Topics in Causal Inference in Epidemiological Studies: Assessing Heterogeneity of Treatment Effects & Addressing Missing Covariates

Hongseok Kim, ScM, KMD

A dissertation submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Epidemiology at Brown University School of Public Health

Providence, Rhode Island

May 2021

This dissertation by Hongseok Kim is accepted in its present from by the Department of Epidemiology as satisfying the dissertation requirement for the degree of Doctor of Philosophy

Date _____ _____

Issa Dahabreh, Ph.D., Advisor

Recommended to the Graduate Council

Date _____ _____

Jon Steingrimsson, Ph.D., Reader

Date _____ _____

David Savitz, Ph.D., Reader

Approved by the Graduate Council

Date _____ _____

Andrew G. Campbell, Ph.D., Dean of the Graduate School

CURRICULUM VITAE

Hongseok Kim, KMD, ScM.
99 Norwood Ave
Cranston, RI 02905
kim.hongseok24@gmail.com

Education

PhD in Epidemiology
Brown University School of Public Health, Providence, RI
Dissertation title: Topics of causal inference in epidemiologic studies: Assessing heterogeneity of treatment effects and addressing missing covariates
Dissertation advisor: Issa Dahabreh, Ph.D.

ScM in Epidemiology
Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD
Thesis title: Longitudinal patterns of HIV medication adherence at an urban HIV clinic
Thesis advisor: Bryan Lau, Ph.D.

B.S. in Information Statistics
Korea National Open University, Seoul, South Korea

KMD (Degree of Korea Medicine Doctor)
Dongguk Univeristy, Gyeongju, South Korea

Professional Positions and Experience

Public Health Doctor (2011-2014)
Suncheon Public Health Center, Suncheon, South Korea

Certification and Licensure

South Korea
SAS Base Programming Specialist

Honors and Awards

Honor scholarship, Dongguk University 2009
Honor scholarship, Dongguk University 2010
Honor scholarship, Korea National Open University 2012
President of Korea Gallup Prize, Statistics competition in Korea Open National University 2012
Third Prize, Brown University Department of Epidemiology Student Abstract Competition 2018

First Prize, Brown University Department of Epidemiology Student Abstract Competition 2019

Teaching Activities

Brown University School of Public Health
Teaching Assistant, Advanced Quantitative Methods in Epidemiologic Research Fall 2019
Teaching Assistant, Introduction to Methods in Epidemiologic Research Fall 2019
Teaching Experience, Intermediate Methods in Epidemiologic Research Spring 2020

Publications

A. Peer-Reviewed Journal Articles

1. Kang AW, Bostom AG, **Kim H**, Eaton CB, Gohh R, Kusek JW, Pfeffer MA, Risica PM, Garber CE. Physical activity and risk of cardiovascular events and all-cause mortality among kidney transplant recipients. Nephrology Dialysis Transplantation. 2020 Aug 1;35(8):1436-43.

2. Monroe AK, Lau B, Mugavero MJ, Mathews WC, Mayer KH, Napravnik S, Hutton HE, **Kim HS**, Jabour S, Moore RD, McCaul ME. Heavy Alcohol Use Is Associated With Worse Retention in HIV Care. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2016 Dec 1;73(4):419-25.

3. **HongSeok Kim**, TaeRim Lee. Age-specific disease network for the Major Diseases. Journal of the Korea Society of Health Informatics and Statistics 2013; 38(1): 66-80.(in Korean)

B. Manuscript in Preparation

Kim HS, Dahabreh IJ. 'Visualizing and quantifying risk heterogeneity in clinical trials'.

Kim HS, Steingrimsson JA, Dahabreh IJ. 'Assessing heterogeneity of treatment effect: a cautionary note about methods that model the outcome in clinical trials'.

Kim HS 'Emulating a target trial using routinely collected observational data to compare treatments for patients with type2 diabetes not responding to metformin therapy', joint work with Tianyu S, Robertson SE, Hernan MA, Dahabreh IJ.

Conference Abstracts

A. Oral Presentations

1. A cautionary note about assessing heterogeneity over outcome scores in randomized trials, 13th International conference on Health Policy Statistics, San Diego, CA, 01/07/2020

B. Poster Presentations

1. A cautionary note about assessing heterogeneity over outcome scores in randomized trials, 53rd Annual Meeting of the Society for Epidemiologic Research 2019,

Minneapolis, MN, JUN 2019, Hongseok Kim, Issa J Dahabreh

2. Physical activity and risk of cardiovascular events and all-cause mortality among kidney transplant recipients., American Heart Association EPI/Lifestyle 2019 Scientific Sessions, Houston, TX, MAR 2019, Augustine W. Kang, Hongseok Kim, Carol E.Garber, Charles B. Eaton, Patricia M. Risica, Andrew G. Bostom

3. Estimating treatment effects after multiple imputation of missing baseline covariate data, 52nd Annual Meeting of the Society for Epidemiologic Research 2018, Baltimore, MD, JUN 2018, Hongseok Kim, Issa J Dahabreh

4. Age-specific disease network for the Major diseases in Korea, 59th World Statistics Congress, HongKong, AUG 2013, Hongseok Kim, Taerim Lee

ACKNOWLEDGEMENTS

First of all, I deeply appreciate my primary advisor, Dr. Issa Dahabreh, for his commitment and support throughout my Ph.D. program. His guidance was always a compass for me whenever I strayed, got lost during the course. His support was not limited to academic activity; personally, emotionally, in every aspect, I got a big help from him. It was truly my privilege to work close and learn from him in past years.

I also thank Dr. Jon Steingrimsson and Dr. David Savitz to be my committee member and review my dissertation. Their advice improved my dissertation and grow me scholarly. I also learned so much from other faculty members of School of Public Health through school curriculum, seminars or journal club. Taking this opportunity, I would like to thank all of them. I apologize that I cannot list all of them in this limited space. I am also indebted a lot to school staff. I appreciate Vickie, Kathleen, Brittany; they spared no effort to help me.

I cannot forget the support from colleagues in School of Public health during the program. I could go through the pandemic thanks for Sarah's support and concern. Without her, I might drop down in the middle. Augustine and Harold were with me whenever I had a hard time. Jason, Sina, Hannah started the program with me at the very first moment. Annie, Will, Courtney, Harry, Marissa, Chris, Nana, Alex, colleagues in Epidemiology department, Shaun, Hace, Mauricio, Elliott, Anthony, ... thank you all of you.

I like to talk about my father and mother. Gratitude is not enough word to express what I feel about their devotion, belief, and love toward me. I have been the luckiest person since I was born. I regret I could not be more affectionate to my grandmother before she passed away although she always said she loves me. I would like to leave an answer here: I love you too.

I owe so many people for helping me get though this journey. I appreciate it and will not forget what I received.

Contents

List of Tables

List of Figures

Introduction

Estimating causal effects is often a main interest of epidemiologic studies. As a subject of epidemiologic studies is a population, a causal question in Epidemiology usually has a form of "What is the difference in outcome if population of interest (target population) received treatment or agent A versus treatment or agent B". A causal estimand to answer this question is average treatment effect (ATE).

There are different ways to estimate ATE of the target population depending on research settings and investigators' preference. I consider three ways of estimating ATE here: 1) Transport inferences of randomized trials to the target population directly when there is no evidence of the existence of heterogeneity of treatment effects, 2) Transport inferences of randomized trials to the target population adjusting different distribution of effect modifiers between trial population and the target population, 3) estimate causal effects in the target population using observational studies. In this dissertation, I deal with a remaining issue of each of aforementioned way. I describe the issue of each way and brief main chapters in the rest of this section.

Randomized controlled trials (RCT) are the gold standard of identifying causal effects. However, study population of RCT is different from the target population and many times it is difficult to broaden study population of RCT because of ethics issue, or resource availability. The one way, maybe the easiest way, to infer causal effects in the target population is directly transporting inferences of RCT to the target population under no heterogeneity of treatment effects (HTE) in RCT. If covariate distribution is overlapped between RCT and the target population, we can assume that the treatment effect is the same across all individuals in the target population and ATE of RCT can be directly represent ATE of the target population. For this direct transportation, we necessarily assess HTE in trial population. Assessing HTE is not a simple task and there are various ways to do and not all of them lead to the correct conclusion; therefore, to avoid a misleading conclusion, we should

understand ways to assess HTE and be cautious about our choice. In chapter 1, we compare one way of assessing HTE (outcome score methods), which is recommended in prestigious medical journals recently, to another approach (effect score methods). We propose that when outcome score and true conditional average treatment effects are independent, outcome score methods do not find HTE even if treatment effects vary over the population significantly and significant amount of heterogeneity exists. However, effect score methods work without any problems. We demonstrate our hypothesis using simulation and an application study (ALLHAT).

If there is heterogeneity of treatment effects and the distribution of effect modifiers is different between RCT and the target population, we need to address this difference; most real-world studies would belong to this case. Missing data is an issue of this study design, adding one more layer of complication. However, there are no studies focusing on this issue fundamentally. In chapter 2, I review the study design and propose how to identify ATE when some covariates have missingness using inverse probability weighting. I focus on the setting when treatment and outcome are not observable for non-participants of a trial and discuss identification and assumptions of missing data mechanism. I illustrate the method using the CASS study.

If randomized trials are not available for a treatment comparison we are interested in or some conditions we need to assume for transportability do not apparently hold, investigators need to resort on observational studies to estimate ATE of the target population. In this case, we cannot avoid unmeasured confounding. Missing covariates is an issue of this study design and complicated with causal inference. In chapter 3, I review missing data methods, causal inference methods and how to combine them for observational analyses from a comprehensive viewpoint. I compare finite sample performance of various estimators that are derived by combining missing data methods and causal inference methods. I also address up-to-date issues of the topic: comparison of Across and Within approach of multiple imputation and augmented inverse probability weighting estimators. The CASS study is used to illustrate the methods.

CHAPTER1

Assessing heterogeneity of treatment effects: a cautionary note about outcome score methods

Abstract

Many randomized trials examine heterogeneity of treatment effects over summary variables, such as outcome scores (i.e., predicted outcomes under no treatment) and effect scores (i.e., predicted conditional average treatment effects). Here, we show that under certain data laws, outcome score methods produce an overwhelming number of false negative results and fail to classify patients into groups that experience differential treatment effects. The pathological laws that adversely affect the performance of outcome score methods may be unusual, but poor performance also occurs under laws that are "close" to the pathological ones. We also show that easy to implement effect score methods can overcome these problems. We illustrate the use of several outcome and effect score methods in a large randomized trial of anti-hypertensive treatments. We conclude that outcome score-based analyses should not be the primary approach for assessing heterogeneity in trials because they can have low power to find HTE and poor classification performance even in the presence of strong effect modification; in contrast, effect score methods perform much better.

1 Introduction

Personalized medicine requires assessment of heterogeneity of treatment effects (HTE) over multiple patient characteristics [1]. The most common approach for assessing HTE in confirmatory randomized trials is one-variable-at-a-time subgroup analysis, where the study sample is stratified over various characteristics (e.g., sex, race, etc.) and treatment effects are compared among strata [2]. This approach, however, does not use the data efficiently and is hard to interpret when the treatment effect is modified by multiple variables [1]. An increasingly popular alternative approach for assessing heterogeneity over multiple variables simultaneously is to construct a *summary score* by aggregating information from multiple baseline covariates and assess HTE over the summary score.

A natural choice of summary score is the estimated conditional average treatment effect at the covariate values of each individual; we refer to this type of summary score as an *effect score*. Effect score methods are appealing because the effect score of each individual is an estimate of the average treatment effect for a group with the same covariate values. Despite the obvious appeal of effect scores, many recent recommendations for practice have focused on examining HTE over the predicted outcome under the "control" treatment or over all treatment groups (but ignoring treatment assignment) [3, 4, 5, 6]; we refer to this type of summary score as an *outcome score*.

Outcome score methods are heuristically motivated by the observation that the outcome under the control treatment is a determinant of the treatment effect. For example, if the event rate of a binary outcome in the absence of treatment is very low then the average treatment effect is at best close to null. More pragmatically, outcome score methods have been justified as a strategy to avoid over-fitting [6, 7], presumably under the belief that such over-fitting is worse for effect score approaches. Yet, by definition, outcome score approaches estimate parameters (aspects of the conditional distribution of the outcome in the absence of treatment) that only partially capture variability in the conditional average treatment effect. This intentional focus on the "wrong target" raises the concern that outcome score approaches may fail to detect HTE even when its magnitude is scientifically interesting or relevant for decision-making. To our knowledge, no study has examined the conditions under which outcome score methods can lead to misleading conclusions about HTE, and whether effect score methods can perform better under the same conditions.

In this paper, we show that outcome score methods can fail to detect strong HTE even in very large trials. The failures depend on trial sample size, model specification, and the strength of the association between the outcome score and the true conditional average treatment effect. We review effect and outcome score methods in the context of simple trials with time-fixed treatments and discuss statistical tests and classification methods for HTE. We use simple theoretical arguments and extensive simulation studies to characterize the conditions under which outcome score methods perform poorly. We illustrate these results using data from the ALLHAT trial of anti-hypertensive treatments. Overall, the theoretical arguments, simulations, and empirical results show that outcome score approaches can lead to misleading conclusions regarding HTE under some data laws, whereas simple effect score approaches perform much better. We conclude that, contrary to recent recommendations in major clinical journals, outcome score approaches should not be the preferred strategy for assessing HTE.

2 Summary scores and the assessment of HTE

2.1 Defining HTE

Suppose we have a randomized trial of the effect of a time-fixed treatment on some outcome of interest, measured at a single point in time. Let A denote the treatment assignment; X the baseline covariates; and Y some outcome of interest assessed at the end of the study. Let Y^a denote the potential (counterfactual) outcome under intervention to assign treatment a from the set of randomized treatments $\mathcal{A}$ [8, 9]. Throughout, we assume that the observed data are realizations of $O = (X, A, Y)$, from a law (distribution) P that belongs in some model $\mathcal{P}$. We use italicized capital letters to denote random variables and corresponding lower case letters to denote realizations. For simplicity, suppose A is binary so that $\mathcal{A} = \{0, 1\}$. We will refer to assignment to $a = 1$ as "treatment" and $a = 0$ as "control," even though both might involve active intervention. Our results generalize readily to the multi-valued discrete treatments, but we focus on binary treatments for simplicity.

A reasonable target of inference in HTE analysis is the *conditional average treatment effect* function, defined as the difference of the potential outcome means under the

treatment and control, conditional on baseline covariates,

$$\delta(X) \equiv \mathrm{E}[Y^1 - Y^0|X].$$

We will say that *HTE exists* when for at least two possible realizations of the covariate vector X, say x_1 and x_2, $x_1 \neq x_2$, the corresponding conditional average treatment effects are unequal, $\delta(x_1) \neq \delta(x_2)$, that is, $\mathrm{E}[Y^1 - Y^0|X = x_1] \neq \mathrm{E}[Y^1 - Y^0|X = x_2]$.

Under the usual identifiability conditions (consistency of potential outcomes; exchangeability of the treated and control groups; and positivity of treatment assignment probabilities [10]), all of which are likely to hold in well-conducted randomized trials, we can identify $\delta(X)$ from randomized trial data using the mean difference between the treatment groups, conditional on baseline covariates,

$$\delta(X) = \mathrm{E}[Y|X, A = 1] - \mathrm{E}[Y|X, A = 0].$$

This result gives an obvious way to assess heterogeneity of treatment effects, by examining how the conditional mean difference, $\mathrm{E}[Y|X, A = 1] - \mathrm{E}[Y|X, A = 0]$, varies over X.

2.2 Assessing HTE over summary scores

The existence of HTE implies that treatment effects are different across subgroups of population or vary over summary scores. For the assessment of HTE, effect score methods attempt to model $\delta(X)$ and examine variation of treatment effects over the estimated values. In contrast, in their simplest form, outcome score methods focus on modeling mean of outcome under control treatment, that is, on modeling $\mathrm{E}[Y|X, A = 0]$ and examining variation of treatment effects over the estimated values of this conditional outcome mean function. Some authors, however, have proposed variants of the outcome score methods that model the mean outcome as a function of baseline covariates *ignoring treatment assignment*, instead of modeling under control treatment.

In this paper, we consider four types of summary scores (two are effect score methods and the other two are outcome score methods):

1. an effect score method based on modeling the outcome under each of the treatments being compared (a variant of the approach described in [11]);

2. an effect score method based on g-estimation of the conditional average treatment effect (as described in [12]);
3. an outcome score method based on modeling the conditional outcome mean under the control treatment; and
4. an outcome scores method based on modeling the conditional outcome mean, ignoring treatment assignment.

We assume that there is no reliable external sources for building models of outcome scores or effect scores so we should resort to internal trial data. We explain how to build those summary scores using parametric models from trial data in the following section.

Building summary scores

Because X is often high-dimensional and randomized trials have modest sample sizes, inference about $\delta(X)$, $\mathrm{E}[Y|X, A = 0]$, or $\mathrm{E}[Y|X]$ often requires the use of statistical models to smooth over the data. In the remainder of this paper we focus on the use of estimates from parametric (finite-dimensional) models to assess HTE, using methods that are popular in applied work.

Modeling the conditional outcome mean under each treatment: A natural idea for modeling $\delta(X)$ is to model each of its component of conditional expectations, $\mathrm{E}[Y|X, A = a]$ for $a \in \{0, 1\}$. Suppose that we posit parametric models for each of these conditional expectations, such as,

$$\mathrm{E}[Y|X, A = 1] = m_1(X; \beta_1) \quad \text{and} \quad \mathrm{E}[Y|X, A = 0] = m_0(X; \beta_0).$$

The model parameters, β_0 and β_1 can be estimated using standard regression approaches (e.g., generalized linear models appropriate for the outcome Y) and then $\delta(X)$ can be estimated as

$$\delta_{TG}(X; \widehat{\beta}) = m_1(X; \widehat{\beta_1}) - m_0(X; \widehat{\beta_0}).$$

If the two conditional expectation models are correctly specified, then $\delta_{TG}(X; \widehat{\beta})$ is a consistent estimator of the conditional average treatment effect.

This approach has a long history in statistics and epidemiology. For example, [11]

described this approach as a first step towards building a classification algorithm, and used the term "virtual twin" to describe obtaining predictions for each observed values of X using models fit in different treatment groups. We refer to this method as "Two GLM" in tables and figures.

Direct modeling of the conditional average treatment effect by g-estimation: Instead of modeling conditional mean outcome of each treatment group, another approach is directly modeling $\delta(X)$ by g-estimation [12]. First, treatment A is encoded as a new variable T (if $A = 1$, $T = 1$, and if $A = 0$, $T = -1$), and covariate values are multiplied by $T/2$. Next, outcome model given modified covariates is fitted, and predictions from the model are computed by plugging original covariate values. These predictions are a consistent estimator of $\delta(X)$ [12]. More efficient estimator of $\delta(X)$ is available using an optimal augmentation term but we only consider the standard estimator here. We refer to this method as "Tian" in tables and figures.

Modeling the conditional outcome mean under the control treatment: These outcome scores are conditional expectation of outcome given covariates for the control group:

$$\mathrm{OS}_1(X) \equiv E[Y|X, A = 0] = m_0(X; \beta_0)$$

The scores are estimated by plugging covariate values into the outcome model, $m_0(X; \beta_0)$.

$$\mathrm{OS}_1(X; \widehat{\beta_0}) = E[Y|X, A = 0; \widehat{\beta_0}] = m_0(X; \widehat{\beta_0})$$

Modeling the conditional outcome mean ignoring treatment assignment: These outcome scores are conditional expectation of outcome given covariates ignoring treatment assignment.

$$\mathrm{OS}_2(X) \equiv E[Y|X] = s(X; \beta_2)$$

, where $s(X; \beta_2)$ is a model for $\mathrm{E}[Y|X]$

The outcome scores are estimated by fitting a model for the outcome on covariates ignoring treatment assignment and taking predictions from the model.

$$\mathrm{OS}_2(X; \widehat{\beta_2}) = E[Y|X; \widehat{\beta_2}] = s(X; \widehat{\beta_2})$$

This approach is "unconventional" in the following sense: when treatment has any effect on the conditional outcome mean (which it usually does), the population model $s(X;\beta_2)$ for $E[Y|X]$ is in general different from either population model $m_1(X;\beta_1)$ or $m_0(X;\beta_0)$ (see Appendix), thus use of $s(X;\widehat{\beta_2})$ intentionally ignores treatment, even if it is an important predictor of the outcome. Be that as it may, it is the approach that has been recommended for use, and in fact some authors have argued that it should be required by journal editors.

Note that we have focused on the concept and estimation of each score. When internal data is used to estimate the summary scores, investigators need to use model validation techniques, such as cross validation or sample splitting, to avoid overfitting. We discuss statistical tests to assess HTE in the section 3 and the classification by summary scores in section 3.2.

3 Detecting HTE and classifying individuals by predicted resonse

3.1 Statistical tests for HTE

We review two ways of testing the presence of HTE using the summary scores: the t-test of an interaction term and the likelihood ratio test comparing a model of constant treatment effect with heterogeneous treatment effect across groups defined by quartiles of the summary scores.

1. *T-test* (Int)

   Consider a model of outcome given any type of the summary scores (S), treatment, and the interaction between them.

$$E[Y|A,S;\gamma] = \gamma_0 + \gamma_1 A + \gamma_2 S + \gamma_3 AS \tag{1.1}$$

   To simplify exposition, we only consider a linear function of summary scores but in practice the model could include non-linear function of S, like splines or square terms. We use the t-test to evaluate the significance of the coefficient γ_3. The null is

$$H_0 : \gamma_3 = 0,$$

and can be tested against the alternative

$$H_A : \gamma_3 \neq 0$$

The rejection of the test means that there is no evidence on linear association between $\delta(X)$ and the summary scores. The test may not find HTE if there is a non-linear relationship between $\delta(X)$ and summary scores.

2. *Likelihood ratio test* (LRT)

   We consider the other approach based on the likelihood ratio test. Suppose we group all the trial participants by quartiles of the summary scores; denote the grouping variable as G. The null model is

$$E[Y|A, G] = \gamma_0' + \gamma_1' A + \gamma_2' G.$$

   indicating the constant treatment effect across the group. The alternative model is

$$E[Y|A, G] = \gamma_0'' + \gamma_1'' A + \gamma_2'' G + \gamma_3'' A^* G.$$

   allowing the treatment effect varies across the group. We compare the goodness of fit of the two models using the likelihood ratio test.

   The null hypothesis is that the null model and the alternative model do not have significant difference in goodness of fit. The rejection means that the alternative model fits data better and that supports heterogeneity of treatment effect across the group based on the quartiles of summary scores. This approach can capture the non-linear relationship better than the t-test but still not perfect. Alternatively, we can increase the number of bins, e.g., using deciles instead of quartiles.

3.2 Classification by predicted response

We also consider how to classify population into two groups based on whether individuals would get benefit from treatment or not. The purpose of HTE analysis to give clinicians useful information on clinical decisions, so the classification by the summary scores might reflect the practical use of them. Assuming no cost and no adverse effects of treatment compared to the control, we define the group of benefiting from treatment as those who have positive conditional treatment effect and define the opposite group as

those who have negative or null conditional treatment effect.

To classify population using the summary scores, we first fit a linear model of conditional outcome given the summary scores, treatment, and the interaction between them, like (1.1). We made two copies of each individual by setting treatment assignment of them as treated ($A = 1$) or under control ($A = 0$). We obtain predictions of outcome risk of these two copies from the fitted model and subtract the prediction of the treated from the prediction of the control and define this quantity as the estimated treatment effect for this individual. The method here is very similar to $\widehat{\delta}_{TG}(X)$ except for that we use one outcome model with the interaction to estimate potential outcome under each treatment option. We classify patients by the sign of the estimated effect. In simulation, we know the true conditional average treatment effect and can classify individuals by a sign of the true treatment effect. We compare the group by the true treatment effect to the group by the estimated treatment effect using two-by-two tables.

4 When do outcome score approaches fail?

4.1 Characterizing the failures

In this section, we discuss when the outcome score methods fail to detect HTE and correctly classify population. We hypothesize that the outcome score methods fail when the limiting values of the outcome scores $(m_0^*(X;\beta_0), s^*(X;\beta_2))$ are independent with the true conditional treatment effect.

$$m_0^*(X;\beta_0) \perp\!\!\!\perp \delta(X) \qquad s^*(X;\beta_2) \perp\!\!\!\perp \delta(X). \tag{1.2}$$

Under this condition the distribution of the estimated outcome scores does not give any information on the true conditional average treatment effect. We also presume that the effect score methods find HTE and classify population without any problems. This condition holds depending on the data laws, which determines the true conditional average treatment effect function and the choices of working models, which determines the limiting value of the estimated outcome scores. Especially, if the above independence holds for true $m_0(X;\beta_0)$ and $s(X;\beta_2)$, we cannot detect HTE even if our working models are correctly specified. Hereafter, we refer to the situation where the independence (1.2) holds as 'the pathological conditions'. Next two sections, we give examples of the

pathological conditions and data laws.

4.2 Examples of failure with semiparametric models

We have considered to use parametric models to estimate HTE so the occurrence of the pathological condition depends on the choice of models. It is noteworthy that the pathological conditions can occur when $m_0(X)$, $s(X)$, and $\delta(X)$ are semiparametrically estimated so the conditions does not necessarily depend on the choice of parametric models. A simple example is when $m_0(X)$ is a constant c and $m_1(X)$ is as a function of X. Then, $\delta(X)$ has a variation over X. Suppose we correctly estimate $m_0(X)$ and $m_1(X)$ but we cannot detect this variation by inspecting $\widehat{m_0}(X)$. Another example is when there are two independent covariates X_1 and X_2 and $m_0(X)$ is a function of only X_1 ($m_0(X) = f(X_1)$) and $m_1(X)$ is a function of X_1 and X_2 such that $m_1(X) = f(X_1)+g(X_2)$, where $f(X_1)$ and $g(X_2)$ are arbitrary functions of each X_1 and X_2. Then, $\delta(X) = g(X_2)$, which is independent with $m_0(X)$. Thus, even if we correctly estimate $m_0(X)$, we cannot infer variation of $\delta(X)$ from the distribution of $\widehat{m_0}(X)$. We can also find a case when functions for $m_1(X)$ is multiplicative of X_1 and X_2. For example, when $m_1(X) = log\{f(X_1)g(X_2)\}$ and $m_0(X) = log\{f(X_1)\}$, $\delta(X) = \log(g(X_2))$. Likewise, we can conceive many cases where condition (1.2) holds in semiparametric models.

4.3 Examples of failure with parametric models

To realize the pathological conditions with correctly specified models easily, throughout this section and simulation study, covariate vector X_i follows bivariate normal distribution $X_i = (X_{1,i}, X_{2,i})' \sim \mathcal{N}(\mu, \Sigma)$, where $\mu = (0, 0)$‘ and $\Sigma = \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} \end{bmatrix}$ with $i = 1, ..., n$. Treatment assignment, A is generated from a Bernoulli distribution with $\Pr[A = 1] = 0.5$. We assume that response Y is generated from the following distribution:

$$Y = tX_1 + uX_2 + \alpha A + vAX_1 + wAX_2 + \epsilon,\ \epsilon \sim \mathcal{N}(0, 1),$$

Then, $\delta(X) = \alpha + vX_1 + wX_2$, $m_0(X) = tX_1 + uX_2$, and $s(X) = tX_1 + uX_2 + 0.5(1+vX_1+wX_2)$ and the limiting value $m^*_0(X; \beta_0)$ converges to $m_0(X)$ and $s^*(X; \beta_2)$ converges to $s(X)$ when the model $m_0(X; \beta_0)$ and $s(X; \beta_2)$ are correctly specified. Because zero covariance implies zero correlation coefficient and zero correlation coefficient

between jointly normal variables implies the independence between them,

$$\mathrm{Cov}[\delta(X), m_0^*(X;\beta_0)] = 0 \implies \delta(X) \perp\!\!\!\perp m_0^*(X;\beta_0)$$

$$\mathrm{Cov}[\delta(X), s^*(X;\beta_0)] = 0 \implies \delta(X) \perp\!\!\!\perp s^*(X;\beta_0)$$

, under this data law. We can derive the covariances as

$$\mathrm{Cov}[\delta(X), m_0^*(X;\beta_0)] = tv\sigma_{X_1}^2 + (tw+uv)\sigma_{X_1X_2} + uw\sigma_{X_2}^2$$

$$\mathrm{Cov}[\delta(X), s^*(X;\beta_2)] = (tv+0.5v^2)\sigma_{X_1}^2 + (tw+vw+uv)\sigma_{X_1X_2} + (uw+0.5w^2)\sigma_{X_2}^2$$

and identified the data laws making those covariance zero. Under those data laws if we estimate the outcome scores using correctly specified models and use them to assess HTE, they will fail to find HTE even when substantial magnitude of HTE exists. We also examined the performance of outcome score methods under data laws of other correlation coefficients, expecting that the performance is worse as the correlation coefficients are closer to zero. The correlation coefficients other than zero were calculated as

$$\rho_{m_0} = \frac{\mathrm{Cov}[\delta(X), m_0^*(X;\beta_0)]}{\sqrt{\mathrm{Var}(\delta(X))}\sqrt{\mathrm{Var}(m_0^*(X;\beta_0))}} \quad \rho_s = \frac{\mathrm{Cov}[\delta(X), s^*(X;\beta_0)]}{\sqrt{\mathrm{Var}(\delta(X))}\sqrt{\mathrm{Var}(s^*(X;\beta_0))}},$$

where ρ_{m_0} is a correlation coefficient between $\delta(X)$ and $m_0^*(X;\beta_0)$ and ρ_s is a correlation coefficient between $\delta(X)$ and $s^*(X;\beta_0)$.

5 Simulation study

We run a series of simulations to assess the statistical power of the tests for HTE detection and the performance of the classification under the pathological conditions. In every simulation, we selected data laws, generated sample from them, built summary scores (Section 2.2) and conducted statistical tests for HTE detection (Section 3), or classified observations using the estimated conditional treatment effects and the true treatment effects (Section 3.2). In this section, we explain the purpose of each simulation scenario and what data laws we selected to accomplish the purpose. In the first simulation, we specified the data laws where $\mathrm{Cor}[\delta(X), m_0(X;\beta_0)]$ or $\mathrm{Cor}[\delta(X), s(X;\beta_2)]$ is 0, 0.2, 0.4, 0.6, and 0.8 and generated data from the data law and correctly esti-

mated the summary scores using parametric models 4.3. We assessed the statistical power of the tests for HTE detection. The data law of zero correlation coefficient for $m_0(X)$ was $Y = 0.15X_1 + 0.15X_2 + 0.15A + 0.125AX_1 - 0.125AX_2 + \epsilon$ and for $s(X)$ was $Y = 0.15X_1 + 0.275X_2 + 0.15A + 0.125AX_1 - 0.125AX_2 + \epsilon$ ($\epsilon \sim \mathcal{N}(0, 1)$). We present all the data laws we chose in appendix.

In the next simulation, we intentionally fitted misspecified models to verify that the problem can also happen when models are misspecified. We used the same data law as in the first simulation but included the square term of X_1 to generate outcome and omitted the square term in the misspecified models. We compared the statistical power of the tests for HTE detection across the summary scores. Because of the square term, $m_0^*(X; \beta_0)$, $s^*(X; \beta_2)$ from the correctly specified models do not lead to the pathological condition but $m_0^*(X; \beta_0)$, $s^*(X; \beta_2)$ from the misspecified models result in the problem as in the previous simulation.

In the third simulation, we conducted a grid search design to examine how the outcome score methods work as the pathological condition is gradually relaxed. We used the same data laws as in the first simulation and varied the coefficient of X_1 by 0.02 in the range of [-0.35, 0.65] and varied the coefficient of AX_1 by 0.01 in the range of [-0.125, 0.375]. Consequently, a correlation coefficient for $m_0(X)$ varied in [-1, 0.853] and a correlation coefficient for $s(X)$ varied in [-0.797, 0.992]. At each point, we assessed statistical power of detecting HTE of the outcome score methods over 1000 replications. We conducted simulation with different sample size (500, 2000, 5000, 10000, 20000) over 1000 replications for those three scenarios.

To get a more complete picture, we conducted another grid search design simulation. Under this design, all the determinants of the pathological condition, the coefficients of X_2, AX_2 and the variances and correlation between X_1 and X_2 as well as the coefficients of X_1, and AX_1, varied in a certain range; we attached the range of parameters in appendix. We specified total 413343 laws of data and at each law, we assessed the statistical power of the tests for HTE detection for outcome score and effect score approaches over 100 replications and compared them. We varied sample size in (500, 2000, 5000, 10000). We classified data laws into two cases according to the presence of HTE; if the coefficients of both AX_1 and AX_2 are 0, we classify those laws into 'no HTE' group, otherwise, we classify data laws into 'HTE group'.

Lastly, we assessed classification performance of each approach under the patholog-

ical condition. We generated data from the data law we used for the first simulation. In addition, We added the case of no average treatment effect, where the coefficient of A is 0 to make positive and negative treatment effect group are evenly distributed; if average treatment effect is positive, more patients are classified into positive group and groups are unbalanced. Under each data law, we generated data with 500, 2000, 5000, 10000, 20000 sample size. In each simulation data, we estimated the outcome and effect scores and classified observations according to the estimated scores and compare classification with classification by true conditional average treatment effect. We repeated the classification over 1000 replications, and computed average proportion of the group over the replications.

In all the simulations, we analyzed data with 5-fold cross validation or sample splitting to avoid overfitting. When we use OS_1 approach with the cross validation, to estimate outcome scores of the treatment group, we built outcome model given covariates for all the controls and predicted risk of outcome of the treatment group from the model. Next, to estimate outcome scores of the control group, we divided control group into five subgroups of approximately equal size. Among them, we left one group and fitted outcome model on the remaining groups and obtained estimated outcome score of the one group left out from the model. We repeated it until all the groups had estimated outcome score, which is the typical procedure of k-fold cross validation. When we combined sample splitting with OS_1 approach, We split control group into two sets: model fitting set and score estimation set. we built outcome model for the model fitting set of control group and estimated the outcome score for all the others, not only for the score estimation set but also for treatment group. We adopted the conventional way of doing cross validation and sample splitting for the other approaches. To reduce computation time, we only conducted sample splitting method in the grid search design. All the statistical analyses were performed using R version 3.5.1 [13]

6 Simulation Results

6.1 Evaluating HTE at the pathological condition

Table 1.1 shows that the statistical power to detect HTE of the summary scores over selected data laws. The outcome score method OS_1 has very low statistical powers When the correlation coefficient between $m_0^*(X;\beta_0)$ and $\delta(X)$ is zero and OS_2 has very

low statistical powers (bold cases in the table) when the correlation coefficient between $s^*(X;\beta_2)$ and $\delta(X)$ is zero. The statistical power of the outcome score methods do not improve as sample size increases but the effect score methods improve. Even with sample size of 20,000 the outcome score methods fail to detect HTE at the pathological conditions, but effect scores methods always find HTE. The outcome score methods have better performance as the correlations coefficients are gradually increased from zero. Both test types, the test for interaction term (Int) and likelihood ratio test (LRT), show the same trends.

Table 1.2 reports the statistical power to detect HTE when the models to construct summary score is misspecified and under the pathological condition. When the model is correctly specified, the statistical power of the outcome score methods improves as sample size increases but this trend is not visible in the case of model misspecification; the power is very low even with large sample size. On the contrary, the model misspecification does not affect the performance of the effect scores, which can be explained by that the model is misspecified in main effects part where cancellation occurs in estimating effect scores. Model misspecification is very common in practice and summary score built from misspecified models still can detect HTE (Appendix 2). However, it cannot avoid the pathological conditions so we need to interpret results cautiously if summary score does not find HTE. We report the results of 5-fold cross validation in the main text, the results of sample splitting is similar and provided in appendix.

6.2 Grid search design

Figure 1.1 and 1.2 describe the results from the first grid search design, illustrating how the statistical power of the outcome score methods changes as the pathological condition (at the center of the plot) is relaxed. The figures show the same pattern as in the previous simulation. When the correlation is very low, the outcome score methods do not detect HTE and increasing sample size does not help. Even with 20,000 sample size, approximately 20% of the area in the figures indicates the statistical power less than 50%. This results strongly suggests that the outcome score methods fail not only at the pathological conditions but also when data laws are close enough to the pathological points.

Figure 1.3 and 1.4 summarize the findings of the big grid search design when test type is a log likelihood test and $n = 500, 10000$. Because of a huge number of data laws

we assessed, we randomly selected 50000 data laws and plotted power under the chosen laws. Also, we divided data laws into fifty groups by the correlation coefficients between $\delta(X)$ and each summary score where the cut points were [-1, -0.96, -0.92, -0.88, ..., 0.92, 0.96, 1] and connected the median (solid lines), 2.5 percentile, and 97.5 percentile (dashed lines) of each group. The figures support our findings; when the correlation coefficients are near zero, the outcome score methods had low statistical power. The effect score methods does not show this pattern.

6.3 Classification

Table 1.3 and 1.4 describe classification results at the pathological condition under positive treatment effect and table 1.5 and 1.6 show the case of the null treatment effect. We define misclassified cases as the cases where the classification by the true and estimated effect do not coincide and report the misclassification rate. Unlike the effect score methods, the misclassification rates of the outcome score methods do not improve anymore at certain points. Under the case of positive treatment effect, the outcome score methods determine that all the patients will benefit from treatment regardless of their baseline covariate values and under the null treatment effect, the misclassification rate of the outcome score methods converges to 0.5 as sample size increases. The effect score methods do not show this pattern. The findings suggest that under the pathological conditions the outcome score methods fail to correctly classify individuals but the effect score approach works well.

7 Application study: the ALLHAT trial

We apply the outcome score and effect score methods to ALLHAT trial data to detect HTE across the study population and classify the subjects according to the estimated conditional average treatment effect. ALLHAT trial is a randomized clinical trial sponsored by the National, Heart, Lung, and Blood Institute (NHLBI) in conjunction with the Department of Veterans' Affairs [14]. All the participants were hypertensive and aged 55 years or older and recruited from 623 centers over the United States, Canada, and Puerto Rico between 1994 February and January 1998. Total 33357 participants were enrolled into the study and follow-up ended on March 31, 2002 [15]. A detailed description of study protocol is provided elsewhere [14, 15].

The ALLHAT study consists of two components and four arms: Antihypertensive and lipid-lowering component; a diuretic (Chlorthalidone), a calcium antagonist (Amlodipine), an ACE inhibitor (Lisinopril), and an α-adrenergic blocker (Doxazosin). We focus on the antihypertensive component and compared the effect of diuretic, a calcium antagonist, and an ACE inhibitor on systolic blood pressure (SBP) 1 year after baseline; follow-up of the doxazosin arm was terminated earlier because Cholorthalidone turned out to be superior to Doxazosin before the end of the study [16, 17]. The analysis was adjusted for a few baseline covariates: SBP, age, sex, smoking status, diabetes prevalence, MI or stroke history, height, total cholesterol, left ventricular hypertrophy by electrocardiogram in past 2 years, using antihypertensive treatment. The covariates were chosen based on the previous study [18] and background knowledge. For simplicity, we excluded the missing cases and conducted complete case analysis; The original analysis [15] handled missing data the same way. We built the outcome and effect scores through 5-fold cross validation and sample splitting method and assessed HTE over the summary scores and classified patients by the estimated conditional average treatment effect. In the main text, we only present the results of 5-fold cross validation and provide the results of sampling splitting in the appendix.

