The results from the simulation study are summarized in Figure 3.1 and Table 3.1. The simulation modified settings for outcome-level noise (“ σ ”), treatment-level noise (“ τ ”), and sample size (“ n ”). For each combination of settings, Figure 3.1 displays box plots of ATE estimates for 100 simulated data sets. For each simulated data set, $\hat{\tau}$ was computed by means of the Bayesian hierarchical approach described in the previous section, as well as via the IPW approach. Meanwhile, Table 3.1 summarizes the 200 different estimates for each scenario in terms of average bias, 95% CI (credible interval for Bayesian method, bootstrap percentile confidence interval for frequentist method) coverage rate, and average 95% CI width for both of the two approaches.

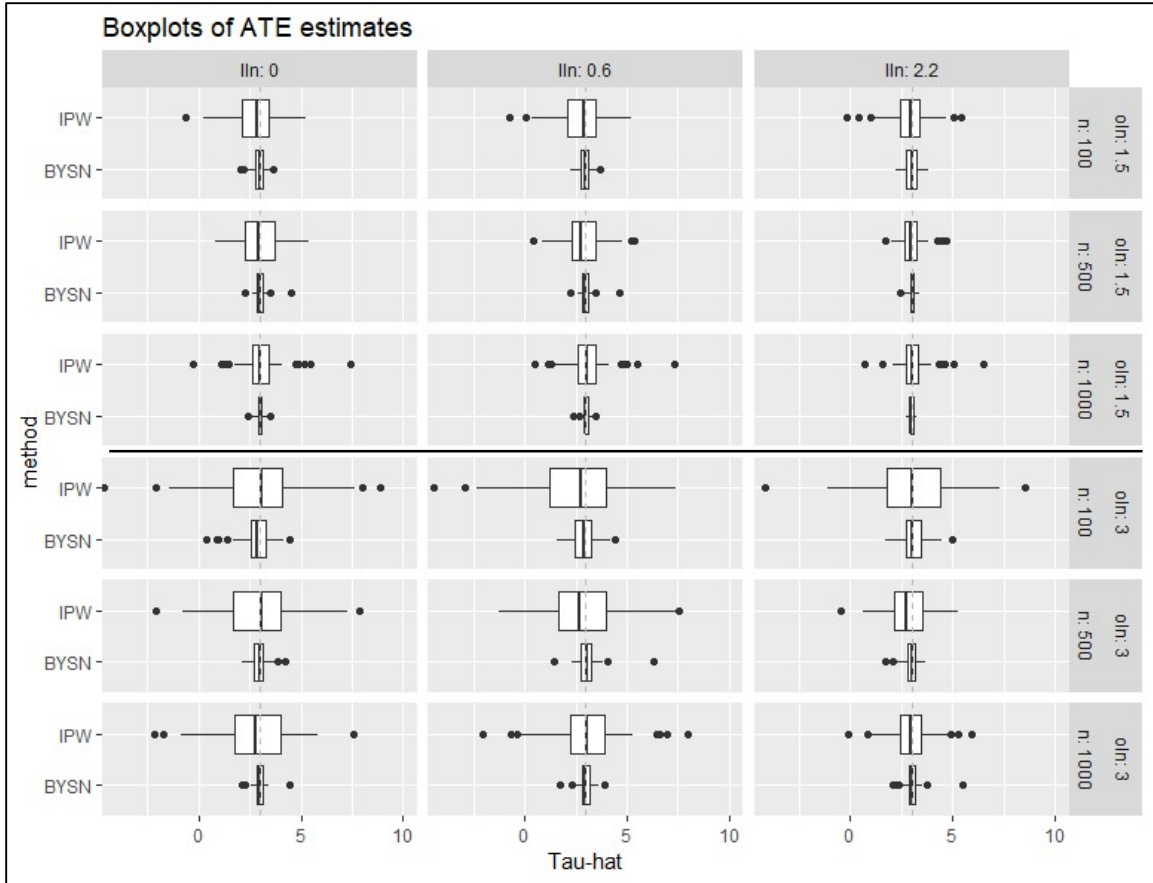


Figure 3.1: Boxplots of simulation study estimates of τ in 18 scenarios. Comparison of Bayesian (BYSN) and Inverse Probability Weighting (IPW) methods for estimating ATE. Simulation scenarios vary by treatment-level noise (“lln,” columns), outcome-level noise (“oln,” top three rows vs. bottom three rows), and sample size (“n”). The tighter, more precise spread of estimates is apparent for the Bayesian method as opposed to IPW, and for the lower setting of outcome-level noise. Treatment-level noise does not noticeably affect the distributions of the estimates.

Several patterns are evident in Figure 3.1. First, the MCMC posterior estimates for τ have much smaller variations (much higher precision) than the corresponding IPW estimates. This is true for every combination of parameter settings. Second, looking across the three columns of the figure, changes in treatment-level noise do not have a noticeable impact on the spread of effect estimates, for either method. Third, in contrast to the previous point, an increase in outcome-level noise does have an impact on the spread of effect estimates, specifically by widening the distributions of estimates. This holds consistent for

all settings of treatment-level noise and sample size, as well as for both methods. Fourth, bias is low, indicating unbiasedness or near unbiasedness of the procedures.

Table 3.1: Comparison of Bayesian and IPW simulation study results by σ_y^2 , n , and σ_p^2 .

			lln=0.0		lln=0.6		lln=2.2	
			BYSN	IPW	BYSN	IPW	BYSN	IPW
oln=1.5	n=100	Bias	-0.049	-0.214	-0.044	-0.236	-0.013	-0.123
		CI cvrg rt.	0.940	0.920	0.950	0.910	0.940	0.910
		Avg. CI width	1.132	2.876	1.130	2.881	1.124	2.511
	n=500	Bias	0.015	-0.052	0.005	-0.119	0.016	-0.017
		CI cvrg rt.	0.870	0.900	0.850	0.860	0.910	0.920
		Avg. CI width	0.492	2.261	0.489	2.220	0.504	1.556
	n=1000	Bias	0.002	0.036	0.011	0.061	-0.017	0.075
		CI cvrg rt.	0.810	0.890	0.850	0.870	0.900	0.940
		Avg. CI width	0.352	1.910	0.359	1.842	0.359	1.309
oln=3.0	n=100	Bias	-0.127	-0.028	-0.086	-0.471	0.076	0.130
		CI cvrg rt.	0.910	0.940	0.950	0.890	0.920	0.940
		Avg. CI width	2.282	5.904	2.252	5.673	2.264	5.018
	n=500	Bias	-0.004	-0.073	0.023	-0.267	-0.001	-0.230
		CI cvrg rt.	0.870	0.840	0.870	0.860	0.930	0.880
		Avg. CI width	0.988	4.231	0.974	4.485	0.994	2.951
	n=1000	Bias	0.005	-0.239	0.011	0.122	0.012	-0.027
		CI cvrg rt.	0.880	0.910	0.850	0.910	0.900	0.900
		Avg. CI width	0.731	3.943	0.714	3.678	0.755	2.608

Notes: oln = outcome-level noise (σ_y^2); n = sample size; lln = treatment-level noise (σ_p^2).

CI: for the Bayesian method, the 95% credible interval for the second 50% of 10,000 MCMC repetitions (the first 50% of posterior samples deleted for burn-in); for the IPW method, 95% confidence interval (table shows average coverages and widths of the empirical 95% bootstrap confidence intervals resulting from 100 bootstrap samples for each repetition within each simulation setting).

The aim for this simulation study as described in the previous sub-section was to evaluate whether the proposed Bayesian hierarchical scheme is a viable approach for sampling and estimating the causal estimand. The results shown in Figure 3.1 and Table 3.1 indicate that the method is indeed successfully estimating the average causal effect, and it is doing so with greater precision than that which is obtained via the IPW approach.

3.4 Case study

To demonstrate how the proposed method may be put into practical use, we apply it to the Lindner dataset contained within the PSAgraphics R package (Helmreich & Pruzek, 2009):

The lindner data contain data on 996 patients treated at the Lindner Center, Christ Hospital, Cincinnati in 1997. Patients received a Percutaneous Coronary Intervention (PCI). The data consists of 10 variables. Two are outcome: lifepres ranges over two values, 11.4 or 0 depending on whether patients survived to six months. Secondly, [cost] contains the costs in 1998 dollars for the first six months... after treatment with the drug abciximab.... The treatment variable is abcix, where 0 indicates standard PCI treatment and 1 indicates standard PCI treatment and additional treatment in some form with abciximab. Covariates include acutemi, 1 indicating a recent acute myocardial infarction and 0 not; ejection for the left ventricle ejection fraction, a percentage from 0 to 90; ves1proc giving the number of vessels (0 to 5) involved in the initial PCI; stent with 1 indicating coronary stent inserted, 0 not; diabetic where 1 indicates that the patient has been diagnosed with diabetes, 0 not; height in centimeters; and female coding the sex of the patient, 1 for female, 0 male.

In this analysis, we focus on the cost outcome variable. The primary research question is whether there is a causal association between use of the drug abciximab and 6-month hospital costs. Higher 6-month hospital costs for those using the drug may well be justified if the drug is effective at extending life spans. However, to understand the cost effectiveness of the drug, it is necessary to quantify the causal difference in hospital costs. As this is observational data, care must be taken to control for possible confounding variables, i.e., other variables that may affect both the outcome (6-month hospital costs) and the probability of treatment. To alleviate issues with skewness and extreme positive outliers, we work with the natural logarithm of cost. We also remove 26 observations for patients who died within 6 months of the procedure, so that we are comparing only patients with full 6-month costs.

Table 3.2 presents the baseline distributions of all covariates, stratified by the treatment. The table also shows the absolute standardized mean difference (ASMD) for each covariate, both before and after adjustment based on the estimated propensity scores. The ASMD is a metric commonly used in propensity score analysis for evaluating covariate balance before and after adjustment on the propensity scores (Austin & Stuart, 2015). The propensity score serves two roles; it is, by definition, the probability of the subject being in the treatment group, conditional on that subject’s covariates. It is also a balancing score, by which is meant that conditional on the propensity score, the distribution of covariates is balanced between the treatment and control groups (Rosenbaum & Rubin, 1983). It is this latter role, the balancing role, that is most important for controlling for confounding bias in observational studies (Shortreed & Ertefaie, 2017). The table demonstrates that the baseline propensity score successfully balances all covariates; the maximum adjusted ASMD is 0.102, which is sufficiently low.

Table 3.2: Baseline characteristics stratified by treatment group, and covariate balance, Lindner dataset

	Stratified by Treatment*		Covariate Balance	
	Control 283	Treatment 687	Unadjusted ASMD	Adjusted ASMD
OUTCOME	Cost (1998 \$s)	14,253 (14,488)		
	ln(Cost)	9.38 (0.53)		
CONTINUOUS COVARIATES	Height (cm)	171.63 (10.50)	0.012	0.035
	Ejection fraction (%)	52.93 (9.62)	0.247	0.046
	# Vessels in procedure	1.20 (0.47)	0.433	0.014
CATEGORICAL COVARIATES		# (%)		
	Stent used (yes)	165 (58.3)	0.256	0.102
	Gender (female)	109 (38.5)	0.117	0.097
	Diabetic (yes)	73 (25.8)	0.132	0.024
	Acute MI within 1 wk (yes)	16 (5.7)	0.380	0.090

* Treatment is use of the drug abciximab.

ASMD: absolute standardized mean difference; MI: myocardial infarction

In the Bayesian approach, after initial estimation via logistic regression of the treatment on the covariates, the vector of treatment probabilities changes with each MCMC step. Consequently, the balance of the covariates between the treatment and control groups changes. Because the central role of the propensity score is to balance the covariates, we impose a constraint to ensure that adequate covariate balance is maintained as we progress through the Markov chain posterior sampling. As mentioned in Section 3.2, the proposed method is flexible in allowing constraints to be incorporated into the prior specification. Here, we impose the constraint by requiring that $Max(ASMD) < 0.2$. If the new sample of the vector of propensity scores results in any covariate having $ASMD > 0.2$, then we reject that sample. As the ASMD metric is a function of the p_i s (that is, a function of the subject-level Bernoulli treatment group probabilities), this constraint is imposed on the prior distribution of the p_i s. Whereas in the statement of the model in Section 3.2, we had:

$$p_i | \alpha_i, \lambda \sim Beta(\lambda \alpha_i, \lambda)$$

we now modify this prior to incorporate the constraint needed for this case study, as:

$$p | \alpha, \lambda \sim \prod_{i=1}^n Beta(\lambda \alpha_i, \lambda) \cdot I\{Max(ASMD) < 0.2\}$$

As shown above in Table 3.2, of the 970 patients who survived to six months, 687 (70.8%) received the treatment drug, and the remaining received only standard care. The average 6-month costs for the treatment group was \$16,008, while for the control group it was \$14,253. This leads to a naïve (unadjusted) estimate for the cost difference of \$1755, with a 95% confidence interval of (-\$62, \$3572). We used the proposed method to estimate the average treatment effect of the use of the drug abciximab on 6-month hospital costs. We ran 10,000 Bayesian MCMC iterations, dropping the first 50% for burn-in. The MCMC

chains showed good convergence even before the 5000th iteration. From this method, we estimated a covariate-adjusted treatment effect (cost difference) of \$2078, with a 95% credible interval of (\$1287, \$2860). This point estimate is nearly 20% higher than the unadjusted/naïve estimate of \$1755, and the credible interval is only 43% as wide as the unadjusted confidence interval.

For further comparison, we also used the frequentist IPW propensity score method (using stabilized weights), with 200 bootstrap samples to estimate the standard error of the estimated treatment effect. The point estimate from this approach was \$1970, with a 95% empirical bootstrap confidence interval of (\$911, \$2916). Figure 3.2, below, illustrates the distributions of the Bayesian posterior estimates and of the IPW bootstrap estimates, as well as the point estimate and confidence interval of the unadjusted approach. We observe that the unadjusted estimate is simply inappropriate in this case: basic assumptions of normally distributed data and independent observations that are involved in estimating the difference in means of two populations are clearly violated. The IPW point estimate is similar to that of the Bayesian method, but the confidence interval is about 30% wider. We believe that the Bayesian credible interval represents a reliable estimate of the true causal difference in 6-month costs between the two treatment groups.

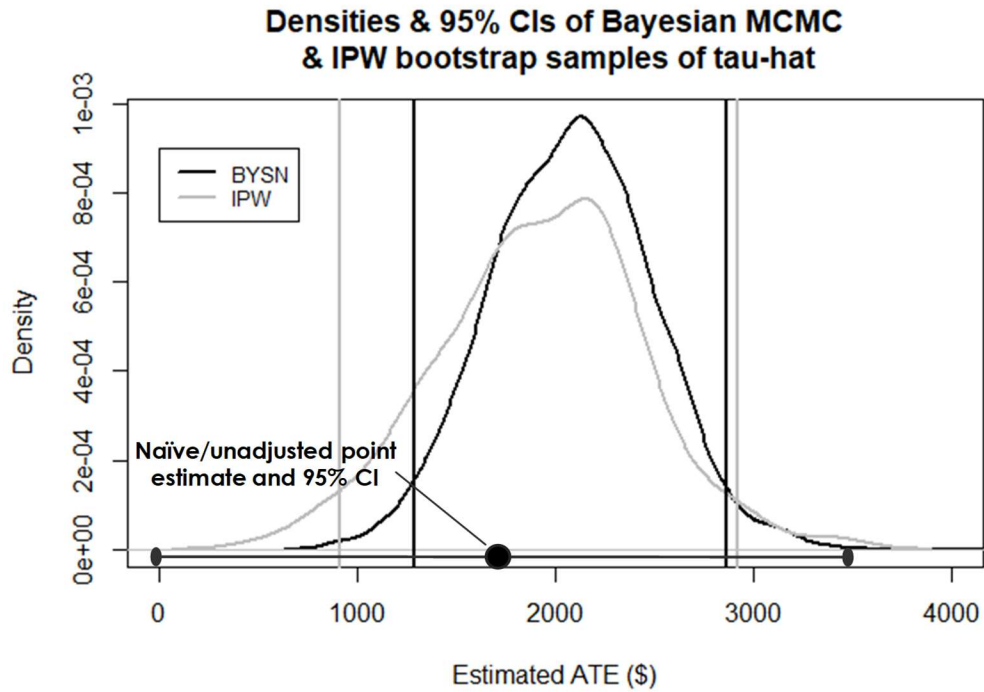


Figure 3.2: Densities of posterior and bootstrap samples, case study. For Bayesian (BYSN) method, density of 5000 posterior values (10K posterior samples, 50% removed for burn-in). For Inverse Probability Weighting (IPW), density of 200 bootstrap samples. Vertical reference lines indicate boundaries for the 95% credible interval (BYSN) and 95% empirical bootstrap confidence interval (IPW). Naïve/unadjusted point estimate and 95% confidence interval also depicted.

3.5 Discussion and conclusions

There are many statistical considerations involved in the analysis of observational data for causal inference. Likewise, there are many potential approaches one may consider in such scenarios. On the frequentist side, a variety of propensity score (and other) methods have been developed over the past several decades. On the Bayesian side, typically Bayesian methods are used for estimating posterior predictive distributions for “missing” counterfactual outcomes or for testing sensitivity to various assumptions. In the approach described in this paper, we attempt to focus directly on the core research question, specifically, what is the average treatment effect of the treatment on the outcome?

The simulation study described above provides a proof-of-concept for the proposed Bayesian approach. It demonstrates that the approach can work, and that it compares favorably against the frequentist IPW method. Further enhancements to the simulation study could test a variety of underlying data structures, as well as scenarios with binary outcomes, or multilevel treatments. As is common in Bayesian analysis, computational time is a consideration. With small to moderate sized datasets, this method performs quite well. With large datasets (say $n > 10,000$ and/or $p > 50$), the computational demands from this method may become onerous.

In this project, we wish to estimate the average causal effect, but we approach the problem from a Bayesian perspective. We use a Bayesian hierarchical structure to conceptualize the data, and then attempt to use posterior sampling to estimate the causal effect parameter directly. The hierarchical model described above is by no means the only way to formulate the problem. In Appendix 4, we provide an alternative formulation and apply a weighted likelihood approach for summarizing the data, which will be investigated in future work.

CHAPTER 4

INNOVATIVE APPROACH FOR SUBGROUP ANALYSIS

4.1 Introduction

Observational studies differ from experimental studies in that assignment of subjects to treatments is not randomized but rather occurs due to natural mechanisms, which are usually hidden from the researchers. Yet objectives of the two studies are frequently the same: identify the treatment effect of some exposure on a population. Furthermore, in both types of studies it is frequently of interest to learn whether treatment effects differ across particular subgroups of subjects, a situation sometimes termed treatment heterogeneity. While these objectives can be achieved directly in an experimental context due to the design imposed on the study, in an observational study special care must be taken to avoid confounding bias in treatment effect estimates, particularly when the number of covariates is large. This research focuses on avoiding confounding bias in estimation of treatment effect, with special focus on identifying effect modifiers. We present a method which efficiently selects effect modifiers from the set of covariates and computes unbiased estimates for subgroups of interest. The goal is to deliver more targeted advice describing circumstances where a treatment may be more beneficial for one or some groups versus others.

There are several motivations for being concerned with subgroup analysis (Lipkovich, Dmitrienko, & B D'Agostino Sr, 2017). First, in the context of a Phase III clinical trial, it may be that the trial fails overall, but the experimental treatment may still offer benefit to some subset(s) of the population. Or, perhaps the Phase III trial is successful, but the sponsor wants to target the experimental treatment to the subset of the population likely to benefit the most. Similarly, but on the negative side, it may be that due to differing safety profiles across the population, regulators deem certain labeling restrictions are needed for a drug on the market. Finally, with ever-increasing attention being given to “personalized medicine,” the goal of identifying optimal treatment regimes from a variety of available treatments often leads to methods involving subgroup analysis (Foster, Taylor, Kaciroti, & Nan, 2015). Any of these motivations, or variations of them, could occur in situations in which observational data, rather than experimental data, is the only – or the most feasible – data available for informing the research question.

A helpful framework for classifying statistical methods related to subgroup analysis is provided by Lipkovich et al. (2017). The authors distinguish four types of methods that fall along a spectrum from purely confirmatory to focused on subgroup discovery. First is confirmatory subgroup analysis, which is concerned with evaluating a small number of pre-defined subgroups. Second is exploratory subgroup evaluation, in which analysis focuses on a relatively small number of subgroups that are pre-specified, and where the focus is mostly on interactions between treatments and covariates, as well as evaluating consistency. Third is post-hoc subgroup evaluation, in which post-hoc assessments are made of the differing treatment effects within a relatively small number of subgroups. This group is more ad hoc than the previous group and would apply to situations dealing with

regulatory inquiries, post-marketing activities, and safety monitoring. Fourth is subgroup discovery, in which the goal is selecting the most promising subgroups out of a potentially large number of candidates. Data mining or machine learning algorithms are likely to be employed here, and the objective is frequently to define subgroups for future analysis in confirmatory studies, such as done by Wang, Schoenfeld, Hoepfner, and Evins (2015).

One extremely common theme in the subgroup analysis literature is that of multiplicity control, i.e., controlling Type I error rate. Lipkovich et al. (2017) reviewed a large amount of literature that provided guidelines that should be followed in conducting subgroup analysis. The authors summarized the “general theme” of the guidelines in six ideas; five of these, arguably, relate to multiple comparison issues. In arguing for “principled data-driven strategies” for conducting subgroup analysis, as opposed to the “guideline-driven approach,” the authors expand on the idea of necessary multiplicity control in an interesting way; for confirmatory subgroup analysis, multiplicity control must be used but must encompass the entire subgroup identification strategy. Furthermore, the multiplicity control should be used in conjunction with “complexity control.” While the former is concerned with controlling Type I error rates, either via strong family-wise error rate control or via limited false discovery rate, the latter is concerned with avoiding data overfitting. As the authors put it, “[a]pplying multiplicity adjustments following subgroup selection is an important but insufficient step, as it would not help find the right covariates ‘after the fact’” (p. 139). The complexity control should be built into the full process of model selection, e.g., via penalized likelihood methods; this lessens the multiplicity burden, making these two ideas complementary concepts that “should be used in combination” (p. 139).

Another common theme in the subgroup analysis literature is the distinction between “black box” methods and those with readily interpretable decision rules. Laber and Zhao (2015) explain that methods based on decision trees (Breiman, 2001) are popular because they fit this latter description, and as such, they are easily implemented in the field. In contrast, regression-based approaches (e.g., Qian and Murphy (2011) and Brinkley, Tsiatis, and Anstrom (2010)) typically face the choice between constructing parsimonious models leading to more interpretable decision rules but subject to model misspecification, or more complex models that may avoid misspecification but result in “unintelligible” treatment rules. Tian, Alizadeh, Gentles, and Tibshirani (2014) propose such a regression-based model; in their approach, they use modified covariates in a regression of the outcome on treatment and treatment-covariate interaction terms and stratify subjects based upon their resulting predicted treatment response profile; while this approach serves to divide up an existing data set into, say, a low score and a high score group (with one or the other of those indicated as benefiting more from the treatment), no particular rules are provided for guiding decisions about future subjects.

In this project, we propose a flexible outcome model, where the control group response profile is captured by a non-parametric function, and treatment heterogeneity is captured by the interaction term between treatment and a linear combination of covariates. Penalized regression and inverse probability weighting (IPW) are applied to select the important variables in the interaction, thus the effect modifiers and subgroups can be identified. The approach is quite flexible, and it can be applied to data from either randomized trials or observational studies.

The structure of the remainder of this chapter is as follows: in Section 4.2, we describe the context and a proposed method for conducting subgroup analysis. This method fits within the fourth category of subgroup analysis described above; it is concerned with identifying the right treatment for a given patient, rather than identifying the right patient for a given treatment. In Section 4.3, we describe the simulation study we executed to test the proposed methodology. In Section 4.4, we illustrate application of the method to a case study data set. Finally, in Section 4.5, we provide further discussion of the results and general conclusions.

4.2 Methods

Let (\mathbf{X}, T, Y) indicate the triplet for the baseline covariates, treatment group and outcome variable, with $T \in \{0,1\}$. Let (\mathbf{X}_i, T_i, Y_i) ($i = 1, \dots, n$) indicate a random sample from a population of interest. The sample could result from either a randomized experimental design or an observational study. The following model (1) has often been used (Fu, Zhou, & Faries, 2016) to examine treatment heterogeneity, identify the subgroup which may benefit from the treatment, and select the optimal treatment regime:

$$E(Y|\mathbf{X}, T) = h(\mathbf{X}) + g(\mathbf{X})T. \quad (1)$$

In particular, the interaction term $g(\mathbf{X})T$ plays an important role in identifying the optimal treatment or identifying the subgroup which receives more benefit from the treatment. We use the concept of potential outcomes (Rubin, 1974) to look at the role of $g(\mathbf{X})$. We denote with $Y^{(0)}$ and $Y^{(1)}$, respectively, the potential outcomes for control and treatment for a given subject with covariates \mathbf{X} . Here we assume that exchangeability and consistency hold (Hernán & Robins, 2020): (i) exchangeability means that $Y^{(a)} \perp T | \mathbf{X}$, and (ii) consistency means that the observed outcome equals the potential outcome

corresponding to the treatment the subject receives, that is, $Y = TY^{(1)} + (1 - T)Y^{(0)}$. If we assume model (1) is correctly specified, then the treatment effect for a subject with covariate \mathbf{X} would be

$$\begin{aligned}
E[Y^{(1)} - Y^{(0)}|\mathbf{X}] &= E[Y^{(1)}|\mathbf{X}] - E[Y^{(0)}|\mathbf{X}] \\
&= E[Y^{(1)}|\mathbf{X}, T = 1] - E[Y^{(0)}|\mathbf{X}, T = 0] && \text{(by exchangeability)} \\
&= E[Y|\mathbf{X}, T = 1] - E[Y|\mathbf{X}, T = 0] && \text{(by consistency)} \\
&= [h(\mathbf{X}) + g(\mathbf{X})] - [h(\mathbf{X})] && \text{(by (1))} \\
&= g(\mathbf{X}).
\end{aligned}$$

We want to identify the group in which subjects have a beneficial treatment effect, that is $\mathcal{S} = \{\mathbf{X}: E[Y^{(1)} - Y^{(0)}|\mathbf{X}] > 0\}$. In other words, $\mathcal{S} = \{\mathbf{X}: g(\mathbf{X}) > 0\}$. Here, the primary interest is to estimate the contrast $g(\mathbf{X})$ to identify the subgroup which benefits more from treatment. To facilitate a clinical decision, we would like the $g(\mathbf{X})$ function to be simple and to capture the main variables for decision making. The $h(\mathbf{X})$ function captures the response profile under control, which is not of direct interest when the research question is to examine whether any subgroup has a differing response to treatment. Thus, we can consider $h(\mathbf{X})$ to be a nuisance function, which is used to facilitate accurate estimation of the $g(\mathbf{X})$ function. The structural nested mean model (SNMM) (Hernán & Robins, 2020) uses a parametric function for $g(\mathbf{X})$ (for example, $g(\mathbf{X}; \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$) to estimate the parameter $\boldsymbol{\beta}$. The first step is to link the SNMM with the observed data; using the assumption of consistency, we have (Hernán & Robins, 2020):

$$Y^{(0)} = Y - Tg(\mathbf{X}; \boldsymbol{\beta}). \tag{2}$$

When only one parameter β is involved, β has been estimated by minimizing the association between $Y^{(0)}$ and T . Specifically, β is obtained by solving the following estimating equations:

$$\sum_{i=1}^n (Y_i - T_i g(X_i; \beta))(T_i - E(T_i|X)) = 0. \quad (3)$$

Here, $E(T|X) = \Pr(T = 1|X)$, the probability, conditional on covariates, of being in the treatment group. In a randomized experiment, this probability is a constant, while in an observational study, it is the propensity score and must be estimated.

When β is a vector, in what is termed the A-learning method (Schulte, Tsiatis, Laber, & Davidian, 2014) the following estimating equations have been proposed (Vansteelandt & Joffe, 2014):

$$\sum_{i=1}^n \left\{ \frac{\partial g(X_i; \beta)}{\partial \beta} \right\} \{Y_i - E(Y_i|X_i) - g(X_i; \beta)(T_i - E(T_i|X))\} \{T_i - E(T_i|X)\} = 0. \quad (4)$$

Note that $E(Y|X) \neq h(X)$, instead

$$\begin{aligned} E(Y|X) &= \int yf(y|x)dy = \iint yf(y, t|x)dtdy = \iint yf(y|x, t)f(t|x)dtdy \\ &= \int yf(y|x, t=0)f(t=0|x)dy + \int yf(y|x, t=1)f(t=1|x)dy \\ &= \int y^{(0)}f(y^{(0)}|x, t=0)f(t=0|x)dy^{(0)} \\ &\quad + \int y^{(1)}f(y^{(1)}|x, t=1)f(t=1|x)dy^{(1)} \\ &= E(Y^{(0)}|X)Pr(T=0|X) + E(Y^{(1)}|X)Pr(T=1|X) \\ &= h(X)(1 - E(T|X)) + (h(X) + g(X))E(T|X) \\ &= h(X) + g(X)E(T|X) = h_0(X). \end{aligned}$$

We define $h_0(\mathbf{X}) = E(Y|\mathbf{X}) = h(\mathbf{X}) + g(\mathbf{X})E(T|\mathbf{X})$. It has been shown (Schulte et al., 2014) that Equation (4) provides consistent estimates for the contrast function if the model for $g(\mathbf{X})$ and the propensity score model are correctly specified. Also, Lu, Zhang, and Zeng (2013) assume $h(X)$ to be a parametric function, for example $h(X; \gamma) = X\gamma$. However, our objective is not to estimate $h(X)$, nor to estimate $h_0(X)$. Instead, we are interested in estimating the interaction function $g(X)$, which captures the treatment heterogeneity. Let us assume that $g(X)$ has a functional form as $g(X_i; \beta)$. The loss function (i.e., the sum of squares for errors) $L_1(\beta) = \frac{1}{n} \sum_{i=1}^n [Y_i - h(X_i) - T_i g(X_i; \beta)]^2$ can be written equivalently as

$$L(\beta) = \frac{1}{n} \sum_{i=1}^n [Y_i - h_0(X_i) - g(X_i; \beta)(T_i - E(T_i|X_i))]^2. \quad (5)$$

Estimating equation (4) can be linked with the loss function in equation (5) (Lu et al., 2013). The parameter β can be estimated by

$$\sum_{i=1}^n \left\{ \frac{\partial g(X_i; \beta)}{\partial \beta} \right\} \{Y_i - h(X_i; \gamma) - T_i g(X_i; \beta)\} \{T_i - E(T_i|X)\} = 0. \quad (6)$$

The derivative of $L_1(\beta)$ with respect to β will not result in equation (6), which is the equation related to A-learning. Thus, we use the objective function in equation (5). We propose using a more flexible and nonparametric model for $h_0(X)$ (for example, the generalized boosted model (McCaffrey, Ridgeway, & Morral, 2004), or the random forest method (Breiman, 2001)), while we use a relatively simple model for $g(X; \beta)$. For example, we take

$$g(X_i; \beta) = \beta_0 + \sum_{j=1}^p X_{ij}\beta_j + \sum_{j=1}^{p-1} \sum_{j'=j+1}^p X_{ij} X_{ij'}\beta_{jj'} = X_{ext}\beta,$$

in other words, a simple linear combination of the variables we are interested in.

We incorporate complexity control into the process by using lasso (Tibshirani, 1996) or the elastic net (Zou & Hastie, 2005) method to select the few important variables to identify the subgroup that receives more benefit from the treatment. We propose to estimate $h(\mathbf{X})$ non-parametrically and estimate $g(\mathbf{X}; \beta)$ using penalized regression, as detailed in the following steps:

- (i) *Obtain $E(T|X) = Pr(T = 1|X)$* : For an observational study, we estimate the probability, perhaps via logistic regression; for a randomized study, $E(T|X)$ is assumed to be a known constant. For example, $E(T|X)$ might equal 0.5 for a randomized trial with equal probability of control or treatment assignment.
- (ii) *Estimate $E(Y|X) = h_0(X)$ nonparametrically*: We fit the model $E(Y|X) = h_0(X)$ using a nonparametric approach (e.g., random forest or generalized boosted model) to obtain an estimate of $h_0(X)$, which we denote as $\hat{h}_0(X)$. For this step, we ignore the treatment variable and fit the outcome model using all observed data, including observations from both control and treatment subjects.
- (iii) *Estimate the contrast function $g(X; \beta)$* : We set $Y^* = Y - \hat{h}_0(X)$, and set $X^* = \text{Diag}(T - E(T|X))X_{ext}$. We then use lasso or the elastic net method to estimate $g(X; \beta)$ by minimizing $L(\beta) = (Y^* - X^*\beta)^T(Y^* - X^*\beta)$. The minimization of the loss function in this step using penalized regression methods results in the selection of variables that interact with the treatment to modify the treatment effect.

- (iv) *Identify subgroup(s) benefiting from treatment:* We define the estimated subgroup as $\hat{\mathcal{S}} = \{X: g(X; \hat{\beta}) > 0\}$, or, depending on the specific application, we may use $\hat{\mathcal{S}}_\delta = \{X: g(X; \hat{\beta}) > \delta\}$ for some clinically meaningful value δ .