Baseline characteristics of the study population and the distribution of each score are described in the appendix. Figure 1.5 (Amlodipine vs. Chlorthalidone) and figure 1.6(Lisinopril vs. Chlorthalidone) depict the smooth curves of the outcome over the summary scores by each arm and their point-wise confidence intervals (The smooth curves are made by generalized Additive Models using the default option of *geom_smooth* function of *ggplot2* package in R). Corresponding to the main findings of the ALLHAT study, Amlodipine group has higher SBP than Chlorthalidone and Lisinopril group. The original study assessed HTE by subgroup analysis but did not find any evidence of HTE (It is not clear whether they did subgroup analysis for SBP because SBP was not a primary outcome). In our current analysis, the smoothing curves intersect in figure 1.5, suggesting the treatment effect of Amlodipine compared to Chlothalidone might qualitatively different in sub-population. All the summary scores claim that HTE exists in the population (figure 1.7) and give similar classification (figure 1.8) in a comparison between Amlodipine and Chlothalidone; the classification among the scores is the same for 90% of the population. The smoothing curves and classification table over Lisinopril and Chlorthalidone group show that Chlorthalidone is almost always

effective in reducting SBP than Lisinopril (figure 1.6, 1.10). The effect score methods indicate quantitative treatment effect heterogeneity (figure 1.9) but the outcome score methods do not. Heterogeneity of treatment effects by effect score methods indicates HTE by covariate values in large sample by the theorem in the appendix and the failure to detect this HTE by outcome score methods can be interpreted as an inefficiency of the outcome score methods. The results of sample splitting method are similar and reported in appendix.

8 Conclusions

In this study, we propose that the outcome score methods can be misleading in HTE analysis under certain conditions and demonstrate poor performance of the outcome score methods. Under those pathological conditions, the outcome score methods do not find HTE even when there exists clinically significant effect heterogeneity, while the alternative approach, using the effect score, does not had the problem. The problem can arise not only at the exact pathological conditions but also in conditions close to the pathological ones depending on sample size, data laws and a choice of working models. We also consider estimating conditional average treatment effect by summary scores and classifying individuals by the estimated treatment effect. The outcome score methods fail in the correct classification of patients as well.

The reason of the failure is that the estimated outcome scores do not associate with the true treatment effects under the pathological conditions so do not give any information about the variability of treatment effects. The risk of outcome under control group is a part of treatment effects, like we need to figure baseline risk to calculate risk difference, so can give a partial information on the treatment effects but not a determinant of it. The outcome score methods can be useful for some research questions like how much the most vulnerable or unhealthy population can gain benefit from treatment. In that case, the outcome scores can classify population the way we want to do (high risk of outcome vs. low risk of outcome) and can answer the research questions properly. Furthermore, for many medical conditions, there exists rich background knowledge about predictors of the outcome in the absence of treatment or under standard-of-care treatments [4]. That makes model building of outcome score methods handy and produces accurate outcome score estimates. However, the ultimate purpose of HTE analysis is to identify a group of patients who receive the most or least benefit from the treatment

(high treatment effects vs. low treatment effects) and utilize this information for clinical purpose. If expected risk of outcome is a crucial factor in deciding treatment options, that could be absorbed into cost function [19], which allows them to be considered in HTE analysis. However, as long as the goal of analysis is to group patients by 'treatment effects', we should consider effect score methods first and be cautious of choosing the outcome score methods as a primary option for HTE analysis.

To make the specification of the pathological conditions easier and avoid unnecessary complications, we assume that data laws have only two normal random variables and are in the form of a linear function. In general, it is not obvious how to find a closed-form solutions of the zero correlation coefficient when data include non-linear forms. Also, if covariates are not generated from jointly multivariate normal distribution, zero correlation coefficient does not imply the independence between variables, so we also need to consider the case of non-linear dependence. That would make our argument unnecessarily complex. Apparently, possible data laws behind health-related outcomes would have much more complicated, including non-linear relationships, and determined by a massive number of factors. Also, investigators might be interested in binary or survival outcome, where the simple linear model is not the primary choice. According to circumstances, maybe much more complicated than our example, we may not encounter the exact pathological conditions or the pathological conditions might not exist. However, our simulation suggests the issue still arises outside of the pathological conditions and worsens as the situation is closer to the pathological conditions. However, the effect score methods have consistent performance regardless of chosen data laws within the same sample size. In practice, we do not know data laws so do not know how much outcome scores and the true conditional average treatment effects are correlated. Investigators, especially if they analyze data of small sample size, should not exclude the possibility that their model choice and the data law, are affected by this issue. In this case, they may need to consider using the alternative approaches that is free from the issue.

Instead of relying on ordinary regression models, investigators may use more flexible techniques, such as random forest [11], neural network or other machine learning techniques. The outcome scores from those approaches would also fail in the assessment of HTE and the classification of patients under the pathological condition because the problem is on inherent data structure not on the estimation techniques.

In conclusion, we demonstrate that when we encounter the certain pathological conditions or close enough to the conditions, the outcome score approach does not find the presence of HTE using statistical tests and does not properly classify individuals according to the true treatment effects. However, the alternative approach that aims at conditional average treatment effect does not have the same problem. We recommend against routine use of outcome score approach for HTE analysis in clinical trials because under certain situation even when there exists extreme effect heterogeneity the approach cannot find the heterogeneity.

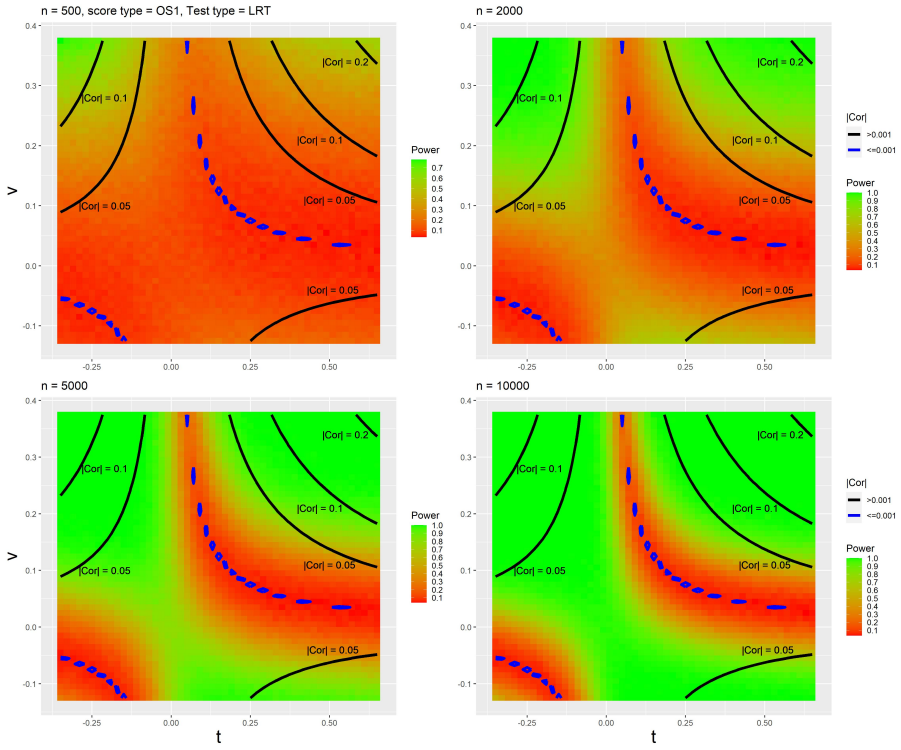


Figure 1.1: Heatmap plot for the statistical power of OS_1 from grid search design varying coeffieicnt a and c; test type: LRT

Table 1.1: Statistical power of detecting HTE by the different sample size (n) and the correlation between $m_0^*(X;\beta_0)$, $s^*(X;\beta_0)$ and $\delta(X)$ when 5-fold Cross validation was used: bold-faced numbers indicate the pathological conditions

			LRT				Int			
n	Cor[$m_0^*(X)$, $\delta(X)$]	Cor[$s^*(X)$, $\delta(X)$]	OS_1	OS_2	Two GLM	Tian	OS_1	OS_2	Two GLM	Tian
500	0.000	0.385	**0.104**	0.104	0.213	0.221	**0.17**	0.141	0.288	0.279
	0.200	0.577	0.151	0.135	0.221	0.217	0.243	0.214	0.29	0.281
	0.400	0.719	0.192	0.175	0.208	0.211	0.345	0.285	0.276	0.276
	0.600	0.828	0.272	0.204	0.2	0.194	0.447	0.368	0.275	0.282
	0.800	0.917	0.325	0.256	0.227	0.229	0.535	0.448	0.29	0.306
	-0.282	0.000	0.068	**0.043**	0.185	0.182	0.1	**0.068**	0.253	0.25
	-0.148	0.200	0.085	0.059	0.19	0.188	0.123	0.09	0.27	0.257
	0.014	0.400	0.088	0.081	0.194	0.201	0.173	0.123	0.27	0.278
	0.229	0.600	0.142	0.14	0.203	0.208	0.256	0.22	0.306	0.309
	0.543	0.800	0.238	0.2	0.196	0.189	0.416	0.332	0.279	0.283
2,000	0.000	0.385	**0.074**	0.2	0.721	0.71	**0.105**	0.327	0.826	0.818
	0.200	0.577	0.166	0.409	0.731	0.73	0.293	0.607	0.828	0.82
	0.400	0.719	0.273	0.568	0.733	0.733	0.489	0.797	0.834	0.839
	0.600	0.828	0.493	0.709	0.737	0.746	0.73	0.894	0.855	0.849
	0.800	0.917	0.709	0.808	0.735	0.731	0.909	0.944	0.844	0.846
	-0.282	0.000	0.113	**0.062**	0.725	0.69	0.169	**0.067**	0.83	0.812
	-0.148	0.200	0.087	0.098	0.728	0.719	0.113	0.166	0.85	0.838
	0.014	0.400	0.084	0.217	0.723	0.712	0.126	0.361	0.836	0.826
	0.229	0.600	0.15	0.419	0.711	0.697	0.272	0.623	0.822	0.813
	0.543	0.800	0.405	0.664	0.723	0.724	0.653	0.871	0.853	0.848
5,000	0.000	0.385	**0.062**	0.455	0.995	0.993	**0.087**	0.656	0.999	0.999
	0.200	0.577	0.334	0.932	0.997	0.996	0.181	0.779	0.995	0.994
	0.400	0.719	0.718	0.987	0.999	0.999	0.497	0.936	0.992	0.992
	0.600	0.828	0.937	1	0.997	0.997	0.819	0.981	0.993	0.992
	0.800	0.917	0.979	0.998	0.994	0.998	0.997	1	1	1
	-0.282	0.000	0.243	**0.05**	0.995	0.994	0.355	**0.043**	0.999	0.999
	-0.148	0.200	0.099	0.162	0.998	0.996	0.141	0.273	0.999	1
	0.014	0.400	0.08	0.473	0.999	0.999	0.103	0.687	1	1
	0.229	0.600	0.23	0.849	0.995	0.992	0.384	0.958	0.999	0.999
	0.543	0.800	0.751	0.98	0.995	0.992	0.916	0.998	0.999	0.999
10,000	0.000	0.385	**0.075**	0.745	1	1	**0.1**	0.896	1	1
	0.200	0.577	0.254	0.982	1	1	0.48	1	1	1
	0.400	0.719	0.757	1	1	1	0.906	1	1	1
	0.600	0.828	0.988	1	1	1	0.999	1	1	1
	0.800	0.917	0.999	1	1	1	1	1	1	1
	-0.282	0.000	0.46	**0.043**	1	1	0.675	**0.059**	1	1
	-0.148	0.200	0.149	0.255	0.999	0.999	0.242	0.463	1	1
	0.014	0.400	0.074	0.783	1	1	0.095	0.93	1	1
	0.229	0.600	0.398	0.984	1	1	0.568	1	1	1
	0.543	0.800	0.95	1	1	1	0.996	1	1	1
20,000	0.000	0.385	**0.082**	0.971	1	1	**0.107**	0.999	1	1
	0.200	0.577	0.48	1	1	1	0.694	1	1	1
	0.400	0.719	0.964	1	1	1	0.995	1	1	1
	0.600	0.828	1	1	1	1	1	1	1	1
	0.800	0.917	1	1	1	1	1	1	1	1
	-0.282	0.000	0.757	**0.062**	1	1	0.907	**0.066**	1	1
	-0.148	0.200	0.265	0.467	1	1	0.425	0.71	1	1
	0.014	0.400	0.062	0.967	1	1	0.096	0.994	1	1
	0.229	0.600	0.597	1	1	1	0.797	1	1	1
	0.543	0.800	0.999	1	1	1	1	1	1	1

Table 1.2: Statistical power of the different methods for detecting HTE when the model is correctly specified or misspecified when 5-fold Cross validation was used: bold-faced numbers indicate the pathological conditions

			OS_1		OS_2		Two GLM		Tian	
n	Scenario	Model	LRT	Int	LRT	Int	LRT	Int	LRT	Int
500	1	C	0.096	0.128	0.092	0.105	0.208	0.271	0.198	0.26
		M	**0.104**	**0.178**	0.087	0.159	0.157	0.201	0.219	0.318
	2	C	0.077	0.095	0.084	0.066	0.205	0.269	0.187	0.27
		M	0.071	0.1	**0.051**	**0.076**	0.156	0.202	0.206	0.286
2000	1	C	0.134	0.079	0.204	0.215	0.68	0.795	0.628	0.727
		M	0.069	0.13	0.202	0.351	0.415	0.553	0.662	0.788
	2	C	0.155	0.131	0.12	0.067	0.652	0.766	0.584	0.682
		M	0.09	0.149	**0.056**	**0.065**	0.399	0.494	0.629	0.739
5000	1	C	0.323	0.084	0.475	0.423	0.998	0.999	0.985	0.995
		M	**0.083**	**0.11**	0.419	0.629	0.821	0.912	0.992	0.998
	2	C	0.399	0.292	0.306	0.055	0.992	1	0.984	0.996
		M	0.244	0.382	**0.049**	**0.062**	0.832	0.897	0.989	0.997
10000	1	C	0.617	0.077	0.79	0.712	1	1	1	1
		M	**0.083**	**0.101**	0.73	0.884	0.996	0.996	1	1
	2	C	0.729	0.506	0.551	0.067	1	1	0.999	1
		M	0.422	0.638	**0.055**	**0.082**	0.988	0.992	1	1
20000	1	C	0.92	0.075	0.976	0.923	1	1	1	1
		M	**0.061**	**0.089**	0.957	0.993	1	1	1	1
	2	C	0.962	0.795	0.86	0.065	1	1	1	1
		M	0.747	0.899	**0.05**	**0.077**	1	1	1	1

Scenario 1: The pathological condition is realized with misspecified model for OS_1.
Scenario 2: The pathological condition is realized with misspecified model for OS_2.
C, Correctly specified model; M, Misspecified model omitting the X^2 term.

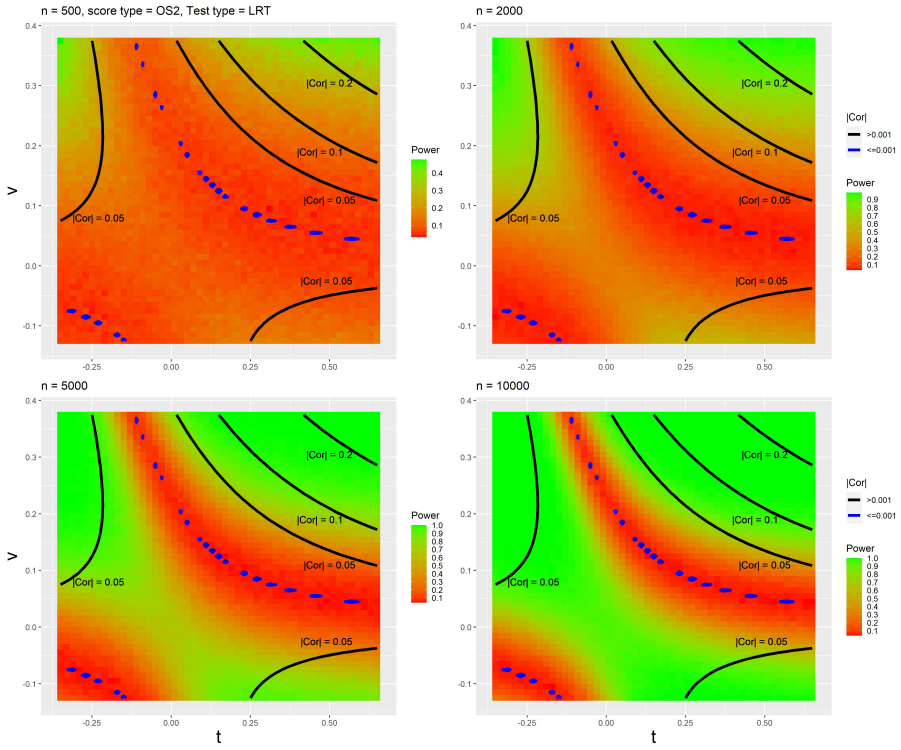


Figure 1.2: Heatmap plot for the statistical power of OS_2 from grid search design varying coeffieicnt a and c; test type: LRT.

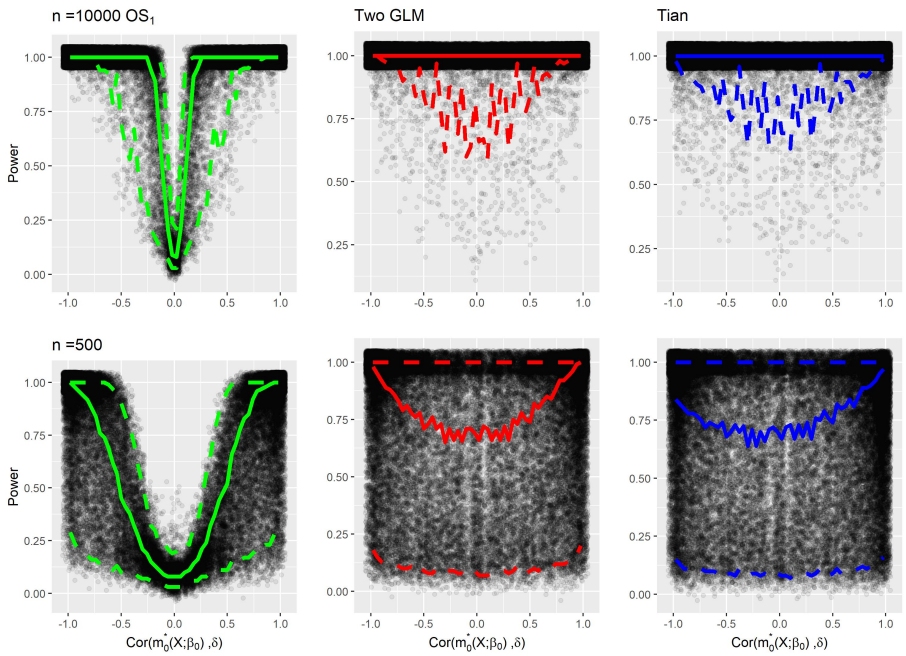


Figure 1.3: A scatter plot of the statistical power and the correlation coefficient for OS_1 in big grid search design when $n = 500$ and 10000. The solid line connects the median of the power for each 50 correlation coefficient group and the dashed lines connect 2.5 percentile and 97.5 percentile of the power

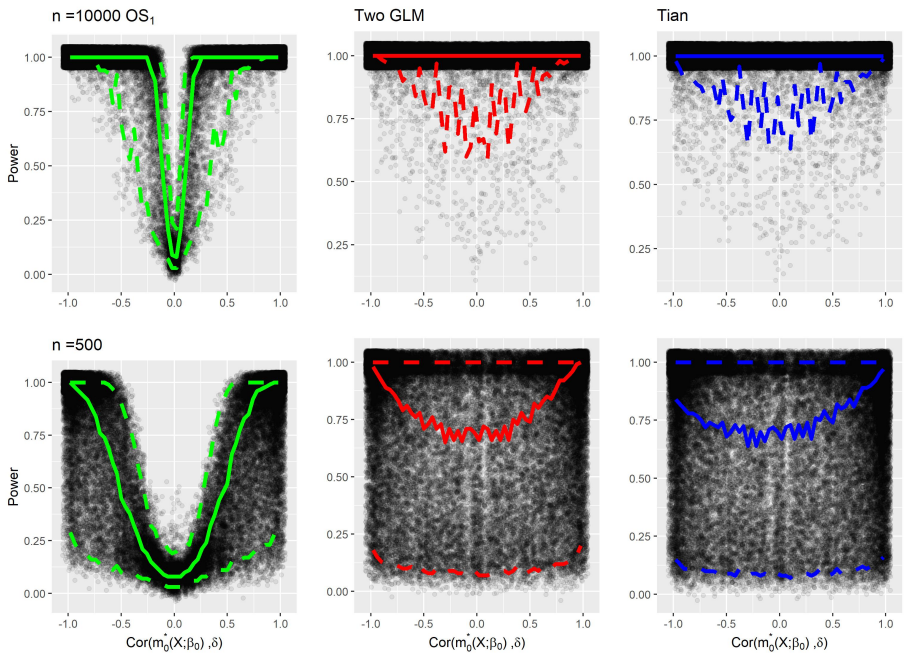


Figure 1.4: A scatter plot of the statistical power and the correlation coefficient for OS_2 in big grid search design when $n = 500$ and 10000. The solid line connects the median of the power for each 50 correlation coefficient group and the dashed lines connect 2.5 percentile and 97.5 percentile of the power

Table 1.3: Classification by estimted and true conditional average treatment effect the coefficient of A, is 0.15 and when Cor$[m_0^*(X), \delta(X)] = 0$ and 5-fold Cross validation was used

			OS_1 Est $\delta(X)$		OS_2 Est $\delta(X)$		Two GLM Est $\delta(X)$		Tian Est $\delta(X)$	
n	Proportion		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
500	True $\delta(X)$	(+)	0.709	0.093	0.720	0.082	0.718	0.084	0.717	0.085
		(-)	0.165	0.033	0.159	0.039	0.121	0.077	0.123	0.075
	Misclassification rate		0.258		0.241		0.205		0.208	
2000	True $\delta(X)$	(+)	0.786	0.016	0.771	0.030	0.770	0.032	0.769	0.032
		(-)	0.193	0.006	0.176	0.022	0.079	0.119	0.082	0.116
	Misclassification rate		0.208		0.206		0.111		0.114	
5000	True $\delta(X)$	(+)	0.800	0.002	0.784	0.018	0.778	0.024	0.778	0.024
		(-)	0.197	0.001	0.183	0.015	0.045	0.153	0.047	0.152
	Misclassification rate		0.199		0.201		0.069		0.071	
10000	True $\delta(X)$	(+)	0.802	0.000	0.789	0.013	0.784	0.018	0.784	0.018
		(-)	0.198	0.000	0.186	0.012	0.029	0.169	0.030	0.168
	Misclassification rate		0.198		0.199		0.047		0.048	
20000	True $\delta(X)$	(+)	0.802	0.000	0.793	0.009	0.789	0.013	0.788	0.014
		(-)	0.198	0.000	0.189	0.010	0.020	0.178	0.020	0.178
	Misclassification rate		0.198		0.198		0.033		0.034	

(+): Positive treatment effect, (-): negative treatment effect

Table 1.4: Classification by estimted and true conditional average treatment effect when the coefficient of A, is 0.15 and and Cor$[s^*(X), \delta(X)] = 0$ when 5-fold Cross validation was used

			OS_1 Est $\delta(X)$		OS_2 Est $\delta(X)$		Two GLM Est $\delta(X)$		Tian Est $\delta(X)$	
n	Proportion		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
500	True $\delta(X)$	(+)	0.716	0.085	0.721	0.080	0.716	0.085	0.709	0.092
		(-)	0.169	0.030	0.176	0.023	0.126	0.073	0.129	0.070
	Misclassification rate		0.254		0.256		0.211		0.221	
2000	True $\delta(X)$	(+)	0.781	0.021	0.790	0.012	0.768	0.034	0.768	0.034
		(-)	0.186	0.012	0.195	0.003	0.079	0.119	0.084	0.114
	Misclassification rate		0.207		0.207		0.113		0.118	
5000	True $\delta(X)$	(+)	0.793	0.009	0.801	0.001	0.778	0.024	0.778	0.024
		(-)	0.192	0.006	0.198	0.000	0.045	0.152	0.048	0.150
	Misclassification rate		0.201		0.199		0.070		0.073	
10000	True $\delta(X)$	(+)	0.798	0.004	0.802	0.000	0.783	0.019	0.782	0.020
		(-)	0.195	0.003	0.198	0.000	0.029	0.169	0.031	0.167
	Misclassification rate		0.199		0.198		0.048		0.050	
20000	True $\delta(X)$	(+)	0.799	0.003	0.802	0.000	0.789	0.013	0.788	0.014
		(-)	0.196	0.002	0.198	0.000	0.020	0.178	0.021	0.177
	Misclassification rate		0.199		0.198		0.034		0.035	

(+): Positive treatment effect, (-): negative treatment effect

Table 1.5: Classification by estimted and true conditional average treatment effect when the coefficient of A, is 0 and when $\text{Cor}[m_0^*(X), \delta(X)] = 0$ and 5-fold Cross validation was used

n	Proportion		OS_1 Est $\delta(X)$ (+)	OS_1 Est $\delta(X)$ (-)	OS_2 Est $\delta(X)$ (+)	OS_2 Est $\delta(X)$ (-)	Two GLM Est $\delta(X)$ (+)	Two GLM Est $\delta(X)$ (-)	Tian Est $\delta(X)$ (+)	Tian Est $\delta(X)$ (-)
500	True $\delta(X)$	(+)	0.258	0.242	0.276	0.224	0.319	0.181	0.317	0.184
		(-)	0.232	0.268	0.223	0.276	0.180	0.320	0.183	0.317
	Misclassification rate		0.474		0.447		0.361		0.366	
2000	True $\delta(X)$	(+)	0.266	0.235	0.302	0.198	0.427	0.074	0.425	0.076
		(-)	0.250	0.249	0.212	0.287	0.083	0.416	0.086	0.414
	Misclassification rate		0.486		0.411		0.157		0.162	
5000	True $\delta(X)$	(+)	0.255	0.245	0.304	0.196	0.459	0.041	0.458	0.042
		(-)	0.246	0.254	0.196	0.304	0.042	0.458	0.044	0.456
	Misclassification rate		0.491		0.393		0.083		0.086	
10000	True $\delta(X)$	(+)	0.260	0.240	0.311	0.189	0.472	0.028	0.471	0.029
		(-)	0.254	0.246	0.193	0.307	0.029	0.471	0.030	0.470
	Misclassification rate		0.494		0.382		0.057		0.058	
20000	True $\delta(X)$	(+)	0.247	0.253	0.310	0.190	0.480	0.020	0.479	0.020
		(-)	0.243	0.257	0.187	0.313	0.019	0.481	0.020	0.480
	Misclassification rate		0.495		0.377		0.039		0.040	

(+): Positive treatment effect, (-): negative treatment effect

Table 1.6: Classification by estimted and true conditional average treatment effect when the coefficient of A, is 0 and when $\text{Cor}[s^*(X), \delta(X)] = 0$ and 5-fold Cross validation was used

n	Proportion		OS_1 Est $\delta(X)$ (+)	OS_1 Est $\delta(X)$ (-)	OS_2 Est $\delta(X)$ (+)	OS_2 Est $\delta(X)$ (-)	Two GLM Est $\delta(X)$ (+)	Two GLM Est $\delta(X)$ (-)	Tian Est $\delta(X)$ (+)	Tian Est $\delta(X)$ (-)
500	True $\delta(X)$	(+)	0.263	0.237	0.258	0.242	0.314	0.186	0.309	0.191
		(-)	0.238	0.261	0.249	0.251	0.187	0.313	0.195	0.305
	Misclassification rate		0.476		0.491		0.374		0.386	
2000	True $\delta(X)$	(+)	0.281	0.219	0.259	0.241	0.425	0.075	0.420	0.080
		(-)	0.235	0.265	0.255	0.245	0.083	0.418	0.088	0.413
	Misclassification rate		0.454		0.496		0.158		0.167	
5000	True $\delta(X)$	(+)	0.285	0.215	0.253	0.247	0.460	0.041	0.457	0.043
		(-)	0.219	0.281	0.250	0.250	0.043	0.457	0.045	0.455
	Misclassification rate		0.433		0.497		0.083		0.088	
10000	True $\delta(X)$	(+)	0.289	0.211	0.247	0.253	0.470	0.030	0.468	0.031
		(-)	0.210	0.290	0.245	0.255	0.028	0.472	0.029	0.471
	Misclassification rate		0.421		0.498		0.058		0.061	
20000	True $\delta(X)$	(+)	0.293	0.207	0.252	0.248	0.480	0.020	0.479	0.021
		(-)	0.207	0.293	0.251	0.249	0.020	0.480	0.021	0.479
	Misclassification rate		0.414		0.499		0.040		0.042	

(+): Positive treatment effect, (-): negative treatment effect

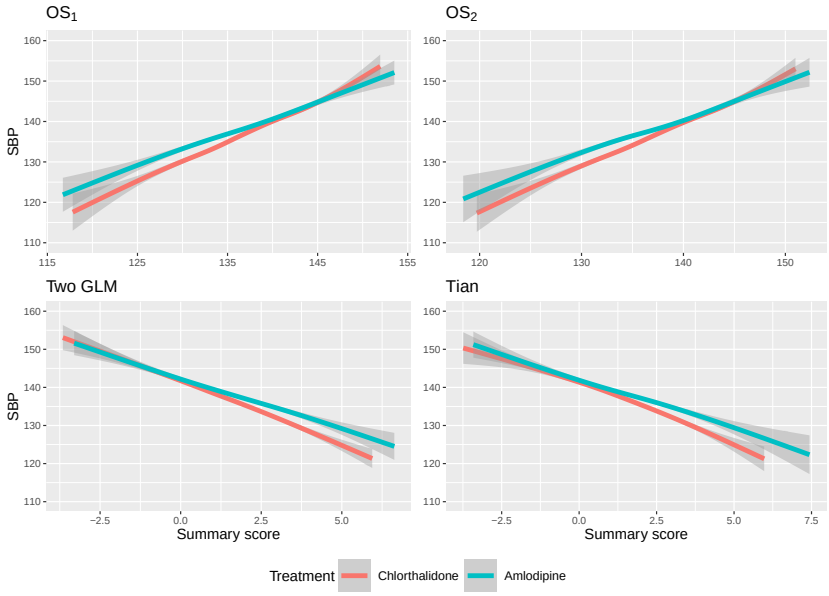


Figure 1.5: Smoothing curves of SBP over the scores by Amlodipine and Chlorthalidone group (5-fold CV)

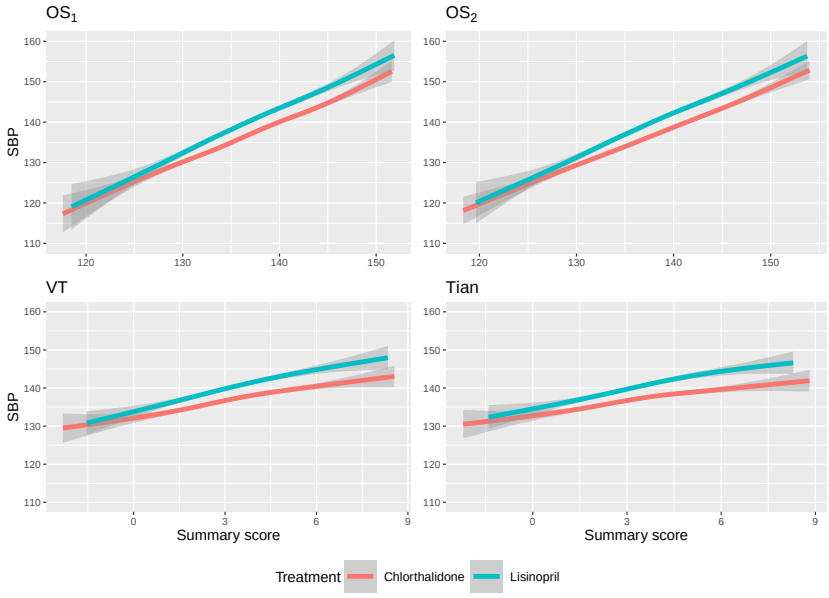


Figure 1.6: Smoothing curves of SBP over the scores by Lisinopril and Chlorthalidone group (5-fold CV)

Table 1.7: P-value of the statistical tests to detect HTE in ALLHAT study between Amlodipine and Chlorthalidone (5-folds CV)

Type of test	OS_1	OS_2	Two GLM	Tian
LRT	0	0	0	0.001
Int	0	0	0	0

Table 1.8: Classification of ALLHAT data by the linear model between Amlodipine and Chlorthalidone (5-fold CV)

		OS_2		Two GLM				Tian	
Proportion		(+)	(-)	(+)	(-)	Proportion		(+)	(-)
OS_1	(+)	0.902	0.008	0.898	0.012	OS_1	(+)	0.899	0.011
	(-)	0.003	0.087	0.047	0.043		(-)	0.065	0.025
		Two GLM		Tian				Tian	
Proportion		(+)	(-)	(+)	(-)	Proportion		(+)	(-)
OS_2	(+)	0.892	0.014	0.893	0.012	Two GLM	(+)	0.938	0.007
	(-)	0.053	0.042	0.070	0.025		(-)	0.026	0.030

Table 1.9: P-value of the statistical tests to detect HTE in ALLHAT study between Lisinopril and Chlorthalidone (5-folds CV)

Type of test	OS_1	OS_2	Two GLM	Tian
LRT	0.204	0.112	0.011	0.014
Int	0.024	0.014	0.008	0.006

Table 1.10: Classification of ALLHAT data by the linear model by Lisinopril and Chlorthalidone group

		OS_2		Two GLM				Tian	
Proportion		(+)	(-)	(+)	(-)	Proportion		(+)	(-)
OS_1	(+)	1	0	1	0	OS_1	(+)	1	0
	(-)	0	0	0	0		(-)	0	0
		Two GLM		Tian				Tian	
Proportion		(+)	(-)	(+)	(-)	Proportion		(+)	(-)
OS_2	(+)	1	0	1	0	Two GLM	(+)	1	0
	(-)	0	0	0	0		(-)	0	0

Chapter 1 Appendix

A Modeling the outcome across treatment groups, ignoring treatment assignment

These variants of the outcome score approach involve modeling

$$\mathrm{E}[Y|X] = \sum_{a=0}^{1} \mathrm{E}[Y|X, A=a] \Pr[A=a|X],$$

instead of $\mathrm{E}[Y|X, A=0]$. Note that if the trial is marginally randomized with treatment probabilities $\Pr[A=a]$ for $a \in \{0,1\}$, then $\mathrm{E}[Y|X] = \sum_{a=0}^{1} \mathrm{E}[Y|X, A=a] \Pr[A=a]$. By the law of total expectation

$$\begin{aligned} E[Y|X] &= \sum_{a} E[Y|X, A=a] \Pr[A=a|X] \\ &= m_1(X;\beta_1)e(X) + m_0(X;\beta_0)(1-e(X)) \\ &= m_0(X;\beta_0) + \big\{m_1(X;\beta_1) - m_0(X;\beta_0)\big\}e(X) \\ &= m_0(X;\beta_0) + \delta(X)e(X). \end{aligned}$$

This result suggests that consistent estimators of $E[Y|X]$ will converge to the limit different from $m_0(X;\beta_0)$ or $\delta(X)$.

B A theorem about summary score: Heterogeneity of treatment effects by summary score implies Heterogeneity of treatment effects by covariate values

Define the conditional average treatment effect given that $S = s$, $\rho(s)$, as

$$\rho(s) \equiv \mathrm{E}[Y^1 - Y^0 | S = s]$$

and recall the definition of $\delta(x)$ (section 2.1) as

$$\delta(x) \equiv \mathrm{E}[Y^1 - Y^0 | X = x]$$

Let X be a high-dimensional covariate vector and S be an one dimensional functional of X, $S = g(X)$, where g is an arbitrary many-to-one function.

We will now state a theorem that formalizes the following idea: if there exists heterogeneity of treatment effects over the summary score S, then there exists heterogeneity of treatment effects over the covariates (as defined in section 2.1).

Theorem 0.1. *If s_1 and s_2 are two values of the summary score, such that $s_1 \neq s_2$ and $\rho(s_1) \neq \rho(s_2)$, then there exist two values x_1 and x_2, such that $x_1 \neq x_2$ and $\delta(x_1) \neq \delta(x_2)$.*

Proof. Then, the statement can be expressed as

$$\exists s_1 \in S, \exists s_2 \in S \text{ such that } s_1 \neq s_2, \rho(s_1) \neq \rho(s_2) \implies \exists x_1 \in X, \exists x_2 \in X \text{ such that } x_1 \neq x_2,\ \delta(x_1) \neq \delta(x_2) \tag{3}$$

We prove this statement by showing that the contrapositive is true,

$$\forall x_1 \in X, \forall x_2 \in X, x_1 = x_2 \text{ or } \delta(x_1) = \delta(x_2) \implies \forall s_1 \in S, \forall s_2 \in S, s_1 = s_2 \text{ or } \rho(s_1) = \rho(s_2) \tag{4}$$

The above statement is true if $x_1 = x_2$, then either $s_1 = s_2$ or $\rho(s_1) = \rho(s_2)$ is true and if $\delta(x_1) = \delta(x_2)$, then either $s_1 = s_2$ or $\rho(s_1) = \rho(s_2)$ is true. We prove that by showing the following both statements are true.

$$\forall x_1 \in X, \forall x_2 \in X, x_1 = x_2 \implies \forall s_1 \in S, \forall s_2 \in S, s_1 = s_2$$

$$\forall x_1 \in X, \forall x_2 \in X, \delta(x_1) = \delta(x_2) \implies \forall s_1 \in S, \forall s_2 \in S, \rho(s_1) = \rho(s_2)$$

The first statement: If every $x \in X$ is equal, X has only one element and necessarily $S = g(X)$ should have only one element and every element of S is equal.

The second statement: $\delta(X)$ is the same for every $x \in X$ means $\delta(X)$ has only one element and no variation. By iterative expectation,

$$\mathrm{E}[Y^1 - Y^0|S] = \mathrm{E}[\mathrm{E}[Y^1 - Y^0|X, S]|S] = \mathrm{E}[\mathrm{E}[Y^1 - Y^0|X]|S] = \mathrm{E}[\delta(X)|S]$$

The second equality follows from that $\mathrm{E}[Y^1 - Y^0|X, g(X)] = \mathrm{E}[Y^1 - Y^0|X]$. We deduced $\delta(X)$ has no variation; therefore, $\rho(S) = \mathrm{E}[\delta(X)|S]$ is also a constant and the same for every $s \in S$.