4.3 Simulation

4.3.1 Simulation procedure

As a proof-of-concept we designed a simulation study to implement the proposed method. In this section we use the ADEMP framework described by Morris et al. (2019) as a methodical approach for planning and describing our simulation study. Our decisions for the various aspects of the study were to a large extent influenced by the simulation study conducted by Lu et al. (2013), which is for randomized controlled trials (RCTs). Our approach is applicable to both RCTs and observational studies.

Aims: What specifically do we want to learn from the simulation study?

Our aims for the simulation study were to demonstrate that our proposed method is effective at identifying subgroups under a variety of underlying data conditions. The outcome was designed such that a higher value indicated a more desirable response.

Data-generating mechanisms: How will simulated data sets be generated?

Our data generating mechanism varied those dimensions that were of greatest importance in determining the conditions under which the proposed method is effective. As in Lu et al. (2013), we used 10 covariates, but whereas those authors solely used a correlated structure, we compared performance under both correlated and independent

covariates. We do advise our method for both experimental and observational data, and so while Lu et al. (2013) treated only experimental data, we also included several different types of observational data scenarios, varying in the degree and type of confounding that was present. Other simulation parameters that were varied included sample size, underlying probability of treatment, signal-to-noise ratio, and treatment effect size.

To describe our data generating mechanism in detail, we use the outcome model $Y = h(\mathbf{X}; \boldsymbol{\gamma}) + Tg(\mathbf{X}; \boldsymbol{\beta}) + \epsilon$, while the probability of treatment is computed via the model $\text{logit}(\Pr[T = 1|\mathbf{X}]) = k(\mathbf{X}; \boldsymbol{\phi})$, where $k(\mathbf{X}; \boldsymbol{\phi}) = \boldsymbol{\phi}^T \tilde{\mathbf{X}}$. We thus have three functions, $h(\cdot)$, $g(\cdot)$, and $k(\cdot)$, that govern the relationships among covariates, treatment, and outcome. The h -function controls the complexity and linearity of the relationship between the outcome Y and the covariates, under no treatment. The g -function controls the nature of the interaction effects between the treatment and covariates; as described in Section 4.2, this is the key function for identifying subgroups. Finally, the k -function specifies the relationship between the covariates and the treatment.

The simulation evaluated the proposed method under scenarios that differed first with respect to the complexity of the response profile under control (i.e., the h -function), and second with respect to the nature and degree of confounding (i.e., the k -function). Table 4.1 shows the exact settings used for the h -, g -, and k -functions.

Table 4.1: Settings for h -, g -, and k -functions in simulation scenarios.

$Y = h(\mathbf{X}; \boldsymbol{\gamma}) + Tg(\mathbf{X}; \boldsymbol{\beta}) + \epsilon$ $X = (X_1, X_2, \dots, X_{10}) \sim MVN(0, \boldsymbol{\Sigma}), \epsilon \sim N(0, 0.5^2)$ $g(\mathbf{X}; \boldsymbol{\beta}) = \boldsymbol{\beta}^T \tilde{\mathbf{X}}, \text{ where } \tilde{\mathbf{X}} \equiv (1, \mathbf{X}^T)^T$ $\boldsymbol{\beta} = (1, 1, \mathbf{0}_7, -0.9, 0.8)^T$ $\mathbf{0}_d \text{ indicates a vector of zeroes of length } d.$	
<p><i>Models for $h(\mathbf{X}; \boldsymbol{\gamma})$:</i></p>	<p>Y1: $h(\mathbf{X}; \boldsymbol{\gamma}) = 1 + \boldsymbol{\gamma}_1^T \mathbf{X}$</p> <p>Y2: $h(\mathbf{X}; \boldsymbol{\gamma}) = 1 + 0.5(\boldsymbol{\gamma}_1^T \mathbf{X})(\boldsymbol{\gamma}_2^T \mathbf{X})$</p> <p>Y3: $h(\mathbf{X}; \boldsymbol{\gamma}) = 1 + 0.5 \sin(\pi \boldsymbol{\gamma}_1^T \mathbf{X}) + 0.25(1 + \boldsymbol{\gamma}_2^T \mathbf{X})^2$ $\boldsymbol{\gamma}_1 = (1, -1, \mathbf{0}_8)^T$ and $\boldsymbol{\gamma}_2 = (1, \mathbf{0}_2, -1, \mathbf{0}_5, 1)^T$</p>
<p><i>Propensity score models:</i></p> <p>$\text{logit}(\Pr[T = 1 \mathbf{X}]) =$ $k(\mathbf{X}; \boldsymbol{\phi}) = \boldsymbol{\phi}^T \tilde{\mathbf{X}}.$</p>	<p>$\boldsymbol{\phi}_A = (1, \mathbf{0}_5, 1, 1, \mathbf{0}_3)^T$;</p> <p>$\boldsymbol{\phi}_B = (1, \mathbf{0}_6, 1, 1, 1, 0)^T$</p> <p>$\boldsymbol{\phi}_C = (1, \mathbf{0}_2, 1, 0, 1, 1, 1, 1, \mathbf{0}_2)^T$</p> <p>$\boldsymbol{\phi}_D = (1, 0, 1, 0, 1, \mathbf{0}_6)^T$</p> <p>$\boldsymbol{\phi}_R = \mathbf{0}_{11}, k(\mathbf{X}; \boldsymbol{\phi}_R) = 0$ (Randomized experiments)</p>

In every case, $g(\mathbf{X}; \boldsymbol{\beta}) = \boldsymbol{\beta}^T \tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}} \equiv (1, \mathbf{X}^T)^T$ and $\boldsymbol{\beta} = (1, 1, \mathbf{0}_7, -0.9, 0.8)^T$; $\mathbf{0}_d$ indicates the zero vector of length d . The three different formulations for the outcome, Y , were obtained by varying the h -function. For Y1, $h(\mathbf{X}; \boldsymbol{\gamma}) = 1 + \boldsymbol{\gamma}_1^T \mathbf{X}$, an ordinary linear combination of the covariates; for Y2, $h(\mathbf{X}; \boldsymbol{\gamma}) = 1 + 0.5(\boldsymbol{\gamma}_1^T \mathbf{X})(\boldsymbol{\gamma}_2^T \mathbf{X})$; and for Y3, $h(\mathbf{X}; \boldsymbol{\gamma}) = 1 + 0.5 \sin(\pi \boldsymbol{\gamma}_1^T \mathbf{X}) + 0.25(1 + \boldsymbol{\gamma}_2^T \mathbf{X})^2$, where $\boldsymbol{\gamma}_1 = (1, -1, \mathbf{0}_8)^T$ and $\boldsymbol{\gamma}_2 = (1, \mathbf{0}_2, -1, \mathbf{0}_5, 1)^T$. These specifications were motivated by the intention of mimicing the simulation approach used in Lu (2013). However, whereas that study considered only experimental data, we also considered observational data where $\Pr(T = 1|\mathbf{X})$ was assumed unknown. We experimented with four different approaches at constructing the $\Pr(T = 1|\mathbf{X})$. In each case, the relationship could be expressed generally as $\text{logit}(\Pr[T = 1|\mathbf{X}]) = k(\mathbf{X}; \boldsymbol{\phi}) = \boldsymbol{\phi}^T \tilde{\mathbf{X}}$. For specifications A, B, C, and D, we used $\boldsymbol{\phi}_A = (1, \mathbf{0}_5, 1, 1, \mathbf{0}_3)^T$; $\boldsymbol{\phi}_B = (1, \mathbf{0}_6, 1, 1, 1, 0)^T$; $\boldsymbol{\phi}_C = (1, \mathbf{0}_2, 1, 0, 1, 1, 1, 1, \mathbf{0}_2)^T$; and $\boldsymbol{\phi}_D = (1, 0, 1, 0, 1, \mathbf{0}_6)^T$, respectively. These choices were made based upon considerations of how the covariates

were used in the h - and g -functions and the structure of confounding that resulted; the variance-covariance matrix for the correlated covariates scenarios, detailed in the next paragraph, also factored into making these decisions. The causal diagrams in Figure 4.1 illustrate these relationships as resulting from the first h -function specification (Y1) and four different propensity score models.

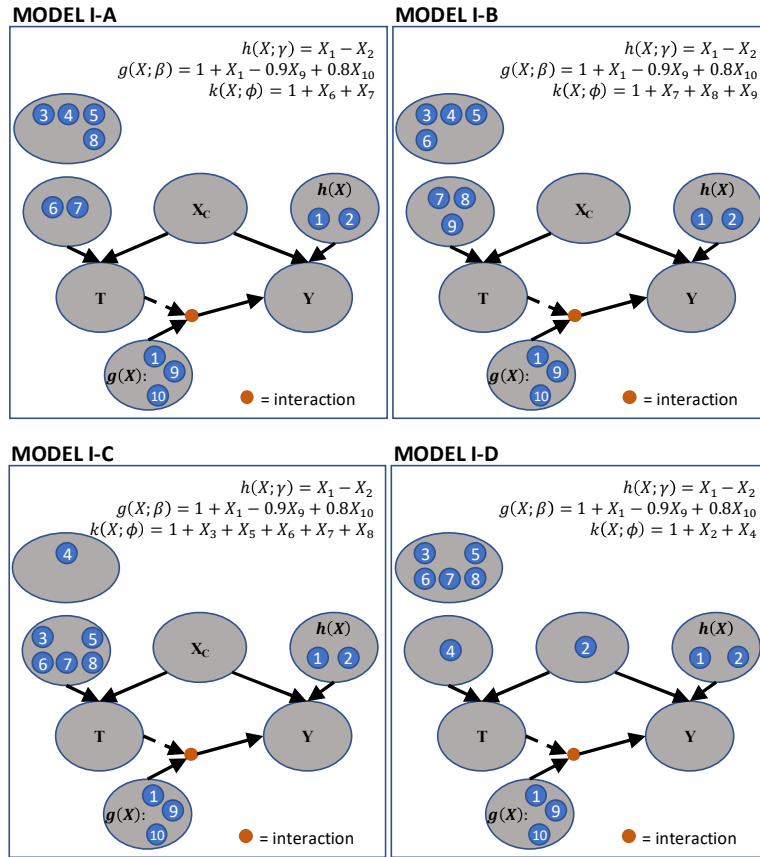


Figure 4.1: Structure of confounding in various simulation scenarios. These causal diagrams show the structure of confounding for the four observational data scenarios for the first specification of the outcome, Y. Small circles with numbers represent individual covariates, of which there are 10. The h -function and the g -function remain the same for these four scenarios; only the k -function, which specifies the relationship between the covariates and the probability of treatment, changes.

Consequently, in our simulation, for each of the three constructions of Y, we tested five constructions of the treatment-covariate relationships: the four observational scenarios

just described, in which during the estimation process we treated the propensity score as unknown, and the one experimental data scenario in which the propensity score was known. This gave 15 versions of the generated outcome values. The process for generating the data was as follows: we first generated values for 10 covariates and one error term, with sample size n . The covariates had a multivariate normal distribution, $\mathbf{X} \sim \text{Multivariate Normal}(0, \mathbf{\Sigma})$, and the error term had the distribution $\epsilon \sim N(0, 0.5^2)$. For the independent covariates scenarios, $\mathbf{\Sigma} = \mathbf{I}_{10}$, while for the correlated covariates scenarios, we used $\text{Corr}(X_j, X_k) = 0.5^{|j-k|}$, in keeping with Lu 2013. (Lu et al., 2013) We then used the values of the covariates \mathbf{X} to compute $\text{Pr}(T = 1|\mathbf{X})$ for the observational data scenarios, or we set $\text{Pr}(T = 1|\mathbf{X}) = 0.5$ for the experimental data scenarios. We then generated the vector of binary treatment indicators according to $T_i \sim \text{Bernoulli}(\text{Pr}(T_i = 1|\mathbf{X}_i))$. Finally, we computed Y using the known values of \mathbf{X} and T and the appropriate h - and g -functions for each specific scenario.

Estimands: What is the target of the study?

The target of this simulation study was accurate prediction. Hence we focused on the more general term “target” rather than “estimand” as described in Morris et al. (2019), Section 3.3.

Methods: What methods are to be tested or compared?

The proposed method, which is the subject of this paper, was the primary method used in this simulation, as we wished to demonstrate that the method works.

Performance measures: By what criteria will the various methods be measured and compared?

Following Lu et al. (2013), we employed several different performance measures. The most important of these was termed Percent Correct Decision (PCD). This could be expressed simply as $PCD = \frac{1}{n} \sum_{i=1}^n I \left[\text{sign} \left(g(X_i; \hat{\beta}) \right) = \text{sign} \left(g(X_i; \beta) \right) \right]$. We constructed Y such that larger values were more desirable; thus, $g(X_i; \hat{\beta}) > 0$ implied that subject i should be prescribed the treatment, while $g(X_i; \hat{\beta}) < 0$ implied that subject i should not be prescribed the treatment. PCD, then, measured how well the predicted decision matched the known best decision. Other performance measures (see Table 4.2) included mean squared error (MSE), measuring the accuracy of coefficient estimates in the g -function; the number of correctly dropped covariates (Corr0); the number of incorrectly dropped covariates (Incorr0); and the average proportion of times that the method selected the exactly correct set of covariates (Exact).

4.3.2 Simulation results

The simulation scenarios varied in terms of sample size, independence versus correlation of covariates, structure of confounding, and nature of the h -function. Table 4.2 presents results for the largest sample size ($n=1000$); similarly structured tables of results for other sample sizes ($n=400, 200$ and 100) are presented in Appendix 5.

From Table 4.2 we observe that the absolute levels of the PCD are typically in the high 90 percents for independent covariates, and in the mid-90 percents for correlated covariates, demonstrating that the proposed method leads, a great majority of the time, to the correct decision regarding whether a particular subject should receive treatment or not.

It is also apparent that the method performs better for independent than for correlated covariates; the ratio of the PCD of the correlated covariates scenarios to that of the independent covariates scenarios is always in the range of 0.95-0.99, i.e., the PCD is about 1% to 5% lower when covariates are correlated. Moreover, the number of correctly dropped covariates is close to seven, the true number in all scenarios, but the independent covariates scenarios performed slightly better than the scenarios with correlated covariates. Correlation engrained in the covariate structure will make it more difficult to accurately de-select those which are not involved in the g -function (see the ratio for the Corr0 metric in Table 4.2).

Comparing the outcome models Y1, Y2, and Y3, the method does best for the least complicated h -function (Y1), then next best for the Y3 h -function, and third best for the Y2 h -function. Performance does not vary markedly under different approaches at specifying the covariate-treatment relationship in the observational data scenarios (specifications A, B, C, and D). For those scenarios, elastic net typically achieves a fractionally higher PCD than lasso, but a fractionally lower Corr0 and Exact; this can probably be explained by the fact that lasso is shrinking the estimated regression parameters more aggressively than is elastic net. These differences seem insubstantial, suggesting that either of these penalized regression approaches should achieve satisfactory results in the observational data scenarios.

The number of incorrectly dropped covariates is 0 in almost all scenarios (Incorr0 in Table 4.2), indicating that our proposed method selects the important variables very well. Although, the proposed method performs slightly better for experimental data (the last two rows of each of the three main vertical table sections) as compared to observational data

(rows 1 through 8 of each of the three main vertical table sections), the proposed method performs well for both experimental data and observational data.

Table 4.2: Simulation results from 1000 Monte Carlo repetitions for each of the 15 scenarios (3 outcome models \times 5 PS models) under independent covariates and correlated covariates (N=1000)

N=1000	Independent covariates					Correlated covariates					Ratios: Corr::Indpndt			Diff: Corr-Indpndt	
	PCD	MSE	Corr0	Incorr0	Exact	PCD	MSE	Corr0	Incorr0	Exact	PCD	MSE	Corr0	Incorr0	Exact
Y1.A.EN	97.7	0.046	6.90	0	0.91	95.8	0.053	6.39	0	0.61	0.98	1.16	0.93	0.00	(0.30)
Y1.A.L	97.5	0.045	6.97	0	0.97	95.6	0.051	6.77	0	0.82	0.98	1.12	0.97	0.00	(0.15)
Y1.B.EN	97.6	0.054	6.97	0	0.98	94.5	0.070	6.38	0	0.55	0.97	1.30	0.92	0.00	(0.43)
Y1.B.L	97.4	0.054	7.00	0	1.00	94.0	0.072	6.75	0	0.79	0.97	1.34	0.96	0.00	(0.21)
Y1.C.EN	97.5	0.045	6.99	0	0.99	94.9	0.057	6.73	0	0.74	0.97	1.27	0.96	0.00	(0.25)
Y1.C.L	97.3	0.045	7.00	0	1.00	94.5	0.056	6.96	0	0.97	0.97	1.26	0.99	0.00	(0.03)
Y1.D.EN	97.9	0.041	6.93	0	0.94	96.3	0.047	6.58	0	0.63	0.98	1.16	0.95	0.00	(0.31)
Y1.D.L	97.7	0.041	6.99	0	0.99	95.9	0.048	6.92	0	0.92	0.98	1.17	0.99	0.00	(0.07)
Y1.EN	98.4	0.036	6.92	0	0.93	97.2	0.039	6.33	0	0.55	0.99	1.09	0.91	0.00	(0.38)
Y1.L	98.3	0.035	6.99	0	0.99	97.0	0.038	6.84	0	0.86	0.99	1.10	0.98	0.00	(0.13)
Y2.A.EN	95.7	0.085	6.94	0	0.94	92.6	0.090	6.60	0	0.64	0.97	1.06	0.95	0.00	(0.30)
Y2.A.L	95.4	0.082	7.00	0	1.00	92.3	0.086	6.83	0	0.84	0.97	1.05	0.98	0.00	(0.16)
Y2.B.EN	94.9	0.099	6.98	0	0.98	91.1	0.112	6.59	0.02	0.65	0.96	1.13	0.94	0.02	(0.33)
Y2.B.L	94.5	0.098	7.00	0	1.00	90.3	0.113	6.78	0.1	0.75	0.96	1.16	0.97	0.10	(0.25)
Y2.C.EN	94.8	0.084	6.99	0	0.99	92.2	0.092	6.64	0	0.68	0.97	1.09	0.95	0.00	(0.31)
Y2.C.L	94.2	0.085	6.99	0	0.99	91.2	0.093	6.88	0.02	0.87	0.97	1.10	0.98	0.02	(0.12)
Y2.D.EN	96.4	0.066	6.96	0	0.96	94.8	0.069	6.59	0	0.67	0.98	1.05	0.95	0.00	(0.29)
Y2.D.L	96.0	0.067	7.00	0	1.00	94.5	0.069	6.83	0	0.85	0.98	1.02	0.98	0.00	(0.15)
Y2.EN	96.7	0.064	6.98	0	0.98	95.3	0.064	6.54	0	0.67	0.99	1.00	0.94	0.00	(0.31)
Y2.L	96.4	0.063	6.99	0	0.99	95.2	0.061	6.83	0	0.90	0.99	0.97	0.98	0.00	(0.09)
Y3.A.EN	96.4	0.070	6.91	0	0.91	93.0	0.085	6.57	0	0.66	0.96	1.22	0.95	0.00	(0.25)
Y3.A.L	96.1	0.070	6.99	0	0.99	92.6	0.084	6.80	0.01	0.81	0.96	1.21	0.97	0.01	(0.18)
Y3.B.EN	96.1	0.082	7.00	0	1.00	91.3	0.110	6.58	0.08	0.62	0.95	1.34	0.94	0.08	(0.38)
Y3.B.L	95.8	0.081	7.00	0	1.00	90.8	0.108	6.77	0.13	0.70	0.95	1.33	0.97	0.13	(0.30)
Y3.C.EN	95.9	0.074	6.98	0	0.98	91.8	0.095	6.67	0	0.76	0.96	1.30	0.96	0.00	(0.22)
Y3.C.L	95.7	0.070	6.98	0	0.98	91.0	0.096	6.83	0	0.87	0.95	1.39	0.98	0.00	(0.11)
Y3.D.EN	96.9	0.062	6.96	0	0.96	94.5	0.076	6.68	0	0.73	0.98	1.22	0.96	0.00	(0.23)
Y3.D.L	96.6	0.062	6.99	0	0.99	94.1	0.076	6.93	0	0.93	0.97	1.22	0.99	0.00	(0.06)
Y3.EN	97.4	0.054	6.93	0	0.94	95.6	0.061	6.49	0	0.62	0.98	1.13	0.94	0.00	(0.32)
Y3.L	97.2	0.053	6.98	0	0.98	95.2	0.062	6.88	0	0.88	0.98	1.16	0.99	0.00	(0.10)

Notes: PCD=Percent Correct Decision; MSE=Mean Squared Error; Corr0=avg. # of covariates correctly estimated as 0 (in g-function); Incorr0=avg. # of covariates incorrectly estimated as 0 (in g-function); Exact=proportion of times the exactly correct set of covariates is selected

Y1/Y2/Y3=different specifications of h -function; A/B/C/D=different specifications of k -function; EN=elastic net; L=lasso

4.4 Case study

4.4.1 Case study background

To demonstrate our proposed method, we applied it to a case study data set to identify the subgroup of patients with heavy alcohol use who may benefit from varenicline treatment (Litten et al., 2013). Alcohol use has been identified as the third-leading risk factor for the global burden of disease and injury. Excessive alcohol consumption is estimated to drive costs of over \$200 billion annually in the USA. Varenicline, which goes

by the trade name Chantix, was approved by the US Food and Drug Administration in 2006 as an aid for smoking cessation. Varenicline is an $\alpha 4\beta 2$ nicotinic acetylcholine agonist. Converging lines of research suggest that both alcohol and nicotine affect nicotinic acetylcholine receptors, which control rewarding effects in the brain. Since varenicline has proven effective at aiding smoking cessation, and since alcohol appears to affect the brain in a similar fashion as nicotine, it seems reasonable that this same drug may show benefit in aiding drinking cessation, or at least drinking reduction.

The target population consisted of adults at least 18 years old who were consistent heavy drinkers (defined as at least 28 drinks per week for females, or at least 35 drinks per week for males). Exclusion criteria included being pregnant, being addicted to any drugs other than alcohol or nicotine, having any psychiatric disorders, and certain other comorbidities. The study design was a Phase II randomized, double-blind, placebo-controlled multisite trial, and the duration of the study period was 13 weeks. Ultimately 99 subjects were enrolled in the treatment arm, and 101 subjects were enrolled in the control arm. Three subjects in the treatment group had insufficient outcome data for analysis. Table 4.3 summarizes the baseline characteristics of subjects, stratified by treatment group.

Table 4.3: Baseline characteristics of patients, varenicline case study

N (%)	Stratified by Treatment				
	Placebo		Varenicline		
	101	(51.3)	96	(48.7)	
CONTINUOUS VARIABLES	Mean		Mean		
	Age	45.0	(12.3)	46.0	(11.0)
	Years education	14.8	(2.7)	14.4	(3.1)
	FTND	3.1	(2.6)	3.0	(2.4)
	Baseline CIWA	1.3	(1.7)	1.3	(1.5)
	Baseline Avg. SDUs	12.5	(8.9)	14.3	(9.3)
	Baseline DPDD	13.6	(9.0)	15.4	(9.6)
	Baseline % Days Abstained	0.1	(0.1)	0.1	(0.1)
	Baseline % HDD	0.9	(0.2)	0.9	(0.2)
	Baseline % VHDD	0.6	(0.4)	0.7	(0.4)
	Baseline PACS	16.7	(6.8)	17.7	(6.2)
LTDH	25.7	(12.6)	27.3	(11.8)	
CATEGORICAL VARIABLES	N		N		
		(%)		(%)	
	Gender				
	Female	32	(31.7)	25	(26.0)
	Male	69	(68.3)	71	(74.0)
	Employment status				
	Unemp./Ret.	20	(19.8)	24	(25.0)
	Part-time	23	(22.8)	17	(17.7)
	Full-time	58	(57.4)	55	(57.3)
	Marital status				
	With partner*	43	(42.6)	46	(47.9)
	Without partner**	58	(57.4)	50	(52.1)
	Race				
	Asian or Other	1	(1.0)	7	(7.3)
	Black	27	(26.7)	30	(31.3)
	White	73	(72.3)	59	(61.5)
	Ethnicity				
	Hispanic/Latino	2	(2.0)	2	(2.1)
	Not Hispanic/Latino	99	(98.0)	94	(97.9)
	Current smoker				
No	60	(59.4)	59	(61.5)	
Yes	41	(40.6)	37	(38.5)	
Goal: Abstinence [#]					
No	73	(72.3)	69	(71.9)	
Yes	28	(27.7)	27	(28.1)	
Family hist. ^{###}					
No	34	(33.7)	27	(28.1)	
Yes	67	(66.3)	69	(71.9)	

FTND: Fagerstrom Test for Nicotine Dependence score; CIWA: Clinical Institute Withdrawal Assessment of Alcohol score; SDU: Standard Drink Unit; DPDD: Drinks Per Drinking Day; HDD: Heavy Drinking Day; VHDD: Very Heavy Drinking Day; PACS: Penn Alcohol Craving Scale score; LTDH: Lifetime Drinking History (years)

*With partner includes legally married and living with partner/cohabiting

**Without partner includes divorced, never married, separated, and widowed.

[#]Goal:Abstinence: subject's indicated alcohol-related goal from the study was abstinence

^{###}Family hist.: subject indicated history of alcohol problems with parent, sibling, or child

4.4.2 Case study results

Figure 4.2 displays, for each study arm, the 7-day moving average of standard drink units (SDUs²) reported by study subjects for the time period ranging from 90 days pre-study through the full 91-day (13 week) study period. It is evident that, on average across all subjects in the respective groups, substantial reductions in drinking occurred during the study period. The fact that the decrease is apparent for the placebo arm as well as the treatment arm is of primary interest and highlights the question of how much of the change is actual treatment effect from varenicline versus placebo effect driven by, for example, heightened attention to consumption quantities or enhanced commitment to drinking reduction goals, both consequences merely of being in the study.

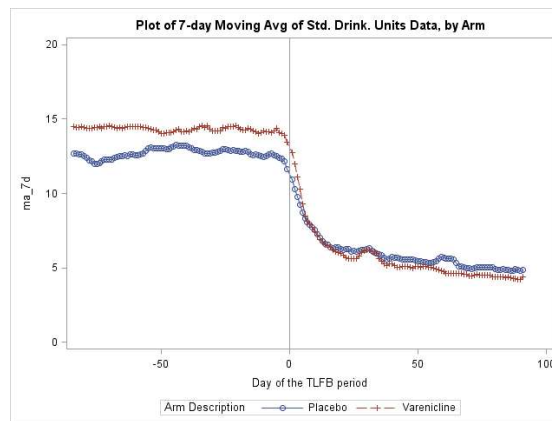


Figure 4.2: Case study outcome data, pre-study and 13-week study period

In the current analysis of the case study data, the outcome measure was the change in average SDUs per day, comparing between the 28-day pre-study period and the three-month post-titration study period. Figure 4.2 suggests a somewhat greater magnitude drop

² The SDU metric standardizes the measurement of alcohol serving across various types of drinks, such as beer, wine, and liquor.

in average SDUs for the treatment group. Indeed, a simple comparison of the difference in the before-after change in average SDUs between the placebo arm and the treatment arm shows a statistically significant difference (estimated group difference = -2.31 SDUs, t-statistic = -2.01, p-value [one-sided test with 95 d.f.] = 0.024).

While this result is of interest, it remains that the magnitude of the treatment effect is fairly modest. An additional research question is whether there exists any subgroup of the study population for which the treatment effect is substantially different from the overall ATE. Thus we applied the method described in Section 4.2 to the case study data in order to identify any such subgroups. For the probability of treatment group membership (step (i) in Section 4.2), we used 0.5, since the data were from a randomized controlled trial and the balance of sample size and covariates across treatment arms was generally quite good. For step (ii), we again used generalized boosted modeling, as in the simulation study. In estimating the g -function (i.e., those covariates that interact with treatment in affecting the outcome, step (iii)), we employed a design matrix that included covariate main effects and 2-way covariate interactions and used elastic net for selecting predictor variables. Two subgroups were identified via the method, one involving the interaction of two binary covariates, the other a single categorical covariate. After the covariates were selected into the g -function, we estimated the treatment effects for the defined subgroups using an ordinary linear regression model with only the treatment arm indicator and the selected covariates.

Table 4.4 summarizes the two groups and the estimated ATE for the groups and their complements, illustrating the differing treatment effects of the group membership. The first identified subgroup was defined based on two binary covariates. In pre-study

screening, patients were interviewed as to their personal goals with respect to alcohol consumption from participation in the study. The first covariate was a binary indicator coding whether the subject stated their goal was to “achieve total abstinence” (N=56, 28.4%) or something else (N=141, 71.6%). The second covariate was a binary indicator coding whether the subject indicated in the pre-study questionnaire that they had had any family member who had a history of alcohol problems (N=136, 69%). This first subgroup, then, was defined as those subjects who *both* had a goal of total abstinence *and* had a family member with a history of alcohol problems (N=48, 24.4%). The estimated treatment effect for members of this subgroup was a before-after drop in SDUs that was 5.8 SDUs greater than subjects not in the group (see Table 4.4; avg. SDU difference of -15.3 vs. -8.6 for treatment and control arms, respectively, for members of the subgroup; avg. SDU difference of -7.0 vs. -6.1 for treatment and control arms, respectively, for subjects not in the subgroup).

Table 4.4: Subgroups identified in varenicline case study

GROUP	# (%) obs in grp	ARM=Placebo Mean (s.d.)	ARM=Varenicline Mean (s.d.)	ATE
A	48 (24%)	-8.6 (4.7)	-15.3 (12.3)	-6.7
A ^c	149 (76%)	-6.1 (8.5)	-7.0 (5.3)	-0.9
B	25 (13%)	-16.3 (15.0)	-9.7 (5.4)	6.6
B ^c	172 (87%)	-5.7 (5.7)	-8.9 (8.7)	-3.2

Notes: Superscript ^c indicates the complement of the group.

Group definitions:

A: Goal_abstain=1 and Fam_alc_hist=1

B: Household income <= \$15,000

The second identified subgroup was defined as those whose household income was less than or equal to \$15,000 per year. In this case, members of the subgroup realized less

benefit from the study drug: the estimated treatment effect in the subgroup was +6.6 SDUs, while for those not in the subgroup it was -3.2 SDUs.

4.5 Conclusions

This chapter presents a method for identifying covariates that interact with a binary treatment to affect the outcome, hence characterizing subgroups of a study population that have differing average treatment effects. The method is capable of handling a large number of covariates, due to the complexity control exerted by way of penalized regression. Also, the method is capable of handling both experimental data and observational data; a simple change from using an assumed known probability of treatment group membership to using a modeled probability is all that is required. In a simulation study, it performed strongly in correctly deciding whether to assign particular subjects to treatment or control, based on their covariates.

This work does have some limitations. So far, it applies only to point-in-time treatments. Extending this work to dynamic treatment regimens would be beneficial, particularly given that individualized treatment regimes are receiving heightened attention in today's push toward personalized medicine. Also, the method should be studied for its effectiveness with binary or categorical outcomes, as the work to date has focused on continuous outcome measures. The characterization of the subgroups resulting from the selected covariates is not automatic, but a separate step following the method. Nevertheless, this approach appears flexible and powerful for selecting predictor variables in order to define subgroups in preparation for a future confirmatory study.

REFERENCES

- Abdia, Y., Kulasekera, K. B., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, *59*(5), 967-985. doi:10.1002/bimj.201600094
- Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*, *26*(16), 3078-3094. doi:10.1002/sim.2781
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*, *46*(3), 399-424. doi:10.1080/00273171.2011.568786
- Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*, *26*(4), 734-753.
- Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*, *34*(28), 3661-3679. doi:10.1002/sim.6607

- Biancari, F., Mikkola, R., Heikkinen, J., & al., e. (2012). Estimating the risk of complications related to re-exploration for bleeding after adult cardiac surgery: A systematic review and meta-analysis. *European Journal of Cardio-Thoracic Surgery (Amsterdam)*, *41*, 50-55.
- Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
- Brinkley, J., Tsiatis, A., & Anstrom, K. J. (2010). A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics*, *66*(2), 512-522.
- Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *Am J Epidemiol*, *163*(12), 1149-1156. doi:10.1093/aje/kwj149
- Cornelissen, H., & Arrowsmith, J. (2006). Preoperative assessment for cardiac surgery. *Continuing Education in Anaesthesia, Critical Care & Pain*, *3*, 109-113.
- Craycroft, J. A., Huang, J., & Kong, M. (2020). Propensity score specification for optimal estimation of average treatment effect with binary response. *Stat Methods Med Res*, *29*(12), 3623-3640.
- Foster, J. C., Taylor, J. M., Kaciroti, N., & Nan, B. (2015). Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. *Biostatistics*, *16*(2), 368-382.

- Franklin, J. M., Eddings, W., Glynn, R. J., & Schneeweiss, S. (2015). Regularized regression versus the high-dimensional propensity score for confounding adjustment in secondary database analyses. *Am J Epidemiol*, *182*(7), 651-659.
- Fu, H., Zhou, J., & Faries, D. E. (2016). Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. *Stat Med*, *35*(19), 3285-3302.
- Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., & Aldridge, M. D. (2014). Methods for constructing and assessing propensity scores. *Health services research*, *49*(5), 1701-1720.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.): CRC press.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 315-331.
- Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*(2), 481-488.
- Hastie, T., & Qian, J. (2014). Glmnet vignette, 1-30. Retrieved from http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf
- Helmreich, J. E., & Pruzek, R. M. (2009). PSAgraphics: An R package to support propensity score analysis. *Journal of Statistical Software*, *29*(6), 1-23.