By presenting the both statements are true, we show that (4) is true.

□

It implies if any summary score based on baseline covariates finds an evidence of the existence of HTE, we can faithfully interpret the results as HTE exists; especially the evidence is concrete in large sample studies. The results hold with any one dimensional functional so it help interpret the second simulation. Regardless of whether summary score is built with a correctly specified model or a misspecified model, if treatment effects are heterogeneous by summary score, there is an evidence of the existence of HTE. The second simulation scenario doing HTE analysis with mis-specified model is not absurd in that sense; rather that would reflect real world practice if we consider model misspecification as the norm of real data analysis. However, that does not mean that we can do HTE analysis with any summary score. Some summary score may report the absence of HTE when clinically significant HTE exists; this study focuses on this case.

C The data laws used in the simulation

Table 1.11: The data laws used in the simulation

Simulation 1		
$\text{Cor}[g_0^*(X),\, \delta(X)]$	0	$Y = 0.15 * X_1 + 0.15 * X_2 + 0.15 * A + 0.125 * AX_1 - 0.125AX_2 + \epsilon$
	0.2	$Y = 0.15 * X_1 + 0.09914375 * X_2 + 0.15 * A + 0.125 * AX_1 - 0.125AX_2 + \epsilon$
	0.4	$Y = 0.15 * X_1 + 0.05885027 * X_2 + 0.15 * A + 0.125 * AX_1 - 0.125AX_2 + \epsilon$
	0.6	$Y = 0.15 * X_1 + 0.02142857 * X_2 + 0.15 * A + 0.125 * AX_1 - 0.125AX_2 + \epsilon$
	0.8	$Y = 0.15 * X_1 - 0.02142857 * X_2 + 0.15 * A + 0.125 * AX_1 - 0.125AX_2 + \epsilon$
$\text{Cor}[s^*(X),\, \delta(X)]$	0	$Y = 0.15 * X_1 + 0.275 * X_2 + 0.15 * A + 0.125 * AX_1 - 0.125AX_2 + \epsilon$
	0.2	$Y = 0.15 * X_1 + 0.20295 * X_2 + 0.15 * A + 0.125 * AX_1 - 0.125AX_2 + \epsilon$
	0.4	$Y = 0.15 * X_1 + 0.14587 * X_2 + 0.15 * A + 0.125 * AX_1 - 0.125AX_2 + \epsilon$
	0.6	$Y = 0.15 * X_1 + 0.09286 * X_2 + 0.15 * A + 0.125 * AX_1 - 0.125AX_2 + \epsilon$
	0.8	$Y = 0.15 * X_1 + 0.03214 * X_2 + 0.15 * A + 0.125 * AX_1 - 0.125AX_2 + \epsilon$
Simulation 2		
Data law 1		$Y = 0.15X_1 + 0.15X_1^2 + 0.15X_2 + 0.15A + 0.125AX_1 - 0.125AX_2 + \epsilon$
Data law 2		$Y = 0.15X_1 + 0.15X_1^2 + 0.275X_2 + 0.15A + 0.125AX_1 - 0.125AX_2 + \epsilon$
Grid search design 1		
$\text{Cor}[m_0^*(X),\, \delta(X)]$		$Y = tX_1 + 0.15X_2 + 0.15A + vAX_1 - 0.125AX_2 + \epsilon$, where $t \in [-0.35, -0.33, ..., 0.63, 0.65]$, $v \in [-0.125, -0.115, ..., 0.365, 0.375]$
$\text{Cor}[s^*(X),\, \delta(X)]$		$Y = tX_1 + 0.275X_2 + 0.15A + vAX_1 - 0.125AX_2 + \epsilon$, where $t \in [-0.35, -0.33, ..., 0.63, 0.65]$, $v \in [-0.125, -0.115, ..., 0.365, 0.375]$
Grid search design 2		
		$Y = tX_1 + uX_2 + \alpha A + vAX_1 + wAX_2 + \epsilon$
		$t \in [-2, -1.5, -1, ..., 1.5, 2]$, $u \in [-2, -1.5, -1, ..., 1.5, 2]$
		$v \in [-1, -0.75, -0.5, ..., 0.75, 1]$, $w \in [-1, -0.75, -0.5, ..., 0.75, 1]$
		$\text{Var}[X_1] \in [0.5, 1, 2]$, $\text{Var}[X_2] \in [0.5, 1, 2]$, $\text{Cor}[X_1, X_2] \in [-0.8, -0.5, -0.2, 0, 0.2, 0.5, 0.8]$
Classification		
$\text{Cor}[m_0^*(X),\, \delta(X)]$		$Y = 0.15X_1 + 0.15X_2 + 0.15A + 0.125AX_1 - 0.125AX_2 + \epsilon$ $Y = 0.15X_1 + 0.15X_2 + 0.125AX_1 - 0.125AX_2 + \epsilon$
$\text{Cor}[s^*(X),\, \delta(X)]$		$Y = 0.15X_1 + 0.275X_2 + 0.15A + 0.125AX_1 - 0.125AX_2 + \epsilon$ $Y = 0.15X_1 + 0.275X_2 + 0.125AX_1 - 0.125AX_2 + \epsilon$

In all scenarios, $\epsilon \sim N(0, 1)$

D Simulation results from sample splitting method

Table 1.13: Statistical power of the different methods for detecting HTE when correctly spdecified misspecified models are used with sample splitting method

			OS_1		OS_2		Two GLM		Tian	
n	Data Law	Model	LRT	Int	LRT	Int	LRT	Int	LRT	Int
500	1	C	0.057	0.086	0.078	0.079	0.105	0.174	0.1	0.169
		M	0.067	0.117	0.063	0.108	0.116	0.203	0.121	0.199
	2	C	0.066	0.082	0.056	0.064	0.121	0.172	0.105	0.168
		M	0.072	0.107	0.075	0.079	0.123	0.205	0.114	0.196
2000	1	C	0.117	0.081	0.138	0.134	0.43	0.627	0.391	0.579
		M	0.076	0.133	0.138	0.235	0.467	0.666	0.45	0.654
	2	C	0.122	0.11	0.072	0.072	0.42	0.621	0.377	0.574
		M	0.098	0.162	0.058	0.083	0.436	0.66	0.421	0.646
5000	1	C	0.243	0.084	0.232	0.22	0.879	0.966	0.86	0.947
		M	0.089	0.128	0.231	0.385	0.896	0.978	0.893	0.975
	2	C	0.293	0.234	0.155	0.059	0.884	0.974	0.864	0.948
		M	0.18	0.339	0.063	0.076	0.893	0.977	0.893	0.972
10000	1	C	0.46	0.073	0.473	0.411	0.999	1	0.997	1
		M	0.077	0.133	0.423	0.642	0.998	1	0.999	1
	2	C	0.545	0.384	0.293	0.059	0.998	1	0.995	1
		M	0.304	0.486	0.066	0.085	0.997	1	0.995	1
20000	1	C	0.772	0.075	0.785	0.703	1	1	1	1
		M	0.077	0.112	0.73	0.907	1	1	1	1
	2	C	0.855	0.633	0.553	0.052	1	1	1	1
		M	0.559	0.78	0.055	0.054	1	1	1	1

Data law 1: $Y = 0.15X_1 + 0.15X_2 + 0.15X_1^2 + 0.15A + 0.125AX_1 - 0.125AX_2$

Data law 2: $Y = 0.15X_1 + 0.275X_2 + 0.15X_1^2 + 0.15A + 0.125AX_1 - 0.125AX_2$

C, Correctly specified model; M, Misspecified model omitting the X^2 term.

Table 1.14: Classification by estimated and true conditional average treatment effect the coefficient of A, is 0.15 and when $\text{Cor}[m_0^*(X;\beta_0),\ \delta(X)] = 0$ and sample splitting method was used

			OS_1Est $\delta(X)$		OS_2 Est $\delta(X)$		Two GLM Est $\delta(X)$		Tian Est $\delta($	
n	Proportion		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-
500	True $\delta(X)$	(+)	0.691	0.109	0.668	0.133	0.684	0.117	0.683	0.1
		(-)	0.155	0.044	0.146	0.053	0.103	0.096	0.106	0.0
	Misclassification rate		0264		0.279		0.221		0.223	
2000	True $\delta(X)$	(+)	0.774	0.028	0.750	0.052	0.747	0.055	0.747	0.0
		(-)	0.187	0.010	0.168	0.029	0.068	0.130	0.070	0.1
	Misclassification rate		0.215		0.221		0.123		0.126	
5000	True $\delta(X)$	(+)	0.796	0.006	0.773	0.029	0.761	0.040	0.761	0.0
		(-)	0.196	0.002	0.177	0.022	0.042	0.157	0.043	0.1
	Misclassification rate		0.202		0.205		0.082		0.083	
10000	True $\delta(X)$	(+)	0.800	0.001	0.785	0.016	0.776	0.026	0.775	0.0
		(-)	0.198	0.000	0.184	0.014	0.030	0.169	0.031	0.1
	Misclassification rate		0.199		0.201		0.056		0.058	
20000	True $\delta(X)$	(+)	0.802	0.000	0.789	0.013	0.782	0.020	0.782	0.0
		(-)	0.198	0.000	0.186	0.012	0.022	0.176	0.022	0.1
	Misclassification rate		0.198		0.199		0.041		0.042	

(+): Positive treatment effect, (-): negative treatment effect

Table 1.15: Classification by estimated and true conditional average treatment effect the coefficient of A, is 0.15 and when $\text{Cor}[s^*(X;\beta_2),\ \delta(X)] = 0$ and sample splitting method was used

			OS_1Est $\delta(X)$		OS_2Est$\delta(X)$		Two GLM Est $\delta(X)$		Tian Est $\delta(X)$	
n	Proportion		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
500	True $\delta(X)$	(+)	0.688	0.114	0.653	0.149	0.671	0.131	0.668	0.134
		(-)	0.155	0.043	0.157	0.041	0.100	0.098	0.102	0.096
	Misclassification rate		0.269		0.306		0.231		0.236	
2000	True $\delta(X)$	(+)	0.765	0.036	0.764	0.038	0.744	0.058	0.744	0.058
		(-)	0.180	0.019	0.187	0.011	0.067	0.131	0.071	0.128
	Misclassification rate		0.216		0.225		0.124		0.128	
5000	True $\delta(X)$	(+)	0.787	0.015	0.793	0.009	0.761	0.041	0.761	0.041
		(-)	0.188	0.010	0.195	0.002	0.040	0.158	0.043	0.155
	Misclassification rate		0.204		0.204		0.081		0.084	
10000	True $\delta(X)$	(+)	0.794	0.008	0.800	0.001	0.775	0.027	0.774	0.028
		(-)	0.193	0.006	0.198	0.000	0.033	0.165	0.035	0.163
	Misclassification rate		0.201		0.199		0.060		0.063	
20000	True $\delta(X)$	(+)	0.798	0.004	0.802	0.000	0.781	0.021	0.780	0.021
		(-)	0.195	0.003	0.198	0.000	0.021	0.177	0.022	0.176
	Misclassification rate		0.199		0.198		0.042		0.043	

(+): Positive treatment effect, (-): negative treatment effect

Table 1.16: Classification by estimated and true conditional average treatment effect the coefficient of A, is 0 and when $\text{Cor}[s^*(X;\beta_2),\ \delta(X)] = 0$ and sample splitting method was used

			OS_1Est$\delta(X)$		OS_2Est$\delta(X)$		Two GLM Est $\delta(X)$		Tian Est $\delta(X)$	
n	Proportion		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
500	True $\delta(X)$	(+)	0.273	0.225	0.274	0.224	0.332	0.166	0.330	0.168
		(-)	0.226	0.274	0.226	0.276	0.172	0.330	0.176	0.326
	Misclassification rate		0.452		0.450		0.339		0.344	
2000	True $\delta(X)$	(+)	0.256	0.244	0.282	0.218	0.409	0.091	0.406	0.094
		(-)	0.235	0.265	0.210	0.290	0.085	0.415	0.086	0.414
	Misclassification rate		0.479		0.427		0.176		0.180	
5000	True $\delta(X)$	(+)	0.252	0.248	0.293	0.207	0.445	0.055	0.444	0.055
		(-)	0.236	0.265	0.198	0.302	0.050	0.450	0.051	0.449
	Misclassification rate		0.484		0.405		0.105		0.106	
10000	True $\delta(X)$	(+)	0.258	0.242	0.300	0.200	0.465	0.035	0.463	0.036
		(-)	0.247	0.252	0.193	0.307	0.033	0.467	0.035	0.466
	Misclassification rate		0.489		0.394		0.068		0.071	
20000	True $\delta(X)$	(+)	0.248	0.252	0.308	0.192	0.474	0.026	0.473	0.027
		(-)	0.240	0.260	0.191	0.309	0.025	0.475	0.026	0.474
	Misclassification rate		0.492		0.383		0.052		0.053	

(+): Positive treatment effect, (-): negative treatment effect

Table 1.17: Classification by the linear model of outcome Y given treatment and risk and effect scores allowing the interaction when average true treatment effect is 0 and when Cor[$s^*(X)$, $\delta(X)$] = 0 and sample splitting method was used

			OS_1Est$\delta(X)$		OS_2Est$\delta(X)$		Two GLM Est $\delta(X)$		Tian Est $\delta(X)$	
n	Proportion		(+)	(-)	(+)	(-)	(+)	(-)	(+)	(-)
500	True $\delta(X)$	(+)	0.270	0.232	0.255	0.247	0.330	0.172	0.329	0.173
		(-)	0.227	0.272	0.239	0.260	0.167	0.332	0.170	0.329
	Misclassification rate		0.459		0.486		0.338		0.342	
2000	True $\delta(X)$	(+)	0.275	0.224	0.246	0.253	0.412	0.088	0.407	0.092
		(-)	0.228	0.273	0.240	0.260	0.088	0.413	0.092	0.409
	Misclassification rate		0.452		0.493		0.175		0.184	
5000	True $\delta(X)$	(+)	0.275	0.225	0.250	0.251	0.446	0.055	0.443	0.057
		(-)	0.212	0.288	0.244	0.256	0.048	0.452	0.051	0.449
	Misclassification rate		0.437		0.494		0.103		0.108	
10000	True $\delta(X)$	(+)	0.288	0.212	0.255	0.244	0.464	0.036	0.462	0.037
		(-)	0.215	0.285	0.253	0.248	0.038	0.463	0.039	0.461
	Misclassification rate		0.428		0.497		0.073		0.077	
20000	True $\delta(X)$	(+)	0.292	0.208	0.250	0.250	0.474	0.026	0.473	0.027
		(-)	0.208	0.292	0.248	0.252	0.025	0.475	0.027	0.473
	Misclassification rate		0.416		0.498		0.051		0.054	

(+): Positive treatment effect, (-): negative treatment effect

Table 1.12: Statistical power of the different approaches for detecting HTE by the different sample size (n) and the correlation between $g_0^*(X;\beta_0)$, $s^*(X;\beta_2)$ and $\delta(X)$ when sample splitting method was used. Bold cases indicate the pathological conditions

			LRT				Int			
n	Cor[$g_0^*(X;\beta_0)$, $\delta(X)$]	Cor[$s^*(X;\beta_2)$, $\delta(X)$]	OS_1	OS_2	Two GLM	Tian	OS_1	OS_2	Two GLM	Tia
	0.000	0.385	**0.071**	0.075	0.101	0.107	**0.087**	0.1	0.182	0.17
	0.200	0.577	0.084	0.092	0.104	0.108	0.099	0.127	0.184	0.18
	0.400	0.719	0.109	0.108	0.119	0.127	0.144	0.176	0.203	0.20
	0.600	0.828	0.128	0.141	0.15	0.142	0.185	0.207	0.224	0.23
	0.800	0.917	0.119	0.132	0.112	0.114	0.204	0.23	0.192	0.19
	-0.282	0.000	0.072	**0.063**	0.105	0.112	0.092	**0.064**	0.153	0.17
	-0.148	0.200	0.074	0.062	0.135	0.141	0.1	0.078	0.198	0.19
	0.014	0.400	0.068	0.068	0.125	0.112	0.091	0.089	0.199	0.1
	0.229	0.600	0.059	0.08	0.112	0.103	0.119	0.121	0.185	0.17
	0.543	0.800	0.107	0.097	0.112	0.113	0.171	0.176	0.19	0.19
	0.000	0.385	**0.062**	0.135	0.475	0.458	**0.089**	0.194	0.709	0.70
	0.200	0.577	0.096	0.221	0.488	0.478	0.158	0.369	0.703	0.6
	0.400	0.719	0.18	0.287	0.462	0.462	0.293	0.484	0.694	0.70
	0.600	0.828	0.335	0.42	0.514	0.509	0.502	0.63	0.731	0.73
	0.800	0.917	0.45	0.471	0.504	0.488	0.651	0.707	0.71	0.69
2,000	-0.282	0.000	0.099	**0.039**	0.467	0.455	0.162	**0.052**	0.696	0.66
	-0.148	0.200	0.078	0.062	0.458	0.454	0.116	0.106	0.714	0.70
	0.014	0.400	0.074	0.133	0.507	0.495	0.118	0.21	0.729	0.72
	0.229	0.600	0.132	0.249	0.508	0.507	0.182	0.401	0.739	0.73
	0.543	0.800	0.254	0.34	0.486	0.474	0.401	0.58	0.701	0.69
	0.000	0.385	**0.077**	0.231	0.915	0.907	**0.094**	0.381	0.984	0.9
	0.200	0.577	0.158	0.491	0.91	0.916	0.247	0.691	0.982	0.98
	0.400	0.719	0.344	0.706	0.932	0.936	0.501	0.883	0.986	0.98
	0.600	0.828	0.625	0.808	0.923	0.92	0.796	0.951	0.986	0.98
	0.800	0.917	0.85	0.879	0.916	0.91	0.956	0.972	0.987	0.98
5,000	-0.282	0.000	0.179	**0.048**	0.889	0.885	0.308	**0.055**	0.981	0.97
	-0.148	0.200	0.088	0.098	0.922	0.922	0.161	0.167	0.991	0.98
	0.014	0.400	0.068	0.211	0.922	0.911	0.107	0.405	0.984	0.98
	0.229	0.600	0.173	0.502	0.908	0.909	0.266	0.732	0.979	0.97
	0.543	0.800	0.557	0.774	0.919	0.928	0.742	0.934	0.985	0.9
	0.000	0.385	**0.068**	0.441	0.999	0.999	**0.103**	0.66	1	
	0.200	0.577	0.225	0.788	0.998	0.998	0.35	0.942	1	
	0.400	0.719	0.554	0.936	0.999	0.999	0.749	0.993	1	
	0.600	0.828	0.895	0.983	0.999	0.999	0.953	0.996	1	
	0.800	0.917	0.985	0.994	1	0.998	0.999	1	1	
10,000	-0.282	0.000	0.313	**0.059**	1	0.999	0.493	**0.06**	1	
	-0.148	0.200	0.131	0.139	0.999	0.999	0.222	0.249	1	
	0.014	0.400	0.072	0.442	0.999	0.999	0.112	0.654	1	
	0.229	0.600	0.259	0.831	1	1	0.389	0.962	1	
	0.543	0.800	0.809	0.977	0.999	0.999	0.928	0.998	1	
	0.000	0.385	**0.075**	0.698	1	1	**0.11**	0.895	1	
	0.200	0.577	0.329	0.132	1	1	0.518	0.206	1	
	0.400	0.719	0.907	0.258	1	1	0.982	0.416	1	
	0.600	0.828	0.998	0.948	1	0.999	1	0.997	1	
	0.800	0.917	1	1	0.999	0.999	1	1	1	
20,000	-0.282	0.000	0.582	**0.054**	1	1	0.795	**0.054**	1	
	-0.148	0.200	0.348	0.987	1	1	0.535	0.999	1	
	0.014	0.400	0.876	0.999	1	1	0.952	1	1	
	0.229	0.600	0.992	1	1	1	1	1	1	
	0.543	0.800	1	1	1	1	1	1	1	

E Additional results on grid search design simulation

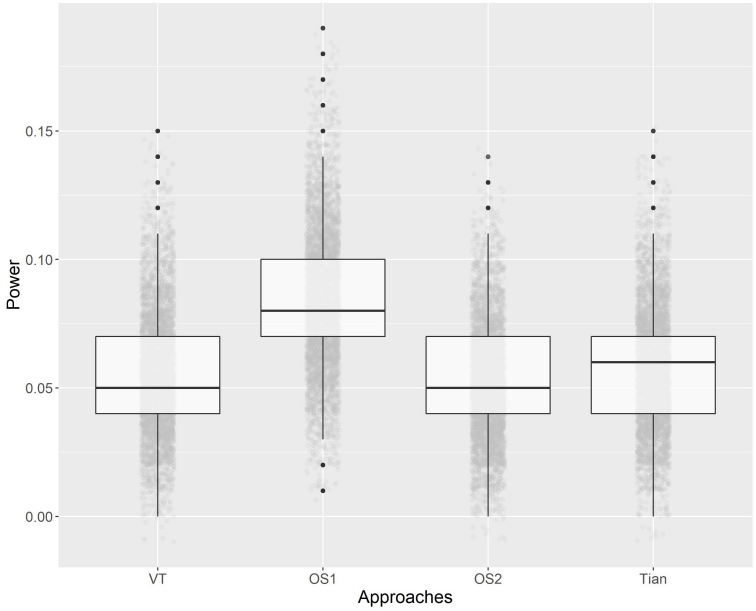


Figure 1.7: Box plot of the power of no HTE cases when n = 10000

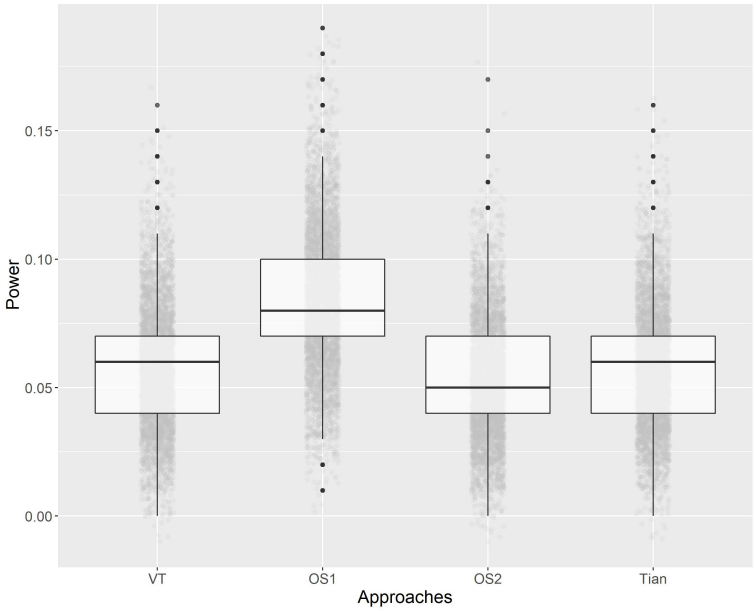


Figure 1.8: Box plot of the power of no HTE cases when n = 5000

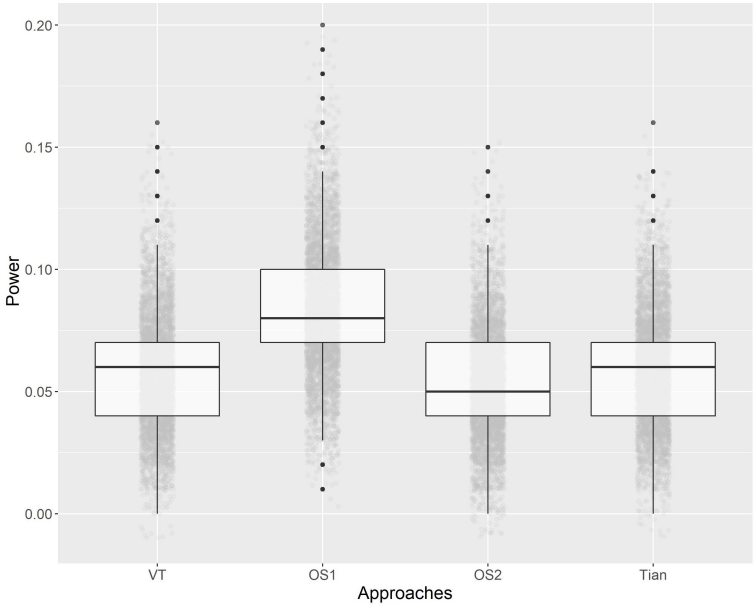


Figure 1.9: Box plot of the power of no HTE cases when n = 2000

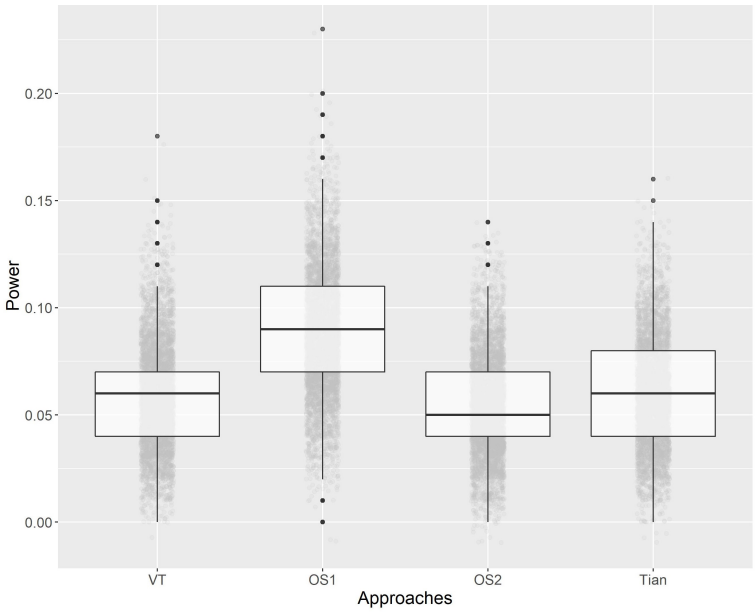


Figure 1.10: Box plot of the power of no HTE cases when n = 500

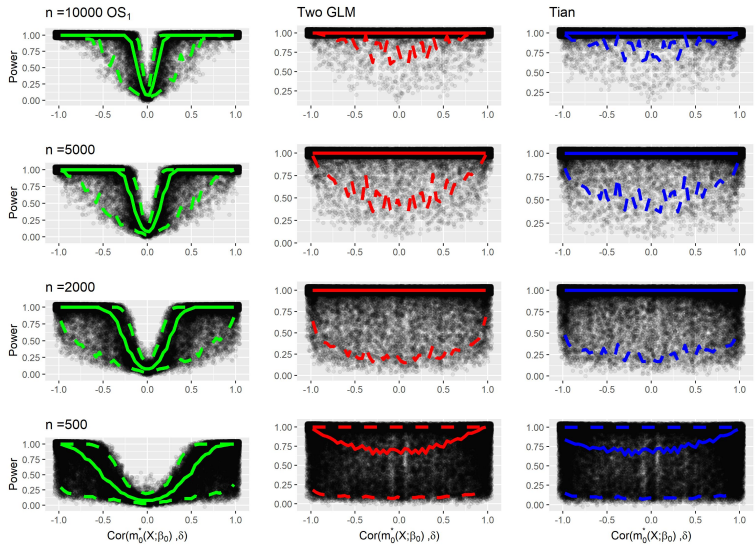


Figure 1.11: A scatter plot of the statistical power and the correlation coefficient for OS_1 in big grid search design by sample size (n = 500, 2000, 5000, 10000). The solid line connects the median of the power for each 50 correlation coefficient group and the dashed lines connect 2.5 percentile and 97.5 percentile of the power

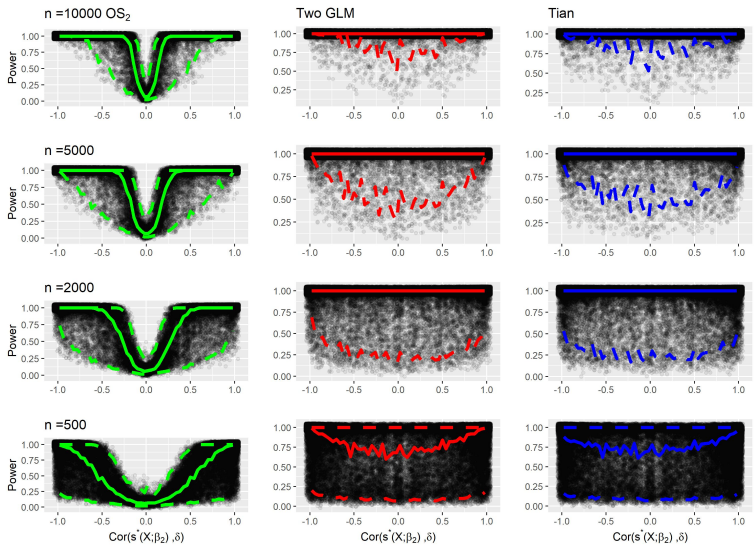


Figure 1.12: A scatter plot of the statistical power and the correlation coefficient for OS_2 in big grid search design by sample size (n = 500, 2000, 5000, 10000). The solid line connects the median of the power for each 50 correlation coefficient group and the dashed lines connect 2.5 percentile and 97.5 percentile of the power

F Additional results on ALLHAT trial analysis

Table 1.18: Missing cases and proportion of the ALLHAT trial data

Variables having missing values	SBP1yr	Total Cholesterol	Height	Smoking status	Antihypertens
Number of missing cases (%)	5364	1594	469	2	1

Table 1.19: Baseline characteristics of missing cases and complete cases of the ALLHAT trial data

	Level	Missing Cases	Complete Cases	p-value
n (%)		6858 (20.6)	26499 (79.4)	
Treatment (%)	Chlorthalidone	3031 (44.2)	12224 (46.1)	0.002
	Amlodipine	1858 (27.1)	7190 (27.1)	
	Lisinopril	1969 (28.7)	7085 (26.7)	
Sex (%)	Male	3115 (45.4)	14604 (55.1)	<0.001
	Female	3743 (54.6)	11895 (44.9)	
Smoking (%)	Current	1563 (22.8)	5740 (21.7)	<0.001
	Past	2378 (34.7)	11070 (41.8)	
	Never	2915 (42.5)	9689 (36.6)	
Diabetes (%)	Y	2676 (39.0)	9387 (35.4)	<0.001
	N	4182 (61.0)	17112 (64.6)	
Antihypertensive drugs (%)	Y	6104 (89.0)	23985 (90.5)	<0.001
	N	753 (11.0)	2514 (9.5)	
MI or Stroke History (%)	Y	1557 (22.7)	6180 (23.3)	0.287
	N	5301 (77.3)	20319 (76.7)	
past LVH (%)	Y	1157 (16.9)	4317 (16.3)	0.256
	N	5701 (83.1)	22182 (83.7)	
SBP1yr (mean (sd))		139.99 (17.39)	138.03 (16.32)	<0.001
Baseline SBP (mean (sd))		147.35 (15.40)	145.99 (15.68)	<0.001
Height (mean (sd))		65.43 (4.11)	66.22 (4.09)	<0.001
Total Cholesterol (mean (sd))		217.78 (45.08)	215.76 (43.05)	0.002
Age (median [IQR])		66.00 [60.00, 73.00]	66.00 [61.00, 72.00]	0.271

Because the distribution of age seems not to follow the normal distribution, we present median and IQR to summarize age. P-values were calculated based on the χ^2 test for categorical variables, ANOVA one-way test form normal continuous variables, and Kruskal-Wallis test for non-normal continuous variables.

Table 1.20: Baseline characteristics of analytic sample (complete cases) of the ALLHAT data by treatment group

		Treatment		
	Level	Chlorthalidone	Amlodipine	Lisinopril
n (%)		12224 (46.1)	7190 (27.1)	7085 (26.7)
Sex (%)	Male	6685 (54.7)	3929 (54.6)	3990 (56.3)
	Female	5539 (45.3)	3261 (45.4)	3095 (43.7)
Smoking (%)	Current	2645 (21.6)	1568 (21.8)	1527 (21.6)
	Past	5124 (41.9)	2963 (41.2)	2983 (42.1)
	Never	4455 (36.4)	2659 (37.0)	2575 (36.3)
Diabetes (%)	Y	4372 (35.8)	2571 (35.8)	2444 (34.5)
	N	7852 (64.2)	4619 (64.2)	4641 (65.5)
Antihypertensive drug (%)	Y	11085 (90.7)	6504 (90.5)	6396 (90.3)
	N	1139 (9.3)	686 (9.5)	689 (9.7)
MI Stroke History (%)	Y	2911 (23.8)	1643 (22.9)	1626 (22.9)
	N	9313 (76.2)	5547 (77.1)	5459 (77.1)
past LVH (%)	Y	1925 (15.7)	1230 (17.1)	1162 (16.4)
	N	10299 (84.3)	5960 (82.9)	5923 (83.6)
SBP 1yr (mean (sd))		136.77 (15.78)	138.29 (14.76)	139.94 (18.42)
Baseline SBP (mean (sd))		145.93 (15.74)	145.90 (15.67)	146.20 (15.59)
Height (mean (sd))		66.18 (4.09)	66.19 (4.09)	66.32 (4.07)
Total_Cholesterol (mean (sd))		215.70 (43.45)	216.29 (43.57)	215.32 (41.81)
Age (median [IQR])		66.00 [61.00, 72.00]	66.00 [61.00, 72.00]	66.00 [61.00, 72.00]

Because the distribution of age seems not to follow the normal distribution, we present median and IQR to summarize age. P-values were calculated based on the χ^2 test for categorical variables, ANOVA one-way test form normal continuous variables, and Kruskal-Wallis test for non-normal continuous variables.

Table 1.21: P-value of the statistical tests to detect HTE in ALLHAT study between Amlodipine and Chlorthalidone (Sample splitting)

	OS_1	OS_2	Two GLM	Tian
LRT	0	0.001	0.029	0.095
Int	0	0	0.004	0.029

Table 1.22: Classification of ALLHAT data by the linear model between Amlodipine and Chlorthalidone (Sample splitting)

Approach		OS2		Two GLM				Tian	
Proportion		(+)	(-)	(+)	(-)	Proportion		(+)	(-)
OS1	(+)	0.895	0.005	0.887	0.013	OS1	(+)	0.895	0.005
	(-)	0.004	0.096	0.071	0.029		(-)	0.085	0.015
Approach		Two GLM		Tian				Tian	
Proportion		(+)	(-)	(+)	(-)	Proportion		(+)	(-)
OS2	(+)	0.882	0.016	0.892	0.006	Two GLM	(+)	0.957	0
	(-)	0.075	0.026	0.088	0.013		(-)	0.023	0.020

Table 1.23: P-value of the statistical tests to detect HTE in ALLHAT study between Lisinopril and Chlorthalidone

Type of test	OS_1	OS_2	Two GLM	Tian
5 folds CV				
LRT	0.204	0.112	0.011	0.014
Int	0.024	0.014	0.008	0.006
Sample splitting				
LRT	0.155	0.283	0.236	0.158
Int	0.050	0.105	0.017	0.005

Table 1.24: Classification of ALLHAT data by the linear model by Lisinopril and Chlorthalidone group

Approach		OS2		Two GLM				Tian	
Proportion		(+)	(-)	(+)	(-)	Proportion		(+)	(-)
OS1	(+)	1	0	1	0	OS1	(+)	1	0
	(-)	0	0	0	0		(-)	0	0
Approach		Two GLM		Tian				Tian	
Proportion		(+)	(-)	(+)	(-)	Proportion		(+)	(-)
OS2	(+)	1	0	1	0	Two GLM	(+)	1	0
	(-)	0	0	0	0		(-)	0	0

In this analysis, 5-fold cross validation and sample splitting method give the same results.

CHAPTER2

Extending inferences from a randomized trial to a target population when baseline covariate data are incomplete

Abstract

Missing covariates is a common problem in epidemiological data but how to address it when extending causal inferences to the target population is not studied much. We consider to use IPW for complete cases to address missing baseline covariates when extending causal inferences from a nested randomized controlled trial to the total eligible population. We focus on identification and assumptions of IPW for complete cases estimators especially when treatment and outcome are not observable for non-randomized population. We illustrate how our ideas are applied using a nested trial comparing coronary artery bypass grafting surgery plus medical therapy versus medical therapy alone on 10-year mortality.

1 Introduction

Randomized controlled trials (RCTs) are regarded as the gold standard for evaluating the causal effect of a treatment. However, inferences from RCTs are not applicable to the population of interest (target population) if the distribution of modifiers is different between trial participants and the target population [20, 21]. To address this issue, investigators have been interested in when and how to generalize (or transport) causal inferences from RCTs to the target population. Previous studies on this topic have not paid attention to missing data problems, even though missing baseline covariates is a common problem in real world studies. So far, only one study has focused on addressing missing baseline covariates in the context of extending causal inference to the target population [22]. The authors utilized multiple imputation to deal with missing data in the target population but how to address this issue using the inverse probability weighting (IPW) approach especially when treatment and outcome are not observable in the target population have not been discussed. In this study, we present the identification and assumptions of IPW for complete cases estimators for average treatment effects when causal inferences are extended from nested trials to the target population. Our results can be applicable to a broad class of estimators for parameters that can be expressed as solutions to estimating equations, but we take inverse probability (IP) weighting estimators as an example to illustrate our approach. We also consider two distinct scenarios on missing data mechanism in the target population: Treatment and outcome are always observable or not observable at all in non-randomized participants.

2 Study design and data

In our paper, we consider that RCTs are nested within cohort studies of trial-eligible individuals [23, 24]. Investigators might want to extend causal inferences from nested RCTs to all trial eligible individuals under this study design. In this setting, the target population of all trial eligible participants is well defined and baseline covariates are collected for them but outcome and treatment information might be omitted for non-

randomized participants. The example of the nested trial design is that a study of which participants are recruited from broad population and baseline information is collected but only who give consent to participation are randomized [25].

In this paper, capital letters represent random variables, lower cases refer to particular values of the corresponding random variables, and individuals are indexed by $i = 1, 2, \ldots, n$. Let A be a binary point treatment and Y be an outcome. The potential (counterfactual) outcome under specific treatment assignment a is denoted as Y^a and an indicator for trial participation among trial-eligible individuals is denoted as S ($S = 1$: trial participants, $S = 0$: non-participants). Baseline covariates, X, are collected for all the trial-eligible individuals but some of X has missing values. We denote full data as O, i.e., $O = (X, S, A, Y)$. Let R_{X_k} be an indicator of whether or not k-th variable of X is observable ($R_{X_k} = 1$: observed, $R_{X_k} = 0$: not observed). Define R_S, R_A, and R_Y the same way. When X consists of total K variables, a missing data pattern can be represented by a random $(K+3)$-vector, $(R_S, R_A, R_Y, R_{X_1}, R_{X_2}, \ldots, R_{X_K})$. Let R be a scalar random variable corresponding to different missing data patterns; conventionally, complete cases are denoted as $R = 1$ and R has integer values between 1 and 2^{K+2}. Let $O_{(r)}$ be the observed part of O under missing data pattern $R = r$ and define $X_{(r)}$ the same way. Missing data mechanism refers to the conditional distribution of R given complete data O. Let $\mathbb{P}_n$ be empirical measure defined as $\mathbb{P}_n f(O) = n^{-1} \sum_{i=1}^{n} f(O_i)$. We assume the situation of full adherence, no measurement error, no interference, and no dropout.