- Hernán, M., & Robins, J. (2020). *Causal inference: what if*. Boca Raton: Chapman & Hill/CRC.
- Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161-1189.
- Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945-960. doi:10.2307/2289064
- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663-685.
- Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 243-263. doi:10.1111/rssb.12027
- Keil, A. P., Daza, E. J., Engel, S. M., Buckley, J. P., & Edwards, J. K. (2018). A Bayesian approach to the g-formula. *Stat Methods Med Res*, *27*(10), 3183-3204.
- Laber, E. B., & Zhao, Y.-Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika*, *102*(3), 501-514.
- Leacy, F. P., & Stuart, E. A. (2014). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Stat Med*, *33*(20), 3488-3508. doi:10.1002/sim.6030

- Lewis, D. (1973). Causation. *The Journal of Philosophy*, 70(17), 556-567.
doi:10.2307/2025310
- Lipkovich, I., Dmitrienko, A., & B D'Agostino Sr, R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med*, 36(1), 136-196.
- Litten, R. Z., Ryan, M. L., Fertig, J. B., Falk, D. E., Johnson, B., Dunn, K. E., . . . Sarid-Segal, O. (2013). A double-blind, placebo-controlled trial assessing the efficacy of varenicline tartrate for alcohol dependence. *Journal of addiction medicine*, 7(4), 277.
- Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health*, 21, 121-145.
- Lu, W., Zhang, H. H., & Zeng, D. (2013). Variable selection for optimal treatment decision. *Stat Methods Med Res*, 22(5), 493-504.
- Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*, 23(19), 2937-2960. doi:10.1002/sim.1903
- McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med*, 32(19), 3388-3414.
doi:10.1002/sim.5753

- McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychol Methods, 9*(4), 403-425.
- McCandless, L. C., Gustafson, P., & Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Stat Med, 28*(1), 94-112.
- McCandless, L. C., Gustafson, P., & Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Stat Med, 26*(11), 2331-2347.
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Stat Med, 38*(11), 2074-2102.
- Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., & Stürmer, T. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiol Drug Saf, 20*(6), 551-559.
- Pearl, J. (2009). *Causality*: Cambridge university press.
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*: John Wiley & Sons.
- Qian, M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics, 39*(2), 1180.

- R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org>
- Ratkovic, M., Imai, K., & Fong, C. (2012). CBPS: R package for covariate balancing propensity score. Retrieved from <http://CRAN.R-project.org/package=CBPS>.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.
doi:10.1093/biomet/70.1.41
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688-701.
- Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 34-58.
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*, *26*(1), 20-36.
- Schulte, P. J., Tsiatis, A. A., Laber, E. B., & Davidian, M. (2014). Q-and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *29*(4), 640.
- Shortreed, S. M., & Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, *73*(4), 1111-1122.

- Tian, L., Alizadeh, A. A., Gentles, A. J., & 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.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*(1), 267-288.
- Vansteelandt, S., & Joffe, M. (2014). Structural nested models and G-estimation: the partially realized promise. *Statistical science*, *29*(4), 707-731.
- Wang, R., Schoenfeld, D. A., Hoepfner, B., & Evins, A. E. (2015). Detecting treatment-covariate interactions using permutation methods. *Stat Med*, *34*(12), 2035-2047.
- Yan, X., Abdia, Y., Datta, S., Kulasekera, K. B., Ugiliweneza, B., Boakye, M., & Kong, M. (2019). Estimation of average treatment effects among multiple treatment groups by using an ensemble approach. *Stat Med*, *38*(15), 2828-2846. doi:10.1002/sim.8146
- Zhu, Y., Schonbach, M., Coffman, D. L., & Williams, J. S. (2015). Variable selection for propensity score estimation via balancing covariates. *Epidemiology*, *26*(2), e14-e15.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418-1429.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320. doi:10.1111/j.1467-9868.2005.00503.x

APPENDICES

A1 Appendix 1: Proofs of theorems in Chapter 2

Proof of Theorem 1 for exchangeability

- i. $T \perp (Y^{(0)}, Y^{(1)})|X_C$, since X_C blocks the backdoor path from T to Y .
- ii. $T \perp (Y^{(0)}, Y^{(1)})|(X_I, X_C)$ can be obtained from the following:

$$\begin{aligned}
 f(Y^{(0)}, Y^{(1)}, T|X_I, X_C) &= \frac{f(Y^{(0)}, Y^{(1)}, T, X_I, X_C)}{f(X_I, X_C)} \\
 &= \frac{f(Y^{(0)}, Y^{(1)}|X_C)f(T, X_I|X_C)f(X_C)}{f(X_I, X_C)} && \text{by } (X_I, T) \perp (Y^{(0)}, Y^{(1)})|X_C \\
 &= \frac{f(Y^{(0)}, Y^{(1)}|X_I, X_C)f(T|X_I, X_C)f(X_I|X_C)f(X_C)}{f(X_I, X_C)} && \text{by } X_I \perp (Y^{(0)}, Y^{(1)})|X_C \\
 &= f(Y^{(0)}, Y^{(1)}|X_I, X_C)f(T|X_I, X_C)
 \end{aligned}$$

- iii. $T \perp (Y^{(0)}, Y^{(1)})|(X_C, X_P)$ can be obtained from the following:

$$\begin{aligned}
 &f(Y^{(0)}, Y^{(1)}, T|X_C, X_P) \\
 &= \frac{f(Y^{(0)}, Y^{(1)}, T, X_C, X_P)}{f(X_C, X_P)} \\
 &= \frac{f(X_P, Y^{(0)}, Y^{(1)}|X_C)f(T|X_C)f(X_C)}{f(X_C, X_P)} && \text{by } T \perp (X_P, Y^{(0)}, Y^{(1)})|X_C
 \end{aligned}$$

$$\begin{aligned}
&= \frac{f(Y^{(0)}, Y^{(1)} | X_C, X_P) f(X_P | X_C) f(T | X_C, X_P) f(X_C)}{f(X_C, X_P)} && \text{by } T \perp X_P | X_C \\
&= f(Y^{(0)}, Y^{(1)} | X_C, X_P) f(T | X_C, X_P)
\end{aligned}$$

iv. $T \perp (Y^{(0)}, Y^{(1)}) | (X_I, X_C, X_P)$ can be obtained from the following:

$$\begin{aligned}
&f(Y^{(0)}, Y^{(1)}, T | X_I, X_C, X_P) \\
&= \frac{f(Y^{(0)}, Y^{(1)}, T, X_I, X_C, X_P)}{f(X_I, X_C, X_P)} \\
&= \frac{f(X_P, Y^{(0)}, Y^{(1)} | X_C) f(X_I, T | X_C) f(X_C)}{f(X_I, X_C, X_P)} && \text{by } (X_I, T) \perp (X_P, Y^{(0)}, Y^{(1)}) | X_C \\
&= \frac{f(Y^{(0)}, Y^{(1)} | X_C, X_P) f(X_P | X_C) f(T | X_C, X_I) f(X_I | X_C) f(X_C)}{f(X_I, X_C, X_P)} \\
&= \frac{f(Y^{(0)}, Y^{(1)} | X_I, X_C, X_P) f(X_P | X_C) f(T | X_I, X_C, X_P) f(X_I | X_C) f(X_C)}{f(X_I, X_C, X_P)} \\
&= f(Y^{(0)}, Y^{(1)} | X_I, X_C, X_P) f(T | X_I, X_C, X_P)
\end{aligned}$$

The second-to-last equation is due to $X_I \perp (Y^{(0)}, Y^{(1)}) | X_C, X_P = f(Y^{(0)}, Y^{(1)} | X_I, X_C, X_P) f(T | X_I, X_C, X_P)$.

Proof of Theorem 1 for unbiasedness of ATE estimators:

Let $X^{(*)}$ represent any one of the adjustment sets (i) X_C ; (ii) X_C, X_I ; (iii) X_C, X_P ; or (iv) X_C, X_I, X_P . Under the assumptions of exchangeability (i.e., $T \perp (Y^{(0)}, Y^{(1)}) | X^{(*)}$) and positivity (i.e., $0 < P(Y = 1 | X^{(*)}) = p(X^{(*)}) < 1$), we claim that the IPW estimator for the ATE is unbiased. To prove it, let us denote the IPW estimator for the ATE as

$$\hat{\tau}^{(*)} = \frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i}{p(\mathbf{X}_i^{(*)})} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - T_i) Y_i}{1 - p(\mathbf{X}_i^{(*)})}.$$

If exchangeability and positivity hold for a set of adjustment variables $\mathbf{X}^{(*)}$, then we have

$$\begin{aligned} E\left(\frac{TY}{p(\mathbf{X}^{(*)})}\right) &= E\left[E\left(\frac{I_{\{T=1\}}Y^{(1)}}{p(\mathbf{X}^{(*)})} \mid \mathbf{X}^{(*)}\right)\right] = E\left[\frac{1}{p(\mathbf{X}^{(*)})} E(I_{\{T=1\}}Y^{(1)} \mid \mathbf{X}^{(*)})\right] \\ &= E\left[\frac{1}{p(\mathbf{X}^{(*)})} E(I_{\{T=1\}} \mid \mathbf{X}^{(*)}) E(Y^{(1)} \mid \mathbf{X}^{(*)})\right] \quad \text{by exchangeability} \\ &= E\left[\frac{1}{p(\mathbf{X}^{(*)})} p(\mathbf{X}^{(*)}) E(Y^{(1)} \mid \mathbf{X}^{(*)})\right] = E[E(Y^{(1)} \mid \mathbf{X}^{(*)})] = E(Y^{(1)}) \end{aligned}$$

Similarly, under exchangeability, we have $E\left(\frac{(1-T)Y}{1-p(\mathbf{X}^{(*)})}\right) = E(Y^{(0)})$.

Thus $E[\hat{\tau}^{(*)}] = E(Y^{(1)}) - E(Y^{(0)}) = \tau$.

Proof of Theorem 2(i): Note that under Proposition 1, $X_I \perp (Y^{(0)}, Y^{(1)}) \mid (X_C, X_P)$, so we have $E\{Y^{(0)} \mid X_I, X_C, X_P\} = E\{Y^{(0)} \mid X_C, X_P\}$, $E\{Y^{(1)} \mid X_I, X_C, X_P\} = E\{Y^{(1)} \mid X_C, X_P\}$, $\sigma_1^2(X_I, X_C, X_P) = \sigma_1^2(X_C, X_P)$, and $\sigma_0^2(X_I, X_C, X_P) = \sigma_0^2(X_C, X_P)$. The first two equations imply that $\tau(X_C, X_P) = \tau(X_I, X_C, X_P)$. Thus,

$$\begin{aligned} E\left[\frac{\sigma_1^2(X_I, X_C, X_P)}{p(X_I, X_C, X_P)}\right] &= E\left\{E\left[\frac{\sigma_1^2(X_I, X_C, X_P)}{p(X_I, X_C, X_P)} \mid X_C, X_P\right]\right\} \\ &= E\left\{\sigma_1^2(X_C, X_P) E\left[\frac{1}{p(X_I, X_C, X_P)} \mid X_C, X_P\right]\right\} \\ &\geq E\left\{\sigma_1^2(X_C, X_P) \left[\frac{1}{E(p(X_I, X_C, X_P) \mid (X_C, X_P))}\right]\right\} \quad \text{(by using Jensen's Inequality)} \end{aligned}$$

$$= E \left[\frac{\sigma_1^2(X_C, X_P)}{p(X_C, X_P)} \right].$$

Similarly, we have

$$E \left[\frac{\sigma_0^2(X_I, X_C, X_P)}{1 - p(X_I, X_C, X_P)} \right] \geq E \left[\frac{\sigma_0^2(X_C, X_P)}{1 - p(X_C, X_P)} \right].$$

In addition, we have

$$E[(\tau(X_C, X_P) - \tau)^2] = E[(\tau(X_I, X_C, X_P) - \tau)^2].$$

Thus, we have proved the first inequality in (i). For the second inequality in (i), note that $p(X_I, X_C, X_P) = p(X_I, X_C)$. Thus,

$$\begin{aligned} E \left[\frac{\sigma_1^2(X_I, X_C, X_P)}{p(X_I, X_C, X_P)} \right] &= E \left\{ E \left[\frac{E((Y^{(1)})^2 | X_I, X_C, X_P) - (E(Y^{(1)} | X_I, X_C, X_P))^2}{p(X_I, X_C, X_P)} \middle| X_I, X_C \right] \right\} \\ &= E \left\{ \frac{E \{ (Y^{(1)})^2 | X_I, X_C \} - E \{ (E(Y^{(1)} | X_I, X_C, X_P))^2 | X_I, X_C \}}{p(X_I, X_C)} \right\} \\ &\leq E \left\{ \frac{E \{ (Y^{(1)})^2 | X_I, X_C \} - \{ E[Y^{(1)} | X_I, X_C] \}^2}{p(X_I, X_C)} \right\} = E \left[\frac{\sigma_1^2(X_I, X_C)}{p(X_I, X_C)} \right]. \end{aligned}$$

The last inequality is from using Jensen's Inequality: $E \left\{ (E(Y^{(1)} | X_I, X_C, X_P))^2 \middle| X_I, X_C \right\} \geq \{E[E(Y^{(1)} | X_I, X_C, X_P) | X_I, X_C]\}^2 = \{E[Y^{(1)} | X_I, X_C]\}^2$.

Similarly, we can prove that $E \left[\frac{\sigma_0^2(X_I, X_C, X_P)}{1 - p(X_I, X_C, X_P)} \right] \leq E \left[\frac{\sigma_0^2(X_I, X_C)}{1 - p(X_I, X_C)} \right]$. In addition, due to omitting the variable X_P in the expectation of outcome, we have

$$E[(\tau(X_I, X_C, X_P) - \tau)^2] \leq E[(\tau(X_I, X_C) - \tau)^2].$$

Thus, the second inequality holds.

Proof of Theorem 2(ii): The first inequality is straightforward, since $p(X_C, X_P) = P(T = 1|X_C, X_P) = P(T = 1|X_C) = p(X_C)$. We also expect $E\sigma_1^2(X_C, X_P) \leq E\sigma_1^2(X_C)$, $\sigma_0^2(X_C, X_P) \leq E\sigma_0^2(X_C)$, and

$$E[(\tau(X_C, X_P) - \tau)^2] \leq E[(\tau(X_C) - \tau)^2].$$

The second inequality in (ii) holds due to the following argument:

$$\begin{aligned} E\left[\frac{\sigma_1^2(X_I, X_C)}{p(X_I, X_C)}\right] &= E\left\{E\left[\frac{E((Y^{(1)})^2|X_I, X_C) - (E(Y^{(1)}|X_I, X_C))^2}{p(X_I, X_C)} \middle| X_C\right]\right\} \\ &= E\left\{E\left[\frac{E((Y^{(1)})^2|X_C) - (E(Y^{(1)}|X_C))^2}{p(X_I, X_C)} \middle| X_C\right]\right\} \\ &= E\left\{\left(E((Y^{(1)})^2|X_C) - (E(Y^{(1)}|X_C))^2\right)E\left(\frac{1}{p(X_I, X_C)} \middle| X_C\right)\right\} \\ &\geq E\left\{\left(E((Y^{(1)})^2|X_C) - (E(Y^{(1)}|X_C))^2\right)\frac{1}{E\{p(X_I, X_C)|X_C\}}\right\} \\ &= E\left[\frac{\sigma_1^2(X_C)}{p(X_C)}\right]. \end{aligned}$$

Similarly, one can show that $E\left[\frac{\sigma_0^2(X_I, X_C)}{1-p(X_I, X_C)}\right] \geq E\left[\frac{\sigma_0^2(X_C)}{1-p(X_C)}\right]$, and $E(\tau(X_I, X_C) - \tau)^2 = E(\tau(X_C) - \tau)^2$. Thus, we complete the proof of Theorem 2.

A2 Appendix 2: Boxplots (all scenarios) and tables (correlated scenarios) of simulation study results

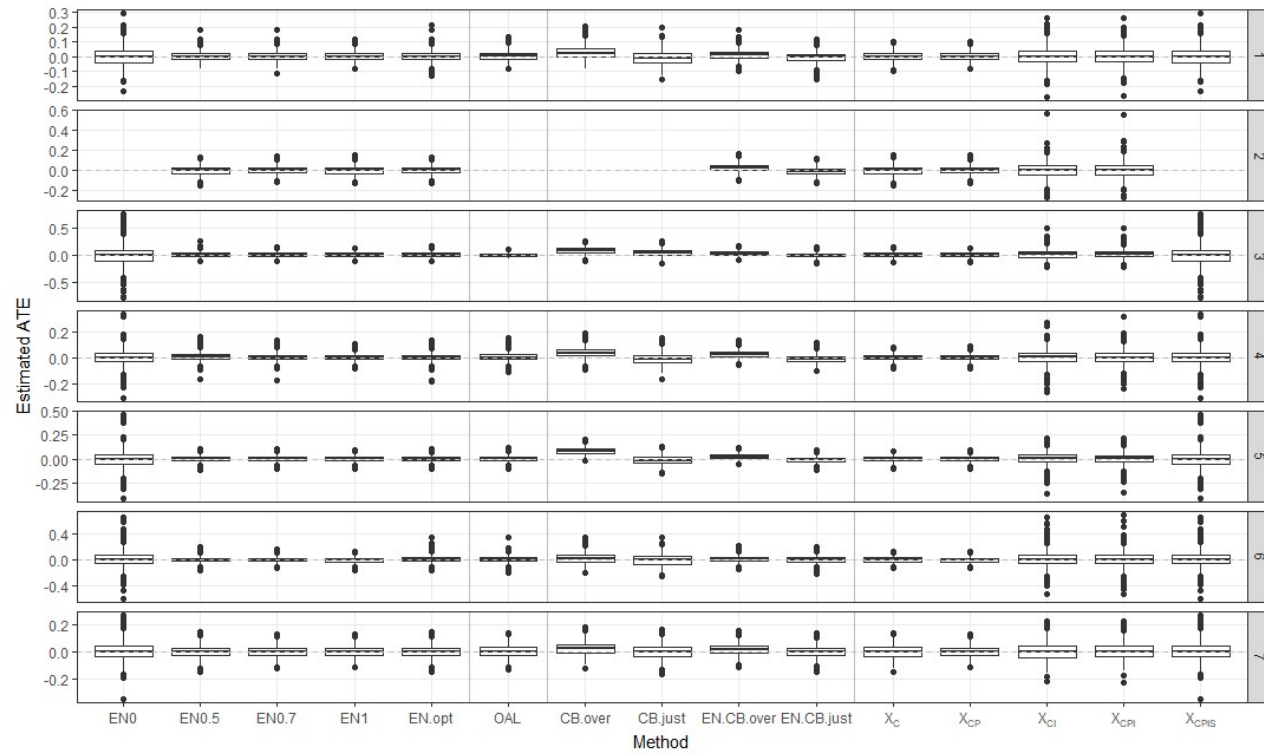
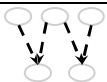
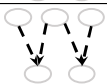
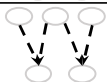


Figure A2.1. Box plots of ATE estimates in simulation study
All covariates independent (cf. Table 2.2 in Chapter 2)

Table A2.1. Bias, Standard Error, and Root MSE for Correlated Simulation Scenario
 $\rho_C = \rho_I = \rho_P = 0.2$

#	SCENARIO		TESTED MODELS										REFERENCE MODELS				
			EN0	EN0.5	EN0.7	EN1	EN.opt	OAL	CB.over	CB.just	EN.CB.o	EN.CB.j	X _C	X _C X _P	X _C X _I	X _C X _P X _I	X _C X _P X _I X _S
8	Model A: 	Bias	0.001	0.003	0.002	0.001	0.003	0.003	0.029	-0.007	0.016	-0.006	0.002	0.001	0.004	0.003	0.001
		SE	0.065	0.032	0.030	0.029	0.034	0.034	0.044	0.048	0.031	0.033	0.032	0.029	0.065	0.063	0.065
		RMSE	0.065	0.033	0.030	0.029	0.034	0.034	0.052	0.049	0.035	0.034	0.032	0.029	0.065	0.063	0.065
9	Model A: 	Bias	*	0.005	0.005	0.005	0.004	*	*	*	0.029	-0.006	0.004	0.004	0.010	0.010	*
		SE	*	0.043	0.042	0.041	0.042	*	*	*	0.040	0.042	0.047	0.042	0.089	0.086	*
		RMSE	*	0.043	0.042	0.042	0.043	*	*	*	0.049	0.042	0.047	0.042	0.090	0.087	*
10	Model A: 	Bias	-0.009	0.004	0.004	0.003	0.004	0.016	0.085	0.054	0.025	-0.006	0.003	0.003	0.004	0.005	-0.009
		SE	0.206	0.044	0.043	0.043	0.043	0.049	0.064	0.064	0.041	0.044	0.047	0.042	0.084	0.082	0.206
		RMSE	0.206	0.045	0.043	0.043	0.043	0.052	0.106	0.083	0.049	0.044	0.047	0.042	0.084	0.082	0.206

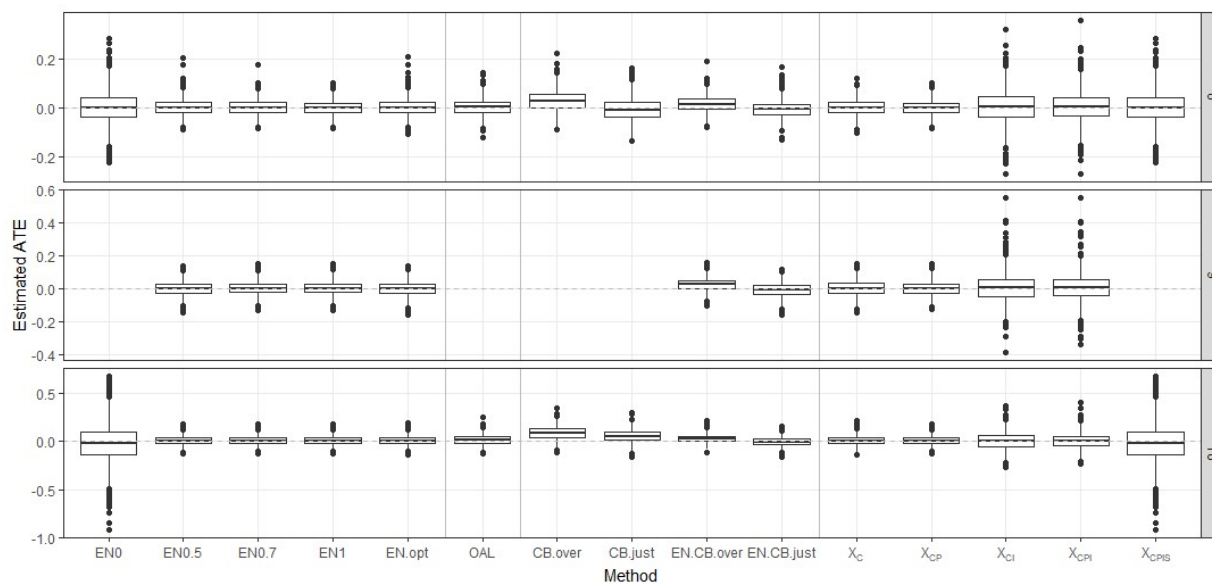


Figure A2.2. Box plots of ATE estimates in simulation study
Moderate correlation: $\rho_C = \rho_I = \rho_P = 0.2$

Table A2.2. Bias, Standard Error, and Root MSE for Correlated Simulation Scenario

$$\rho_C = \rho_I = \rho_P = 0.5$$

#	SCENARIO		TESTED MODELS										REFERENCE MODELS				
			ENO	ENO.5	ENO.7	EN1	EN.opt	OAL	CB.over	CB.just	EN.CB.o	EN.CB.j	X_C	$X_C X_P$	$X_C X_I$	$X_C X_P X_I$	$X_C X_P X_I X_S$
11	Model A: 	Bias	0.009	0.007	0.006	0.006	0.007	0.009	0.036	-0.006	0.022	-0.002	0.007	0.006	0.011	0.010	0.009
		SE	0.077	0.035	0.032	0.030	0.038	0.035	0.049	0.052	0.033	0.033	0.034	0.029	0.075	0.073	0.077
		RMSE	0.078	0.035	0.033	0.031	0.039	0.036	0.061	0.052	0.039	0.033	0.034	0.030	0.076	0.074	0.078
12	Model A: 	Bias	*	0.005	0.006	0.005	0.005	*	*	*	0.031	-0.006	0.006	0.005	0.016	0.015	*
		SE	*	0.045	0.044	0.043	0.044	*	*	*	0.043	0.044	0.049	0.043	0.098	0.095	*
		RMSE	*	0.046	0.045	0.044	0.045	*	*	*	0.053	0.045	0.049	0.043	0.099	0.097	*
13	Model A: 	Bias	-0.008	0.005	0.005	0.004	0.005	0.017	0.102	0.069	0.028	-0.007	0.002	0.003	0.015	0.016	-0.008
		SE	0.217	0.045	0.044	0.041	0.045	0.049	0.065	0.062	0.041	0.044	0.046	0.041	0.094	0.093	0.217
		RMSE	0.217	0.045	0.045	0.042	0.046	0.052	0.121	0.093	0.050	0.044	0.046	0.041	0.095	0.095	0.217

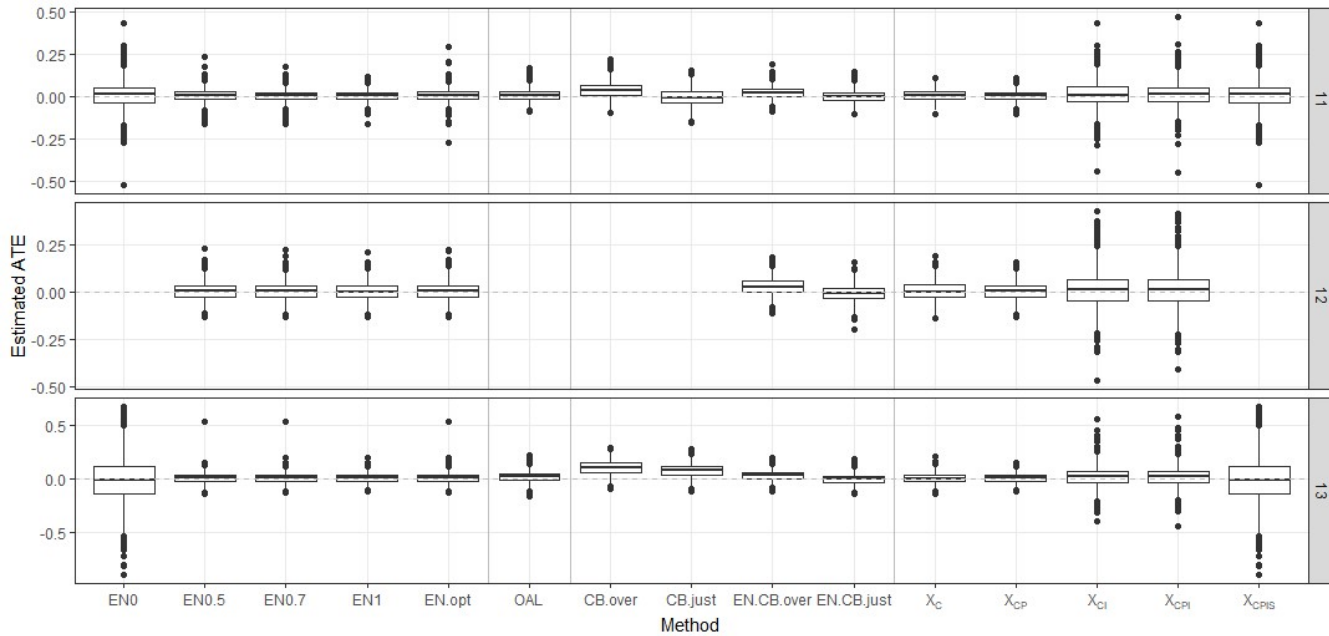


Figure A2.3. Box plots of ATE estimates in simulation study

Strong correlation: $\rho_I = \rho_C = \rho_P = 0.5$

A3 Appendix 3: Derivation of posterior conditional distributions, Chapter 3

This portion specifies the model, all prior distributions, the full joint posterior distribution, and the conditional posterior distribution for each parameter.

Model:

$$Y_i|T_i \sim \begin{cases} \text{Normal}(\theta + \tau, p_i\sigma^2), & \text{if } T_i = 1 \\ \text{Normal}(\theta, (1 - p_i)\sigma^2), & \text{if } T_i = 0 \end{cases}$$

$$T_i|p_i \sim \text{Bernoulli}(p_i)$$

$$\tau \sim \text{Normal}(0, \sigma_\tau^2)$$

$$\theta \sim \text{Normal}(0, \sigma_\theta^2)$$

$$\sigma^2 \sim \text{Inv. Gamma}(a, b)$$

$$p_i|\alpha_i, \lambda \sim \text{Beta}(\lambda\alpha_i, \lambda)$$

$$\alpha_i|\beta, v^2 \sim \text{Lognormal}(x_i^T \beta, v^2)$$

$$\beta \sim \text{Multivar. Normal}(0, \Sigma_\beta)$$

$$v^2 \sim \text{Inv. Gamma}(c, d)$$

$$\lambda \sim \text{Gamma}(f, g)$$

Now we detail the full joint posterior distribution for the data and all of the parameters:

$$\text{Data} = \mathcal{D} = \{y, x, t\}$$

$$\text{Parameters} = \theta = \{\tau, \theta, \sigma^2, p, \alpha, \beta, \nu^2, \lambda\}$$

$$\text{Fixed settings/hyperparameters} = \mathcal{P} = \{\sigma_\tau^2, \sigma_\theta^2, \Sigma_\beta, a, b, c, d, f, g\}$$

$$\text{Full joint posterior: } f(y | t, \theta, \mathcal{D})f(t | \theta, \mathcal{D})\pi(\theta)$$

$$f(y | T, \theta, \mathcal{D}) = \frac{\exp\left\{-\sum_{i=1}^n \left[\frac{T_i(y_i - \tau - \theta)^2}{2p_i\sigma^2} + \frac{(1 - T_i)(y_i - \theta)^2}{2(1 - p_i)\sigma^2} \right]\right\}}{\sigma^n (\sqrt{2\pi})^n \prod_{i=1}^n (1 - p_i)^{\frac{1 - T_i}{2}} p_i^{\frac{T_i}{2}}}$$

$$f(T | p) = \prod_{i=1}^n p_i^{T_i} (1 - p_i)^{(1 - T_i)}$$

$$\pi(\tau) = \frac{1}{\sqrt{2\pi\sigma_\tau^2}} \exp\left\{-\frac{\tau^2}{2\sigma_\tau^2}\right\}$$

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} \exp\left\{-\frac{\theta^2}{2\sigma_\theta^2}\right\}$$

$$\pi(\sigma^2) = \frac{b^a}{\Gamma(a)} (\sigma^2)^{-a-1} \exp\left\{-\frac{b}{\sigma^2}\right\}$$

$$\pi(p | \alpha, \lambda) = \prod_{i=1}^n \left[\frac{\Gamma(\lambda\alpha_i + \lambda)}{\Gamma(\lambda\alpha_i)\Gamma(\lambda)} \right] p_i^{(\lambda\alpha_i - 1)} (1 - p_i)^{(\lambda - 1)}$$

$$\pi(\alpha | \beta, \nu^2) = \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi\nu^2}} (\alpha_i)^{-1} \right] \exp\left\{-\frac{(\log\alpha_i - x_i^T \beta)^2}{2\nu^2}\right\}$$

$$\pi(\beta) = \left(\frac{1}{\sqrt{2\pi}}\right)^{(k+1)} \Sigma_{\beta}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\beta^T \Sigma_{\beta}^{-1}\beta\right\}$$

$$\pi(v^2) = \frac{d^c}{\Gamma(c)} (v^2)^{-c-1} \exp\left\{-\frac{d}{v^2}\right\}$$

$$\pi(\lambda) = \frac{1}{\Gamma(f)g^f} \lambda^{f-1} \exp\left\{-\frac{\lambda}{g}\right\}$$

Now we detail the conditional posterior distributions for each of the parameters:

$\pi(\tau | \cdot) \propto \text{Normal}(M_{\tau.post}, V_{\tau.post})$, where

$$V_{\tau.post} = \left(\frac{1}{\sigma_\tau^2} + \sum \frac{T_i}{p_i \sigma^2} \right)^{-1} \text{ and}$$

$$M_{\tau.post} = V_{\tau.post} \left(\sum \frac{y_i - \theta}{p_i \sigma^2} T_i \right).$$

$\pi(\theta | \cdot) \propto \text{Normal}(M_{\theta.post}, V_{\theta.post})$, where

$$V_{\theta.post} = \left(\frac{1}{\sigma_\theta^2} + \sum \frac{T_i}{p_i \sigma^2} + \frac{(1 - T_i)}{(1 - p_i) \sigma^2} \right)^{-1} \text{ and}$$

$$M_{\theta.post} = V_{\theta.post} \left(\sum \frac{T_i(y_i - \tau)}{p_i \sigma^2} + \frac{(1 - T_i)y_i}{(1 - p_i) \sigma^2} \right).$$

$$\pi(\sigma^2 | \cdot) \propto \text{Inv. Gamma} \left(\frac{n}{2} + a, b + \sum_{i=1}^n \left[\frac{T_i(y_i - \tau - \theta)^2}{2p_i} + \frac{(1 - T_i)(y_i - \theta)^2}{2(1 - p_i)} \right] \right).$$

$$\pi(p_i | \cdot) \propto p_i^{\left(\frac{T_i}{2} + \lambda \alpha_i - 1\right)} (1 - p_i)^{\left(\frac{(1 - T_i)}{2} + \lambda - 1\right)} \exp \left\{ - \left[\frac{T_i(y_i - \tau - \theta)^2}{2\sigma^2 p_i} + \frac{(1 - T_i)(y_i - \theta)^2}{2\sigma^2 (1 - p_i)} \right] \right\}.$$

$$\pi(\alpha_i | \cdot) \propto \frac{\Gamma(\lambda \alpha_i + \lambda)}{\Gamma(\lambda \alpha_i) \Gamma(\lambda)} p_i^{(\lambda \alpha_i - 1)} \alpha_i^{-1} \exp \left\{ - \frac{(\log \alpha_i - x_i^T \beta)^2}{2\nu^2} \right\}.$$

$$\begin{aligned}\pi(\beta | \cdot) &\propto \text{MVN}(\hat{\beta}, V_{\beta, \text{post}}), \text{ where} \\ V_{\beta, \text{post}} &= [X^T V_{\alpha}^{-1} X + \Sigma_{\beta}^{-1}]^{-1}, \\ V_{\alpha} &= v^2 I_n, \text{ and} \\ \hat{\beta} &= V_{\beta, \text{post}} (X^T V_{\alpha}^{-1} \alpha).\end{aligned}$$

$$\pi(v^2 | \cdot) \propto \text{Inv. Gamma} \left(\frac{n}{2} + c, d + \sum_{i=1}^n \left[\frac{(\log \alpha_i - x_i^T \beta)^2}{2} \right] \right).$$

$$\pi(\lambda | \cdot) \propto \prod_{i=1}^n \left[\frac{\Gamma(\lambda \alpha_i + \lambda)}{\Gamma(\lambda \alpha_i) \Gamma(\lambda)} p_i^{(\lambda \alpha_i - 1)} (1 - p_i)^{(\lambda - 1)} \right] \lambda^{f-1} \exp \left\{ -\frac{\lambda}{g} \right\}.$$

The posterior conditional distributions for five of these parameters (τ , θ , σ^2 , β , and v^2) are known distributions, and hence we may obtain posterior samples via Gibbs sampling. The other three parameters (p , α , and λ) are complex, unknown distributions. Hence we will use the Metropolis-Hastings (MH) algorithm to obtain posterior condition samples for these. For computational purposes, we will use the log-posteriors, shown on the next page.