A causal quantity we consider is the mean potential outcome under treatment a for all the trial-eligible individuals, $\mathrm{E}[Y^a]$. After identifying mean potential outcome for each treatment arm, it is easy to figure out typical target causal estimands, such as the average causal risk difference, $\mathrm{E}[Y^{a=1}] - \mathrm{E}[Y^{a=0}]$. Here, we consider extending inferences from RCTs to all the trial-eligible individuals but the results can be easily applied to the target population of trial-eligible but non-randomized individuals [25, 26].

3 Identification

We first review identifiability conditions and identification results of the causal quantity of full data. Next, we examine additional assumptions about missing data mechanism and identification results for IPW for complete cases when treatment and outcome are observed for non-randomized participants. In this scenario, treatment and outcome are complete in both randomized and non-randomized participants. This is a natural extension of existing literature [27, 28, 29] to causal inferences of our target estimand. Last, we consider a situation where treatment and outcome information are unavailable for non-randomized participants. In this scenario, randomized participants have complete treatment, outcome information and partially observable covariate information but non-randomized participants have only covariate information. We explain a complication of this scenario making hard to use the previous identification results and suggest assumptions about missing data and following identification results to overcome the complication.

3.1 Identification with full data

Identifiability conditions: The mean potential outcome of all trial-eligible participants under treatment assignment a, $\mathrm{E}[Y^a]$, is identifiable under the following conditions [23]:

i) Consistency of potential outcomes: if $A_i = a$, then $Y_i = Y_i^a$, for every $a \in \{0, 1\}$. In other words, the observed outcome Y_i for those who received treatment a is equal to the potential outcome under treatment a.

ii) Mean exchangeability in the trial: $\mathrm{E}[Y^a|A = a,\ S = 1,\ X] = \mathrm{E}[Y^a|S = 1,\ X]$, for every $a \in \{0, 1\}$.

iii) Positivity of the probability of treatment assignment in the trial: $\Pr[A = a|X = x,\ S = 1] > 0$, for every $a \in \{0, 1\}$ and for all x such that $f_{X|S}(x|S = 1) > 0$.

iv) Mean generalizability: $\mathrm{E}[Y^a|S = 1,\ X] = \mathrm{E}[Y^a|X]$, for every $a \in \{0, 1\}$.

v) Positivity of the probability of trial participation: $\Pr[S = 1|X = x] > 0$ for all x such that $f_X(x) > 0$.

In randomized trials with well-defined interventions, we expect condition (i), (ii), and (iii) hold [10].

Identification: Using the identifiability condition (i) through (v), we have previously shown that [23]

$$\mathrm{E}[Y^a] = \mathrm{E}[\mathrm{E}\left[Y|X, S = 1, A = a\right]].$$

This result means that, provided the identifiability conditions hold, the functional of observed data

$$\psi(a) \equiv \mathrm{E}[\mathrm{E}\big[Y|X, S = 1, A = a\big]] \tag{2.1}$$

has a causal interpretation as the mean of potential outcome under intervention to set treatment to a.

Furthermore, $\psi(a)$ has an equivalent re-expression using inverse probability treatment weighting (IPTW) [23],

$$\psi(a) = \mathrm{E}\left[\frac{I(S = 1, A = a)Y}{\Pr[S = 1|X]\Pr[A = a|X, S = 1]}\right] \tag{2.2}$$

When X is a high dimensional random vector, it is hard to estimate $\psi(a)$ nonparametrically so estimation of $\psi(a)$ usually entails model specification of a finite dimensional parameter and solving an estimating equation for it. Let the population estimating equation for $\psi(a)$ as

$$\mathrm{E}\left[W(O; a, \theta, \psi(a))\right] = 0, \tag{2.3}$$

where θ is a nuisance parameter that we need to know. The exact formulation of the estimating function $W(O; a, \theta, \psi(a))$ depends on the type of estimators. Later, we give the formulation of $W(O; a, \theta, \psi(a))$ with respect to IPTW estimators but now $W(O; a, \theta, \psi(a))$ is a general term, embracing outcome modelling, doubly robust estimators, and other consistent estimators.

3.2 Identification with missing baseline covariates, but treatment and outcome data available from all participants

Missing data assumptions: Throughout, we assume that the identifiability conditions (i - v) hold so (2.3) is well defined and the solution of it has a causal meaning given full data. We also assume missing data mechanism is everywhere MAR [30] or missing-always-at-random [31], which means that the probability of having missing data pattern r only depends on observed variables under the missing data pattern r,

$$\Pr[R = r|O = o] = \Pr[R = r|O = o^*], \quad \text{for all } r, o, o^* \text{ such that } o_{(r)} = o^*_{(r)} \tag{2.4}$$

In the above equation, o and o^* are specific values of O. Full data O consists of $O_{(r)}$, which is observable part of O under missing data pattern r, and unobservable part of O. The equation means if the value of observable part is the same, the conditional probability of having the missing data pattern is the same regardless of values of unobservable part under the corresponding missing data pattern.

We assume the positivity condition holds on $\Pr[R = 1|O]$,

$$\Pr[R = 1|O = o] > 0 \quad \text{for all } o \text{ in the support of } O \tag{2.5}$$

MAR missing data mechanism (2.4) allows us to identify $\Pr[R = 1|O]$ from observed data and the positivity condition (2.5) ensures that IPW for complete cases estimators have finite asymptotic variances.

IPW for complete cases estimating equation: Suppose A and Y are complete and X has missing values among all the trial-eligible individuals and we want to use parametric models to estimate $\psi(a)$. If the population estimating equation of full data,

$$\mathrm{E}\left[W(O; a, \theta, \psi(a))\right] = 0,$$

has an unique solution $\psi(a)$ and the solution is uniformly convergent to the parameter, the solution of empirical versions of the following population IPW estimating equations for complete cases,

$$\mathrm{E}\left[\frac{I(R=1)}{\Pr[R=1|O]}W(O;a,\theta,\psi(a))\right]=0 \tag{2.6}$$

give consistent estimate of $\psi(a)$. It can be easily proven by the law of total expectation. If we already know $\Pr[R=1|O]$, (2.6) is identifiable from data by itself but in usual cases, we need to identify and estimate $\Pr[R=1|O]$ from observed data.

Define a functional of observed data $\pi_1(O)$ as

$$\pi_1(O) \equiv 1-\sum_{k=2}^{K}\Pr[R=k|O_{(k)}], \tag{2.7}$$

when there are K missing data patterns in total. Then, $\Pr[R=1|O]$ can be identified as $\pi_1(O)$ under the MAR mechanism and the following estimating equation can be identified from data

$$\mathrm{E}\left[\frac{I(R=1)}{\pi_1(O)}W(O;a,\theta,\psi(a))\right]=0 \tag{2.8}$$

In most cases, we estimate $\pi_1(O)$ using finite-dimensional parametric models. We revisit the estimation of $\pi_1(O)$ in the next section.

3.3 Identification with missing baseline covariates, and systematically missing treatment and outcome data from non-randomized individuals

Data structure and complications: We suppose treatment and outcome information is not collected at all for non-randomized participants. The data is realized as

$$(X_{(R_i),i}, S_i, I(S_i=1)A_i, I(S_i=1)Y_i, R_i)\ i=1,2,\ldots,n \tag{2.9}$$

or we can think the data structure as stratified by S:

$$(X_{(R_x),i}, A_i, Y_i, R_i) \;\; i = 1, 2, \ldots, n_1 \quad \text{for } S = 1$$

$$(X_{(R_x),i}, R_i) \;\; i = 1, 2, \ldots, n_0 \quad \text{for } S = 0,$$

where $n_1 = \sum_i^n I(S = 1)$ and $n_0 = \sum_i^n I(S = 0)$. This data structure can be observed when investigators do not assign treatment and do not ascertain outcome for non-randomized participants to reduce the cost of studies or other reasons. Our methods also can apply when some covariates are not collected among non-randomized participants.

Under this data structure, we cannot identify IPW estimating function (2.6) because we can only observe complete cases among randomized participants, that means $\Pr[R = 1|S = 0] = 0$ and that leads to the violation of the positivity condition (2.5). To resolve this issue, we suggest decomposing the estimating function by S and weighting each subpopulation by the inverse of conditional probability of being complete cases, where complete cases are defined in each subpopulaton.

We introduce additional notation here. We denote complete cases of randomized participants and non-randomized participants as $R = 1$ and $R = 0$, respectively. That means,

$$(R_{X_1} = 1, R_{X_2} = 1, \ldots, R_{X_K} = 1, R_A = 1, R_Y = 1) \implies R = 1$$

$$(R_{X_1} = 1, R_{X_2} = 1, \ldots, R_{X_K} = 1, R_A = 0, R_Y = 0) \implies R = 0$$

For the other missing data patterns, R has scalar values as we defined before. We denote r_1 as a specific value of R corresponding to missing data patterns only appear in $S = 1$ and r_0 as a value of R corresponding to missing data patterns only appear in $S = 0$ and $\mathcal{R}_1$ and $\mathcal{R}_0$ refer to the set of all r_1 and r_0 respectively. For example, if there are

only two covariates, (X_1, X_2), $\mathcal{R}_1$ corresponds to the following missing data patterns:

$$(R_{X_1} = 1, R_{X_2} = 1, R_A = 1, R_Y = 1)$$
$$(R_{X_1} = 1, R_{X_2} = 0, R_A = 1, R_Y = 1)$$
$$(R_{X_1} = 0, R_{X_2} = 1, R_A = 1, R_Y = 1)$$
$$(R_{X_1} = 0, R_{X_2} = 0, R_A = 1, R_Y = 1).$$

In the same way, $\mathcal{R}_0$ corresponds to the following patterns:

$$(R_{X_1} = 1, R_{X_2} = 1, R_A = 0, R_Y = 0)$$
$$(R_{X_1} = 1, R_{X_2} = 0, R_A = 0, R_Y = 0)$$
$$(R_{X_1} = 0, R_{X_2} = 1, R_A = 0, R_Y = 0)$$
$$(R_{X_1} = 0, R_{X_2} = 0, R_A = 0, R_Y = 0).$$

Also, we denote the complete data conditional on S as

$$O_1 = (X, S = 1, A, Y), \quad O_0 = (X, S = 0)$$

In this setting, $1 \in \mathcal{R}_1$ and $0 \in \mathcal{R}_0$ exclusively and we know the follwing equalities hold

$$\Pr[R = r_1 | S = 0] = 0 \text{ for all } r_1 \in \mathcal{R}_1 \tag{2.10}$$
$$\Pr[R = r_0 | S = 1] = 0 \text{ for all } r_0 \in \mathcal{R}_0,$$

which is the given information about missing data mechanism of our data structure.

Assumptions on missing data: To separate the estimating function by S and identify the probability of complete cases in each subgroup, we assume that missing data mechanism follows a stratified version of the MAR (2.4):

$$\Pr[R = r_s | o_s] = \Pr[R = r_s | o_s^*], \text{ for all } r_s, s, o_s, o_s^* \text{ such that } o_{s,(r_s)} = o^*_{s,(r_s)} \tag{2.11}$$

The stratified MAR is weaker than the unstratified MAR (2.4) because it is built upon the given information about missing data mechanism (2.10). Actually, combining the given information and the stratified MAR, we impose a stronger restriction on missing data mechanism than the unstratified MAR.

We assume a stratified version of positivity, which means that the positivity condition holds on each sub-population not on the entire population.

$$\begin{aligned} &\Pr[R=1|O_1=o_1]>0 \quad \text{for all } o_1 \\ &\Pr[R=0|O_0=o_0]>0 \quad \text{for all } o_0 \end{aligned} \tag{2.12}$$

We discuss the relationship between "conventional" MAR assumptions and the stratified MAR used in this paper, in the Appendix.

Under the stratified MAR mechanism (2.11), we can identify $\Pr[R=1|O_1]$ and $\Pr[R=0|O_0]$ from observed data. Define $\pi_1(O_s)$ as

$$\pi_1(O_s) \equiv 1 - \sum_{r_s \in \mathcal{R}_s} \Pr[R=r_s|O_{s,(r_s)}], \text{ where } s \in \{0,1\}.$$

Under the stratified MAR mechanism, $\Pr[R=1|O_1]$ can be identified as $\pi_1(O_1)$ and $\Pr[R=0|O_0]$ can be identified as $\pi_1(O_0)$. The stratified MAR mechnism allows $\Pr[R=1|O_1]$ and $\Pr[R=0|O_0]$ to be idenfitable and the stratified positivity ensures that the IPW estimators have a finite asymptotic variance.

Identification: The left side of the full data estimating equation (2.2), $\mathrm{E}[W(O;a,\theta,\psi(a))]$, can be decomposed by S:

$$\begin{aligned} \mathrm{E}[W(O;a,\theta,\psi(a))] &= \mathrm{E}[I(S=1)W(O;a,\theta,\psi(a))] + \mathrm{E}[I(S=0)W(O;a,\theta,\psi(a))] \\ &= \mathrm{E}[W(O;a,\theta,\psi(a))|S=1]\Pr[S=1] + \mathrm{E}[W(X;a,\theta,\psi(a))|S=0]\Pr[S=0]. \end{aligned}$$

Using the law of total expectation,

$$
\begin{aligned}
&\mathrm{E}[W(O;a,\theta,\psi(a))|S=1] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[I(R=1)|O_1]}{\Pr[R=1|O_1]}W(O;a,\theta,\psi(a))\middle|S=1\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\frac{I(R=1)}{\Pr[R=1|O_1]}W(O;a,\theta,\psi(a))\middle|O_1\right]\middle|S=1\right] \\
&= \mathrm{E}\left[\frac{I(R=1)W(O;a,\theta,\psi(a))}{\Pr[R=1|O_1]}\middle|S=1\right]
\end{aligned}
$$

In the same way,

$$
\begin{aligned}
&\mathrm{E}[W(O;a,\theta,\psi(a))|S=0] \\
&= \mathrm{E}\left[\frac{\mathrm{E}[I(R=0)|O_0]}{\Pr[R=0|O_0]}W(O;a,\theta,\psi(a))\middle|S=0\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\frac{I(R=0)}{\Pr[R=0|O_0]}W(O;a,\theta,\psi(a))\middle|O_0\right]\middle|S=0\right] \\
&= \mathrm{E}\left[\frac{I(R=0)W(O;a,\theta,\psi(a))}{\Pr[R=0|O_0]}\middle|S=0\right].
\end{aligned}
$$

The third equation follows from the fact that A and Y of trial non-participants are not necessary to know to identify $\mathrm{E}[Y^a]$ so we can write down

$$
\mathrm{E}[W(O;a,\theta,\psi(a))|S=0] = \mathrm{E}[W(S,X;a,\theta,\psi(a))|S=0],
$$

and

$$
\begin{aligned}
\mathrm{E}[W(O;a,\theta,\psi(a))] &= \mathrm{E}\left[\frac{I(R=1)W(O;a,\theta,\psi(a))}{\Pr[R=1|O_1]}\middle|S=1\right]\Pr[S=1] \\
&\quad + \mathrm{E}\left[\frac{I(R=0)W(O;a,\theta,\psi(a))}{\Pr[R=0|O_0]}\middle|S=0\right]\Pr[S=0] \\
&= \mathrm{E}\left[\frac{I(R=1,S=1)W(O;a,\theta,\psi(a))}{\Pr[R=1|O_1]} + \frac{I(R=0,S=0)W(O;a,\theta,\psi(a))}{\Pr[R=0|O_0]}\right]
\end{aligned}
$$

Therefore, the population estimating equation

$$\mathrm{E}[W(O;a,\theta,\psi(a))]=0$$

and

$$\mathrm{E}\left[\frac{I(R=1,S=1)W(O;a,\theta,\psi(a))}{\Pr[R=1|O_1]}+\frac{I(R=0,S=0)W(O;a,\theta,\psi(a))}{\Pr[R=0|O_0]}\right]=0 \quad (2.13)$$

have the same solution when the denominators are strictly larger than 0 and the estimator from empirical versions of (2.13) is a consistent estimator of $\psi(a)$ given standard regularity conditions.

To identify the estimating function in (2.13), we need to identify $\Pr[R=1|O_1]$ and $\Pr[R=0|O_0]$ from observed data separately in $S=1$ and $S=0$. Assuming missing data mechanism is stratified MAR, we can identify (2.13) as

$$\mathrm{E}\left[\frac{I(R=1,S=1)W(O;a,\theta,\psi(a))}{\pi_1(O_1)}+\frac{I(R=0,S=0)W(O;a,\theta,\psi(a))}{\pi_1(O_0)}\right]=0 \quad (2.14)$$

4 Example: estimation using IP weighting

In this section, we discuss the estimation of the target estimand in each scenario focusing on IPTW estimators for the potential outcome mean because this approach and IPW for complete cases estimators share the same schematic procedure, so it might be the primary choice for estimating the mean potential outcome while missing data is addressed by IPW for complete cases.

4.1 Estimation with full data

As we mentioned in the previous section, parametric models are often used for the estimation. Let $p(X;\beta)$ be a model for $\Pr[S=1|X]$ and $e(X;a,\gamma)$ be a model for $\Pr[A=a|X,S=1]$, where β and γ are finite dimensional parameters. Define a column

vector $W(O; a, \psi(a), \beta, \gamma)$ as

$$W(O; a, \psi(a), \beta, \gamma) \equiv \begin{pmatrix} h(X)\big(S - p(X;\beta)\big) \\ I(S=1)h'(X)\big(I(A=a) - e(X;a,\gamma)\big) \\ w_a(X,S,A;\beta,\gamma)Y - \psi(a) \end{pmatrix},$$

where $w_a(X,S,A;\beta,\gamma) \equiv \dfrac{I(S=1,A=a)}{p(X;\beta)e(X;a,\gamma)}$. The first row is the estimating function for β, the second row is the estimating function for γ, and the last row is the estimating function for $\psi(a)$ with the weights, $w_a(X,S,A;\beta,\gamma)$.

When data have no missingness and realized as $O_i = (X_i, S_i, A_i, Y_i)$ with n independent and identically distributed (i.i.d.) samples, $\psi(a)$ can be estimated by solving the following sample estimating equations jointly.

$$\mathbb{P}_n \begin{pmatrix} \{h(X)\big(S - p(X;a,\beta)\big)\} \\ \{I(S=1)h'(X)\big(I(A=a) - e(X;a,\gamma)\big)\} \\ \{w_a(X,S,A;\beta,\gamma)Y - \psi_n(a)\} \end{pmatrix} = 0.$$

To be more concrete, if a model $p(X;\beta)$ is in a form of the logistic function and $h(X) = X$, solving the first row of the sample estimating equation is solving the score equation of ordinary logistic regression of trial participation for all trial-eligible participants.

Especially, if the models are correctly specified in data, the estimates, $\widehat{\beta}$ and $\widehat{\gamma}$, converge to β and γ in large sample.

$$p(X;\widehat{\beta}) \to p(X;\beta^*) = \Pr[S=1|X]$$

$$e(X;a,\widehat{\gamma}) \to e(X;a,\gamma^*) = \Pr[A=a|X,S=1],$$

where β^*, and γ^* are limiting values of $\widehat{\beta}$, and $\widehat{\gamma}$ and $\to$ means convergence in probability. In this case, $\psi_n(a)$ is a consistent estimator of $\psi(a)$ and under identifiability conditions, $\psi_n(a)$ is a consistent estimator of the mean potential outcome $\psi(a)$.

4.2 Estimation with missing baseline covariates, but treatment and outcome data available from all participants

Estimation of the probability of complete cases: We can estimate $\psi(a)$ by solving a sample analogue of (2.8) and we need to estimate $\pi_1(O)$ from data. We often use parametric models to estimate $\pi_1(O)$ because complete data O is high dimensional in many cases. We denote the model for $\Pr[R = r|O_{(r)}]$ as $\pi_r(O_{(r)}; \delta_r)$, where δ_r is a finite dimensional parameter for the model. Let δ be the collective expression of δ_r, such as $\delta = (\delta_2, \ldots, \delta_K)$ if there are total K missing data patterns in data. Let the model for $\pi_1(O)$ be $\pi_1(O;\ \delta)$, which is

$$\pi_1(O;\ \delta) = 1 - \sum_{k=2}^{K} \pi_k(O_{(k)};\ \delta_k), \tag{2.15}$$

If the models are correctly specified, $\pi_r(O_{(r)}; \widehat{\delta_r})$ converges to $\Pr[R = r|O_{(r)}]$ for every r and $\pi_1(O;\ \widehat{\delta})$ converges to $\Pr[R = 1|O]$ under the MAR mechanism.

In practice, it is hard to estimate $\pi_1(O)$ from (2.15) while assuming missing data mechanism no more restrictive than MAR and not violating numerical constraints ($\sum_{r=2}^{K} \pi_r(O_{(r)}; \delta_r) < 1$ for all the observed O). We can identify $\Pr[R = 1|O]$ from different forms depending on missingness patterns and the other forms can facilitate parametric modeling. We review how to estimate $\pi_1(O; \delta)$ under non-monotone missingness pattern here, which is the most general missingness pattern. We present an overview of missingness patterns and the estimation of $\pi_1(O)$ under the other missingness patterns to the appendix.

Non-monotone missingness is the most unorganized missingness pattern and can represent the most general case of missingness. Under non-monotone missingness pattern, we cannot reexpress $\pi_1(O)$ to other forms so should use (2.15). We can fit a simple logistic regression of an indicator of $I(R = r)$ among individuals having complete data on $O_{(r)}$ to build a model for $\pi_r(O_{(r)}; \delta_r)$ and obtain the estimate of δ through maximizing the joint likelihood of the models for each missing data patterns to estimate $\pi_1(O; \delta)$

[29]. By the positivity assumption, $\pi_1(O;\delta)$ must be larger than 0 so the constraint

$$1 - \sum_{k=2}^{K} \pi_k(O_{(k)};\ \delta_k) > 0$$

should be satisfied for complete cases and the model fails to converge if the constraint is not satisfied for at least one observation [29]. Sun and Tchetgen Tchetgen [29] utilized Constrained Bayesian estimation [32] to overcome this convergence issue. They set the uninformative prior distribution of δ and leave complete cases violating the constraint out of the posterior distribution. They used mode (or mean) of the posterior distribution of δ from the remaining data as an estimate of δ. The details can be found in their paper [29].

Estimation of the causal quantity: After estimating $\pi_1(O;\delta)$, we can solve a sample analogue of (2.8) to obtain the IPW for complete cases estimator of $\psi(a)$. We take the IPTW estimating function as full data estimating function for $\psi(a)$. For the IPW for complete case estimator, the estimating function is evaluated among complete cases only and they are weighted by the inverse of the probability of complete cases. Let $w_a(O;\beta,\gamma,\delta) \equiv \frac{I(S=1,A=a,R=1)}{p(X;\beta)e(X;a,\gamma)\pi_1(O;\delta)}$. Then, $\psi(a)$ can be estimated by solving the following sample estimating equation.

$$\mathbb{P}_n \begin{pmatrix} \frac{I(R=1)h(X)}{\pi_1(O;\delta)}\big(S - p(X;a,\beta)\big) \\ \frac{I(S=1,R=1)h'(X)}{\pi_1(O;\delta)}\big(I(A=a) - e(X;a,\gamma)\big) \\ w_a(O;\beta,\gamma,\delta)Y - \psi_{n,IPW}(a) \end{pmatrix} = 0, \tag{2.16}$$

If all the models are correctly specified and all the aforementioned assumptions and conditions hold, the estimator $\psi_{n,IPW}(a)$ is a consistent estimator of $\psi(a)$.

4.3 Estimation with missing baseline covariates, and systematically missing treatment and outcome data from non-randomized individuals

The estimation procedure is in missing data case 1 but now it is done in $S = 1$ and $S = 0$ separately to estimate $\pi_1(O_1)$ and $\pi_1(O_0)$. Suppose the missingness pattern is

non-monotone and let $\pi_1(O_1;\delta_1)$ and $\pi_1(O_0;\delta_0)$ be the model for $\pi_1(O_1)$ and $\pi_1(O_0)$ using parametric models (series of logistic models or Bayesian estimation). If all the involved models are correctly specified, $\pi_1(O_1;\widehat{\delta_1})$ and $\pi_1(O_0;\widehat{\delta_0})$ converge to $\pi_1(O_1)$ and $\pi_1(O_0)$.

When i.i.d. sample of data, whose structure is the same as (2.9), is given and after estimating $\pi_1(O_1;\delta_1)$ and $\pi_1(O_0;\delta_0)$, we can estimate $\psi(a)$ by solving the following sample estimating,

$$\mathbb{P}_n \begin{pmatrix} w_\pi(O;\delta_0,\delta_1)h(X)(S-p(X;a,\beta)) \\ \frac{I(S=1,R_1=1)h'(X)}{\pi_1(O_1;\delta_1)}\big(I(A=a)-e(X;a,\gamma)\big) \\ w_a(O;\beta,\gamma,\delta_1)Y-\psi_{n,IPW2}(a) \end{pmatrix} = 0, \tag{2.17}$$

where $w_\pi(O;\delta_0,\delta_1) \equiv \frac{I(R=1,S=1)}{\pi_1(O_1;\delta_1)} + \frac{I(R=0,S=0)}{\pi_1(O_0;\delta_0)}$ and $w_a(O;\beta,\gamma,\delta_1) \equiv \frac{I(S=1,A=a,R=1)}{p(X;a,\beta)e(X;a,\gamma)\pi_1(O_1;\delta_1)}$. If all the models are correctly specified and all the aforementioned assumptions and conditions hold, the estimator $\psi_{n,IPW2}(a)$ is a consistent estimator of $\psi(a)$. In the application, we utilize Constrained Bayesian estimation approach to estimate $\pi_1(O_1;\delta_1)$ and $\pi_1(O_0;\delta_0)$ because the underlying missingness pattern of the data is non-monotone.

5 Implementing the methods in CASS

We implemented the methods discussed in the previous sections to extend a causal inference from RCT samples to all the trial-eligible population in CASS (Coronary artery surgery study). CASS was conducted to assess the effectiveness of coronary artery bypass grafting surgery with medication (henceforth we call it surgery) versus medication only on reducing mortality rates for patients with stable ischemic heart disease. 2099 patients were eligible to the study and 780 patients gave consent to participation and were randomly assigned to each treatment group. We additionally excluded 6 individuals from analysis because they do not have data on treatment assignment or died right away after randomization. All the participants of the trial were followed in the first 10 years of the study. Baseline covariates were collected in the same manner among randomized and non-randomized participants. The details about study protocol is available

elsewhere [33, 34].

5.1 Statistical methods

The target estimand was the 10-year mortality risk difference comparing surgery versus medical therapy. Inverse probability treatment weighting (IPTW) was used to obtain the target estimates. We use logistic regression models to estimate the probability of participation in the trial and the probability of treatment among RCT samples with the following baseline covariates: age, angina status, ejection fraction, percent obstruction of the proximal left anterior descending artery, left ventricular wall motion score, history of previous myocardial infarction, number of diseases proximal vessels, number of pack-years of smoking cigarettes. We chose those variables based on the previous study [35]. Compared to the previous study [23], we considered the number of pack-years as another covariate to make additional missing data patterns. We used the same set of variables to fit the participation and treatment models and used restricted cubic splines to fit age, ejection fraction, and the number of pack-years in those models. We used IPW for complete cases to address missing covariates and the Constrained Bayesian approach was used to estimate the missingness models. We utilized treatment and outcome information of non-randomized participants to fit the missing models. Next time, we did not use the treatment and outcome information. We implemented estimation method in the previous section to extend an inference from the RCT sample to all the trial-eligible individuals and show results of complete case analysis for comparison. We used bootstrap resampling with 500 samples to make inferences about the estimators. Normal-distribution based 95% confidence intervals are reported in the results.

5.2 Results

Table 2.1 describes baseline characteristics of the entire population and complete cases by participation to RCT and the treatment arms. Table 2.2 summarizes missing data information of the data. There are 10 missing data patterns including complete cases. To make analysis simple, we reshaped missing data patterns as follows: 1) complete casese, 2) missing only on pack-years, 3) missing only on ejection fraction, 4)

missing both on ejection fraction and pack-years, 5) missing on variables other than pack-years and ejection fraction, and use those patterns in the analysis.

Table 2.3 presents causal estimates (mean potential outcome under surgery, mean potential outcome under medication, causal risk difference) for RCT samples and the entire population by missing data methods (complete case analysis, IPW for missingness with or without outcome and treatment information of a non-randomized part of CASS). All the estimates of the risk difference are in the range of [-3.9, -5] and are not statistically significant. Addressing missing covariates affects the estimates for all the eligible participants more than RCT samples. The risk difference on all the eligible participants is -3.94 percentage points when complete case analysis is conducted. After addressing missing data, the risk difference is -4.91 similar to the results of RCT samples. Our results are not qualitatively different from the previous study [23] that used the same data but the estimates are little different because we introduced one additional variable with a high proportion of missingness.

6 Discussion

We describe formal missing data assumptions for identification of potential outcome means and treatment effects when extending causal inferences from randomized trials to a target population and missing baseline covariates are addressed by the IPW for complete cases. Especially, when some covariates are totally unobservable in the non-randomized participants, investigators need to be cautious about building models for missingness. In that situation, we propose to build separate models for missingness by trial participation, and weight each sub-population accordingly. Mean potential outcomes for the target population can be obtained by integrating two weighted population.

We like to leave a remark that the IPW approach we described here is not the most efficient way to estimate causal estimates. The standard IPW approach only uses complete cases to estimate parameters for mean potential outcomes, although incomplete cases are used to estimate models for missing data process. The efficiency can be improved by augmenting information in incomplete cases to the standard IPW

estimating equation [27]. Sun and Tchetgen Tchetgen [29] proposed augmented IPW for non-monotone MAR missing data pattern and showed that estimators from augmented IPW is more efficient than the standard IPW approach.

The other popular way to handle missing data is multiple imputation (MI). Multiple imputation creates more than one imputation data sets by replacing each missing value with multiple corresponding imputed values which are drawn from the conditional distribution of the missing variable given observed variables. In the context of extending causal inferences, there are no studies about how to implement MI if both trial participants and the target population have missing covariates and their missing data patterns are radically different, like the situation of this study. In view of our study, we would recommend stratifying data by trial participation and make a separate imputation model within the stratified data. The two approaches, MI and IPW, rely on complementary modeling assumption [36]; IPW requires correct specification of models for missing data process and MI requires correct specification of full data models to obtain a consistent estimator. A series of studies [36, 37, 38] did a blind test to compare the performance of MI using a multivariate joint normal model and Sun and Tchtgen Tchtgen's IPW approach on the prediction of odds ratio under different missing data mechanisms. The studies showed that augmented IPW and MI gave similar results in terms of point estimates and the width of 95% confidence interval. However, the differences between using MI versus IPW with regard to extending inferences from RCTs are not fully investigated and future studies are needed for it.

Table 2.1: Baseline covariates in CASS of the entire population and by complete cases.

		Entire Population			
		Overall	RCT		Obs
	Level		Surgery	Medication	
n		2093	389	389	1315
Angina (%)	Yes	1649 (78.8)	304 (78.1)	305 (78.4)	1040 (79.1)
	No	444 (21.2)	85 (21.9)	84 (21.6)	275 (20.9)
Previous MI (%)	Yes	1218 (58.2)	222 (57.1)	244 (62.7)	752 (57.2)
	No	874 (41.8)	267 (42.9)	145 (37.3)	562 (42.8)
The number of diseased vessels (%)	≥ 1	1322 (63.3)	235 (60.4)	249 (64.2)	838 (64.0)
	0	765 (36.7)	154 (39.6)	139 (35.8)	472 (36.0)
Age (mean (SD))		51.09 (7.61)	51.48 (7.26)	50.97 (7.37)	51.01 (7.78)
Pack-year (mean (SD))		38.75 (21.25)	39.73 (20.67)	41.4 (22.97)	37.65 (20.81)
Ejection Fraction, % (mean (SD))		60.24 (12.58)	60.86 (13.04)	59.86 (12.76)	60.14 (12.35)
Lvscor (median (SD [IQR]))		6 (2.94 [5, 9])	6 (2.92 [5, 9])	6 (2.89 [5, 9])	6 (2.96 [5, 9])
LAD % obstruction (median (SD [IQR]))		30 (38.71 [0, 80])	25 (37.74 [0, 75])	25 (37.24 [0, 70])	30 (39.36 [0, 80])
		Complete cases			
n		1372	294	299	779
Angina (%)	Yes	1054 (76.8)	224 (76.2)	227 (75.9)	603 (77.4)
	No	318 (23.2)	70 (23.8)	72 (24.1)	176 (22.6)
Previous MI (%)	Yes	844 (61.5)	177 (60.2)	194 (64.9)	473 (60.7)
	No	512 (37.3)	117 (39.8)	105 (35.1)	306 (39.3)
The number of diseased vessels (%)	≥ 1	860 (62.7)	176 (59.9)	185 (61.9)	499 (64.1)
	0	512 (37.3)	118 (40.1)	114 (38.1)	280 (35.9)
Age (mean(SD))		50.41 (7.47)	50.91 (7.13)	50.62 (7.35)	50.14 (7.63)
Pack-year (mean(SD))		38.63 (20.60)	39.98 (20.83)	41.25 (22.72)	37.12 (19.52)
Ejection Fraction, % (mean(SD))		59.6 (12.62)	60.38 (13.01)	59.05 (12.93)	59.51 (12.34)
Lvscor (median (SD [IQR]))		6 (2.81 [5, 9])	6 (2.9 [5, 9])	6 (2.81 [5, 9])	6 (2.77 [5, 9])
LAD % obstruction (median (SD [IQR]))		30 (37.61 [0, 70])	25 (37.09[0, 70])	20 (35.41 [0, 70])	30 (38.5 [0, 80])

LAD : left anterior descending coronary artery, Lvscor: Left ventricular score

Table 2.2: Descriptive statistics of missing covariates in CASS.

Missing data by variables			
	Variable	# missing	% missing
	Age	0	0
	Angina	0	0
	Previous MI	1	0.04
	Pack-year	410	19.6
	The number of diseased vessels	6	0.3
	Ejection fraction	372	17.8
	Lvscor	56	2.7
	LAD % obstruction	0	0
Variable with missing data per person			
	# Variables with missing data	# observations	%
	0	1372	65.6
	1	604	28.9
	2	110	5.3
	3	7	0.3
Missing data pattern			
R	# Variables with missing data	# observations	%
1	No missing data	1372	65.6
2	Pack-year	314	15
3	Ejecfr	264	12.6
4	Ejecfr +Pack-year	80	3.8
5	Lvscor	20	1
	The number of diseased vessels	5	0.2
	Previous MI	1	0.04
	The number of diseased vessels + Pack-year	1	0.04
	Lvscor + Ejecfr	21	1
	Lvscor + Pack-year	8	0.4
	Lvscor + Ejecfr + Pack-year	7	0.3
	Total	63	3

R : An indicator variable for missing data pattern used in the analysis
Lvscor: Left ventricular wall motion score

Table 2.3: Estimated 10-year mortality risk for surgery and medication arm and causal risk difference of the CASS study.

Mean Potential outcome under surgery ($E[Y^1]$; %)						
	Population	RCT Sample		Extending to all the eligible population		
	Missing data method	CC	IPWC	CC	IPWC (Use A and Y)	IPWC (Not use A and Y)
Estimator	Ordinary IPTW (95% CI; SD)	17.33 (12.23, 21.01; 2.24)	16.76 (12.94, 21.11; 2.09)	16.43 (11.86, 20.73; 2.26)	17.07 (12.66, 21.19; 2.18)	17.19 (12.68, 21.22; 2.18)
	Normalized IPTW (95% CI; SD)	17.18 (12.24, 20.98; 2.23)	16.74 (12.94, 21.08; 2.08)	16.40 0.164 (11.87, 20.67; 2.25)	17.04 (12.71, 21.18; 2.16)	17.2 (12.71, 21.19; 2.16)
Mean Potential outcome under medication ($E[Y^0]$; %)						
Estimator	Ordinary IPTW (95% CI; SD)	22.31 (16.64, 25.55; 2.27)	21.09 (17.38, 25.87; 2.17)	20.21 (15.73, 24.66; 2.28)	21.61 (17.18, 25.81; 2.2)	21.76 (17.19, 25.83; 2.2)
	Normalized IPTW (95% CI; SD)	22.24 (16.64, 25.56; 2.27)	21.11 (17.42, 25.94; 2.17)	20.26 (15.8, 24.59; 2.24)	21.64 (17.24, 25.95; 2.22)	21.91 (17.22, 25.94; 2.22)
Causal risk difference ($E[Y^1]$ - $E[Y^0]$; pp)						
Estimator	Ordinary IPTW (95% CI; SD)	-4.97 (-10.5, 1.57; 3.08)	-4.33 (-10.63, 1.44; 3.08)	-3.79 (-9.97, 2.16; 3.09)	-4.54 (-10.74, 1.6; 3.15)	-4.60 (-10.84, 1.54; 3.16)
	Normalized IPTW (95% CI; SD)	-5.07 (-10.49, 1.52; 3.06)	-4.37 (-10.69, 1.35; 3.07)	-3.86 (-9.91, 2.06; 3.06)	-4.57 (-10.74, 1.62; 3.15)	-4.71 (-10.83, 1.56; 3.16)

CC : Complete case analysis, IPWC: IPW for complete cases, Use A and Y: Used treatment and outcome information of non-randomized participants, Not use A and Y: Not used that information

Chapter 2 Appendix

A Missingness patterns

Definition

Missingness patterns represent a specific structure of missing data patterns and there are three missingness patterns (uniform, monotone, and non-monotone missingness) [39]. The first pattern is uniform missingness. That implies only one missingness pattern, except for complete cases, exists because all the partially observed variables have missing values together or not. For example, suppose V_1, V_2, and V_3 are the partially observed variables and denote $R_{V,i}$ as a missing indicator for the variable V and individual i. Then, there are only two missingness patterns for (V_1, V_2, and V_3) under uniform missingness: complete cases ($R_{V_1,i} = 1$, $R_{V_2,i} = 1$, $R_{V_3,i} = 1$) and missing data in all the partially observed variables ($R_{V_1,i} = 0$, $R_{V_2,i} = 0$, $R_{V_3,i} = 0$). The next pattern is monotone missingness which implies that there exists specific underlying structure in missing data. For example, if V_1, V_2, and V_3 are the same baseline characteristic measured at time 1, time 2, and time 3. Individuals dropped out of the study at time 1 so having a missing value on V_1 necessarily have missing values on V_2 and V_3 as well. In the same way, individuals dropped out of the study at time 2 and having missing values on V_2 necessarily have a missing value on V_3. In this case, $R_{V_1,i} = 1$ implies ($R_{V_2,i} = 1$, $R_{V_3,i} = 1$) and $R_{V_2,i} = 1$ implies $R_{V_3,i} = 1$, but the missingness pattern, such as ($R_{V_1,i} = 1$, $R_{V_2,i} = 0$, $R_{V_3,i} = 0$) is not allowed under monotone missingness. Like this example, if a missing value on one variable implies missing values on subsequent variables on the same observation, that missing ness pattern is called monotone missingness. The last missingness pattern, non-monotone missingness, refers to any missingness patterns do not belong to uniform or monotone missingness. Contrary to monotone missingness, non-monotone missingness does not have the structure like monotone missingness. Under non-monotone missingness, the patterns, such as ($R_{V_1,i} = 1$, $R_{V_2,i} = 0$, $R_{V_3,i} = 0$) and ($R_{V_1,i} = 1$, $R_{V_2,i} = 1$, $R_{V_3,i} = 0$), are allowed. Figure 2.1 graphically presents the missing data patterns.

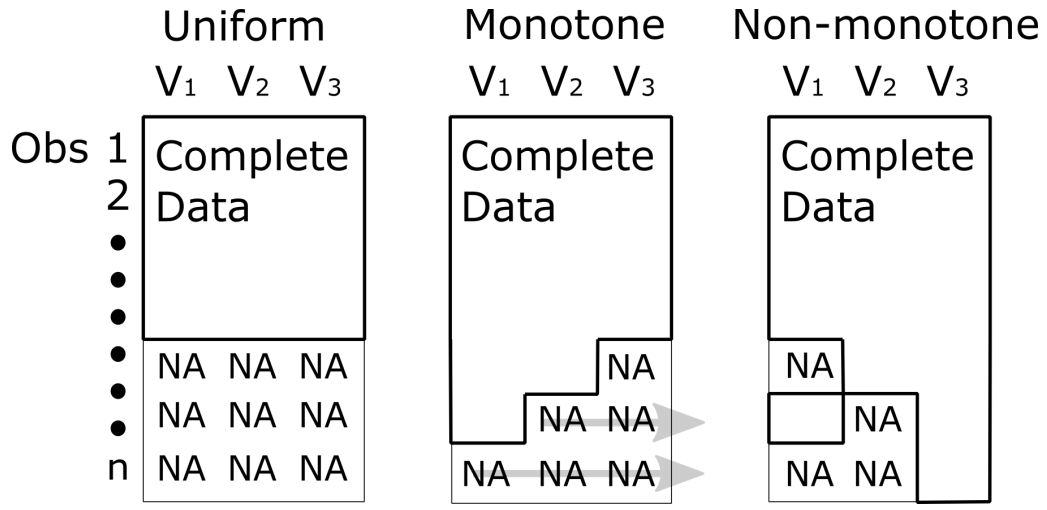


Figure 2.1: Description of the different Missing Data Patterns; gray arrows imply the underlying structure of missingness pattern

Identification under uniform missingness

Considering the structure of uniform missingness and assuming MAR missing data mechanism, missing data mechanism under uniform missingness pattern is thought as a Bernoulli distribution and identified from (2.7), which is:

$$f(R|O) = \begin{cases} \Pr[R=1|O] = 1 - \Pr[R=2|O_{(2)}] \\ \Pr[R=2|O] = \Pr[R=2|O_{(2)}] \end{cases}$$

Using a parametric model, missing data process can be estimated from the logistic regression and $1 - \pi_2(O_{(2)}; \widehat{\delta}_2)$ is an estimator of $\Pr[R=1|O]$. If the model is correctly specified, the estimator is a consistent estimator of $\Pr[R=1|O]$.

Identification under monotone missingness

Missing data mechanism under monotone missingness pattern can be identified through a discrete hazard function of missing data patterns. [28]. To apply it, we need to arrange total K missing data patterns from complete case ($R=1$) to ($R=K$) in order of the closeness to complete case. For example, when data is realized with three missing data patterns, like (X_1, X_2, X_3), (X_1, X_2), (X_1), complete case is (X_1, X_2, X_3),

and $R = 2$ should be (X_1, X_2), and $R = 3$ should be (X_1) because (X_1, X_2) is closer to the complete case than (X_1).

After arranging missing data patterns, the hazard of having missing data pattern k, λ_k, is defined as:

$$\lambda_k \equiv \begin{cases} \Pr[R = k | R \leq k, O], & \text{if } 2 \leq k \leq K, \\ 1, & \text{if } k = 1 \end{cases}$$

Assuming MAR mechanism, we can identify the hazard function as

$$\lambda_k = \lambda_k(O_{(k)}) \begin{cases} \Pr[R = k | R \leq k, O_{(k)}], & \text{if } 2 \leq k \leq K, \\ 1, & \text{if } k = 1 \end{cases}$$

Then, missing data mechanism can be identified using the hazard function, which is

$$f(R|O) = \begin{cases} \Pr[R = k|O] = \lambda_k O_{(k)} \prod_{j=k+1}^{m} (1 - \lambda_j(O_{(j)})) & \text{if } 1 \leq k \leq K-1 \\ \Pr[R = K|O] = \lambda_K(O_{(K)}) \end{cases}$$

To estimate $\Pr[R = 1|O]$, we need to estimate the hazard for each missing data pattern and compose the conditional probabilities of each missing data pattern from the estimated hazards. In practice, the hazard function can be estimated from a series of logistic regression. To estimate $\lambda_k(O_k)$ from data, we need to set an indicator variable of $I(R = k)$ and run a logistic regression of the indicator on $O_{(k)}$ among individuals who have R less than k, in other words, closer to complete cases than missing data pattern represented by $R = k$. Let the estimated hazard for missing data pattern k be $\lambda_k(O_{(k)}; \widehat{\zeta_k})$, where ζ_k is a finite dimensional parameter for a model $\lambda_k(O_k)$. The estimate of $\Pr[R = 1|O]$ is $\prod_{k=2}^{K}(1 - \lambda_k(O_{(k)}; \widehat{\zeta_k}))$ and when all the models for the hazard function are correctly specified, which is a consistent estimator for $\Pr[R = 1|O]$.

B Discussion on the unstratified MAR (2.4) and the stratified MAR (2.11)

Let me give an example first. Suppose there are only four variables in data $O = (A, Y, S, X)$, we know missing data pattern r_0 only appears in $S = 0$, and $O_{r_0} = \{S, X\}$. That means under missing data pattern r_0, we observe $\{S = 0, X\}$. Assuming missing data mechanism is unstratified MAR means the following equation holds for r_0.

$$\Pr[R = r_0|S = s, X = x] = \Pr[R = r_0|S = s, X = x, A, Y]$$

It can be divided to the two equations by the value of S.

$$\Pr[R = r_0|S = 1, X = x] = \Pr[R = r_0|S = 1, X = x, A, Y] \tag{B.1}$$

$$\Pr[R = r_0|S = 0, X = x] = \Pr[R = r_0|S = 0, X = x, A, Y] \tag{B.2}$$

The stratified MAR only assumes (B.2) holds because r_0 appears in $S = 0$ and but does not say anything about (B.1). However, from the given information (2.10), we know $\Pr[R = r_0|S = 1] = 0$ so we can deduct

$$\Pr[R = r_0|S = 1, X = x] = \Pr[R = r_0|S = 1, X = x, A, Y] = 0, \tag{B.3}$$

which is stronger than (B.1).

In general, the unstratified MAR mechanism assumes the following two equations hold:

$$\Pr[R = r_s|o_s] = \Pr[R = r_s|o_s^*] \text{ for all } r_s, o_s, \text{ and } o_s^* \text{ such that } o_{s,(r_s)} = o^*_{s,(r_s)}$$

$$\Pr[R = r_s|o_{1-s}] = \Pr[R = r_s|o^*_{1-s}] \text{ for all } s, r_s, o_{1-s}, \text{ and } o^*_{1-s} \text{ such that } o_{1-s,(r_s)} = o^*_{1-s,(r_s)}.$$

The stratified MAR assumes the upper equation holds and the lower equation can be deducted from the given information. To sum up, the stratified MAR itself is weaker than the unstratified MAR because it is built upon the given information about missing

data mechanism. In this data structure and the information is given, stratified MAR is sufficient to make a valid inference.

The positivity condition of the unstratified (2.5) and the stratified version (2.12) do not have logical implications between them. The stratified positivity condition means the following inequalities hold.

$$\Pr[R = 1|o_1] > 0 \quad \text{for all } o_1 \text{ in the support of } O_1$$
$$\Pr[R = 0|o_0] > 0 \quad \text{for all } o_0 \text{ in the support of } O_0$$

The unstratified positivity implies the upper inequality holds but it does not imply the lower inequality holds. Conversely to deduct the unstratified positivity from the stratified version, we need the condition $\Pr[R = 1|O_o] > 0$ upon the upper inequality, which is not a part of the stratified positivity condition.

CHAPTER3

Addressing missing covariates in causal inference: observational studies

Abstract

Missing data is a common issue in epidemiologic studies and especially missing covariates almost always happen and give special challenges to causal inference studies. Previous studies have left some issues on the topic (causal inference with missing covariates data) and this study addressed these issues: 1) G-formula outcome modelling and doubly robust estimators have not been studied, 2) Multiple imputation (MI) and inverse probability of complete cases weighting (IPCCW) have not been compared, 3) controversies about which one is better between the two types of MI: Across and Within 4) It is not clear how to implement augmented IPCCW under univariate and non-monotone missingness.

1 Introduction

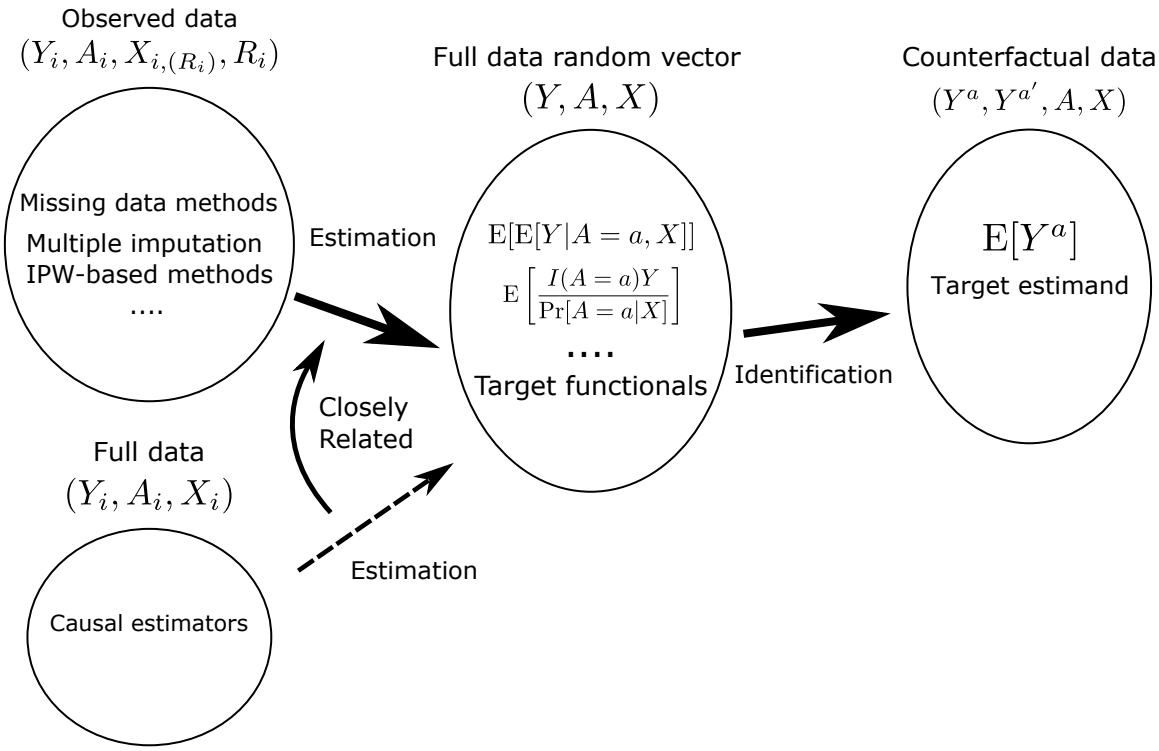
Missing data is a common problem in public health research. Studies on causal inference are not free from it and even more complicated. Causal inference methods and missing data methods have different goals and they have been developed for their own purpose. Causal inference methods focus on the identification of some aspects of counterfactual data, usually causal effects, through full data random vectors and the estimation of the full data functionals reflecting target causal estimands from data, while missing data methods focus on the estimation of some aspects of full data distribution from observed data (Figure 3.1) [1]. Both sides have various approaches to achieve their own goal and investigators choose their approach based on what assumptions they can make confidently on their data. When investigators want to estimate causal effects from observed data [2], they need to consider the both sides, asking what approach we should select, and how to connect the two approaches.

However, previous studies often provide an incomplete scope of the topic considering only a small portion of all the available missing data methods and causal inference methods [40, 41, 42]. They applied propensity score analysis (matching, inverse probability weighting, stratification) to obtain causal effect estimates while missing data is addressed by multiple imputation but they did not consider g-formula outcome modelling and augmented inverse probability weighting (doubly robust methods) as causal inference approaches. In addition, all those studies consider multiple imputation as a missing data method not considering inverse probability weighting (IPW) based approach. Also, some open issues remain in this topic. Depending on how we aggregate information of imputed data set, multiple imputation in causal inference problem can be divided into Across and Within approach [40]. It is not obvious under what cases which one is better between the two approaches. As for IPW-based approach it is not clear how to implement augmented IPW for complete cases estimators to handle

[1] Identification of full data from observed data is also an important area but we do not mention it here because missing data methods we consider here (Multiple imputation and IPW-based approach) focus on estimation

[2] Not all observed data have missingness but when we refer to observed data here that means data with missing values

Figure 3.1: Schematic diagram of missing data methods and causal inference methods



missing covariates and estimate causal effects; especially, when missingness pattern is non-monotone, augmented IPW estimators have no closed form.

In this paper, we limited our scope to the situation when baseline covariates are missing because it is more common and challenging [43] than the case of missing outcome or treatment. We review techniques of both sides (causal inference, missing data methods) and how they are combined and used together. We select combinations of two techniques as inclusive to all the commonly used methods but exclusive to some for which we think it is not reasonable and hard to implement (Table 3.1). In this way, we can maintain a comprehensive scope on the topic and also avoid using ill-fitted methods. Those combinations are compared in simulation studies as we address the aforementioned issues, including the challenging case of non-monotone missingness pattern. We illustrate the methods using the CASS dataset.

2 Notation, setting

We consider a study of estimating average causal effects of binary point treatment A ($a \in \{0, 1\}$) on binary or continuous outcome Y for the entire population (ATE) where the causal mean difference ($\mathrm{E}[Y^1] - \mathrm{E}[Y^0]$) is the measure of causal effects. We denote a random variable Y^a as a potential (counterfactual) outcome under specific treatment assignment a. Our target of inference is mean potential outcome $\mathrm{E}[Y^a]$, because we can easily calculate the causal effect estimates after estimating mean potential outcome. Let X be a random vector for baseline covariates. We model full data as an i.i.d. sample of a random vector, $(X_i, A_i, Y_i, R_i), i = 1, 2, \ldots, n$. Let R be a scalar random variable corresponding to each missing data pattern in data and we set complete cases as $R = 1$. For simple notation, we denote complete data as O_i, $O_i = (X_i, A_i, Y_i)$. Let $O_{(r)}$ be the observed part of O under missing data pattern $R = r$ and define $X_{(r)}$ the same way. On the contrary, let $O^c_{(r)}$ and $X^c_{(r)}$ be the missing part of O and X under missing data pattern $R = r$. We consider the situation when only baseline covariates have missing values and A and Y are fully observable so $O_{(r)} = (X_{(r)}, A, Y)$. Observed data is realization of a random vector $(X_{(R_i,i)}, A_i, Y_i, R_i), i = 1, 2, \ldots, n$. Missing data mechanism refers to the conditional distribution of R given complete data O, $f(R|O)$. Our goal in the study is to estimate mean potential outcome $\mathrm{E}[Y^a]$ from the observed data with missingness on X. Throughout this paper, we assume the situation of no measurement error, no interference, and no dropout.

3 Causal inference

3.1 Identification

The first step of causal inference is to identify target parameters; informally speaking, to find estimable functionals of full data random vectors that can reflect a target parameter of unobservable potential outcome (Figure 3.1). When a target estimand is ATE, the common target of inference is mean potential outcome, $\mu_a \equiv \mathrm{E}[Y^a]$. The following identifiability conditions should hold to identify μ_a.

(i) Consistency of potential outcomes: if $A_i = a$, then $Y_i = Y_i^a$, for every $a \in \{0, 1\}$.

(ii) Mean exchangeability: $\mathrm{E}[Y^a|A = a,\ X] = \mathrm{E}[Y^a|X]$, for every $a \in \{0, 1\}$.

Mean exchangeability is sufficient to identify mean potential outcome but the following stronger independence condition is widely used in literature:

$$Y^a \perp\!\!\!\perp A|X$$

(iii) Positivity of the probability of treatment assignment: $\Pr[A = a|X = x] > 0$, for every $a \in \{0, 1\}$ and for all x such that $f_X(x) > 0$.

Mean exchangeability and consistency conditions are not testable so we need to ruminate on whether they are plausible in our data and research settings. Assuming these identification conditions hold, mean potential outcome can be expressed as a functional of full data random vectors, such as [44]:

$$\mu_a = \mathrm{E}[\mathrm{E}[Y|A = a,\ X]]. \tag{3.1}$$

The full data functional μ_a is the target of inference in 'data analysis' and the object for which missing data methods make an inference from observed data. There are other representations of μ_a that can be derived from semiparametric theory [27]. For example, μ_a can be expressed the other way using iterative expectation [45]:

$$\mu_a = \mathrm{E}\left[\frac{I(A = a)Y}{\Pr[A = a|X]}\right]. \tag{3.2}$$

Those functionals are just different representations of μ_a but their estimating strategies and variances are different, which allows us to devise a number of valid estimators for μ_a; henceforth, we refer to those functionals including (3.1), collectively as μ_a because they are mathematically equivalent and to use notation economically and call them the target functionals.

3.2 Estimation

We explain the estimation of μ_a from the perspective of the estimating equation framework. The framework fits well with this estimation problem and connects this section with inverse probability weighting approach to handle missing data (Section 5.2). Suppose we want to estimate μ_a using IPTW functional, which is representation (3.2). When X is high-dimensional, it is difficult to conceive function $\Pr[A = a|X]$ non-parametrically so we posit parametric models to $\Pr[A = a|X]$ and let $p(X;\alpha)$ be the model for $\Pr[A = a|X]$ (treatment model), where α is a finite dimensional parameter of the model. Consider the case when $p(X;\alpha)$ are known. Then, a solution of the following estimating equation is a consistent estimator for μ_a when standard regularity conditions hold [46].

$$\sum_{i=1}^{n} M(O_i;\mu_a,\alpha) = 0, \text{ where } M(O;\mu_a,\alpha) = \left\{ \frac{I(A=a)Y}{p(X;\alpha)} - \mu_a \right\}. \tag{3.3}$$

We refer to $M(O;\mu_a,\alpha)$ as the inverse probability treatment weighting (IPTW) estimating function and refer to the solution of (3.3) as the IPTW estimator.

In the same way, we can consider other estimators from different representations. Let $g_a(X;\theta)$ be the model for $\mathrm{E}[Y|A = a,\ X]$ (outcome model). Then g-formula outcome model-based (OM) estimator is a solution to the estimating equation,

$$\sum_{i=1}^{n} M(O_i;\mu_a,\theta) = 0, \text{ where } M(O;\mu_a,\theta) = g_a(X;\theta) - \mu_a$$

and augmented IPTW (AIPTW) estimator is the solution to the estimating equation,

$$\sum_{i=1}^{n} M(O_i;\mu_a,\alpha,\theta) = 0, \text{ where } M(O;\mu_a,\alpha,\theta) = \frac{I(A=a)}{p(X;\alpha)}(Y - g_a(X;\theta)) + g_a(X;\theta) - \mu_a \tag{3.4}$$

Those estimators are also consistent estimators for μ_a. We refer to $M(O_i;\mu_a,\theta)$ as g-formula outcome model estimating function and $M(O;\mu_a,\alpha,\theta)$ as the AIPTW estimating function.

In many cases, we do not know true nuisance models, $p(X;\alpha)$ and $g_a(X;\theta)$ so need

to estimate them. The estimation of nuisance parameters can be incorporated into the estimating equation framework and that adds additional estimating equations to solve. We need to solve estimating equations for the nuisance parameters first and plug the estimates into the estimating equation for μ_a. For example, with regard to the AIPTW estimating equation, when treatment model and outcome model are unknown, we need to solve the following stacked estimating equations to estimate nuisance parameter α and θ as well as μ_a,

$$\begin{pmatrix} \sum_{i=1}^{n} M_g(O_i;\theta) \\ \sum_{i=1}^{n} M_p(A_i, X_i;\alpha) \\ \sum_{i=1}^{n} M(O_i;\mu_a,\alpha,\theta) \end{pmatrix} = \mathbf{0}, \tag{3.5}$$

where

$$M_g(O;a,\theta) = h(X)I(A=a)(Y - g_a(X;\theta)) \tag{3.6}$$

$$M_p(A,X;\alpha) = h'(X)(I(A=a) - p(X;\alpha)), \tag{3.7}$$

and $M(O;\mu_a,\alpha,\theta)$ is defined as (3.4) and $h(X)$ and $h'(X)$ are arbitrary functions of X. For example, if a logistic regression model is used to estimate $p(X;\alpha)$, $p(X;\alpha) = \exp(X\alpha)/(1+\exp(X\alpha))$ and $h'(X) = X$. We need to estimate θ and α first by solving the first and second row of the equations and plug the estimates $p(X_i;\widehat{\alpha})$ and $g_a(X_i;\widehat{\theta})$ into the last estimating equation. The estimate $\widehat{\mu}_a$ is obtained by solving the last row.

The AIPTW estimator obtained from the above procedure has a doubly robust property that assures consistency of the estimator when either outcome model or treatment model is correct [47]. We also use a normalized version of IPTW and AIPTW estimator where weighted terms are divided by the sum of weights $\left(\sum_{i=1}^{n} \frac{I(A_i=a)}{p(X_i;\widehat{\alpha})}\right)$. Normalization makes IPTW estimators bounded and less variable in finite sample [48]. We provide formulas for normalized version of estimators and other details in appendix 0.1.

4 Missing data methods

The goal of missing data methods is to estimate some aspects of full data distribution from observed data. The most popular way to handle missing data is complete cases analysis [49] that only analyzes complete cases. However, complete cases analysis can be biased unless missing mechanism is missing completely at random (MCAR) [50, 51] and in many cases MCAR is not plausible. In addition, complete cases analysis is statistically inefficient because it discards the information of incomplete data [50, 51]. In order to overcome the drawbacks of complete cases analysis, other principled approaches have been developed.

The most well known principled approaches are multiple imputation (MI) and IPW based approach [36]. We focus on the concept and underlying assumptions of these two methods this section. MI is a simulation-based approach that creates $m > 1$ imputation data sets by replacing each missing value with m corresponding imputation values that are drawn from the conditional distribution of the missing variable given observed variables (imputation model). Standard statistical analyses are done in each m data set and the results are aggregated. The other principled approach is IPW-based approach. The reasoning of it is to weight complete cases according as how many incomplete cases they can represent given that the complete cases and incomplete cases have the same baseline characteristics. A complete case with the weight two means that this case can reflect one incomplete case as well as itself because this complete case is considered as a random sample between two. Practically, the IPW-based approach upweights complete cases by the inverse of the probability of complete cases given full data, $\Pr[R = 1|O]$, which we denote by $\pi_1(O)$ and in the same way we denote $\Pr[R = r|O]$ by $\pi_r(O)$. We refer to IPW for missingness as inverse probability of complete cases weighting (IPCCW).

Both approaches assumes that missing data mechanism is missing at random (MAR). MAR missing data mechanism means that the probability of being missing data pattern

r only depends on observed variables under missing data pattern r,

$$\Pr[R = r|O = o] = \Pr[R = r|O = o^*], \quad \text{for all } r, o, o^* \text{ such that } o_{(r)} = o^*_{(r)},$$

where, o and o^* are specific values of O. In our setting, it can be expressed as

$$\Pr[R = r|A, Y, X_{(r)}, X^c_{(r)}] = \Pr[R = r|A, Y, X_{(r)}], \quad \text{for all } r$$

We assume that missing data mechanism is everywhere MAR [30] or missing-always-at-random [52] that is the proper MAR mechanism to do maximum likelihood-based inference for full data parameters [30]. The MAR missing data mechanism implies

$$\Pr[X^c_{(r)}|R = r, A, Y, X_{(r)}] = \Pr[X^c_{(r)}|A, Y, X_{(r)}] \quad \text{for all } r. \tag{3.8}$$

That allows us to fit an imputation model from whole data not conditioned on the missingness indicator R. Also, the MAR mechanism allows $\pi_1(O)$ to be identifiable from observed data, such as

$$\pi_1(O) = 1 - \sum_{r=2}^{K} \Pr[R = r|O_{(k)}], \tag{3.9}$$

where K is the total number of missing data patterns in observed data[3]. IPW-based approach also requires positivity of $\pi_1(O)$,

$$\pi_1(O = o) > 0 \quad \text{for all } o \text{ in the support of } O, \tag{3.10}$$

to make IPW estimating equation identifiable.

MI and IPW-based approach rely on complimentary modelling assumptions. MI needs the specification of underlying full data model but does not impose restrictions to missing data mechanism. IPCCW needs the model specification of missing data mechanism but full data model can be unrestricted except for target and nuisance

[3]It is possible to identify missing data mechanism with weaker conditions. Recently, Bhattacharya et al. [53] showed that missing data mechanism can be identified given no self-censoring edge and no colluder and Nabi et al.[54] provided the completeness of the algorithm.

models we need to estimate, e.g., outcome model or treatment model. In the next section, we discuss how to combine causal inference methods and missing data methods to estimate the target functionals from observed data.

5 Causal inference with missing data methods

Identifying causal effects by an estimable full data functionals is an area of causal inference and after we find such full data functionals, missing data methods play a role of estimating them from observed data. However, causal inference models take a huge share of estimation procedure because estimation from observed data is closely related to estimation from full data (Figure 3.1) and that is the venue where causal inference and missing data methods are combined. Missing data methods are usually used to solve one-step estimation problem, such as, estimating population mean or regression coefficients, which are common cases in data analysis but causal estimators are essentially multi-step estimators in that we need to estimate nuisance parameters first and estimate the target parameters later. This multi-step nature of causal estimators induces complications when combined with missing data methods. In the rest of this section, we explain how to combine each of MI and IPCCW approach with the causal estimators and practical complications arisen from it.

5.1 Combination of MI and causal inference methods

Estimating the target functionals via MI is relatively straightforward. After we have m imputed datasets, we can analyze each dataset as they are complete data. The estimation procedure is essentially the same as Section 3.2. Two types of MI procedure is possible to estimate the target functionals according to whether MI is used for the estimation of the target functional or nuisance parameters. We introduce a concept of both approaches, provide an example, and review literature about it.

Following the principle of MI, we can estimate μ_a by solving stacked estimating equations in each imputed data set, and aggregating the solutions across imputed data sets. This approach is called 'Within' approach in literature [40]. On the other hand,

MI procedure can be used to estimate nuisance parameter over imputed data sets and the aggregated nuisance parameter is plugged into the estimating equation for μ_a. This approach is called 'Across' approach [40]. For example, MI-Within approch with IPTW estimating equation (MI-Within-IPTW) can be implemented by solving the following estimating equation in each m imputed dataset

$$\begin{pmatrix} \sum_{i=1}^{n} M_p(A_i, X_{i,m}; \alpha) \\ \\ \sum_{i=1}^{n} M(O_{i,m}; \mu_a, \alpha) \end{pmatrix} = \mathbf{0}, \tag{3.11}$$

where $M_p(A, X; \alpha)$ is defined as (3.7), $M(O; \mu_a, \alpha)$ is defined as (3.6) and $X_{i,m}$ is covariate X of individual i of m-th imputed dataset. We suppress the index m for A because A is assumed to be complete. Let $(\widehat{\alpha}_m, \widehat{\mu}_{a,m})^T$ be the solution of the estimating equation from m-th imputed dataset. We can aggregate the estimates by taking a mean or median of $\widehat{\mu}_{a,m}$ over total M imputed data sets and we consider mean or median as the MI-Within-IPTW estimator for μ_a. MI-Across approach also requires to solve the upper row of (3.11) to estimate α but after that $p(X_{i,m}; \widehat{\alpha}_m)$ is aggregated over imputed datasets and plugged into the estimating equation for μ_a. We can consider the following estimating equation at the end and the MI-Across-IPTW estimator is a soultion to it,

$$\sum_{i=1}^{n} \left\{ \frac{I(A_i = a) Y_i}{1/M \sum_{m=1}^{M} p(X_{i,m}; \widehat{\alpha}_m)} - \mu_a \right\} = 0.$$

Note that information in the imputed dataset is aggregated to estimate $p(X; \alpha)$.

Several studies have compared MI-Across with MI-Within approach. Mitra and Reiter [40] reported that Across approach reduce bias more than Within approach and provoked controversy. The following studies [41, 42, 55, 56] claimed that Mitra's results are driven by the selection of the imputation model not to include outcome and the superiority of Across approach was limited to matching estimators. They showed the contrary by simulation studies. Granger et al. [56] showed that Across approach has more biased than the Within approach with matching estimators as well. However, the previous studies limited their scope to propensity score analysis and do not implement doubly robust estimator, which can achieve semiparametric efficiency bound when all

the working models are correctly specified and has the appealing property. The other issue of Mitra's simulation scenario is that data is generated under the sharp causal null hypothesis, which may not be an interesting and plausible situation and it can affect performance of causal estimators. In the main text, we adjusted the outcome process and missing data mechanism of Mitra's scenario and use it for simulation study. we present results from Mitra's scenario in appendix 0.3 for a comparison.

5.2 Combination IPW-based approach and causal inference methods

In this section, We describe estimation of the target functionals using IPW-based approach (IPCCW and AIPCCW estimators). The IPCCW estimators are not efficient because it only uses complete cases to fit a model. The AIPCCW estimators can improve efficiency by incorporating information on incomplete cases. We start with the description of IPCCW estimators. That section is not directly related with causal inference but underlies the following section, which is IPCCW with causal inference methods. Next, we provide general description of AIPCCW estimators and propose three AIPCCW estimators for μ_a depending on how to estimate augmentation term and the type of missingness patterns.

Description of the IPCCW estimators

Suppose we want to make inferences about parameter β and β is the unique solution of the following full data population estimating equation

$$\mathrm{E}[M(O;\beta)] = 0 \tag{3.12}$$

When some parts of data are missing, we may not identify (3.12) from data, motivating us to use the following IPW estimating equation

$$\mathrm{E}\left[\frac{I(R=1)}{\pi_1(O)} M(O;\beta)\right] = 0. \tag{3.13}$$

If we know missing data mechanism and true $\pi_1(O)$, under a certain regularity condition and when the positivity of $\pi_1(O)$ (3.10) holds, the estimates obtained by solving (3.13) is unbiased estimates of β [46] and unbiasedness can be derived from iterative expectation.

In practice, we do not have information on missing data mechanism so need to identify and estimate $\pi_1(O)$ from data. MAR missing data mechanism allows us to identify $\pi_1(O)$ as (3.9). However, when complete data O is high dimensional, we may not estimate $\pi_1(O)$ non-parametrically and need to rely on parametric models. The strategy of modelling $\pi_1(O)$ depends on the type of missingness patterns. Especially, under non-monotone missingness pattern it is hard to estimate $\pi_1(O)$ using parametric models but recently, Sun and Tchetgentchetgen [29] used Constrained Bayesian analysis to estsimate $\pi_1(O)$ under non-monotone missingness [4]. The details about missingness patterns and modelling strategies can be found elsewhere [39]. Henceforth, we denote models for $\pi_1(O)$ by $\pi_1(O;\gamma)$ with a finite dimensional parameter γ regardless of whether it is estimated by parametric models or Bayesian analysis. The model $\pi_1(O;\gamma)$ represents the models for $\pi_2(O_2;\gamma_2), \pi_3(O_3;\gamma_3), \ldots, \pi_r(O_r;\gamma_r)$ collectively. From now, we postulate that $\pi_1(O;\gamma)$ is estimated and $\pi_1(O_i;\widehat{\gamma})$ is an unbiased estimate of $\pi_1(O;\gamma)$.

Then, the IPCCW estimator for β is the solution of the following estimating equation,

$$\sum_{i=1}^{n} \frac{I(R_i = 1)}{\pi_1(O_i;\widehat{\gamma})} M(O_i;\beta) = 0. \tag{3.14}$$

The estimator is consistent when the model $\pi_1(O;\gamma)$ is correctly specified and under the standard regularity conditions. For example, suppose we are interested in regression coefficients β^* that comes from ordinary linear model $Y = X\beta^* + \epsilon$, where $\mathrm{E}[\epsilon] = 0$. The IPCCW estimator for β^* is the solution of the following estimating equation

$$\sum_{i=1}^{n} \frac{I(R_i = 1)}{\pi_1(O_i;\widehat{\gamma})} X_i^T (Y_i - X_i\beta^*) = 0.$$

Henceforth, we denote the models for other missing data patterns as $\pi_r(O_{(r)};\gamma)$, not indexing r to γ for simplicity and assume that the missingness model $\pi_r(O;\gamma)$ is not

[4] In appendix 0.2, we show how multinomial logistic models impose more strict constraints to missing data mechanism than MAR

known so need to be estimated, and model $\pi_r(O_{(r)};\gamma)$ is correctly specified.

Combination of the IPCCW estimators and causal inference methods

The IPCCW estimators can be used to estimate target functionals μ_a from observed data by regarding estimating functions for the target functionals as full data estimating function, $M(O;\beta)$, in IPCCW estimating equation (3.14). For example, if we know treatment model $p(X;\alpha)$, we can construct IPCCW estimating equation with IPTW estimating function (3.3) by plugging $M(O;\mu_a,\alpha)$ into (3.14), which is

$$\sum_{i=1}^{n}\left\{\frac{I(R_i=1,A_i=a)Y_i}{p(X_i;\alpha)\pi_1(O_i;\widehat{\gamma})}-\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}\mu\right\}=0. \tag{3.15}$$

We refer to a solution of the above IPCCW estimating equation as IPCCW-IPTW estimator [5]. The soultion is

$$\widehat{\mu}_a=\sum_{i=1}^{n}\left\{\frac{I(R_i=1,A_i=a)Y_i}{p(X_i;\alpha)\pi_1(O_i;\widehat{\gamma})}\right\}\Big/\sum_{i=1}^{n}\left\{\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}\right\}$$

and that can be thought as a normalized version of the IPCCW estimator. In large sample, $\sum_{i=1}^{n}\left\{\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}\right\}$ converges to n so we can also consider a non-normalized version of the IPCCW estimator

$$\widehat{\mu}_a=\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(R_i=1,A_i=a)Y_i}{p(X_i;\alpha)\pi_1(O_i;\widehat{\gamma})}\right\}.$$

If we do not know nuisance models, we need to estimate them as well as the target functionals by solving stacked estimating equations. The key of utilizing IPCCW approach to our stacked estimating equations environment is to weight all the estimating equations. As an example, suppose we want to solve the stacked AIPTW estimating equation (3.5) to estimate μ_a but data have missing covariates. Because covariates X

[5] meaning we estimate μ_a from the IPCCW estimating equation wirh IPTW estimating functiona as a full data estimating function

are incomplete, the first and second row of the estimating equation (3.5) is not identifiable from observed data so we need to use IPCCW estimating equation to estimate α and θ and the last row also should be upweighted. The resulting IPCCW estimating equations are

$$\begin{pmatrix} \sum_{i=1}^{n} \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})} M_g(O_i;\theta) \\ \sum_{i=1}^{n} \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})} M_p(A_i,X_i;\alpha) \\ \sum_{i=1}^{n} \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})} M(O_i;\mu_a,\alpha,\theta) \end{pmatrix} = \mathbf{0}, \tag{3.16}$$

The estimating procedure is the same as (3.5). When model for $\pi_1(O;\gamma)$ and $p(X;\alpha)$ or $g_a(X;\theta)$ are correctly specified and the other positivity and regularity conditions hold, the solution $\widehat{\mu}_a$ of the estimating equation is a consistent estimator of μ_a. Note that we end up weighting the estimating equation for nuisance parameters as well as the target parameter.

Augmented inverse probability of complete cases weights (AIPCCW) estimators

The limitation of IPCCW estimators is that it does not efficiently utilize information of incomplete cases so in general, IPCCW estimators are not as efficient as MI-based estimators [36]. The efficiency of IPCCW estimators can be improved by augmenting information of incomplete cases to the estimation procedure, leading to the development of an augmented IPCCW (AIPCCW) estimators. We leave details about the derivation of AIPCCW estimators to other literature [29, 27, 28]. The AIPCCW estimator is a solution to the following estimating equation

$$\sum_{i=1}^{n} \left\{ \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})} M(O_i;\beta) + B(R_i, O_{(i,R_i)};\widehat{\gamma}) \right\} = 0, \tag{3.17}$$

where $M(O;\beta)$ and $\pi_1(O;\gamma)$ are defined as before and $B(R_i, O_{(i,R_i)};\gamma)$ is called an augmentation term and represents an element of set $\mathbb{B}$ which consists of a span of all

scores of the missing data mechanism and have a form of

$$\left\{ \sum_{r \neq 1} \left[\frac{I(R_i = 1)}{\pi_1(O_i; \gamma)} - \frac{I(R_i = r)}{\pi_r(O_{i,(r)}; \gamma)} \right] t_r(O_{i,(r)}) \right\}, \tag{3.18}$$

where $t_r(O_{(r)})$ is an arbitrary function of observed data $O_{(r)}$ under missing data pattern $R = r$. When the augmentation term is zero, the estimating equation (3.14) and (3.17) are the same so we can think IPCCW estimators as a subclass of AIPCCW estimators. Among AIPCCW estimators, the most efficient one achieves semiparametric efficiency bound and often we are interested in estimating it. Constructing the most efficient AIPCCW estimator requires the selection of the optimal full data estimating function $M(O; \beta)$ in the space of orthogonal complement of full-data nuisance tangent space $(\Lambda^{F\perp})$and the optimal augmentation term $B^{opt}(R, O_{(R)}; \gamma)$ in the augmentation space Λ_2 represented by (3.18) this case. However, deriving the optimal choice of $M(O; \beta)$ and $B(R, O_{(R)}; \gamma)$ requires solving complicated integral equations that we need to solve numerically and it is often computationally prohibitive to find numerical solutions [29, 57, 28]. Instead of delving into impractical cases, we focus on AIPCCW estimators under fixed $M(O; \beta)$. Details on the ways to construct efficient IPCCW estimator and discussion on the challenges with regard to it can be found in [28] and a summarized version is provided in chapter 8 of [57].