For $\pi(p_i | \cdot)$, we have

$$\log[\pi(p_i | \cdot)] \propto \left(\frac{T_i}{2} + \lambda\alpha_i - 1\right) \log p_i + \left[\frac{(1 - T_i)}{2} + \lambda - 1\right] \log(1 - p_i) - \frac{T_i(y_i - \tau - \theta)^2}{2\sigma^2 p_i} - \frac{(1 - T_i)(y_i - \theta)^2}{2\sigma^2(1 - p_i)}.$$

For $\pi(\alpha_i | \cdot)$, we have

$$\log[\pi(\alpha_i | \cdot)] \propto \log[\Gamma(\lambda\alpha_i + \lambda)] - \log[\Gamma(\lambda\alpha_i)] - \log[\Gamma(\lambda)] + (\lambda\alpha_i - 1)\log p_i - \log\alpha_i - \frac{(\log\alpha_i - x_i^T \beta)^2}{2\nu^2}.$$

Finally for $\pi(\lambda | \cdot)$, we have

$$\log[\pi(\lambda | \cdot)] \propto \sum_{i=1}^n \{\log[\Gamma(\lambda\alpha_i + \lambda)] - \log[\Gamma(\lambda\alpha_i)] - \log[\Gamma(\lambda)] + (\lambda\alpha_i - 1)\log p_i + (\lambda - 1)\log(1 - p_i)\} + (f - 1)\log\lambda - \frac{\lambda}{g}.$$

A4 Appendix 4: Alternative Bayesian model formulation using weighted likelihood

Here we present an alternative formulation of the Bayesian model. This approach has the appeal that it is somewhat less complex than the approach described in the main body of the article, with less parameterizations involved. This approach uses a weighted likelihood.

We assume the following distributions for the data Y_i and T_i :

$$Y_i|T_i \sim \begin{cases} \text{Normal}(\theta + \tau, \sigma^2), & \text{if } T_i = 1 \\ \text{Normal}(\theta, \sigma^2), & \text{if } T_i = 0 \end{cases}$$

$$T_i|p_i \sim \text{Bernoulli}(\text{expit}(x_i^T \beta)),$$

where $\text{expit}(\cdot)$ is the inverse logit function, i.e., $\text{expit}(a) = \frac{\exp(a)}{\exp(a)+1}$.

$$\tau \sim \text{Normal}(0, \sigma_\tau^2),$$

$$\theta \sim \text{Normal}(0, \sigma_\theta^2),$$

$$\sigma^2 \sim \text{Inv. Gamma}(a, b),$$

and

$$\beta | \nu^2 \sim \text{Multivar. Normal}(0, \nu^2 * I_p).$$

Then we express a weighted likelihood for the data as

$$L(\tau, \theta) = \frac{\exp\left\{-\sum_{i=1}^n \frac{1}{2\sigma^2} \left[\frac{T_i(y_i - \tau - \theta)^2}{p_i} + \frac{(1-T_i)(y_i - \theta)^2}{(1-p_i)} \right]\right\}}{(\sigma\sqrt{2\pi})^{\sum_{i=1}^n \frac{T_i}{p_i} + \frac{1-T_i}{1-p_i}}}.$$

A5 Appendix 5: Tables of additional simulation results, Chapter 4

Table A5.1 Subgroup analysis simulation study results (N=400)

N=400	Independent covariates					Correlated covariates					Ratios: Corr::Indpndt					Diff: Corr-Indpndt				
	PCD	MSE	Corr0	Incorr0	Exact	PCD	MSE	Corr0	Incorr0	Exact	PCD	MSE	Corr0	Incorr0	Exact	PCD	MSE	Corr0	Incorr0	Exact
	Y1.A.EN	95.76	0.102	6.89	0	0.90	92.7	0.118	6.29	0	0.48	0.97	1.16	0.91	0.00	(0.42)	0.97	1.13	0.95	0.02
Y1.A.L	95.36	0.102	6.99	0	0.99	92.4	0.115	6.64	0.02	0.70	0.97	1.13	0.95	0.02	(0.29)	0.96	1.23	0.93	0.14	(0.40)
Y1.B.EN	94.68	0.121	6.90	0	0.91	90.5	0.149	6.40	0.14	0.51	0.96	1.23	0.96	0.19	(0.35)	0.96	1.23	0.96	0.19	(0.35)
Y1.B.L	94.35	0.117	6.94	0	0.96	90.2	0.144	6.68	0.19	0.61	0.96	1.23	0.96	0.19	(0.35)	0.96	1.22	0.94	0.04	(0.35)
Y1.C.EN	95.64	0.102	6.89	0	0.92	91.6	0.124	6.46	0.04	0.57	0.96	1.22	0.96	0.06	(0.29)	0.98	1.13	0.94	0.00	(0.34)
Y1.C.L	95.19	0.100	6.99	0	0.99	91.0	0.123	6.72	0.06	0.70	0.98	1.12	0.96	0.00	(0.26)	0.98	1.12	0.96	0.00	(0.26)
Y1.D.EN	95.72	0.097	6.82	0	0.85	93.7	0.110	6.38	0	0.51	0.98	1.04	0.90	0.00	(0.42)	0.98	1.04	0.90	0.00	(0.42)
Y1.D.L	95.43	0.096	6.92	0	0.93	93.5	0.107	6.63	0	0.67	0.98	1.06	0.97	0.00	(0.19)	0.98	1.06	0.97	0.00	(0.19)
Y1.EN	96.74	0.090	6.79	0	0.83	95.1	0.093	6.09	0	0.41	0.98	1.03	0.94	0.35	(0.51)	0.98	1.03	0.94	0.35	(0.51)
Y1.L	96.48	0.088	6.91	0	0.93	94.7	0.094	6.69	0	0.74	0.98	1.03	0.94	0.35	(0.51)	0.98	1.03	0.94	0.35	(0.51)
Y2.A.EN	89.42	0.169	6.93	0.08	0.89	87.9	0.175	6.34	0.34	0.39	0.98	1.07	0.95	0.46	(0.56)	0.98	1.07	0.95	0.46	(0.56)
Y2.A.L	89.34	0.165	6.96	0.08	0.90	87.5	0.171	6.55	0.43	0.39	0.98	1.07	0.95	0.46	(0.56)	0.98	1.07	0.95	0.46	(0.56)
Y2.B.EN	85.52	0.194	6.84	0.45	0.63	82.6	0.215	6.52	0.99	0.17	0.97	1.11	0.95	0.54	(0.46)	0.97	1.11	0.95	0.54	(0.46)
Y2.B.L	84.9	0.192	6.90	0.43	0.65	84.6	0.208	6.69	1.02	0.19	1.00	1.08	0.97	0.59	(0.46)	1.00	1.08	0.97	0.59	(0.46)
Y2.C.EN	87.89	0.175	6.90	0.18	0.81	86.3	0.187	6.56	0.64	0.25	0.98	1.07	0.95	0.46	(0.56)	0.98	1.07	0.95	0.46	(0.56)
Y2.C.L	87.04	0.173	6.93	0.22	0.80	85.6	0.190	6.80	0.76	0.37	0.98	1.10	0.98	0.54	(0.43)	0.98	1.10	0.98	0.54	(0.43)
Y2.D.EN	92.52	0.142	6.82	0.01	0.83	90.5	0.152	6.47	0.12	0.50	0.98	1.07	0.95	0.11	(0.33)	0.98	1.07	0.95	0.11	(0.33)
Y2.D.L	91.9	0.138	6.92	0.03	0.89	90.0	0.150	6.73	0.14	0.63	0.98	1.08	0.97	0.11	(0.26)	0.98	1.08	0.97	0.11	(0.26)
Y2.EN	93.17	0.140	6.88	0	0.88	92.0	0.134	6.23	0.06	0.42	0.99	0.98	0.94	0.07	(0.32)	0.99	0.98	0.94	0.07	(0.32)
Y2.L	92.61	0.137	6.98	0.01	0.97	91.4	0.134	6.57	0.08	0.65	0.99	0.98	0.94	0.07	(0.32)	0.99	0.98	0.94	0.07	(0.32)
Y3.A.EN	92.39	0.144	6.85	0.02	0.85	87.9	0.173	6.36	0.41	0.30	0.95	1.20	0.93	0.39	(0.55)	0.95	1.20	0.93	0.39	(0.55)
Y3.A.L	91.53	0.143	6.90	0.01	0.91	87.4	0.172	6.67	0.51	0.37	0.95	1.20	0.97	0.50	(0.54)	0.95	1.20	0.97	0.50	(0.54)
Y3.B.EN	89.78	0.168	6.90	0.06	0.86	83.1	0.217	6.65	0.96	0.24	0.93	1.29	0.96	0.90	(0.62)	0.93	1.29	0.96	0.90	(0.62)
Y3.B.L	88.68	0.168	6.96	0.14	0.86	84.5	0.212	6.82	1.03	0.23	0.95	1.26	0.98	0.89	(0.63)	0.95	1.26	0.98	0.89	(0.63)
Y3.C.EN	89.9	0.157	6.92	0.11	0.86	85.9	0.190	6.76	0.58	0.38	0.96	1.21	0.98	0.47	(0.48)	0.96	1.21	0.98	0.47	(0.48)
Y3.C.L	88.62	0.155	6.95	0.18	0.84	84.9	0.189	6.81	0.73	0.35	0.96	1.22	0.98	0.55	(0.49)	0.96	1.22	0.98	0.55	(0.49)
Y3.D.EN	92.94	0.134	6.78	0	0.82	88.9	0.169	6.42	0.32	0.40	0.96	1.26	0.95	0.32	(0.42)	0.96	1.26	0.95	0.32	(0.42)
Y3.D.L	92.47	0.130	6.91	0.01	0.92	88.7	0.161	6.70	0.28	0.59	0.96	1.24	0.97	0.27	(0.33)	0.96	1.24	0.97	0.27	(0.33)
Y3.EN	94.25	0.127	6.86	0.01	0.87	91.6	0.137	6.21	0.06	0.43	0.97	1.08	0.91	0.05	(0.44)	0.97	1.08	0.91	0.05	(0.44)
Y3.L	93.85	0.125	6.91	0	0.92	91.2	0.134	6.55	0.08	0.65	0.97	1.07	0.95	0.08	(0.27)	0.97	1.07	0.95	0.08	(0.27)

Table A5.2 Subgroup analysis simulation study results (N=200)

N=200	Independent covariates					Correlated covariates					Ratios: Corr::Indpndt					Diff: Corr-Indpndt				
	PCD	MSE	Corr0	Incorr0	Exact	PCD	MSE	Corr0	Incorr0	Exact	PCD	MSE	Corr0	Incorr0	Exact	PCD	MSE	Corr0	Incorr0	Exact
	Y1.A.EN	88.88	0.183	6.68	0.14	0.65	87.7	0.198	5.94	0.38	0.18	0.99	1.08	0.89	0.24	(0.47)	0.99	1.08	0.89	0.24
Y1.A.L	88.73	0.181	6.82	0.17	0.74	86.7	0.199	6.27	0.55	0.26	0.98	1.10	0.92	0.38	(0.48)	0.98	1.10	0.92	0.38	(0.48)
Y1.B.EN	87.92	0.202	6.72	0.23	0.62	81.0	0.243	6.47	1.25	0.12	0.92	1.20	0.96	1.02	(0.50)	0.92	1.20	0.96	1.02	(0.50)
Y1.B.L	87.7	0.196	6.81	0.22	0.70	78.9	0.242	6.55	1.43	0.09	0.90	1.24	0.96	1.21	(0.61)	0.90	1.24	0.96	1.21	(0.61)
Y1.C.EN	89.42	0.185	6.84	0.09	0.79	84.8	0.216	6.53	0.77	0.25	0.95	1.17	0.95	0.68	(0.54)	0.95	1.17	0.95	0.68	(0.54)
Y1.C.L	88.58	0.180	6.85	0.15	0.76	84.2	0.215	6.71	0.89	0.27	0.95	1.19	0.98	0.74	(0.49)	0.95	1.19	0.98	0.74	(0.49)
Y1.D.EN	91.54	0.180	6.68	0.07	0.70	88.2	0.198	6.19	0.42	0.25	0.96	1.10	0.93	0.35	(0.45)	0.96	1.10	0.93	0.35	(0.45)
Y1.D.L	90.78	0.179	6.78	0.11	0.73	88.1	0.192	6.43	0.43	0.38	0.97	1.07	0.95	0.32	(0.35)	0.97	1.07	0.95	0.32	(0.35)
Y1.EN	93.48	0.155	6.67	0	0.74	91.5	0.166	5.76	0.07	0.31	0.98	1.07	0.86	0.07	(0.43)	0.98	1.07	0.86	0.07	(0.43)
Y1.L	93.02	0.153	6.78	0.01	0.81	90.9	0.167	6.27	0.14	0.44	0.98	1.09	0.92	0.13	(0.37)	0.98	1.09	0.92	0.13	(0.37)
Y2.A.EN	62.72	0.286	6.83	1.9	0.24	74.6	0.263	6.44	1.6	0.10	1.19	0.92	0.94	(0.30)	(0.14)	1.19	0.92	0.94	(0.30)	(0.14)
Y2.A.L	60.88	0.285	6.93	2.07	0.27	73.6	0.262	6.61	1.71	0.08	1.21	0.92	0.95	(0.36)	(0.19)	1.21	0.92	0.95	(0.36)	(0.19)
Y2.B.EN	55.54	0.295	6.83	2.41	0.09	58.2	0.299	6.79	2.53	0.04	1.05	1.01	0.99	0.12	(0.05)	1.05	1.01	0.99	0.12	(0.05)
Y2.B.L	54.55	0.294	6.91	2.5	0.16	58.0	0.296	6.79	2.58	0.02	1.06	1.01	0.98	0.08	(0.14)	1.06	1.01	0.98	0.08	(0.14)
Y2.C.EN	63.41	0.286	6.92	1.93	0.21	64.7	0.288	6.74	2.21	0.07	1.02	1.01	0.97	0.28	(0.14)	1.02	1.01	0.97	0.28	(0.14)
Y2.C.L	59.82	0.288	6.96	2.33	0.16	62.7	0.287	6.78	2.37	0.07	1.05	0.99	0.97	0.04	(0.09)	1.05	0.99	0.97	0.04	(0.09)
Y2.D.EN	72.22	0.261	6.74	1.29	0.29	78.7	0.245	6.39	1.14	0.17	1.09	0.94	0.95	(0.15)	(0.12)	1.09	0.94	0.95	(0.15)	(0.12)
Y2.D.L	71.7	0.254	6.80	1.33	0.29	80.1	0.241	6.57	1.17	0.16	1.12	0.95	0.97	(0.16)	(0.13)	1.12	0.95	0.97	(0.16)	(0.13)
Y2.EN	81.91	0.235	6.86	0.68	0.52	85.7	0.221	6.05	0.6	0.18	1.05	0.94	0.88	(0.08)	(0.34)	1.05	0.94	0.88	(0.08)	(0.34)
Y2.L	81.47	0.226	6.90	0.65	0.56	85.4	0.217	6.34	0.74	0.18	1.05	0.96	0.92	0.09	(0.38)	1.05	0.96	0.92	0.09	(0.38)
Y3.A.EN	70.91	0.265	6.85	1.54	0.26	64.5	0.285	6.69	2.16	0.05	0.91	1.07	0.98	0.62	(0.21)	0.91	1.07	0.98	0.62	(0.21)
Y3.A.L	72.22	0.259	6.91	1.59	0.26	59.9	0.288	6.86	2.42	0.04	0.83	1.11	0.99	0.83	(0.22)	0.83	1.11	0.99	0.83	(0.22)
Y3.B.EN	62.03	0.281	6.86	2.01	0.23	51.1	0.306	6.85	2.79	0.06	0.82	1.09	1.00	0.78	(0.17)	0.82	1.09	1.00	0.78	(0.17)
Y3.B.L	61.08	0.282	6.88	2.08	0.22	50.1	0.307	6.90	2.91	0.05	0.82	1.09	1.00	0.83	(0.17)	0.82	1.09	1.00	0.83	(0.17)
Y3.C.EN	70.4	0.268	6.89	1.54	0.26	56.7	0.299	6.80	2.6	0.02	0.81	1.12	0.99	1.06	(0.24)	0.81	1.12	0.99	1.06	(0.24)
Y3.C.L	67.95	0.262	6.94	1.7	0.26	56.6	0.299	6.86	2.67	0.02	0.83	1.14	0.99	0.97	(0.24)	0.83	1.14	0.99	0.97	(0.24)
Y3.D.EN	78.36	0.238	6.65	0.9	0.33	61.7	0.281	6.64	2.02	0.04	0.79	1.18	1.00	1.12	(0.29)	0.79	1.18	1.00	1.12	(0.29)
Y3.D.L	77.82	0.233	6.76	0.86	0.38	62.1	0.282	6.76	2.14	0.06	0.80	1.21	1.00	1.28	(0.32)	0.80	1.21	1.00	1.28	(0.32)
Y3.EN	85.7	0.216	6.75	0.41	0.59	78.8	0.242	6.24	1.2	0.13	0.92	1.12	0.92	0.79	(0.46)	0.92	1.12	0.92	0.79	(0.46)
Y3.L	84.52	0.21																		