When full data estimating function $M(O; \beta)$ is fixed, the key is to find the optimal augmentation term. The optimal choice for $B(R, O_{(R)}; \gamma)$ is

$$B^{opt}(R, O_{(R)}; \gamma) = -\Pi \left\{ \frac{I(R = 1) M(O; \beta)}{\pi_1(O; \gamma)} \middle| \Lambda_2 \right\},$$

where $\Pi(x|H)$ means the projection of x onto space H. In general, the projection does not have closed form solutions but it is feasible under univariate missingness and monotone missingness pattern. When missingness pattern is univariate, such as $O = (Y, A, X_o, X_m)$ and X_o is always observable and X_m is potentially missing, the optimal augmentation term is

$$B^{opt}(R, O_{(R)}; \gamma) = - \left[\frac{I(R = 1) - \pi_1(O; \gamma)}{\pi_1(O; \gamma)} \right] \mathrm{E}[M(O; \beta) | Y, A, X_o].$$

The AIPCCW estimator for β when $M(O;\beta)$ is fixed and under univariate missingness is a solution to the estimating equation in our case,

$$\sum_{i=1}^{n}\left\{\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}M(O_i;\beta)-\left[\frac{I(R_i=1)-\pi_1(O_i;\widehat{\gamma})}{\pi_1(O_i;\widehat{\gamma})}\right]\mathrm{E}[M(O;\beta)|Y_i,A_i,X_{i,o}]\right\}=0.$$

The optimal choice of the augmentation term under monotone missingness pattern can be found in [28]. We focus on the case of univariate and non-monotone missingness pattern next section and simulation studies.

The AIPWCC estimators with the causal inference methods

We start this section with a simple case where we know all the nuisance models, $p(X;\alpha)$ and $g_a(X;\theta)$ and develop an idea from it. Let $M(O;\mu_a)$ denotes the fixed estimating function for μ_a. Then, the AIPCCW estimtor with the corresponding causal estimating function, $M(O;\mu_a)$ is a solution of the estimating equation,

$$\sum_{i=1}^{n}\left\{\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}M(O_i;\mu_a)+B(R_i,O_{(i,R_i)};\widehat{\gamma})\right\}=0. \tag{3.19}$$

Under univariate missingness, the optimal augmentation term, $B^{opt}(R,O_{(R)};\gamma)$, has a closed form but not in general (non-monotone missingness pattern). It is theoretically interesting to examine this AIPCCW estimators under univariate missingness because we can construct the exact or very similar estimators to theoretical one and compare their performance to other existing methods. Thus, we focus on a case of univariate missingness in simulation studies and apply univariate missingness analysis to applied data analysis as well. For non-monotone missingness, we implement a restricted class of AIPCCW estimator which restricts the search for optimal full data estimating function and augmentation term to subspaces of $\Lambda^{F\perp}$ and Λ_2 represented by linear basis functions. This method is introduced by Tsiatis [28] and used by Sun and Tchetgentchetgen [29]. We implement this method to handle univariate missingness in simulation studies

and non-monotone missingness in applied data analysis.

Univariate missingness

Consider the setting of the equation (3.19). When, missingness pattern is univariate, the optimal augmentation term B^{opt} is

$$B^{opt}(R, O; \gamma) = -\left[\frac{I(R=1) - \pi_1(O;\gamma)}{\pi_1(O;\gamma)}\right] \mathrm{E}[M(O;\mu_a)|Y, A, X_o].$$

and an AIPCCW estimator is the solution to the equation

$$\sum_{i=1}^{n} \left\{ \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})} M(O_i;\mu_a) - \left[\frac{I(R_i=1) - \pi_1(O_i;\widehat{\gamma})}{\pi_1(O_i;\widehat{\gamma})}\right] \mathrm{E}[M(O;\mu_a)|Y_i, A_i, X_{o,i}] \right\} = 0. \tag{3.20}$$

The issue here is that we do not know the functional $\mathrm{E}[M(O;\mu_a)|Y, A, X_o]$ and need to estimate it. There are different strategies to estimate $\mathrm{E}[M(O;\mu_a)|Y, A, X_o]$; we can compute $M(O;\mu_a)$ for complete cases and regress it on (Y, A, X_o) [57] or we can posit a model for conditional distribution of X_m given (Y, A, X_o) and utilize it [58, 43]. We follow the latter strategy, utilizing a model for the conditional distribution of X_m and elaborate our strategy later. We refer the conditional distribution of X_m as the imputation model and denote it as $t(X_m|Y, A, X_o; \psi)$, where ψ is a finite dimensional parameter.

Now, let us move to the case of unknown nuisance parameters and when we need to estimate them. This is the focus of our paper and more plausible situation in real data analysis. We give an example of an AIPCCW estimator with AIPTW full data estimating function to explain this case. Define $M_g(O;\theta)$, $M_p(A, X;\alpha)$, $M(O;\mu_a,\alpha,\theta)$ as in (3.5), (3.7), (3.6). Extending the idea of building the IPCCW-AIPTW estimating equation (3.16), weighting all the estimating equations, we may naturally think thr efficient AIPCCW estimating equations by adding the optimal augmentation terms to each estimating equation, such as

$$\begin{pmatrix} \sum_{i=1}^{n} \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})} M_g(O_i;\theta) - \frac{I(R_i=1)-\pi_1(O_i;\widehat{\gamma})}{\pi_1(O_i;\widehat{\gamma})} \mathrm{E}\left[M_g(O) \middle| Y_i, A_i, X_{i,o}; \widehat{\psi} \right] \\ \sum_{i=1}^{n} \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})} M_p(A_i, X_i;\alpha) - \frac{I(R_i=1)-\pi_1(O_i;\widehat{\gamma})}{\pi_1(O_i;\widehat{\gamma})} \mathrm{E}\left[M_p(A,X) \middle| Y_i, A_i, X_{i,o}; \widehat{\psi} \right] \\ \sum_{i=1}^{n} \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})} M(O_i;\mu_a,\alpha,\theta) - \frac{I(R_i=1)-\pi_1(O_i;\widehat{\gamma})}{\pi_1(O_i;\widehat{\gamma})} \mathrm{E}\left[M(O) \middle| Y_i, A_i, X_{i,o}; \widehat{\psi} \right] \end{pmatrix} = \mathbf{0},$$

where ψ in the conditional expectation means that the imputation model is used to compute the conditional expectation.

This intuition is a base of Williamson et al.'s AIPCCW estimators for μ_a [58]. They constructed and solved the above estimating equation through their proposed algorithm and claimed their estimator has a multiple robust property. However, Evans et al. [43] pointed out that the estimator of Williamson et al. cannot achieve multiple robustness property because they did not take into account model compatibility between the imputation model and nuisance models (treatment and outcome model); besides, Williamson et al's algorithm requires statistical software that can handle negative weights, which is not common. Two conditional distributions, $g(y|x)$ and $h(x|y)$ are compatible if a such joint distribution $f(x,y)$ exists. In our case, treatment model and imputation model, outcome model and imputation model are not guaranteed to be compatible; then, joint full data model may not exist. Also, two models are incompatible means one of them are not true, which implies fitting one of them gives an implicit restriction to choices of other model to make sure the models are compatible. If not, true model cannot be a candidate from our model choices. This implicit restriction threatens the assumption of variation independent parameters as well [43]. Considering this issue, Evans et al. reparametrized conditional likelihood, $f(Y, X_m|X_o, A)$ and $f(A|X_o, X_m)$, to show model dependencies among the working models and derive variationally independent components. They proposed the conditions and strategies to build the efficient AIPCCW estimators using standard parametrization that we are using in this paper; it is hard to directly model variationally independent parameters from their reparametrization. Exposition on model compatibility and variation independence among treatment, outcome, and imputation models and Evans' algorithm to build efficient AIPCCW estimators can

be found in their paper [43] and Arnold and Press's paper [59] delivers a concept and examples of compatible conditional distributions.

Our approach to build AIPCCW estimators is based on Evans et al's approach but we adjusted their approach two ways. One thing is that we do not stick to making all the working models compatible. We analyze Mitra's simulation scenario to examine whether we can make treatment, outcome, and imputation model compatible in appendix 0.4. Even under this simple simulation scenario, with only two continuous covariates and modelled via ordinary linear regression, very restricted cases can satisfy compatibility among the models and if we add more variables, it will be harder to obtain compatible models. In real data analysis, it would not be desirable to limit our model choices based on the compatibility; rather it is better to broaden our model choices and increase goodness of fit of our models. This issue also arises in multiple imputation; however this can be minor if imputation model fits data well [60]. Even if the conditional distributions (imputation models) are incompatible, imputation-based estimators are still consistent if the conditional distributions are correctly specified [61]. Also, in our simulation and the other study [62], even working models are not guaranteed to be compatible, AIPCCW estimators has much better performance than IPCCW estimators. However, because of this issue, we may not claim that our AIPCCW estimators achieve semiparametric efficiency bound. Next, Evans et al. utilized a moment generating function to compute $\mathrm{E}[M(O;\mu_a)|Y,A,X_o]$, which is only feasible when the normal distribution is used to model covariates X. To build AIPCCW estimating equations in more general cases, we modified this part of Evans et al.'s algorithm and estimate $\mathrm{E}[M(O;\mu_a)|Y,A,X_o]$ either via parametric modeling or non-parametric Monte Carlo method. We describe our algorithms to obtain AIPCCW estimators with the AIPTW estimating function as an example.

In the description, define $\eta(O)$ as

$$\eta(O;\alpha,\theta) \equiv M(O;\alpha,\theta,\mu_a) + \mu_a,$$

and other notation is defined as previous sections.

Algorithm 1. Estimating $\mathbf{E}[\eta(O)|A, Y, X_o]$ from parametric modelling

1) Fit parametric models or Bayesian analysis to obtain estimates of conditional probabilities of being complete cases for each subject, $\pi_1(O; \widehat{\gamma})$.

2) Specify a model for the conditional density of X_m given the other variables from complete cases. The MAR missing data mechanism assures that the density in complete cases is the same as the density in the full data: $f(X_m|A, Y, X_o) = f(X_m|A, Y, X_o, R = 1)$. Let $t(X_m|A, Y, X_o; \hat{\psi})$ be the specified imputation model.

3) Create M duplicate of our dataset and stack the datasets to make a long dataset.

4) In the long dataset, fill in missing X_m from random draws from the imputation model. That means if we fit $t(X_m|A, Y, X_o; \hat{\psi})$ using an ordinary linear regression model with homoscedascity. We get random sample from $N \sim (\hat{\psi_0} + \hat{\psi_1}A + \hat{\psi_2}Y + \hat{\psi_3}X_o,\ \widehat{\sigma}^2)$, where $\psi_0, \psi_1, \psi_2, \psi_3$ are coefficients and $\widehat{\sigma}^2$ is the mean squared error of the imputation model.

5) Build a parametric model of $p(X; \alpha)$ for complete cases in original dataset using a weighted logistic regression model where the weights are $1/\pi_1(O; \widehat{\gamma})$. Let $p(X; \widehat{\alpha}_\pi)$ be the estimates from this model from complete cases. We index parameter α with π to emphasize that the consistency of the estimates from the model depends on the model for $\pi_1(O; \gamma)$

6) Build a parametric model for $g_a(X; \theta)$ in the long dataset, get estimates from the model for every observation in the long dataset. Let $g_a(X_{i,m}; \widehat{\theta}_t)$ denote the estimates for m-th duplicate of i-th individual from this model. We index parameter θ with t to emphasize that the consistency of the model depends on the imputation model, $t(X_m|A, Y, X_o; \psi)$.

7) Obtain estimates of $p(X; \alpha)$ for every observation in the long dataset. For complete cases, it is easy to obtain estimates from model $p(X; \alpha_\pi)$ but for missing cases, it's not obvious because their predictor X_m is missing. For missing cases, use imputed X_m to get the estimates. Let $p(X_{i,m}; \widehat{\alpha}_{\pi,t})$ be the estimates for m-th duplicate of i-th individual from this procedure. Because the coefficients are from the model for $p(X; \alpha_\pi)$ but also we used imputed X_m drawn from $t(X_m|A, Y, X_o; \hat{\psi})$, we index parameter α with π and t.

8) Use $p(X;\widehat{\alpha}_\pi)$, $p(X;\widehat{\alpha}_{\pi,t})$, and $g_a(X;\widehat{\gamma}_t)$ from the previous steps, estimate $\eta(O;\alpha,\theta)$ of every observation in the long dataset. Estimated $\eta(O;\alpha,\theta)$ for individual i is

$$\eta(O_{i,m};\widehat{\alpha}_{\pi,t}{}^{6},\widehat{\theta}_t) = \frac{I(A_i=a)Y_i}{p(X_{i,m};\widehat{\alpha}_{\pi,t})}{}^{7} - \frac{I(A_i=a)g_a(X_{i,m};\widehat{\theta}_t)}{p(X_{i,m};\widehat{\alpha}_{\pi,t})} + g_a(X_{i,m};\widehat{\theta}_t).$$

9) Fit a model for $\mathrm{E}[\eta(O;\widehat{\alpha}_{\pi,t},\widehat{\theta}_t)|Y,A,X_o]$ in the long dataset and get predictions from the model for every observations and denote the prediction by

$$\mathrm{E}[\eta(O_{i,m};\widehat{\alpha}_{\pi,t},\widehat{\theta}_t)|Y_i,A_i,X_{i,o};\widehat{\zeta}]$$

for m-th duplicate of i-th individual. Practically, that means fitting a regression model of $\eta(O;\widehat{\alpha}_{\pi,t},\widehat{\theta}_t)$ against (Y,A,X_o) and ζ refers to a parameter of this model. Next, taking an average of the individuals' estimates over their duplicates. Let $\mathrm{E}[\eta(O_i;\widehat{\alpha}_{\pi,t},\widehat{\theta}_t)|Y_i,A_i,X_{i,o};\widehat{\zeta}]$ be the average of the predictions for individual i,

$$\mathrm{E}[\eta(O_i;\widehat{\alpha}_{\pi,t},\widehat{\theta}_t)|A_i,Y_i,X_{i,o};\widehat{\zeta}] = \frac{1}{M}\sum_{m=1}^{M}\mathrm{E}[\eta(O_{i,m};\widehat{\alpha}_{\pi,t},\widehat{\theta}_t)|Y_i,A_i,X_{i,o};\widehat{\zeta}]$$

10) Construct the following estimating equation referring to (3.20). All the terms of the equation can be obtained via previous steps. The solution to the estimating equation is an estimator for μ_a.

$$\sum_{i=1}^{n}\Big\{\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}\left(\frac{I(A_i=a)Y_i}{p(X_i;\widehat{\alpha_\pi})} - \frac{I(A_i=a)g_a(X_i;\widehat{\theta}_t)}{p(X_i;\widehat{\alpha_\pi})} + g_a(X_i;\widehat{\theta}_t)\right)$$
$$- \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}\mathrm{E}[\eta(O_i;\widehat{\alpha}_{\pi,t},\widehat{\gamma})|A_i,Y_i,X_{i,o};\widehat{\zeta}] + \mathrm{E}[\eta(O_i;\widehat{\alpha}_{\pi,t},\widehat{\gamma})|A_i,Y_i,X_{i,o};\widehat{\zeta}] - \mu_a\Big\} = 0$$

We refer to the solution of the equation as an AIPCCW-AIPTW(Model) estimator for μ_a because we obtain a AIPCCW estimator with the AIPTW estimating function as full data estimating function and use a parametric model to estimate $\mathrm{E}[\eta(O)|A,Y,X_o]$.

[6] we use $p(X;\widehat{\alpha}_\pi)$ for complete cases and $p(X;\widehat{\alpha}_{\pi,t})$ for incomplete cases here. However, we didn't distinguish them for simplicity. Next step, this term will be a dependent variable of regression model to estimate $\mathrm{E}[\eta(O)|A,Y,X_o]$. Predictions from the regression model will be affected by both $t(X_m|A,Y,X_o;\hat{\psi})$ and $\pi_1(O;\widehat{\gamma})$ anyway so the distinguishment does not give useful information here.

[7] We suppress index m for outcome and treatment because they are complete and no variation across duplicates

The estimates can be affected by model specification for $\mathrm{E}[\eta(O)|A, Y, X_o]$.

Algorithm 2. Estimating $\mathbf{E}[\eta(O)|A, Y, X_o]$ from Monte Carlo method

Step 1) through 6) are the same as algorithm 1. Remind that Each individual has M complete duplicated records through Step 1) to Step 4) and we can obtain estimates $p(X_{i,m}; \widehat{\alpha}_\pi)$ and $g_a(X_{i,m}; \widehat{\theta}_t)$ in step 5) and step 6).

7) In the long dataset, fit a model for $p(X; \alpha)$ and get estimates from the model. Let $p(X_{i,m}; \widehat{\alpha}_t)$ be the estimates of m-th duplicate for i-th individual.

8) Compute estimates of $\eta(O; \alpha, \theta)$ of m-th duplicate for i-th individual, using $p(X_{i,m}; \widehat{\alpha}_t)$ and $g_a(X_{i,m}; \widehat{\theta}_t)$,

$$\eta(O_{i,m}; \widehat{\alpha}_t, \widehat{\theta}_t)) = \frac{I(A_i = a)Y_i}{p(X_{i,m}; \widehat{\alpha}_t)} - \frac{I(A_i = a)g_a(X_{i,m}; \widehat{\theta}_t)}{p(X_{i,m}; \widehat{\alpha}_t)} + g_a(X_{i,m}; \widehat{\theta}_t)$$

9) Taking an average of $\eta(O_{i,m}; \widehat{\alpha}_t, \widehat{\theta}_t)$ of individual i over M duplications and set the average as an estimate of $\mathrm{E}[\eta(O)|A, Y, X_o]$ for individual i, which is denoted by $\widehat{\mathrm{E}}[\eta(O_i; \widehat{\alpha}_t, \widehat{\theta}_t)|A_i, Y_i, X_{i,o}]$,

$$\widehat{\mathrm{E}}[\eta(O_i; \widehat{\alpha}_t, \widehat{\theta}_t)|A_i, Y_i, X_{i,o}] = \frac{1}{M} \sum_{m=1}^{M} \eta(O_{i,m}; \widehat{\alpha}_t, \widehat{\theta}_t)$$

10) Construct the following estimating equation referring to (3.19). All the terms of the equation can be obtained via previous steps and the solution to the estimating equation is an estimator for μ_a.

$$\sum_{i=1}^{n} \Big\{ \frac{I(R_i = 1)}{\pi_1(O_i; \widehat{\gamma})} \left(\frac{I(A_i = a)Y_i}{p(X_i; \widehat{\alpha_\pi})} - \frac{I(A_i = a)g_a(X_i; \widehat{\theta}_t)}{p(X_i; \widehat{\alpha_\pi})} + g_a(X_i; \widehat{\theta}_t) \right) - \frac{I(R_i = 1)}{\pi_1(O_i; \widehat{\gamma})} \widehat{\mathrm{E}}[\eta(O_i; \widehat{\alpha}_t, \widehat{\theta}_t)|A_i, Y_i, X_{i,o}] + \widehat{\mathrm{E}}[\eta(O_i; \widehat{\alpha}_t, \widehat{\theta}_t)|A_i, Y_i, X_{i,o}] - \mu_a \Big\} = 0$$

We refer to the solution of the equation as an AIPCCW-AIPTW(MC) estimator for μ_a because we use a Monte Carlo method to estimate $\mathrm{E}[\eta(O)|A, Y, X_o]$.

This way does not rely on a parametric model to estimate $\mathrm{E}[\eta(O)|A, Y, X_o]$ but consistency depends on the imputation model. It may be hard to derive asymptotic

variances of those estimators to accommodate the uncertainty of random draws from the imputation model; we recommend using non-parametric bootstrap to estimate the variance. We illustrated how to build a AIPCCW estimator with AIPTW estimating function but this example can be extended to estimating other types of the target functionals. We provide the other types of estimators in appendix 0.1.

Non-monotone missingness

Univariate missingness pattern is not common in real data setting. Most of real data is under non-monotone missingness pattern but there is no closed form of the augmentation term and finding the optimal choices of the augmentation term and full data estimating function is often computationally infeasible. We introduce one way to approximate the optimal choices by restricting searching space to linear subspaces represented by linear basis functions [28, 29] and propose a way to apply this method to build AIPCCW estimating equations for the target functionals.

Description of the restricted AIPCCW estimators

Suppose we estimate q-dimensional full data parameter β and $M(O;\beta)$ is a specified and fixed estimating function for β. Under non-monotone missingness (as a general form), AIPCCW estimator of β is the solution of estimating equation (3.17). The challenging thing is to find optimal choices of $M(O;\beta)$ among the set of full estimating functions for β in $\Lambda^{F\perp}$ and $B(R, O_{(R)};\gamma)$ among the elements of augmentation space Λ_2 (See section 5.4.3); in most cases, they are computationally infeasible. Instead of searching the optimal choices through the entire space of $\Lambda^{F\perp}$ or Λ_2, we can restrict our search to subspaces where computation to find optimal choices is approachable. Tsiatis [28] suggested the way to find optimal choices of $M(O;\beta)$ and $B(R, O_{(R)};\gamma)$ within the subspaces of $\Lambda^{F\perp}$ or Λ_2 represented by linear basis functions. Sun and Tchetgentchetgen [29] used this method when they handle non-monotone missing data pattern.

When we utilize this method to estimate q-dimensional parameter β, the estimating

equation has the following form (Tsiatis, 2006 page 297)

$$\sum_{i=1}^{n}\left[\frac{I(R_i=1)B^{F^{q\times t_1}}m^*(O_i,\beta)}{\pi_1(O_i;\gamma)}+B_2^{q\times t_2}J_2\{R_i,O_{i,(R_i)};\widehat{\gamma}\}\right]=0, \tag{3.21}$$

where $m^*(O,\beta)$ is a t_1-dimensional full data estimating function, $J_2\{R,O_{(R)}\}$ is a t_2-dimensional function of $\{R,O_{(R)}\}$, $B^{F^{q\times t_1}}$ and $B_2^{q\times t_2}$ are arbitrary constant matrices. Here, optimal full data estimating function is found in t_1 dimensional space and optimal augmentation term is found in t_2 dimensional space; t_1 elements are linearly independent each other and t_2 are linearly independent each other and $t_1 > q$. The optimal constant matrix $B^{F^{q\times t_1}}$ and $B_2^{q\times t_2}$ can be figured through matrix computation to find a projection. The details on how to find optimal $B^{F^{q\times t_1}}$ and $B_2^{q\times t_2}$ are provided in [28, 29]. The performance of the method depends on how close the t_1 or t_2 dimensional subspaces are to the entire spaces. In other words, the optimal choice derived from the subspace can be similar to the optimal choice from the entire space. We explain how to apply this method to build AIPCCW estimating equations for the target functionals next subsection.

Restricted AIPCCW estimators in causal inference

Estimation of the target functionals entails estimating nuisance parameters in usual cases and that requires solving stacked estimating equations. However, as far as we know, the restricted AIPCCW estimators are not used in this context so we propose one way to obtain the restricted AIPCCW estimators with stacked estimating equation aiming at the target functionals. The point of our approach is to fix the full data estimating function for μ_a, even this approach allows us to approximate an optimal full data estimating function in a larger class of full data estimating functions. Finding the optimal full data estimating function for μ_a is computationally more complicated than finding it for nuisance parameters because the class of estimators for μ_a already has an augmentation term [27] so we need to find the optimum of three terms dealing with two projections.

Our approach to obtain a AIPCCW estimator for μ_a has the following steps. First, build the restricted AIPCCW estimating equation for nuisance parameter using (3.21).

Next, obtain an improved estimates of nuisance parameters by solving the estimating equations. Plug those estimates into the estimating function for μ_a and solving the restricted AIPCCW estimating equation while we fix full data estimating function. We give an example of estimating the IPTW functional by our approach.

Algorithm 3. Restricted AIPCCW estimators for the target functionals

1) Fit parametric models or Bayesian analysis to obtain estimates of conditional probabilities of being complete cases for each subject, $\pi_1(O;\widehat{\gamma})$.

2) Consider the following estimating equation[8]:

$$\begin{pmatrix} \sum_{i=1}^{n}\left\{\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}C_1h^*(X_i)\big(I(A_i=a)-p(X_i;\alpha)\big)+C_2B^*(R_i,O_{i,(R)};\widehat{\gamma})\right\} \\ \sum_{i=1}^{n}\left\{\frac{I(A_i=a,R_i=1)Y_i}{p(X_i;\alpha)\pi_1(O_i;\widehat{\gamma})}+C_3B^*(R_i,O_{i,(R)};\widehat{\gamma})-\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}\mu_a\right\} \end{pmatrix}=0. \tag{3.22}$$

In the first estimating equation, $h^*(X)$ is a user-defined t_1 dimensional function of X, $B^*(R_i,O_{(R_i)};\gamma)$ is a t_2 dimensional vector which is represented as

$$\left\{\left[\frac{I(R_i=1)}{\pi_1(O_i;\gamma)}-\frac{I(R_i=r)}{\pi_r(O_{i,(r)};\gamma)}\right]t_r(O_{i,(r)}):r=2,3,\ldots,K\right\}, \tag{3.23}$$

where K is the number of distinct missing data patterns and $t_r(O_{(r)})$ is a user defined function of $O_{(r)}$ for each missing data pattern, C_1 is a $q\times {t_1}$[9] constant matrix, and C_2 is a $q\times t_2$ constant matrix. First, we need to define $h^*(X)$ and $t_r(O_{(r)}:r=2,\ldots,K)$ as a function of linearly independent vectors that characterize our searching space for the optimal choices.

3) Find the optimal constant matrices C_1 and C_2; the algorithm for the task can be found in [29].

4) Plug the optimal constant matrices and solve the upper row. We can obtain the estimate of α as a solution of the estimating equation, which is denote by $\widehat{\alpha}_{AP}$. Plug the estimate $\widehat{\alpha}_{AP}$ into the lower estimating equation.

5) Because we fix our full data estimating function, we only need to find the optimal

[8]We denote both augmentation spaces of the upper row and the lower row using the same notation (A^*) but they do not have to be the same.

[9]Suppose nuisance parameter α for $p(X_i;\alpha)$ is q-dimensional.

augmentation term. We can find the optimal constant matrix C_3^{opt} by solving the following equation,

$$\prod \left[\frac{I(R=1)I(A=a)Y}{\pi_1(O;\widehat{\gamma})p(X;\widehat{\alpha}_{AP})} \middle| \mathcal{B}^* \right] = -C_{3opt}B^*(R, O_{(R)};\widehat{\gamma}),$$

where $\mathcal{B}^*$ is the linear subspace of Λ_2 consisting of the span of $B^*(R_i, O_{(R)};\widehat{\gamma})$, such as

$$\mathcal{B}^* = \{D^{1\times t_2}B^*(R, O_{(R)};\widehat{\gamma}) \text{ for all constant matrices } D^{1\times t_2}\}.$$

The calculation of C_3^{opt} reduces to find a projection of an one-dimensional vector to a space.

6) Plug C_3^{opt} into the estimating equation. The restricted AIPCCW estimator is a solution to the estimating equation.

We adopted this approach to handle non-monotone missingness pattern but it can be used for the other missingness patterns. However, under univariate missingness pattern, this approach is less efficient than the previous approaches because the previous approaches utilize the closed form of the optimal augmentation term but this approach searches the optimal one from restricted space. Because IPCCW estimators can belong to the AIPCCW estimators, We expect that it is more efficient than the IPCCW estimators when all the working models are correctly specified and all the conditions hold.

6 Simulation study

We investigate finite sample performance of proposed estimators under univariate missingness in simulation studies. Covariate $X = (X_1, X_2)$ are drawn from a bivariate normal distribution where $\mu = (10, 10)^T$, $\text{Var}(X_1) = \text{Var}(X_2) = 5$, and $\text{Cov}(X_1, X_2) = 2.5$. The binary treatment variable A is generated from the logistic model,

$$\text{logit}\Pr[A = 1|X] = -5 + 0.25X_1 + 0.25X_2.$$

The continuous outcome variable Y is generated from the following distribution,

$$Y = -X_1 - X_2 + 5A + \theta_1 X_1 A + \theta_2 X_2 A + \epsilon, \quad \epsilon \sim N(0,1).$$

We used two different sets of θ, $(\theta_1, \theta_2 : \theta_1 = \theta_2 = 2, \theta_1 = \theta_2 = 0.5)$, to change the degree of the interaction between A and X. After complete data (X, A, Y) is generated, we induced missing values on X_2 by missingness model:

$$\text{logit}\Pr[R = 2|X_1, A] = \gamma_0 + 0.6X_1 + A - 0.3X_1 A,$$

where R is an indicator variable for missing data patterns and $R = 2$ corresponds to missing data pattern where X_2 have missing values. We adjust γ_0 to have a proportion of missing cases to 0.15 ($\gamma_0 = -7.060$) or 0.3 ($\gamma_0 = -5.918$). We generated 10000 replicates of each scenario with $n = 2000, 5000, 10000$, and implemented the estimators in each replicate. Models of missingness, treatment and outcome are correctly specified. For the imputation model, we cannot compute the closed form of the true model. Instead, we fit a linear regression model for X_2 against X_1, A, Y among complete cases where we use natural spline terms with four inner knots of X_1 and Y. Even if it may not be the true model for X_2 conditional on (X_1, A, Y), we cannot find any issues of residual plots and consider it as very close to the true model. This practice might be closer to real data analysis in that investigators have much more knowledge on missingness, treatment, and outcome process than the conditional distribution of missing covariates so they would utilize flexible models to fit the model. Evans et al. [43] tweaked data generation process from commonly used one and made a scenario where all the nuisance models are true but we use the common data generation process to make it possible to compare our study to other studies and to make our simulation study more interpretable and understanable to readers. We included complete cases analysis results for a comparison. As for MI procedure, we implemented multivariate imputation by chained equations (MICE), multivariate normal imputation, and predictive mean matching with the number of imputation data sets $m = 5, 10, 20, 50$. We used 'Amelia' package in R to execute multivariate normal imputation, 'mice' package with 'norm'

option (Bayesian linear regression) to execute MICE, and 'mice' package with 'pmm' option to execute predictive mean matching. We followed Sun and Tchetgentchetgen's way [29] of the choice of basis functions for the restricted AIPCCW approach, including linear, quadratic, and pairwise interaction terms and set the number of duplicates for AIPCCW-model and AIPCCW-MC to $M = 100$. We compute estimates of mean potential outcome and of each group and causal mean difference and compute absolute average bias, variance, and mean squared error (MSE) across 10000 iterations for each estimator and compare their performance.

The simulation results are summarized in Table 3.2, 3.3, 3.4 for the scenario of $\theta_1 = \theta_2 = 2$ and the proportion of complete cases in data 0.7. The results of MI are based on multivariate normal imputation and when the number of imputed dataset is 50 ($m = 50$) but the other approaches (MICE and predictive mean matching) and different options of m gave similar results. All the estimators are not biased except for complete cases analysis (CC) estimators and Across MI estimators. CC estimators are substantially biased and Across MI estimators are biased slightly. Both estimators do not improve as sample size increases. In terms of variance, AIPCCW-IPTW estimators show unstable behavior in some cases (Table 3.3). However, when AIPTW estimating function is used instead of IPTW, this instability disappears and the AIPCCW estimators are more efficient than the IPCCW estimators, and the efficiency improvement is substantial. Among the AIPCCW estimators, AIPCCW-MC is slightly more efficient than AIPCCW-Model approach but the difference is small. Restricted AIPCCW approach is less efficient than the other two AIPCCW approaches. However, the restricted AIPCCW estimators are still more efficient than the IPCCW estimators. Comparing MI and IPW-based estimators, the IPCCW estimators are less efficient than MI-based estimators but AIPCCW estimators are comparable to MI-based estimators. CC estimators have high mean squared error because of its substantial bias and the IPCCW estimators have the second highest MSE because they are less efficient than the other approaches. MI-based estimators and the AIPCCW estimators have similar MSEs. Simulation results of a different degree of interaction (different θ) and different missing proportion are provided in appendix 0.5.

7 Application example

The empirical application concerns the effectiveness of coronary artery bypass grafting surgery with medication (henceforth we call it surgery) versus medication only on reducing mortality rates for patients with stable ischemic heart disease in the CASS (Coronary artery surgery study). CASS can be divided into two parts, a randomized controlled trial and an observational study. We included all 2099 eligible participants to our study regardless of participation into the trial. We excluded 6 individuals from analysis because of data issues. Baseline covariates were collected in the same manner among randomized and non-randomized participants. The details about study protocol is available elsewhere [33, 34]. The target estimand is 10-year mortality risk difference between the surgery and medication arm. Missing data pattern of the data is summarized in Table 3.7. We collapsed few missing data patterns because very few observations belong to these patterns and we formed four missing data patterns (non-monotone pattern). To illustrate AIPCCW-model and MC approaches, we also formed data with univariate missing data pattern by only preserving complete cases and missing cases only on ejection fraction ($R = 2$ in Table 3.7) (univariate pattern).

We implemented the estimators as in the simulation except for the AIPCCW-model and MC in non-monotone pattern analysis. Logistic regression is used to fit the treatment model and linear regression is used to fit the outcome model with the following variables chosen by the previous study [35]: an indicator for participation to the trial, age, ejection fraction, percent obstruction of the proximal left anterior descending artery, left ventricular wall motion score, history of previous myocardial infarction, number of diseases proximal vessels. We made a categorized age and ejection fraction variables by dividing them into four evenly distributed groups and used the categorized variables for the analysis of non-monotone missingness pattern because it helps to stabilize the restricted AIPWCC estimator. Multiple imputation assuming multivariate normality with $m = 20$ was used in the analysis and the missingness model was fit by logistic regression model for univariate analysis and Constrained Bayesian analysis was used to estimate missing data mechanism in non-monotone analysis with the same variables

as the outcome and treatment model. We used a linear regression model to fit the imputation model for AIPWCC estiatmors in univariate analysis. Outcome model and missingness model are built separately by treatment arm. Non-parametric bootstrap resampling with 500 samples is used to make inferences about the estimates. Normal-distribution based 95% confidence intervals are reported in the results.

In univariate analysis, all the estimators addressing missing data produced similar results (Table 5). The estimates of 10-year mortality risk difference between surgery and medication arm is in the range of [-0.013, -0.016] when we handled missing ejection fraction with MI or IPW based approaches. The estimates of the mortality risk of complete cases analysis is different, around -0.0195 and the confidence interval is wider than MI and IPW-based approaches. The 95% confidence interval of IPCCW estimates are slightly wider than AIPCCW estimates but difference is very small. This small difference, unlike the simulation study, might result from the small causal effects in the data. The confidence interval of the restricted AIPWCC estimator is wider than any other estimators. In our experience, as more covariates are included in analysis, the restricted AIPCCW estimator comes to unstable. Results of non-monotone analysis is similar to univariate analysis. The confidence interval of IPW-based estimators are wider and thr restricted AIPWCC estimator are more unstable compared to the univariate analysis. two missing data patterns of non-monotone analysis have few observations (Table 3.7). The inverse of the probabilities of being those rare missing data patterns is directly used to compute the augmentation term of the restricted AIPCCW estimator so it affects the variability of the restricted AIPCCW estimator.

8 Discussion

We studied several topics on estimating mean potential outcome using observed data. Our contribution can be summarized to 1) compare AIPTW approach with the other approaches of causal inference in the context of combinations with missing data methods, 2) compare MI-based and IPW-based missing data methods when they are used with causal inference methods, 3) compare Across-MI and Within-MI when they

are used with diverse causal inference methods, 4) propose AIPWCC estimators for mean potential outcome applicable under univariate and non-monotone missingness patterns.

Our results have few discrepancies with theory. The AIPCCW-IPTW estimators are less efficient than IPCCW-IPTW estimator in simulation with small sample size. We think that this results are from inherent numerical instability of IPTW estimators. This numerical instability is not mitigated by adding the augmentation term for missingness; rather the augmentation worsen the situation. However, the AIPCCW estimators are better with large sample size or combined with AIPTW estimating function where the augmentation term effectively handles numerical instability. The other discrepancy is that the restricted AIPCCW estimator is less efficient than IPCCW and other estimators in application study. We expected that at least the restricted AIPCCW estimator is more efficient than IPCCW estimators because IPCCW estimators belong to the class of the restricted AIPCCW estimator with zero augmentation term. One possible reason for it is that missingness models in the application study is misspecified or the assumption of MAR missing data mechanism does not hold. Under these conditions, the estimated conditional probabilities can be apart from the true values and the augmentation space formed with these misspecified probabilities may be wrongly fit with the true augmentation space we should consider. The projection of IPW estimating function to the linear subspace of that 'wrongly fitted space', which is the geometric meaning of the optimal augmentation term, would be very different from the projection to the true augmentation space and it is possible that this wrongly fit augmentation term is worse than zero augmentation term (IPCCW estimator) with respect to the efficiency gain. Moreover, our formulation of the linear subspace contains the inverse of the probabilities of having rare missing data patterns that can add numerical instability to the augmentation term. We follow the previous study [29] to formulate the augmentation term but that is not the only way. In future studies, more robust formulation to model misspecification can be considered to improve the restricted AIPWCC estimators.

The limitations of this study connect to future research topics. In this study, we

Table 3.1: The combinations we studied

		Causal inference methods		
		g-formula	IPTW	AIPTW
Missing data methods	MI-across	✓	✓	✓
	MI-Within	✓	✓	✓
	IPCCW	✓	✓	✓
	AIPCCW - model		✓	✓
	AIPCCW - MC			
	Restricted AIPCCW		✓	✓

rely on traditional parametric approaches to estimate nuisance parameters although machine learning methods have been a popular to obtain more accurate predictions in high dimensional settings. Recently, the use of machine learning approaches along with modern semiparametric methods is formally discussed [63] and this general theory has applied to the estimation of causal effects [63, 64]. However, still there is not much literature on the application of machine learning approaches to estimation problems in missing data, which will be the area of future studies.

Table 3.2: **Bias** of the estimators where $\theta_1 = \theta_2 = 2$ and the proportion of complete cases is 0.7.

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	$E[Y^1]$	-0.4028	-0.0472	-0.5942	-0.4026	-0.4026	-0.0002	-0.0011	0.0010	-0.0001	-0.0001
	$E[Y^0]$	0.4024	1.4066	0.6885	0.4026	0.4026	0.0008	0.0015	0.0015	0.0008	0.0008
	$E[Y^1 - Y^0]$	-0.8052	-1.4538	-1.2827	-0.8052	-0.8053	-0.0010	-0.0026	-0.0005	-0.0009	-0.0009
n = 5,000	$E[Y^1]$	-0.4031	-0.0478	-0.5927	-0.4030	-0.4030	-0.0018	-0.0001	-0.0022	-0.0017	-0.0017
	$E[Y^0]$	0.4029	1.4175	0.6938	0.4029	0.4029	0.0038	0.0271	0.0069	0.0039	0.0039
	$E[Y^1 - Y^0]$	-0.8061	-1.4653	-1.2865	-0.8059	-0.8059	-0.0032	-0.0056	-0.0053	-0.0030	-0.0030
n = 2,000	$E[Y^1]$	-0.4020	-0.0522	-0.5847	-0.4020	-0.4020	-0.0005	0.0013	0.0013	-0.0002	-0.0002
	$E[Y^0]$	0.4016	1.4140	0.6955	0.4017	0.4017	-0.0002	0.0079	0.0035	-0.0003	-0.0003
	$E[Y^1 - Y^0]$	-0.8037	-1.4663	-1.2802	-0.8038	-0.8038	-0.0003	-0.0066	-0.0023	0.0002	0.0002

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	$E[Y^1]$	-0.0341	0.0029	-0.0110	-0.0110	0.0121	-0.0029	0.0012	0.0003
	$E[Y^0]$	0.2131	0.0335	-0.0545	-0.0550	0.0601	0.0235	-0.0008	0.0000
	$E[Y^1 - Y^0]$	-0.2472	-0.0306	0.0435	0.0440	-0.0480	-0.0263	0.0020	0.0003
n = 5,000	$E[Y^1]$	-0.0344	0.0000	-0.0107	-0.0126	0.0229	-0.0073	0.0000	-0.0003
	$E[Y^0]$	0.2199	0.0358	-0.0538	-0.0544	0.1177	0.0442	-0.0007	-0.0005
	$E[Y^1 - Y^0]$	-0.2543	-0.0358	0.0412	0.0417	-0.0948	-0.0515	0.0007	0.0002
n = 2,000	$E[Y^1]$	-0.0238	0.0028	-0.0077	-0.0077	0.0459	-0.0178	-0.0057	-0.0016
	$E[Y^0]$	0.1901	0.0323	-0.0457	-0.0461	0.2361	0.0860	0.0041	0.0010
	$E[Y^1 - Y^0]$	-0.2590	-0.0326	0.0442	0.0447	-0.1901	-0.1038	-0.0098	-0.0026

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	$E[Y^1]$	0.0111	0.0003	0.0002	0.0002	0.0119	0.0003	0.0005	0.0005	0.0011	-0.0018
	$E[Y^0]$	0.0460	-0.0001	0.0000	0.0000	0.0368	0.0000	0.0035	0.0035	-0.0212	0.0016
	$E[Y^1 - Y^0]$	-0.0349	0.0004	0.0001	0.0001	-0.0250	0.0003	-0.0030	-0.0030	0.0223	-0.0034
n = 5,000	$E[Y^1]$	0.0062	-0.0005	-0.0004	-0.0004	0.0229	-0.0003	0.0000	0.0000	0.0019	-0.0043
	$E[Y^0]$	0.0383	0.0003	0.0002	0.0002	0.0799	-0.0005	0.0042	0.0042	-0.0231	0.0026
	$E[Y^1 - Y^0]$	-0.0321	-0.0008	-0.0006	-0.0006	-0.0570	0.0002	-0.0041	-0.0041	0.0249	-0.0069
n = 2,000	$E[Y^1]$	0.0407	-0.0016	0.0000	0.0000	0.0462	-0.0016	0.0004	0.0004	0.0066	-0.0119
	$E[Y^0]$	0.1403	0.0007	-0.0011	-0.0011	0.1747	0.0010	0.0053	0.0053	-0.0664	0.0084
	$E[Y^1 - Y^0]$	-0.0996	-0.0023	0.0011	0.0010	-0.1284	-0.0026	-0.0049	-0.0049	0.0730	-0.0203

Table 3.3: **Variance** of the estimators where $\theta_1 = \theta_2 = 2$ and the proportion of complete cases is 0.7.

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	$E[Y^1]$	0.0025	0.0169	0.0131	0.0026	0.0026	0.0019	0.0040	0.0040	0.0019	0.0019
	$E[Y^0]$	0.0027	0.1053	0.0152	0.0029	0.0029	0.0019	0.0130	0.0035	0.0020	0.0020
	$E[Y^1 - Y^0]$	0.0096	0.0903	0.0377	0.0098	0.0098	0.0069	0.0135	0.0118	0.0069	0.0069
n = 5,000	$E[Y^1]$	0.0051	0.0345	0.0254	0.0053	0.0053	0.0036	0.0083	0.0078	0.0037	0.0037
	$E[Y^0]$	0.0056	0.1983	0.0282	0.0059	0.0059	0.0038	0.0271	0.0069	0.0039	0.0039
	$E[Y^1 - Y^0]$	0.0196	0.1735	0.0716	0.0202	0.0202	0.0133	0.0267	0.0237	0.0134	0.0134
n = 2,000	$E[Y^1]$	0.0125	0.0840	0.0587	0.0131	0.0131	0.0091	0.0216	0.0185	0.0092	0.0092
	$E[Y^0]$	0.0136	0.5344	0.0721	0.0145	0.0145	0.0092	0.0703	0.0172	0.0092	0.0092
	$E[Y^1 - Y^0]$	0.0480	0.4723	0.1740	0.0497	0.0497	0.0324	0.0694	0.0567	0.0326	0.0326

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	$E[Y^1]$	0.0040	0.0040	0.0019	0.0019	0.0091	0.0090	0.0143	0.0027
	$E[Y^0]$	0.0123	0.0033	0.0020	0.0020	0.2040	0.0320	0.0117	0.0037
	$E[Y^1 - Y^0]$	0.0131	0.0116	0.0070	0.0070	0.2196	0.0653	0.0501	0.0112
n = 5,000	$E[Y^1]$	0.0082	0.0078	0.0037	0.0037	0.0193	0.0170	0.0279	0.0054
	$E[Y^0]$	0.0258	0.0067	0.0040	0.0040	0.3051	0.0477	0.0229	0.0072
	$E[Y^1 - Y^0]$	0.0261	0.0234	0.0135	0.0135	0.3448	0.1066	0.0980	0.0221
n = 2,000	$E[Y^1]$	0.0217	0.0181	0.0091	0.0091	0.0579	0.0404	0.0735	0.0137
	$E[Y^0]$	0.0674	0.0166	0.0094	0.0094	0.5369	0.0815	0.0594	0.0175
	$E[Y^1 - Y^0]$	0.0680	0.0560	0.0329	0.0329	0.6654	0.2077	0.2570	0.0552

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	$E[Y^1]$	0.0110	0.0027	0.0021	0.0021	0.0093	0.0027	0.0018	0.0018	0.0077	0.0027
	$E[Y^0]$	0.1033	0.0037	0.0031	0.0031	0.2953	0.0037	0.0022	0.0022	0.6814	0.0224
	$E[Y^1 - Y^0]$	0.0786	0.0112	0.0088	0.0088	0.2596	0.0112	0.0069	0.0069	0.5967	0.0113
n = 5,000	$E[Y^1]$	0.0101	0.0040	0.0038	0.0038	0.0201	0.0054	0.0037	0.0037	0.0163	0.0054
	$E[Y^0]$	0.0833	0.0046	0.0044	0.0044	0.6963	0.0072	0.0045	0.0045	1.1583	0.0077
	$E[Y^1 - Y^0]$	0.0740	0.0154	0.0145	0.0145	0.6048	0.0221	0.0140	0.0140	1.0015	0.0230
n = 2,000	$E[Y^1]$	0.0633	0.0137	0.0106	0.0106	0.0587	0.0137	0.0091	0.0091	0.0498	0.0138
	$E[Y^0]$	0.2802	0.0176	0.0143	0.0142	2.8853	0.0175	0.0123	0.0122	3.1474	0.0195
	$E[Y^1 - Y^0]$	0.1770	0.0552	0.0424	0.0423	2.4980	0.0552	0.0358	0.0356	2.6950	0.0574

Table 3.4: **MSE** of the estimators where $\theta_1 = \theta_2 = 2$ and the proportion of complete cases is 0.7.

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	$E[Y^1]$	0.1647	0.0191	0.3662	0.1647	0.1647	0.0019	0.0040	0.0040	0.0019	0.0019
	$E[Y^0]$	0.1647	2.0839	0.4892	0.1650	0.1650	0.0019	0.0130	0.0035	0.0020	0.0020
	$E[Y^1 - Y^0]$	0.6580	2.2039	1.6830	0.6583	0.6583	0.0069	0.0136	0.0118	0.0069	0.0069
n = 5,000	$E[Y^1]$	0.1676	0.0368	0.3766	0.1677	0.1677	0.0036	0.0083	0.0078	0.0037	0.0037
	$E[Y^0]$	0.1679	2.2076	0.5096	0.1682	0.1682	0.0039	0.0278	0.0069	0.0039	0.0039
	$E[Y^1 - Y^0]$	0.6694	2.3206	1.7266	0.6696	0.6696	0.0133	0.0268	0.0237	0.0134	0.0134
n = 2,000	$E[Y^1]$	0.1741	0.0867	0.4006	0.1748	0.1747	0.0091	0.0216	0.0185	0.0092	0.0092
	$E[Y^0]$	0.1749	2.5339	0.5558	0.1758	0.1759	0.0092	0.0704	0.0172	0.0092	0.0092
	$E[Y^1 - Y^0]$	0.6939	2.6223	1.8128	0.6957	0.6957	0.0324	0.0694	0.0567	0.0326	0.0326

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	$E[Y^1]$	0.0052	0.0040	0.0020	0.0020	0.0092	0.0090	0.0143	0.0027
	$E[Y^0]$	0.0578	0.0044	0.0050	0.0051	0.2076	0.0326	0.0117	0.0037
	$E[Y^1 - Y^0]$	0.0742	0.0126	0.0089	0.0090	0.2219	0.0660	0.0501	0.0112
n = 5,000	$E[Y^1]$	0.0094	0.0078	0.0038	0.0038	0.0198	0.0171	0.0279	0.0054
	$E[Y^0]$	0.0742	0.0079	0.0069	0.0069	0.3190	0.0497	0.0229	0.0072
	$E[Y^1 - Y^0]$	0.0907	0.0247	0.0152	0.0153	0.3537	0.1092	0.0980	0.0221
n = 2,000	$E[Y^1]$	0.0223	0.0181	0.0091	0.0091	0.0600	0.0407	0.0735	0.0137
	$E[Y^0]$	0.1035	0.0177	0.0115	0.0115	0.5927	0.0889	0.0594	0.0175
	$E[Y^1 - Y^0]$	0.1351	0.0571	0.0348	0.0349	0.7015	0.2185	0.2571	0.0552

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	$E[Y^1]$	0.0111	0.0027	0.0021	0.0021	0.0095	0.0027	0.0018	0.0018	0.0077	0.0027
	$E[Y^0]$	0.1054	0.0037	0.0031	0.0031	0.2967	0.0037	0.0022	0.0022	0.6819	0.0224
	$E[Y^1 - Y^0]$	0.0798	0.0112	0.0088	0.0088	0.2602	0.0112	0.0069	0.0069	0.5972	0.0114
n = 5,000	$E[Y^1]$	0.0102	0.0040	0.0038	0.0038	0.0206	0.0054	0.0037	0.0037	0.0163	0.0055
	$E[Y^0]$	0.0848	0.0046	0.0044	0.0044	0.7027	0.0072	0.0045	0.0045	1.1588	0.0077
	$E[Y^1 - Y^0]$	0.0750	0.0154	0.0145	0.0145	0.6081	0.0221	0.0140	0.0140	1.0021	0.0230
n = 2,000	$E[Y^1]$	0.0650	0.0137	0.0106	0.0106	0.0609	0.0137	0.0091	0.0091	0.0498	0.0139
	$E[Y^0]$	0.2999	0.0176	0.0143	0.0142	2.9158	0.0175	0.0123	0.0122	3.1518	0.0196
	$E[Y^1 - Y^0]$	0.1869	0.0552	0.0424	0.0423	2.5145	0.0552	0.0358	0.0357	2.7004	0.0578

Table 3.5: Empirical univariate

Missing data methods		Complete Cases analysis				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
Potential Outcome Under Surgery	Point estimate	0.1793	0.1794	0.1797	0.1795	0.1795
	95% CI	[0.1539 0.2048]	[0.1541 0.2048]	[0.1543 0.2051]	[0.1541 0.2049]	[0.1541 0.2049]
	Width of CI	0.0509	0.0508	0.0508	0.0508	0.0508
Potential Outcome Under Medication	Point estimate	0.1989	0.1991	0.1989	0.1990	0.1990
	95% CI	[0.1742 0.2236]	[0.1742 0.224]	[0.174 0.2238]	[0.1743 0.2237]	[0.1743 0.2237]
	Width of CI	0.0494	0.0498	0.0498	0.0494	0.0494
Causal Risk Difference	Point estimate	-0.0196	-0.0197	-0.0192	-0.0195	-0.0195
	95% CI	[-0.0545 0.0154]	[-0.0547 0.0153]	[-0.0542 0.0159]	[-0.0544 0.0154]	[-0.0544 0.0154]
	Width of CI	0.0699	0.0701	0.0701	0.0699	0.0699
Missing data methods		Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
Potential Outcome Under Surgery	Point estimate	0.1847	0.1848	0.1853	0.1850	0.1850
	95% CI	[0.1602 0.2091]	[0.1603 0.2093]	[0.1608 0.2098]	[0.1605 0.2094]	[0.1605 0.2094]
	Width of CI	0.0490	0.0489	0.0490	0.0490	0.0490
Potential Outcome Under Medication	Point estimate	0.1997	0.2001	0.1997	0.1998	0.1998
	95% CI	[0.1768 0.2225]	[0.177 0.2232]	[0.1767 0.2228]	[0.177 0.2227]	[0.177 0.2227]
	Width of CI	0.0457	0.0462	0.0462	0.0457	0.0457
Causal Risk Difference	Point estimate	-0.0150	-0.0153	-0.0145	-0.0149	-0.0148
	95% CI	[-0.0483, 0.0183]	[-0.0486, 0.0181]	[-0.0478, 0.0189]	[-0.0481, 0.0184]	[-0.0481, 0.0184]
	Width of CI	0.0666	0.0667	0.0668	0.0666	0.0666
Missing data methods		Across MI				
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	
Potential Outcome Under Surgery	Point estimate	0.1848	0.1853	0.1851	0.1851	
	95% CI	[0.1603 0.2092]	[0.1607 0.2098]	[0.1605 0.2096]	[0.1605 0.2096]	
	Width of CI	0.0489	0.0490	0.0490	0.0490	
Potential Outcome Under Medication	Point estimate	0.2000	0.1997	0.1996	0.1996	
	95% CI	[0.1769 0.2231]	[0.1766 0.2228]	[0.1768 0.2224]	[0.1768 0.2224]	
	Width of CI	0.0462	0.0461	0.0457	0.0457	
Causal Risk Difference	Point estimate	-0.0153	-0.0145	-0.0145	-0.0145	
	95% CI	[-0.0486, 0.0181]	[-0.0478, 0.0189]	[-0.0478, 0.0188]	[-0.0478, 0.0188]	
	Width of CI	0.0667	0.0668	0.0666	0.0666	
Missing data methods		IPCCW				
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	
Potential Outcome Under Surgery	Point estimate	0.1863	0.1869	0.1865	0.1866	
	95% CI	[0.1616 0.211]	[0.1621 0.2116]	[0.1618 0.2112]	[0.1619 0.2114]	
	Width of CI	0.0494	0.0495	0.0494	0.0494	
Potential Outcome Under Medication	Point estimate	0.2002	0.2000	0.1997	0.1999	
	95% CI	[0.1767 0.2237]	[0.1765 0.2235]	[0.1765 0.223]	[0.1766 0.2232]	
	Width of CI	0.0470	0.0470	0.0465	0.0465	
Causal Risk Difference	Point estimate	-0.0139	-0.0131	-0.0133	-0.0133	
	95% CI	[-0.0475, 0.0197]	[-0.0468, 0.0205]	[-0.0468, 0.0203]	[-0.0468, 0.0203]	
	Width of CI	0.0672	0.0673	0.0670	0.0671	

Missing data methods		AIPCCW (Model)	AIPCCW (MC)	Restricted AIPCCW
Causal estimators		AIPTW	AIPTW	AIPTW
Potential Outcome Under Surgery	Point estimate	0.1853	0.1851	0.1974
	95% CI	[0.1608 0.2099]	[0.1606 0.2096]	[0.1708 0.2241]
	Width of CI	0.0491	0.0490	0.0532
Potential Outcome Under Medication	Point estimate	0.1990	0.1998	0.2061
	95% CI	[0.176 0.222]	[0.1606 0.2096]	[0.1812 0.2311]
	Width of CI	0.0460	0.0458	0.0499
Causal Risk Difference	Point estimate	-0.0137	-0.0147	-0.0087
	95% CI	[-0.0471, 0.0197]	[-0.0481, 0.0186]	[-0.0464, 0.0289]
	Width of CI	0.0668	0.0667	0.0713

Table 3.6: Empirical non-monotone

Missing data methods		Complete Cases analysis				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
Potential Outcome Under Surgery	Point estimate	0.1793	0.1794	0.1797	0.1795	0.1795
	95% CI	[0.1539 0.2048]	[0.1541 0.2048]	[0.1543 0.2051]	[0.1541 0.2049]	[0.1541 0.2049]
	Width of CI	0.0509	0.0508	0.0508	0.0508	0.0508
Potential Outcome Under Medication	Point estimate	0.1989	0.1991	0.1989	0.1990	0.1990
	95% CI	[0.1742 0.2236]	[0.1742 0.224]	[0.174 0.2238]	[0.1743 0.2237]	[0.1743 0.2237]
	Width of CI	0.0494	0.0498	0.0498	0.0494	0.0494
Causal Risk Difference	Point estimate	-0.0196	-0.0197	-0.0192	-0.0195	-0.0195
	95% CI	[-0.0545 0.0154]	[-0.0547 0.0153]	[-0.0542 0.0159]	[-0.0544 0.0154]	[-0.0544 0.0154]
	Width of CI	0.0699	0.0701	0.0701	0.0699	0.0699
Missing data methods		Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
Potential Outcome Under Surgery	Point estimate	0.1871	0.1868	0.1868	0.1868	0.1868
	95% CI	[0.1631, 0.2112]	[0.1627, 0.2109]	[0.1627, 0.2109]	[0.1627, 0.2109]	[0.1627, 0.2109]
	Width of CI	0.0481	0.0482	0.0481	0.0481	0.0481
Potential Outcome Under Medication	Point estimate	0.2008	0.1999	0.1999	0.2005	0.2005
	95% CI	[0.1778, 0.2238]	[0.1768, 0.2230]	[0.1769, 0.2230]	[0.1776, 0.2235]	[0.1776, 0.2235]
	Width of CI	0.0460	0.0462	0.0461	0.0459	0.0459
Causal Risk Difference	Point estimate	-0.0137	-0.0131	-0.0131	-0.0137	-0.0137
	95% CI	[-0.0468, 0.0195]	[-0.0462, 0.0200]	[-0.0462, 0.0200]	[-0.0468, 0.0193]	[-0.0468, 0.0193]
	Width of CI	0.0663	0.0663	0.0662	0.0661	0.0661
Missing data methods		Across MI				
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	
Potential Outcome Under Surgery	Point estimate	0.1864	0.1867	0.1871	0.1871	
	95% CI	[0.1624, 0.2105]	[0.1627, 0.2108]	[0.1630, 0.2112]	[0.1630, 0.2112]	
	Width of CI	0.0481	0.0481	0.0482	0.0482	
Potential Outcome Under Medication	Point estimate	0.1996	0.1999	0.2001	0.2001	
	95% CI	[0.1765, 0.2226]	[0.1768, 0.2229]	[0.1771, 0.2231]	[0.1771, 0.2231]	
	Width of CI	0.0461	0.0461	0.0459	0.0459	
Causal Risk Difference	Point estimate	-0.0131	-0.0131	-0.0130	-0.0130	
	95% CI	[-0.0462, 0.0199]	[-0.0462, 0.0200]	[-0.0462, 0.0201]	[-0.0462, 0.0201]	
	Width of CI	0.0661	0.0661	0.0663	0.0663	
Missing data methods		IPCCW				Restricted AIPCCW
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	AIPTW
Potential Outcome Under Surgery	Point estimate	0.1825	0.1854	0.1827	0.1852	0.2096
	95% CI	[0.1562, 0.2089]	[0.1588, 0.2120]	[0.1562, 0.2092]	[0.1585, 0.2119]	[0.1544, 0.2649]
	Width of CI	0.0527	0.0532	0.0530	0.0534	0.1105
Potential Outcome Under Medication	Point estimate	0.2018	0.2042	0.2015	0.2043	0.2135
	95% CI	[0.1775, 0.2261]	[0.1798, 0.2287]	[0.1775, 0.2256]	[0.1800, 0.2285]	[0.1545, 0.2725]
	Width of CI	0.0486	0.0489	0.0482	0.0486	0.1180
Causal Risk Difference	Point estimate	-0.0193	-0.0188	-0.0188	-0.0191	-0.0039
	95% CI	[-0.0565, 0.0179]	[-0.0564, 0.0187]	[-0.0560, 0.0184]	[-0.0566, 0.0185]	[-0.0740, 0.0662]
	Width of CI	0.0744	0.0750	0.0744	0.0751	0.1402

Table 3.7: Descriptive statistics of missing covariates in CASS.

Missing data by variables

Variable	# missing	% missing
Age	0	0
Previous MI	1	0.04
The number of diseased vessels	6	0.3
Ejection fraction	372	17.8
Lvscor	56	2.7
LAD % obstruction	0	0

Missing data pattern

R	# Variables with missing data	# observations	%
1	No missing data	1686	80.6
2	Ejecfr	344	16.4
3	Lvscor	28	1.3
4	The number of diseased vessels	6	0.2
	Previous MI	1	0.04
	Lvscor + Ejecfr	28	1.3
	Total	35	1.7

R : An indicator variable for missing data pattern used in the analysis
Lvscor: Left ventricular wall motion score

0.1 Formulation of IPW-based estimators

We list IPCCW-based estimators under univariate missingness in the solution form of estimating equations.

IPCCW-IPTW estimator

$$\widehat{\mu}_a = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A_i = a, R_i = 1)Y_i}{p(X_i;\widehat{\alpha})\pi_1(O_i;\widehat{\gamma})}$$

Normalized IPCCW-IPTW estimator

$$\widehat{\mu}_a = \frac{\sum_{i=1}^{n} \frac{I(A_i = a, R_i = 1)Y_i}{p(X_i;\widehat{\alpha})\pi_1(O_i;\widehat{\gamma})}}{\sum_{i=1}^{n}\{p(X_i;\widehat{\alpha})\pi_1(O_i;\widehat{\gamma})\}^{-1}}$$

IPCCW-AIPTW estimator

$$\widehat{\mu}_a = \frac{1}{n}\sum_{i=1}^{n} \frac{I(A = a)Y}{p(X;\widehat{\alpha})\pi_1(O_i;\widehat{\gamma})} - \frac{I(A = a)g_a(X;\widehat{\theta})}{p(X;\widehat{\alpha})\pi_1(O_i;\widehat{\gamma})} + \frac{g_a(X;\widehat{\theta})}{\pi_1(O_i;\widehat{\gamma})}$$

Normalized IPCCW-AIPTW estimator

$$\widehat{\mu}_a = \frac{\sum_{i=1}^{n} \frac{I(A=a)Y}{p(X;\widehat{\alpha})\pi_1(O_i;\widehat{\gamma})} - \frac{I(A=a)g_a(X;\widehat{\theta})}{p(X;\widehat{\alpha})\pi_1(O_i;\widehat{\gamma})}}{\sum_{i=1}^{n}\{p(X_i;\widehat{\alpha})\pi_1(O_i;\widehat{\gamma})\}^{-1}} + \frac{\sum_{i=1}^{n} \frac{g_a(X;\widehat{\theta})}{\pi_1(O_i;\widehat{\gamma})}}{\sum_{i=1}^{n}\{\pi_1(O_i;\widehat{\gamma})\}^{-1}}$$

AIPCCW-IPTW estimator

$$\widehat{\mu}_a = \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{I(R_i = 1)}{\pi_1(O_i;\widehat{\gamma})} U_1(O_i;\widehat{\alpha}) - \frac{I(R_i = 1)}{\pi_1(O_i;\widehat{\gamma})} \mathrm{E}[U_1(O;\widehat{\alpha})|O_{i,o}] + \mathrm{E}[U_1(O;\widehat{\alpha})|O_{i,o}] \right\},$$

where $U_1(O;\alpha) = \dfrac{I(A = a)Y}{p(X;\alpha)}$, and $O_o = (A, Y, X_o)$.

Normalized AIPCCW-IPTW estimator

$$\widehat{\mu}_a = \frac{\sum_{i=1}^{n} \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})} U_1(O_i;\widehat{\alpha})}{\sum_{i=1}^{n}\{p(X_i;\widehat{\alpha})\pi_1(O_i;\widehat{\gamma})\}^{-1}} - \frac{\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}\mathrm{E}[U_1(O;\widehat{\alpha})|O_{i,o}]}{\sum_{i=1}^{n}\{\pi_1(O_i;\widehat{\gamma})\}^{-1}} + \mathrm{E}[U_1(O;\widehat{\alpha})|O_{i,o}],$$

where $U_1(O;\alpha) = \dfrac{I(A=a)Y}{p(X;\alpha)}$, and $O_o = (A, Y, X_o)$.

AIPCCW-AIPTW estimator

$$\widehat{\mu}_a = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}U_2(O_i;\widehat{\alpha},\widehat{\theta}) - \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}\mathrm{E}[U_2(O;\widehat{\alpha},\widehat{\theta})|O_{i,o}] + \mathrm{E}[U_2(O;\widehat{\alpha},\widehat{\theta})|O_{i,o}]\right\},$$

where $U_2(O;\alpha,\theta) = \dfrac{I(A=a)Y}{p(X;\alpha)} - \dfrac{I(A=a)}{p(X;\alpha)}g_a(X;\theta) + g_a(X;\theta)$, and $O_o = (A, Y, X_o)$.

Normalized AIPCCW-AIPTW estimator

$$\widehat{\mu}_a = \frac{\sum_{i=1}^{n} \frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})} U_2(O_i;\widehat{\alpha})}{\sum_{i=1}^{n}\{p(X_i;\widehat{\alpha})\pi_1(O_i;\widehat{\gamma})\}^{-1}} - \frac{\frac{I(R_i=1)}{\pi_1(O_i;\widehat{\gamma})}\mathrm{E}[U_2(O;\widehat{\alpha})|O_{i,o}]}{\sum_{i=1}^{n}\{\pi_1(O_i;\widehat{\gamma})\}^{-1}} + \mathrm{E}[U_2(O;\widehat{\alpha})|O_{i,o}],$$

where $U_2(O;\alpha,\theta) = \dfrac{I(A=a)Y}{p(X;\alpha)} - \dfrac{I(A=a)}{p(X;\alpha)}g_a(X;\theta) + g_a(X;\theta)$, and $O_o = (A, Y, X_o)$.

0.2 Failure of the multinomial logistic regression to model non-monotone missingness under MAR missing data mechanism

One of the important assumptions of the multinomial logistic model is the assumption of independence of irrelevant alternatives [65, 66]. That means that the odds of choosing one option to another option does not change regardless of the absence or presence of the third alternative choices. This assumption is not flexible enough to convey MAR mechanism. Suppose data is realization of a random vector $O = (X_1, X_2)$ and there are three missing data pattern $R = 1, 2, 3$. Complete cases are indicated by $R = 1$ and only X_2 is fully observable under

$R = 2$ and only X_1 is observable under $R = 3$. If we model the missing data mechanism using the multinomial logistic regression, the conditional probabilities of each missing data pattern is

$$\Pr[R = 1|X] = \frac{1}{1 + \exp(\beta_{01} + \beta_{11}X_1 + \beta_{21}X_2) + \exp(\beta_{02} + \beta_{12}X_1 + \beta_{22}X_2)}$$

$$\Pr[R = 2|X] = \frac{\exp(\beta_{01} + \beta_{11}X_1 + \beta_{21}X_2)}{1 + \exp(\beta_{01} + \beta_{11}X_1 + \beta_{21}X_2) + \exp(\beta_{02} + \beta_{12}X_1 + \beta_{22}X_2)}$$

$$\Pr[R = 3|X] = \frac{\exp(\beta_{02} + \beta_{12}X_1 + \beta_{22}X_2)}{1 + \exp(\beta_{01} + \beta_{11}X_1 + \beta_{21}X_2) + \exp(\beta_{02} + \beta_{12}X_1 + \beta_{22}X_2)},$$

where all the β are coefficients. Linear predictor $\beta_{01} + \beta_{11}X_1 + \beta_{21}X_2$ estimates log odds of $R = 2$ against $R = 1$ and $\beta_{02} + \beta_{12}X_1 + \beta_{22}X_2$ estimates log odds of $R = 3$ against $R = 1$, which means

$$\ln \frac{\Pr[R = 2|X]}{\Pr[R = 1|X]} = \beta_{01} + \beta_{11}X_1 + \beta_{21}X_2 \tag{.24}$$

$$\ln \frac{\Pr[R = 3|X]}{\Pr[R = 1|X]} = \beta_{02} + \beta_{12}X_1 + \beta_{22}X_2$$

The assumption of independence of irrelevant alternatives is manifested by that the log odds (linear predictor) is invariant among the probabilities of $R = 1, 2,$ and 3. This constraint is too strict to reflect MAR missing data mechanism. For example, under missing data pattern $R = 2$, $\Pr[R = 2|X]$ should be only dependent on X_2 to satisfy MAS missing data mechanism so $\beta_{11} = \beta_{12} = 0$ from (.24). However, when we set $\beta_{11} = \beta_{12} = 0$, that affects the choice of the other missing data pattern; $\Pr[R = 3|X]$ does not depend on X_1 even though X_1 is observable under missing data pattern $R = 3$. That violates the assumption of MAR missing data mechanism.

0.3 Results from Mitra's simulation scenario

At first, we implemented our methods to Mitra's [40] simulation scenario for a comparison to their results. However, we encountered unexpected results

that are AIPCCW estimators improve the efficiency of mean potential outcome under treatment ($A = 1$) or control ($A = 0$) substantially but do not improve the efficiency of causal mean difference. We figured that this unusual phenomenon is caused by their choice of simulation scenario, specifically holding the sharp causal null condition and adjusted their scenario and made new one. However, we think it is worthwhile to present the results from Mitra's scenario to show how our methods work under another situation. It is also educational to show how this issue arises from the sharp causal null condition.

Data generation of Mitra's scenario

Covariates X and outcome Y were generated as

$$(X_{i,o}, X_{i,m}) \sim N((10, 10)^T, \Sigma), \text{ where } \Sigma = \begin{bmatrix} 5 & 2.5 \\ 2.5 & 5 \end{bmatrix}$$

$$Y_i = X_{i,o} + X_{i,m} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, 1)$$

The treatment A has three scenarios about generation:

$$\text{(S1) } \operatorname{logit} \Pr[A = 1|x_o, x_m] = -7.8 + 0.5 * x_o$$

$$\text{(S2) } \operatorname{logit} \Pr[A = 1|x_o, x_m] = -7.8 + 0.5 * x_m$$

$$\text{(S3) } \operatorname{logit} \Pr[A = 1|x_o, x_m] = -7.8 + 0.225 * x_o + 0.225 * x_m$$

Two missing data mechanisms were studied: (i) Only control units have missing x_m and missing data mechanism is

$$\operatorname{logit}\{\Pr[R = 2|A = 0, x_o]\} = -10.1 + 0.9x_o,$$

where $R = 2$ means missing x_m, (ii) control units have missingness according to the same mechanism as (i) and additionally 30 % of treated units are randomly chosen to have missing x_m (MCAR). As a result, in scenario (i), approximately 30% of control units have missing x_m. In their original simulation. they did

not include outcome to MI models and that provokes a controversy. We fit a imputation model of MI or AIPCCW estimators with or without outcome also to make a corresponding version to IPCCW and AIPCCW estimators, we made a missingness model with or without outcome; for AIPCCW estimators we did not use outcome information to missingness and imputation models. We found that working models with outcome has better performance than without outcome here so we included outcome in all the imputation and missingness models in the simulation of the main text but present the comparison this section. We generated data with $n = 1100$ and approximated 100 units are treated ($A = 1$) and 1000 units are control units ($A = 0$).

Table 3.8: Bias of the estimators in S1 and missingness in only control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.005				
		IPTW	-0.037	-0.024	1.385	1.057	-0.028	3.225
		N_IPTW	0.099	0.099	-0.143	-0.003	0.028	
		AIPTW	0.006	0.006	-0.018	0.010	0.004	-0.021
		N_AIPTW	0.007	0.007	-0.004	0.011	0.007	
	$E[Y^0]$	g-formula		-0.002				
		IPTW	-0.004	-0.003	-0.126	-0.025	-0.003	-0.140
		N_IPTW	-0.003	-0.002	-0.050	-0.017	0.030	
		AIPTW	-0.002	0.005	-0.029	-0.002	-0.002	-0.032
		N_AIPTW	-0.002	0.005	-0.017	-0.002	-0.002	
	$E[Y^1 - Y^0]$	g-formula		0.007				
		IPTW	-0.033	-0.021	1.511	1.082	-0.025	3.365
		N_IPTW	0.102	0.101	-0.093	0.014	-0.002	
		AIPTW	0.008	0.001	0.011	0.012	0.006	0.011
		N_AIPTW	0.009	0.002	0.013	0.014	0.009	
Without Y	$E[Y^1]$	g-formula		0.004				
		IPTW	-0.067	-0.008	1.389	1.070	-0.028	3.243
		N_IPTW	0.093	0.092	-0.144	0.000	0.021	
		AIPTW	0.005	0.005	-0.020	0.017	0.001	-0.021
		N_AIPTW	0.006	0.006	-0.004	0.015	0.002	
	$E[Y^0]$	g-formula		-0.002				
		IPTW	-0.010	-0.004	-0.129	-0.030	-0.003	-0.142
		N_IPTW	-0.004	-0.003	-0.050	-0.017	0.032	
		AIPTW	-0.002	0.004	-0.032	-0.002	-0.002	-0.032
		N_AIPTW	-0.002	0.004	-0.017	-0.002	-0.002	
	$E[Y^1 - Y^0]$	g-formula		0.006				
		IPTW	-0.056	-0.004	1.518	1.100	-0.025	3.385
		N_IPTW	0.097	0.094	-0.094	0.017	-0.011	
		AIPTW	0.007	0.001	0.012	0.020	0.003	0.011
		N_AIPTW	0.008	0.002	0.013	0.018	0.004	

Table 3.9: Variance of the estimators in S1 and missingness in only control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.035				
		IPTW	1.620	1.623	7.627	5.812	1.636	32.918
		N_IPTW	0.580	0.580	1.252	0.092	2.882	
		AIPTW	0.050	0.050	0.426	0.084	0.053	0.165
		N_AIPTW	0.045	0.045	0.091	0.061	0.045	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.017	0.017	0.439	0.035	0.017	0.243
		N_IPTW	0.015	0.015	0.052	0.045	0.157	
		AIPTW	0.014	0.035	0.364	0.015	0.014	0.045
		N_AIPTW	0.014	0.035	0.045	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.025				
		IPTW	1.631	1.672	8.991	5.886	1.687	34.848
		N_IPTW	0.567	0.567	1.069	0.055	3.282	
		AIPTW	0.041	0.013	0.075	0.074	0.044	0.130
		N_AIPTW	0.036	0.008	0.054	0.052	0.035	
Without Y	$E[Y^1]$	g-formula		0.037				
		IPTW	1.585	1.596	7.637	5.819	1.612	32.615
		N_IPTW	0.578	0.578	1.255	0.113	2.817	
		AIPTW	0.052	0.052	0.419	0.108	0.096	0.166
		N_AIPTW	0.046	0.046	0.093	0.075	0.082	
	$E[Y^0]$	g-formula		0.014				
		IPTW	1.585	0.017	0.432	0.035	0.017	0.234
		N_IPTW	0.578	0.015	0.053	0.046	0.146	
		AIPTW	0.052	0.037	0.357	0.015	0.014	0.046
		N_AIPTW	0.046	0.037	0.046	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.027				
		IPTW	1.669	1.641	9.018	5.894	1.662	34.545
		N_IPTW	0.567	0.567	1.070	0.070	3.214	
		AIPTW	0.039	0.013	0.074	0.100	0.087	0.128
		N_AIPTW	0.034	0.008	0.054	0.067	0.074	

Table 3.10: MSE of the estimators in S1 and missingness in only control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.036				
		IPTW	1.621	1.624	9.544	6.929	1.637	43.320
		N_IPTW	0.590	0.590	1.272	0.092	2.882	
		AIPTW	0.050	0.050	0.426	0.084	0.053	0.166
		N_AIPTW	0.045	0.045	0.091	0.061	0.045	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.017	0.017	0.455	0.036	0.017	0.262
		N_IPTW	0.015	0.015	0.055	0.045	0.157	
		AIPTW	0.014	0.035	0.365	0.015	0.014	0.046
		N_AIPTW	0.014	0.035	0.045	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.025				
		IPTW	1.670	1.672	11.274	7.057	1.688	46.173
		N_IPTW	0.577	0.577	1.078	0.055	3.282	
		AIPTW	0.039	0.013	0.075	0.074	0.044	0.130
		N_AIPTW	0.034	0.008	0.055	0.052	0.036	
Without Y	$E[Y^1]$	g-formula		0.037				
		IPTW	1.585	1.596	9.567	6.964	1.613	43.130
		N_IPTW	0.578	0.586	1.276	0.113	2.818	
		AIPTW	0.052	0.052	0.420	0.108	0.096	0.166
		N_AIPTW	0.046	0.046	0.093	0.075	0.082	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.017	0.017	0.448	0.036	0.017	0.254
		N_IPTW	0.015	0.015	0.055	0.046	0.147	
		AIPTW	0.014	0.037	0.358	0.015	0.014	0.047
		N_AIPTW	0.014	0.037	0.046	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.027				
		IPTW	1.634	1.642	11.322	7.105	1.663	46.003
		N_IPTW	0.577	0.576	1.079	0.071	3.214	
		AIPTW	0.041	0.013	0.074	0.101	0.087	0.128
		N_AIPTW	0.036	0.008	0.055	0.068	0.074	