Table A5.3 Subgroup analysis simulation study results (N=100)

N=100	Independent covariates					Correlated covariates					Ratios: Corr::Indpndt				
	PCD	MSE	Corr0	Incorr0	Exact	PCD	MSE	Corr0	Incorr0	Exact	PCD	MSE	Corr0	Incorr0	Exact
Y1.A.EN	61.69	0.293	6.70	2.22	0.11	54.0	0.303	6.53	2.51	0.01	0.88	1.03	0.97	0.29	(0.10)
Y1.A.L	61.53	0.288	6.73	2.25	0.12	55.0	0.300	6.70	2.59	0.03	0.89	1.04	1.00	0.34	(0.09)
Y1.B.EN	52.75	0.308	6.76	2.73	0.10	41.2	0.324	6.78	3.2	0.02	0.78	1.05	1.00	0.47	(0.08)
Y1.B.L	52.08	0.305	6.80	2.73	0.08	39.1	0.326	6.81	3.35	0.01	0.75	1.07	1.00	0.62	(0.07)
Y1.C.EN	57.51	0.293	6.85	2.29	0.13	48.4	0.312	6.70	2.84	0.02	0.84	1.06	0.98	0.55	(0.11)
Y1.C.L	58.8	0.292	6.86	2.34	0.09	46.2	0.309	6.78	2.91	0.02	0.79	1.06	0.99	0.57	(0.07)
Y1.D.EN	65.07	0.281	6.62	1.9	0.11	48.5	0.310	6.78	2.83	0.04	0.75	1.10	1.02	0.93	(0.07)
Y1.D.L	64.77	0.276	6.60	1.87	0.15	49.6	0.304	6.81	2.75	0.05	0.77	1.10	1.03	0.88	(0.10)
Y1.EN	81.74	0.241	6.37	0.61	0.33	79.0	0.258	6.15	1.12	0.13	0.97	1.07	0.97	0.51	(0.20)
Y1.L	81.9	0.235	6.52	0.62	0.41	78.0	0.255	6.36	1.34	0.09	0.95	1.08	0.98	0.72	(0.32)
Y2.A.EN	36.72	0.330	6.89	3.51	0.02	40.3	0.319	6.71	3.18	0.01	1.10	0.96	0.97	(0.33)	(0.01)
Y2.A.L	36.34	0.332	6.95	3.56	0.02	40.0	0.318	6.73	3.22	0.00	1.10	0.96	0.97	(0.34)	(0.02)
Y2.B.EN	32.39	0.335	6.93	3.65	0.03	30.2	0.336	6.88	3.68	0.00	0.93	1.00	0.99	0.03	(0.03)
Y2.B.L	31.14	0.335	6.92	3.73	0.00	30.2	0.336	6.91	3.74	0.00	0.97	1.00	1.00	0.01	0.00
Y2.C.EN	34.17	0.333	6.96	3.59	0.02	35.5	0.329	6.83	3.48	0.01	1.04	0.99	0.98	(0.11)	(0.01)
Y2.C.L	34.99	0.330	6.96	3.54	0.03	35.2	0.327	6.86	3.47	0.00	1.01	0.99	0.99	(0.07)	(0.03)
Y2.D.EN	41.98	0.320	6.82	3.27	0.01	43.8	0.319	6.67	3.06	0.03	1.04	1.00	0.98	(0.21)	0.02
Y2.D.L	40.9	0.321	6.87	3.33	0.02	40.7	0.322	6.71	3.24	0.01	1.00	1.00	0.98	(0.09)	(0.01)
Y2.EN	48.62	0.313	6.79	2.78	0.06	63.7	0.293	6.44	2.19	0.02	1.31	0.94	0.95	(0.59)	(0.04)
Y2.L	49.55	0.310	6.79	2.79	0.06	60.1	0.293	6.54	2.32	0.02	1.21	0.94	0.96	(0.47)	(0.04)
Y3.A.EN	37.21	0.329	6.87	3.39	0.05	38.2	0.320	6.64	3.33	0.00	1.03	0.97	0.97	(0.06)	(0.05)
Y3.A.L	37.22	0.326	6.92	3.4	0.05	37.2	0.319	6.73	3.41	0.01	1.00	0.98	0.97	0.01	(0.04)
Y3.B.EN	34.66	0.327	6.80	3.46	0.04	28.1	0.340	6.92	3.77	0.00	0.81	1.04	1.02	0.31	(0.04)
Y3.B.L	33.36	0.328	6.91	3.54	0.03	26.7	0.339	6.90	3.83	0.00	0.80	1.03	1.00	0.29	(0.03)
Y3.C.EN	34.98	0.332	6.95	3.56	0.04	30.6	0.338	6.92	3.72	0.00	0.87	1.02	1.00	0.16	(0.04)
Y3.C.L	35.24	0.330	6.95	3.52	0.06	30.7	0.339	6.96	3.73	0.00	0.87	1.03	1.00	0.21	(0.06)
Y3.D.EN	41.81	0.323	6.86	3.2	0.02	31.1	0.331	6.79	3.51	0.01	0.74	1.02	0.99	0.31	(0.01)
Y3.D.L	39.51	0.323	6.86	3.29	0.03	33.4	0.326	6.79	3.44	0.00	0.85	1.01	0.99	0.15	(0.03)
Y3.EN	51.59	0.303	6.73	2.56	0.09	50.2	0.314	6.85	2.75	0.04	0.97	1.04	1.02	0.19	(0.05)
Y3.L	52.7	0.299	6.77	2.57	0.08	50.9	0.308	6.86	2.7	0.06	0.97	1.03	1.01	0.13	(0.02)

CURRICULUM VITA

Name: John A. Craycroft

Address: Department of Bioinformatics and Biostatistics
University of Louisville
485 E Gray St
Louisville, KY 40292

Birth place: Louisville, KY

Telephone: (502) 528-6292

Internet: john.craycroft@louisville.edu
john_craycroft@yahoo.com
LinkedIn: <https://www.linkedin.com/in/john-craycroft-1a6a581>

ACADEMIC BACKGROUND:

Anticipated
PhD (2020)
Biostatistics

School of Public Health and Information Sciences,
University of Louisville, Louisville, KY, USA
Coursework: completed Apr., 2018
Research topics: Causal inference, observational studies,
variable selection, longitudinal studies, penalized
regression
Advisor, Dr. Maiying Kong

M. Sc. (2016)
Biostatistics

School of Public Health and Information Sciences,
University of Louisville, Louisville, KY, USA
Coursework: completed Dec., 2015, cum. GPA 3.87
Master's Thesis: "Using Propensity Scores in Estimating
Treatment Effects with Observational Data";
Master's thesis advisor, Dr. Maiying Kong, April
2016

MBA (2004) Decision Information Analysis	Goizueta Business School, Emory University, Atlanta, GA, USA <i>Dean's List all 4 semesters (top 10%)</i> <i>80% tuition scholarship</i>
B.Sc. (1998) Statistics	George Washington University, Washington, DC, USA <i>Magna cum laude</i> Kullback prize in Statistics
BA (1998) Philosophy	George Washington University, Washington, DC, USA <i>Magna cum laude</i>

BRIEF CHRONOLOGY OF EMPLOYMENT:

Dec 2012 – Sep 2015	Revenue Science Fellow International Strategic Market Analysis division FedEx Services, Memphis, TN
July 2011 – Dec 2012	Strategic Marketing Fellow International Strategic Market Analysis division FedEx Services, Memphis, TN
July 2007 – July 2011	Sr. Strategic Marketing Analyst International Strategic Market Analysis division FedEx Services, Memphis, TN
July 2004 – July 2007	Strategic Marketing Analyst International Strategic Market Analysis division FedEx Services, Memphis, TN
Oct 2001 – Aug 2002	Survey Statistician National Opinion Research Center (NORC) University of Chicago
June 2000 – Sep 2001	Associate Management Consultant The Wexford Group, International Budapest, Hungary
May 1998 – June 2000	Staff Consultant Quantitative Economics and Statistics (QUEST) division Ernst & Young LLP, Washington, DC
Jan 1998 – May 1998	Statistics and Economics Consulting Intern Quantitative Economics and Statistics (QUEST) division Ernst & Young LLP, Washington, DC

Spring 1998

Teaching Assistant for Statistical Computing with SAS
undergraduate statistics class (for Dr. Reza Modarres)

PUBLICATIONS:

Craycroft, J., Huang, J., Kong, M. (2019). "Propensity score specification for optimal estimation of average treatment effect with binary response." Statistical Methods in Medical Research. (2020): 0962280220934847.

Vatsalya, V., Cave, M. C., Kong, M., Gobejishvili, L., Falkner, K. C., Craycroft, J., ... & Radaeva, S. (2019). "Keratin 18 is a diagnostic and prognostic factor for acute alcoholic hepatitis." *Clinical Gastroenterology and Hepatology*.

Rai, S. N., X. Wu, D. K. Srivastava, J. A. Craycroft, J. P. Rai, S. Srivastava, R. F. James, M. Boakye, A. Bhatnagar and R. Baumgartner (2018). "Review: propensity score methods with application to the HELP clinic clinical study." Open Access Medical Statistics(8): 11-23.

James, Robert F., Khattar, Nicolas K., Aljuboori, Zaid S., Page, Paul S., Shao, Elaine Y., Carter, Lacey M., Meyer, Kimberly S., Daniels, Michael W., Craycroft, John, Gaughen, John R., Chaudry, M. Imran, Rai, Shesh N., Everhart, D. Erik, Simard, J. Marc. (2018). "Continuous infusion of low-dose unfractionated heparin after aneurysmal subarachnoid hemorrhage: a preliminary study of cognitive outcomes." Journal of Neurosurgery. May 11, 2018.

Craycroft, J. (2016). "Propensity score methods: a simulation and case study involving breast cancer patients." Master's of Science in Biostatistics thesis.

AWARDS & HONORS:

June, 2019 M. Clinton Miller III Outstanding Poster Award, at the Summer Research Conference of the Southern Regional Council on Statistics (SRCOS)

Spring, 2016 University Fellowship, University of Louisville
Department of Bioinformatics and Biostatistics
(Full tuition, stipend, and health benefits, "...the most prestigious award for graduate students at the University of Louisville.")

April, 2015 Revenue Sciences Achievement Award,
FedEx Services

Spring 2014 Rising Star, Teamwork Award,
FedEx Global Marketing

2004 – 2015 12 BZ Awards ("Bravo Zulu," general FedEx award for outstanding performance, service, collaboration, etc.)

Spring 1998 Kullback Prize in Statistics, The George Washington University

ADDITIONAL CONSULTING AND COLLABORATION EXPERIENCE :

Spring, 2019 “Intraoperative Nasal, Bladder, and Blood Temperature during Cardiac Surgery.” Exploratory, time series, multiple regression, and correlation analyses for poster presentation by Roshan Babu and Jiapeng Huang, University of Louisville, May 2019

Fall, 2015 Time Series analysis and ANCOVA for effect of vegetation level of environment on cardiovascular health, research in process for Environmental Health PhD student Ray Yeager, with Dan Riggs and Dr. Shesh Rai

Spring, 2015 “Absorption capability of knitted and braided retraction cords.” ANOVA analysis for Dr. Gustavo Oliveira (University of Louisville School of Dentistry) and Dr. Guy Brock (University of Louisville, School of Public Health and Information Sciences, Department of Biostatistics and Bioinformatics)

Spring, 2015 Baseline demographics, ANOVA, and regression analysis for effect of heparin use on brain aneurysm patients for Dr. Robert James IV (Dept. of Neurological Surgery) and Dr. Shesh Rai

COMPUTING AND SOFTWARE EXPERIENCE:

- 15+ years experience with SAS, including Macros and SAS Enterprise Guide; **Certified in Advanced SAS Programming**
- Additional experience in R, MINITAB, and Tableau
- Experienced with DOS/WINDOWS and UNIX on PC workstation
- Experienced with SQL for accessing Oracle/Teradata-based data warehouses

PRESENTATIONS:

- “Propensity score specification for optimal estimation of treatment effect with binary response,” poster presentation at the Southern Regional Council on Statistics Summer Research Conference, June 2019, and at Research! Louisville, Sep. 2019.
- “The Role of the Statistical Consultant,” panelist on webinar sponsored by the

American Statistical Association's Committee for Career Development, April 30, 2019.

- "Causal Inference: Introduction to Directed Acyclic Graphs," and "Causal Inference: Confounders," guest lecture for 2 sessions of Causal Inference class, University of Louisville School of Public Health, Spring 2019.
- "SAS Basics," an introduction to SAS seminar, presented to U of L's School of Public Health, covering SAS environment; data input, output, and manipulation; syntax; concepts regarding data sets and PROCs; and testing and debugging SAS code. February 9, 2018
- "Bringing Value: Market Share Analysis That Goes Deeper," poster presentation, American Statistical Association Conference on Statistical Practice, San Diego, CA, February 18, 2016
- "Effect of Tumor DNA Profile on Survival Prognosis for Patients with Cancer of the Tongue," delivered to Survival Analysis class, December 10, 2015
- "Introduction to SAS Topics," guest lecturer presentation to Biostatistical Methods I classes, September, 2014, and October, 2015
- "Effect of Sample Design on Precision of Estimates," delivered to FedEx Strategic Market Analysis division, March 12, 2008

VOLUNTEER EXPERIENCE:

- St. Xavier High School Class of 1994 25-year Anniversary Endowment Committee
- President, University of Louisville Biostatistics Club, 2018-19 academic year; also Secretary, 2017-18; Treasurer, 2016-17; and inaugural Vice President, 2015-16
- Little League (Spring 2015), T-ball (Spring 2014), and soccer (Fall 2015) coach
- Corporate chairman, FedEx Operation Feed, in support of the Memphis Food Bank, 2005-2007 (*raised average of \$165,000 per year, plus thousands of pounds of food donations each summer for the Food Bank*)
- Teaching Assistant, Decision Information Analysis, Goizueta Business School, 2003-04 academic year
- Honors Council, Goizueta Business School, 2003-04 academic year
- Reader for the Kentucky Recording for the Blind and Dyslexic
- Junior Achievement teacher
- Habitat for Humanity

PROFESSIONAL MEMBERSHIP:

- American Statistical Association (ASA), since 2013

HONOR SOCIETIES:

- Phi Beta Kappa
- Beta Gamma Sigma