Table 3.11: Bias of the estimators in S1 and missingness in both case and control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.008				
		IPTW	-0.066	-0.023	1.356	1.037	-0.031	3.006
		N_IPTW	0.103	0.102	-0.190	-0.005	0.074	
		AIPTW	0.006	0.006	-0.016	0.008	0.005	0.001
		N_AIPTW	0.007	0.008	-0.005	0.009	0.007	
	$E[Y^0]$	g-formula		-0.002				
		IPTW	-0.005	-0.003	-0.124	-0.027	-0.002	-0.120
		N_IPTW	-0.003	-0.002	-0.049	-0.016	0.028	
		AIPTW	-0.002	0.008	-0.027	-0.002	-0.002	-0.012
		N_AIPTW	-0.002	0.008	-0.016	-0.002	-0.002	
	$E[Y^1 - Y^0]$	g-formula		0.010				
		IPTW	-0.061	-0.020	1.480	1.064	-0.028	3.127
		N_IPTW	0.106	0.105	-0.141	0.011	0.047	
		AIPTW	0.008	-0.002	0.011	0.010	0.007	0.012
		N_AIPTW	0.010	0.000	0.011	0.011	0.009	
Without Y	$E[Y^1]$	g-formula		0.009				
		IPTW	-0.235	-0.005	1.358	1.019	-0.038	3.019
		N_IPTW	0.102	0.096	-0.191	-0.005	0.062	
		AIPTW	0.005	0.005	-0.021	0.008	0.004	0.000
		N_AIPTW	0.006	0.007	-0.005	0.008	0.005	
	$E[Y^0]$	g-formula		-0.002				
		IPTW	-0.014	-0.004	-0.128	-0.030	-0.002	-0.123
		N_IPTW	-0.005	-0.003	-0.050	-0.017	0.030	
		AIPTW	-0.002	0.009	-0.032	-0.002	-0.002	-0.013
		N_AIPTW	-0.002	0.009	-0.017	-0.002	-0.002	
	$E[Y^1 - Y^0]$	g-formula		0.012				
		IPTW	-0.222	-0.001	1.486	1.049	-0.036	3.143
		N_IPTW	0.107	0.099	-0.141	0.011	0.031	
		AIPTW	0.007	-0.004	0.011	0.010	0.006	0.012
		N_AIPTW	0.008	-0.003	0.011	0.010	0.007	

Table 3.12: Variance of the estimators in S1 and missingness in both case and control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.037				
		IPTW	1.653	1.671	5.558	3.890	1.604	19.190
		N_IPTW	0.578	0.578	1.078	0.078	0.758	
		AIPTW	0.053	0.052	0.410	0.058	0.051	0.110
		N_AIPTW	0.046	0.045	0.078	0.047	0.045	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.017	0.017	0.442	0.034	0.017	0.280
		N_IPTW	0.015	0.015	0.052	0.045	0.162	
		AIPTW	0.014	0.037	0.367	0.015	0.014	0.045
		N_AIPTW	0.014	0.037	0.045	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.028				
		IPTW	1.704	1.722	6.842	3.968	1.654	21.046
		N_IPTW	0.565	0.565	0.887	0.040	1.172	
		AIPTW	0.043	0.014	0.052	0.048	0.040	0.073
		N_AIPTW	0.036	0.007	0.040	0.037	0.035	
Without Y	$E[Y^1]$	g-formula		0.061				
		IPTW	1.618	1.654	5.555	3.856	1.584	19.119
		N_IPTW	0.592	0.591	1.073	0.079	0.751	
		AIPTW	0.096	0.096	0.400	0.057	0.052	0.112
		N_AIPTW	0.084	0.083	0.079	0.047	0.047	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.017	0.017	0.432	0.033	0.017	0.270
		N_IPTW	0.015	0.015	0.053	0.046	0.156	
		AIPTW	0.014	0.061	0.357	0.015	0.014	0.046
		N_AIPTW	0.014	0.061	0.046	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.052				
		IPTW	1.662	1.703	6.854	3.935	1.632	20.960
		N_IPTW	0.582	0.582	0.882	0.040	1.151	
		AIPTW	0.088	0.031	0.051	0.047	0.041	0.074
		N_AIPTW	0.076	0.018	0.040	0.037	0.036	

Table 3.13: MSE of the estimators in S1 and missingness in both case and control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.037				
		IPTW	1.658	1.672	7.398	4.966	1.605	28.228
		N_IPTW	0.589	0.589	1.114	0.078	0.764	
		AIPTW	0.053	0.052	0.411	0.058	0.051	0.110
		N_AIPTW	0.046	0.046	0.078	0.047	0.045	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.0166	0.017	0.458	0.034	0.017	0.294
		N_IPTW	0.0147	0.015	0.055	0.045	0.163	
		AIPTW	0.0143	0.037	0.368	0.015	0.014	0.045
		N_AIPTW	0.0143	0.037	0.045	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.028				
		IPTW	1.707	1.722	9.032	5.100	1.655	30.821
		N_IPTW	0.577	0.576	0.907	0.040	1.174	
		AIPTW	0.043	0.014	0.052	0.048	0.040	0.073
		N_AIPTW	0.036	0.007	0.040	0.037	0.035	
Without Y	$E[Y^1]$	g-formula		0.061				
		IPTW	1.674	1.655	9.063	4.894	1.585	28.235
		N_IPTW	0.602	0.601	0.902	0.079	0.755	
		AIPTW	0.096	0.096	0.051	0.058	0.052	0.112
		N_AIPTW	0.084	0.083	0.040	0.047	0.047	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.0169	0.017	0.448	0.034	0.017	0.285
		N_IPTW	0.0148	0.015	0.055	0.046	0.157	
		AIPTW	0.0144	0.061	0.358	0.015	0.014	0.046
		N_AIPTW	0.0144	0.061	0.046	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.052				
		IPTW	1.712	1.703	9.063	5.036	1.649	30.837
		N_IPTW	0.593	0.591	0.902	0.040	1.152	
		AIPTW	0.088	0.031	0.051	0.047	0.041	0.074
		N_AIPTW	0.076	0.018	0.040	0.037	0.036	

Table 3.14: Bias of the estimators in S2 and missingness in only control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		-0.012				
		IPTW	-0.157	-0.137	0.304	0.082	-0.079	1.867
		N_IPTW	0.109	0.108	-0.070	-0.032	0.104	
		AIPTW	-0.014	-0.014	-0.056	-0.010	-0.005	-0.044
		N_AIPTW	-0.014	-0.014	-0.032	-0.010	-0.003	
	$E[Y^0]$	g-formula		-0.009				
		IPTW	-0.195	-0.005	-0.088	0.005	-0.007	-0.154
		N_IPTW	-0.099	-0.008	-0.045	-0.030	0.017	
		AIPTW	-0.033	-0.013	-0.054	-0.009	-0.010	-0.045
		N_AIPTW	-0.033	-0.013	-0.030	-0.009	-0.010	
	$E[Y^1 - Y^0]$	g-formula		-0.003				
		IPTW	-0.108	-0.132	0.392	0.077	-0.072	2.021
		N_IPTW	0.128	0.116	-0.025	-0.002	0.087	
		AIPTW	-0.019	-0.001	-0.002	-0.001	0.005	0.002
		N_AIPTW	-0.019	-0.002	-0.002	-0.001	0.007	
Without Y	$E[Y^1]$	g-formula		-0.010				
		IPTW	-0.260	-0.192	0.297	0.092	0.316	1.870
		N_IPTW	0.144	0.139	-0.069	-0.001	0.746	
		AIPTW	-0.012	-0.012	-0.057	0.000	0.502	-0.043
		N_AIPTW	-0.012	-0.012	-0.032	0.025	0.502	
	$E[Y^0]$	g-formula		-0.058				
		IPTW	-0.195	-0.070	-0.089	-0.002	-0.071	-0.155
		N_IPTW	-0.099	-0.077	-0.045	-0.031	-0.040	
		AIPTW	-0.033	-0.034	-0.055	-0.011	-0.082	-0.046
		N_AIPTW	-0.033	-0.034	-0.030	-0.011	-0.082	
	$E[Y^1 - Y^0]$	g-formula		0.047				
		IPTW	-0.064	-0.122	0.386	0.094	0.387	2.025
		N_IPTW	0.242	0.216	-0.024	0.030	0.786	
		AIPTW	0.021	0.022	-0.002	0.011	0.584	0.002
		N_AIPTW	0.021	0.022	-0.002	0.036	0.584	

Table 3.15: Variance of the estimators in S2 and missingness in only control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.038				
		IPTW	1.254	1.269	4.184	4.215	1.164	28.884
		N_IPTW	0.488	0.490	0.981	0.103	3.055	
		AIPTW	0.044	0.044	0.604	0.068	0.045	0.130
		N_AIPTW	0.044	0.044	0.103	0.061	0.044	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.016	0.015	0.550	0.034	0.016	0.144
		N_IPTW	0.014	0.014	0.057	0.054	0.055	
		AIPTW	0.014	0.038	0.531	0.015	0.015	0.054
		N_AIPTW	0.014	0.038	0.054	0.015	0.015	
	$E[Y^1 - Y^0]$	g-formula		0.025				
		IPTW	1.267	1.289	3.894	4.535	1.186	29.843
		N_IPTW	0.478	0.479	0.858	0.049	3.183	
		AIPTW	0.031	0.007	0.056	0.056	0.033	0.075
		N_AIPTW	0.030	0.007	0.049	0.049	0.032	
Without Y	$E[Y^1]$	g-formula		0.040				
		IPTW	1.173	1.224	4.153	4.233	0.926	29.262
		N_IPTW	0.454	0.462	0.994	0.120	1.669	
		AIPTW	0.046	0.046	0.602	0.108	0.069	0.132
		N_AIPTW	0.046	0.046	0.105	0.075	0.059	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.016	0.015	0.549	0.032	0.015	0.145
		N_IPTW	0.014	0.014	0.059	0.056	0.047	
		AIPTW	0.014	0.040	0.528	0.015	0.014	0.056
		N_AIPTW	0.014	0.040	0.056	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.027				
		IPTW	1.168	1.233	3.890	4.551	0.941	30.244
		N_IPTW	0.449	0.455	0.863	0.066	1.865	
		AIPTW	0.033	0.007	0.056	0.098	0.062	0.076
		N_AIPTW	0.033	0.006	0.049	0.065	0.052	

Table 3.16: MSE of the estimators in S2 and missingness in only control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.038				
		IPTW	1.279	1.288	4.277	4.222	1.171	32.371
		N_IPTW	0.500	0.502	0.986	0.104	3.066	
		AIPTW	0.045	0.045	0.607	0.068	0.045	0.132
		N_AIPTW	0.044	0.044	0.104	0.061	0.044	
	$E[Y^0]$	g-formula		0.015				
		IPTW	0.018	0.015	0.558	0.034	0.016	0.168
		N_IPTW	0.015	0.015	0.059	0.055	0.055	
		AIPTW	0.015	0.038	0.533	0.015	0.015	0.056
		N_AIPTW	0.015	0.038	0.055	0.015	0.015	
	$E[Y^1 - Y^0]$	g-formula		0.025				
		IPTW	1.278	1.306	4.047	4.541	1.191	33.929
		N_IPTW	0.494	0.493	0.858	0.049	3.191	
		AIPTW	0.031	0.007	0.056	0.056	0.033	0.075
		N_AIPTW	0.031	0.007	0.049	0.049	0.032	
Without Y	$E[Y^1]$	g-formula		0.040				
		IPTW	1.240	1.261	4.241	4.242	1.025	32.761
		N_IPTW	0.475	0.481	0.999	0.120	2.225	
		AIPTW	0.046	0.046	0.605	0.108	0.322	0.134
		N_AIPTW	0.046	0.046	0.106	0.076	0.311	
	$E[Y^0]$	g-formula		0.017				
		IPTW	0.054	0.019	0.557	0.032	0.020	0.169
		N_IPTW	0.024	0.020	0.061	0.057	0.049	
		AIPTW	0.015	0.041	0.531	0.015	0.021	0.058
		N_AIPTW	0.015	0.041	0.057	0.015	0.021	
	$E[Y^1 - Y^0]$	g-formula		0.030				
		IPTW	1.172	1.248	4.038	4.560	1.091	34.346
		N_IPTW	0.508	0.501	0.864	0.067	2.483	
		AIPTW	0.034	0.008	0.056	0.098	0.404	0.076
		N_AIPTW	0.033	0.007	0.049	0.066	0.392	

Table 3.17: Bias of the estimators in S2 and missingness in both case and control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		-0.005				
		IPTW	-0.969	-0.109	0.355	0.142	-0.104	1.751
		N_IPTW	0.137	0.111	-0.115	-0.035	0.050	
		AIPTW	-0.091	-0.009	-0.060	-0.013	-0.014	-0.030
		N_AIPTW	-0.095	-0.009	-0.035	-0.013	-0.014	
	$E[Y^0]$	g-formula		-0.009				
		IPTW	-0.051	-0.006	-0.090	0.003	-0.006	-0.156
		N_IPTW	-0.020	-0.008	-0.045	-0.030	0.017	
		AIPTW	0.005	-0.005	-0.055	-0.009	-0.010	-0.027
		N_AIPTW	0.005	-0.005	-0.030	-0.009	-0.010	
	$E[Y^1 - Y^0]$	g-formula		0.005				
		IPTW	-0.918	-0.103	0.445	0.139	-0.098	1.908
		N_IPTW	0.157	0.120	-0.070	-0.005	0.033	
		AIPTW	-0.096	-0.004	-0.005	-0.004	-0.004	-0.003
		N_AIPTW	-0.100	-0.004	-0.005	-0.004	-0.004	
Without Y	$E[Y^1]$	g-formula		0.485				
		IPTW	-2.623	0.305	0.363	0.132	-0.231	1.770
		N_IPTW	0.518	0.601	-0.116	-0.035	0.063	
		AIPTW	0.183	0.476	-0.060	-0.013	-0.020	-0.030
		N_AIPTW	0.132	0.476	-0.035	-0.013	-0.020	
	$E[Y^0]$	g-formula		-0.058				
		IPTW	-0.199	-0.074	-0.091	-0.004	-0.068	-0.157
		N_IPTW	-0.100	-0.078	-0.045	-0.031	-0.041	
		AIPTW	-0.034	0.462	-0.055	-0.011	-0.082	-0.027
		N_AIPTW	-0.033	0.462	-0.030	-0.011	-0.082	
	$E[Y^1 - Y^0]$	g-formula		0.543				
		IPTW	-2.423	0.378	0.453	0.135	-0.163	1.926
		N_IPTW	0.618	0.680	-0.071	-0.004	0.104	
		AIPTW	0.216	0.014	-0.005	-0.003	0.062	-0.003
		N_AIPTW	0.165	0.014	-0.005	-0.003	0.062	

Table 3.18: Variance of the estimators in S2 and missingness in both case and control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.038				
		IPTW	1.092	1.103	3.106	3.096	1.314	13.130
		N_IPTW	0.450	0.447	0.849	0.088	0.776	
		AIPTW	0.046	0.044	0.584	0.049	0.045	0.097
		N_AIPTW	0.046	0.043	0.088	0.046	0.044	
	$E[Y^0]$	g-formula		0.015				
		IPTW	0.015	0.016	0.550	0.033	0.015	0.151
		N_IPTW	0.015	0.015	0.056	0.054	0.055	
		AIPTW	0.015	0.038	0.531	0.015	0.014	0.054
		N_AIPTW	0.015	0.038	0.054	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.027				
		IPTW	1.098	1.122	2.873	3.371	1.335	13.956
		N_IPTW	0.440	0.436	0.718	0.033	0.940	
		AIPTW	0.034	0.006	0.036	0.036	0.032	0.043
		N_AIPTW	0.034	0.005	0.033	0.033	0.031	
Without Y	$E[Y^1]$	g-formula		0.049				
		IPTW	1.213	0.940	3.172	3.150	1.216	13.955
		N_IPTW	0.437	0.374	0.862	0.089	0.745	
		AIPTW	0.066	0.069	0.581	0.048	0.047	0.099
		N_AIPTW	0.064	0.065	0.089	0.045	0.046	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.016	0.015	0.548	0.031	0.014	0.151
		N_IPTW	0.014	0.014	0.058	0.056	0.048	
		AIPTW	0.014	0.049	0.528	0.014	0.014	0.055
		N_AIPTW	0.014	0.049	0.056	0.014	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.041				
		IPTW	1.176	0.954	2.964	3.421	1.225	14.799
		N_IPTW	0.431	0.367	0.725	0.034	0.928	
		AIPTW	0.057	0.015	0.037	0.035	0.034	0.043
		N_AIPTW	0.055	0.013	0.034	0.032	0.033	

Table 3.19: MSE of the estimators in S2 and missingness in both case and control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.038				
		IPTW	2.031	1.115	3.232	3.116	1.325	16.198
		N_IPTW	0.469	0.459	0.862	0.089	0.778	
		AIPTW	0.054	0.044	0.587	0.049	0.045	0.098
		N_AIPTW	0.055	0.043	0.089	0.046	0.045	
	$E[Y^0]$	g-formula		0.015				
		IPTW	0.056	0.016	0.559	0.033	0.016	0.175
		N_IPTW	0.024	0.015	0.058	0.055	0.056	
		AIPTW	0.016	0.038	0.534	0.015	0.015	0.054
		N_AIPTW	0.016	0.038	0.055	0.015	0.015	
	$E[Y^1 - Y^0]$	g-formula		0.027				
		IPTW	1.941	1.133	3.071	3.391	1.345	17.595
		N_IPTW	0.465	0.450	0.723	0.033	0.941	
		AIPTW	0.043	0.006	0.036	0.036	0.032	0.043
		N_AIPTW	0.044	0.005	0.033	0.033	0.031	
Without Y	$E[Y^1]$	g-formula		0.284				
		IPTW	8.090	1.032	3.304	3.167	1.269	17.086
		N_IPTW	0.705	0.735	0.876	0.091	0.749	
		AIPTW	0.099	0.295	0.584	0.048	0.047	0.100
		N_AIPTW	0.081	0.292	0.091	0.046	0.047	
	$E[Y^0]$	g-formula		0.018				
		IPTW	0.056	0.020	0.556	0.031	0.019	0.176
		N_IPTW	0.024	0.020	0.060	0.057	0.050	
		AIPTW	0.016	0.262	0.531	0.015	0.021	0.056
		N_AIPTW	0.016	0.262	0.056	0.015	0.021	
	$E[Y^1 - Y^0]$	g-formula		0.336				
		IPTW	7.047	1.097	3.170	3.439	1.251	18.510
		N_IPTW	0.813	0.829	0.730	0.034	0.939	
		AIPTW	0.104	0.015	0.037	0.035	0.038	0.043
		N_AIPTW	0.083	0.013	0.034	0.032	0.037	

Table 3.20: Bias of the estimators in S3 and missingness in only control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.000				
		IPTW	-0.070	-0.057	0.742	0.480	-0.047	35.993
		N_IPTW	0.109	0.109	-0.091	-0.003	0.110	
		AIPTW	0.005	0.005	-0.003	0.006	0.009	0.106
		N_AIPTW	0.004	0.004	-0.004	0.004	0.008	
	$E[Y^0]$	g-formula		0.000				
		IPTW	-0.014	0.001	-0.066	-0.001	0.000	0.214
		N_IPTW	-0.003	0.000	-0.030	-0.007	0.013	
		AIPTW	0.009	0.000	-0.008	0.000	0.000	0.052
		N_AIPTW	0.009	0.000	-0.007	0.000	0.000	
	$E[Y^1 - Y^0]$	g-formula		0.000				
		IPTW	-0.056	-0.057	0.809	0.481	-0.047	38.114
		N_IPTW	0.113	0.108	-0.062	0.004	0.097	
		AIPTW	-0.005	0.005	0.005	0.006	0.009	0.063
		N_AIPTW	-0.005	0.004	0.003	0.004	0.008	
Without Y	$E[Y^1]$	g-formula		0.005				
		IPTW	-0.133	-0.078	0.748	0.497	0.209	36.228
		N_IPTW	0.136	0.133	-0.089	0.016	0.403	
		AIPTW	0.009	0.009	-0.001	0.009	0.280	0.109
		N_AIPTW	0.009	0.009	-0.002	0.025	0.279	
	$E[Y^0]$	g-formula		-0.026				
		IPTW	-0.098	-0.042	-0.065	-0.006	-0.044	0.221
		N_IPTW	-0.056	-0.045	-0.028	-0.006	-0.026	
		AIPTW	-0.016	-0.015	-0.006	-0.002	-0.046	0.056
		N_AIPTW	-0.016	-0.015	-0.006	-0.002	-0.046	
	$E[Y^1 - Y^0]$	g-formula		0.031				
		IPTW	-0.036	-0.036	0.813	0.503	0.252	38.388
		N_IPTW	0.192	0.177	-0.060	0.023	0.429	
		AIPTW	0.025	0.024	0.005	0.011	0.326	0.064
		N_AIPTW	0.025	0.024	0.004	0.027	0.325	

Table 3.21: Variance of the estimators in S3 and missingness in only control units; N_IPTW: normalized IPTW, N_AIPTW: normalized AIPTW, MIA: MI across, MIW: MI within, IPC: IPCCW, AIPM: AIPCCW-model, AIPMC: AIPCCW-Monte Carlo method, RAIPC: restricted AIPCCW

With or Without Y			Missing data					
	Estimand	Causal	MIA	MIW	IPC	AIPM	AIPCMC	RAIPC
With Y	$E[Y^1]$	g-formula		0.036				
		IPTW	0.615	0.618	2.982	2.596	0.605	35.993
		N_IPTW	0.397	0.397	0.937	0.090	1.655	
		AIPTW	0.040	0.040	0.529	0.058	0.043	0.106
		N_AIPTW	0.040	0.040	0.089	0.053	0.043	
	$E[Y^0]$	g-formula		0.015				
		IPTW	0.02	0.016	0.540	0.035	0.016	0.214
		N_IPTW	0.01	0.015	0.058	0.052	0.112	
		AIPTW	0.01	0.036	0.492	0.015	0.015	0.052
		N_AIPTW	0.01	0.036	0.052	0.015	0.015	
	$E[Y^1 - Y^0]$	g-formula		0.023				
		IPTW	0.636	0.640	3.426	2.827	0.624	38.114
		N_IPTW	0.376	0.376	0.756	0.044	1.939	
		AIPTW	0.027	0.005	0.047	0.047	0.031	0.063
		N_AIPTW	0.027	0.004	0.043	0.042	0.031	
Without Y	$E[Y^1]$	g-formula		0.037				
		IPTW	0.573	0.587	2.980	2.720	0.567	36.228
		N_IPTW	0.389	0.393	0.963	0.107	1.414	
		AIPTW	0.042	0.042	0.526	0.078	0.067	0.109
		N_AIPTW	0.041	0.041	0.093	0.066	0.065	
	$E[Y^0]$	g-formula		0.014				
		IPTW	0.02	0.015	0.538	0.033	0.015	0.221
		N_IPTW	0.01	0.014	0.062	0.056	0.107	
		AIPTW	0.01	0.037	0.490	0.015	0.014	0.056
		N_AIPTW	0.01	0.037	0.056	0.015	0.014	
	$E[Y^1 - Y^0]$	g-formula		0.024				
		IPTW	0.586	0.604	3.447	2.964	0.583	38.388
		N_IPTW	0.369	0.372	0.765	0.056	1.717	
		AIPTW	0.029	0.005	0.047	0.066	0.056	0.064
		N_AIPTW	0.029	0.005	0.043	0.054	0.054	

0.4 Analyzing model compatibility of Mitra's scenario

Definition of compatible conditional models

According to section 4 of (Arnold & Press 1989) [59], "Suppose that (X, Y) is to be absolutely continuous with respect to the product measure $\mu_1 \times \mu_2$ on $S_x \times S_y$. Denote the families of candidate conditional densities (with respect to μ_1 and μ_2) by

$$a(x, y) = f_{X|Y}(x|y),\ x \in S_x,\ y \in S_y$$

and

$$b(x, y) = f_{Y|X}(y|x),\ x \in S_x,\ y \in S_y$$

Let $N_a = \{(x, y) : a(x, y) > 0\}$ and $N_b = \{(x, y) : b(x, y) > 0\}$. Then, A joint density $f(x, y)$, with $a(x, y)$ and $b(x, y)$ as its conditional densities, will exist iff (i) $N_a = N_b = N$, and (ii) for all $(x, y) \in N$, there exist functions $u(x)$ and $v(y)$ such that

$$a(x, y)/b(x, y) = u(x)v(y)$$

where $\int u(X)d\mu_1(x) < \infty$. The condition $\int u(X)d\mu_1(x) < \infty$ is equivalent to the condition $\int [1/v(y)]d\mu_2(y) < \infty$, and only one condition must be checked in practice."

Model compatibility in AIPCCW estimators

Utilizing the definition of compatible models to our situation, we articulate the conditions under which all the working models of AIPCCW estimators are compatible. Consider the univariate missingness case and complete data $O = (A, Y, X_o, X_m)$ where X_m has missingness. The joint density of full data is factorized as

$$\begin{aligned} f(A, Y, X_o, X_m) &= \boldsymbol{f(Y|A, X_o, X_m) f(A|X_o, X_m)} f(X_o, X_m) \\ &= \boldsymbol{f(X_m|A, Y, X_o)} f(A, Y, X_o) \end{aligned}$$

In the above equation, the bold-faced part is what we estimate using paramet-

ric models. Let $N_a = \{(a, y, x_o, x_m) : f(y|A, x_o, x_m)f(a|x_o, x_m) > 0\}, N_b = \{(a, y, x_o, x_m) : f(x_m|a, y, x_m))0\}$, and $N = \{(a, y, x_o, x_m) : \forall(a, y, x_o, x_m) \in (A, Y, X_o, X_m)\}$. Then, we can deduct the following results from the previous section. The conditional densities $f(Y|A, X_o, X_m)f(A|X_o, X_m)$ and $f(X_m|A, Y, X_o)$ are compatible iif (i) $N_a = N_b = N$, and for all $(a, y, x_o, x_m) \in N$, (ii) there exists function $u(X_o, X_m)$ and $v(A, Y, X_o)$ such that

$$\frac{f(Y = y|A, X_o, X_m)f(A = a|X_o, X_m)}{f(X_m = x_m|A, Y, X_o)} = u(X_o, X_m)v(A, Y, X_o)$$

where $\int v(A, Y, X_o)d\mu_{A,Y,X_o} < \infty$ or $\int 1/u(X_o, X_m)d\mu_{X_o,X_m} < \infty$.

Assessment of model compatibility of Mitra's [40] simulation scenario

Covariates X and outcome Y were generated as

$$(X_o, X_m) \sim N((10, 10)^T, \Sigma), \text{ where } \Sigma = \begin{bmatrix} 5 & 2.5 \\ 2.5 & 5 \end{bmatrix}$$

$$Y = X_o + X_m + \epsilon, \text{ where } \epsilon \sim N(0, 1)$$

The treatment A has three scenarios:

$$\text{(i) } \operatorname{logit} \Pr[A = 1|x_o, x_m] = -7.8 + 0.5 * x_o$$

$$\text{(ii) } \operatorname{logit} \Pr[A = 1|x_o, x_m] = -7.8 + 0.5 * x_m$$

$$\text{(iii) } \operatorname{logit} \Pr[A = 1|x_o, x_m] = -7.8 + 0.225 * x_o + 0.225 * x_m$$

Suppose we fit a model for the conditional densities $f(Y|A, X_o, X_m)$ and $f(X_m|A, Y, X_o)$ [10] using the linear regression (normal distribution) such as

$$f(Y = y|a, x_o, x_m; \theta) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{1}{2\sigma_Y^2}(y - (\theta_0 + \theta_1 x_o + \theta_2 a + \theta_3 x_m))^2\right)$$

[10] We cannot derive the exact form of the conditional distribution of X_m. However, since the marginal distribution is normal, it would be natural to use a normal distribution to fit the conditional density.

$$f(X_m = x_m|a, y, x_o; \psi) = \frac{1}{\sqrt{2\pi\sigma^2_{X_m}}} \exp\left(-\frac{1}{2\sigma^2_{X_m}}(x_m - (\psi_0 + \psi_1 x_o + \psi_2 a + \psi_3 y))^2\right)$$

and $\Pr[A = 1|X_o, X_m]$ using a logistic regression mode, such as

$$\text{logit}\{\Pr[A = 1|x_o, x_m; \alpha]\} = \alpha_0 + \alpha_1 x_o + \alpha_2 x_m = \alpha' x$$

Those models are outcome model, imputation model, treatment model using the terminology of the main text. We assumed the simple functional forms of each model without non-linear terms. Actually if two conditional densities of normal random variables have non-linear terms, they are not compatible [59]. First examine the condition $N = N_a = N_b$, the conditional density $f(Y|A, X_o, X_m)$ and $f(X_o, X_m)$ are modelled via univariate and bivariate normal distribution, which has positive density over all the domain, $\mathbb{R}$ for univaraite and $\mathbb{R}^2$ for bivariate distribution. Thus, we can conclude $f(X_o, X_m) > 0$ for all (x_o, x_m) and $f(Y|A, X_o, X_m) > 0$ for all (a, x_o, x_m) such that $f(a, x_o, x_m) > 0$. Inverse logit function is mapping $(-\infty, \infty)$ to $(0, 1)$, therefore $\Pr[A = a|x_o, x_m]$ ¿ 0 for all (x_o, x_m) and combined with the case of $f(X_o, X_m)$, we can conclude $f(a, x_o, x_m) > 0$ for all (a, x_o, x_m). This is sufficient to show that $f(a, y, x_0, x_m) > 0$ $\forall(a, y, x_o, x_m)$ and we can deduct $N = N_a = N_b$ from it. Next, we arrange the density functions to examine the second condition.

$$
\begin{aligned}
&\frac{f(Y=y|a,x_o,x_m;\theta)\Pr(A=1|x_o,x_m;\alpha)}{f(X_m=x_m|a,y,x_o;\psi)} \\
&= \frac{\sigma_{X_m}}{\sigma_Y}\exp\left(\frac{1}{2\sigma^2_{X_m}}(x_m-(\psi_0+\psi_1x_o+\psi_2a+\psi_3y)^2)-\frac{1}{2\sigma^2_Y}(y-(\theta_0+\theta_1x_o+\theta_2a+\theta_3x_m))^2\right) \\
&\times\left(\frac{\exp(\alpha'x)}{1+exp(\alpha'x)}\right)^a\left(\frac{1}{1+exp(\alpha'x)}\right)^{1-a} \\
&= \frac{\sigma_{X_m}}{\sigma_Y}\exp\left(\frac{1}{2\sigma^2_{X_m}}(x_m^2+(\psi_0+\psi_1x_o)^2+(\psi_2a)^2+(\psi_3y)^2)-2x_m(\psi_0+\psi_{X_o}+\psi_2a+\psi_3y)+2\psi_0(\psi_1x_o+\psi_2a+\psi_3y)\right) \\
&\times\exp\left(\frac{1}{2\sigma^2_{X_m}}(2\psi_1x_o(\psi_2a+\psi_3y)+2\psi_2\psi_3ay)\right)\times\exp\left(-\frac{1}{2\sigma^2_Y}(y^2+\theta_0^2+(\theta_1x_o)^2+(\theta_2a)^2+(\theta_3x_m)^2)\right) \\
&\times\exp\left(-\frac{1}{2\sigma^2_Y}(-2y(\theta_0+\theta_1x_o+\theta_2a+\theta_3x_m)+2\theta_0(\theta_1x_o+\theta_2a+\theta_3x_m)+2\theta_1x_o(\theta_2a+\theta_3x_m)+2\theta_2\theta_3x_m\right) \\
&\times\frac{\exp(\alpha_0a+\alpha_1ax_m+\alpha_2ax_o)}{1+\exp(\alpha'x)} \\
&= \frac{\sigma_{X_m}}{\sigma_Y(1+\exp(\alpha'x))}\exp\left(\frac{1}{2\sigma^2_{X_m}}(x_m^2+\psi_0^2+(\psi_1x_o)^2-2x_m(\psi_0+\psi_1x_o)+2\psi_0\psi_1x_o)\right) \\
&\times\exp\left(-\frac{1}{2\sigma^2_Y}((\theta_3x_m)^2+2\theta_0\theta_3x_m+2\theta_1\theta_3x_ox_m)\right) \\
&\times\exp(\alpha_0a+\alpha_2ax_o)\exp\left(\frac{1}{2\sigma^2_{X_m}}((\psi_2a)^2+(\psi_3y)^2+2\psi_0(\psi_2a+\psi_3y)+2\psi_1x_o(\psi_2a+\psi_3y)+2\psi_2\psi_3ay\right) \\
&\times\exp\left(-\frac{1}{2\sigma^2_Y}(y^2+\theta_0^2+(\theta_1x_o)^2+(\theta_2a)^2-2y(\theta_0+\theta_1x_o+\theta_2a+\theta_3x_m)+2\theta_0(\theta_1x_o+\theta_2a)+2\theta_1\theta_2x_oa)\right) \\
&\times\exp\left(\frac{1}{2\sigma^2_{X_m}}(-2x_m(\psi_2a+\psi_3y))-\frac{1}{2\sigma^2_Y}(-2\theta_3yx_m+2\theta_2\theta_3ax_m)+\alpha_2ax_m\right)
\end{aligned}
$$

Define function $g(x_o,x_m)$ and $h(x_o,a,y)$ as

$$
\begin{aligned}
g(x_o,x_m) =& \frac{\sigma_{X_m}}{\sigma_Y(1+\exp(\alpha'x))}\exp\left(\frac{1}{2\sigma^2_{X_m}}(x_m^2+\psi_0^2+(\psi_1x_o)^2-2x_m(\psi_0+\psi_1x_o)+2\psi_0\psi_1x_o)\right) \\
&\times\exp\left(-\frac{1}{2\sigma^2_Y}((\theta_3x_m)^2+2\theta_0\theta_3x_m+2\theta_1\theta_3x_ox_m)\right) \\
h(x_o,a,y) =& \exp(\alpha_0a+\alpha_2ax_o)\exp\left(\frac{1}{2\sigma^2_{X_m}}((\psi_2a)^2+(\psi_3y)^2+2\psi_0(\psi_2a+\psi_3y)+2\psi_1x_o(\psi_2a+\psi_3y)+2\psi_2\psi_3ay\right) \\
&\times\exp\left(-\frac{1}{2\sigma^2_Y}(y^2+\theta_0^2+(\theta_1x_o)^2+(\theta_2a)^2-2y(\theta_0+\theta_1x_o+\theta_2a+\theta_3x_m)+2\theta_0(\theta_1x_o+\theta_2a)+2\theta_1\theta_2x_oa)\right)
\end{aligned}
$$

The last term of the derivation is

$$
g(x_o,x_m)h(x_o,a,y)\exp\left(yx_m\left(\frac{\theta_3}{\sigma^2_Y}-\frac{\psi_3}{\sigma^2_{X_m}}\right)+ax_m\left(\alpha_1-\frac{\theta_2\theta_3}{\sigma^2_Y}-\frac{\psi_2}{\sigma^2_{X_m}}\right)\right)
$$

Therefore, if

$$
\frac{\theta_3}{\sigma^2_Y}-\frac{\psi_3}{\sigma^2_{X_m}}=0,\ \ \alpha_1-\frac{\theta_2\theta_3}{\sigma^2_Y}-\frac{\psi_2}{\sigma^2_{X_m}}=0 \tag{.25}
$$

$$\frac{f(Y=y|a,x_o,x_m;\theta)\Pr(A=1|x_o,x_m;\alpha)}{f(X_m=x_m|a,y,x_o;\psi)} = g(x_o,x_m)h(x_o,a,y)$$

for any (x_o, x_m, a, y). We generated data with $n = 10000$ and 500 iterations and fitted all the working models and obtained sample estimates of coefficients and standard errors to assess the fitted models satisfy (.25). We calculated the quantity A and B from the estimates,

$$A = \frac{\widehat{\theta}_3}{\widehat{\sigma}^2_Y} - \frac{\widehat{\psi}_3}{\widehat{\sigma}^2_{X_m}}, \quad B = \widehat{\alpha}_1 - \frac{\widehat{\theta}_2\widehat{\theta}_3}{\widehat{\sigma}^2_Y} - \frac{\widehat{\psi}_2}{\widehat{\sigma}^2_{X_m}}.$$

We calculated mean and 95% confidence interval over the iterations based on percentiles, considering this procedure is non-parametric bootstrap. We first utilized this procedure to Evan's et al's data generation [43], where all those models are compatible.

$$\text{mean of } A : -6.0e-16 \quad \text{mean of} B; [[95\% CI]] : 0.0002; [-0.0015, 0.0019]$$

We omitted 95% CI of quantity A but 95% CI is on the multiple decimal points and includes 0. Because those models are compatible, limiting value of A and B are expected to 0, and this test procedure indicates it correctly. When we generated data according Mitra's scenario 1, 2, and 3 did the same procedure, the results were

	Scenario1	Scenario2	Scenario3
A (mean)	-1.6e-16	-2.5e-16	3.7e-16
B (mean; [95% CI])	0.0002; [-0.0057, 0.0062]	0.0364; [0.0258, 0.0478]	0.0086; [0.0018, 0.0162]

We can conclude that all the models are compatible only under scenario 1. It does not mean that all the models in scenario 1 are correctly specified because compatible models do not imply true models. This analysis shows that it is difficult to make simulation scenarios with compatible working models. Evans and her colleagues made a scenario of compatible models but their data generation mechanism was unusual for studies on causal inference. In her simulation, X_m was generated as a function of A and X_o. If we view their data generation mechanism as any structural causal models, such as finest fully randomized causally interpretable structured tree graph [44] or non-parametric structural equation model [67], their scenario is interpreted as that A is a cause of X_m. Consequently, adjusting X_m blocks a causal pathway from A to Y rather than avoiding confounding. For this reason, we did not adopt their data generation mechanism and followed a commonly used generation mechanism that reflects causal structure better. However, we do not stick to making all the models compatible because it is difficult to make and forces us to limit our data generation mechanism unnecessarily, i.e., without

interaction term or more independence structure.

0.5 Results from different scenarios according to missing proportion and the magnitude of interaction terms

In the following results, AIPCCW-MC estimators have a little bias in the scenario of $\theta_1 = \theta_2 = 0.5$. The reason for the bias is that the imputation model is not flexible enough in that scenario. We modified the imputation model to include interaction term between spline terms of X_o and A and report the results at the end.

Table 3.22: **Bias** of the estimators where $\theta_1 = \theta_2 = 2$ and the proportion of complete cases is 0.85.

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	-0.2265	0.0297	-0.3673	-0.2264	-0.2264	0.0014	0.006	0.0033	0.0013	0.0013
	Y0	0.2262	0.9316	0.4449	0.2263	0.2263	0.0000	0.0016	0.0010	0.0000	0.0000
	MD	-0.4527	-0.9019	-0.8122	-0.4526	-0.4526	-0.0003	-0.00248	-0.00078	-0.0004	-0.0004
n = 5,000	Y1	-0.2273	0.0296	-0.3676	-0.2272	-0.2272	-0.0016	0.0000	-0.0022	-0.0015	-0.0015
	Y0	0.2267	0.9395	0.4485	0.2268	0.2268	0.0060	0.0013	0.0013	0.0037	0.0037
	MD	-0.4540	-0.9099	-0.8161	-0.4540	-0.4540	-0.0030	-0.0062	-0.0056	-0.0028	-0.0028
n = 2,000	Y1	-0.2264	0.0263	-0.3621	-0.2263	-0.2263	-0.0006	0.0002	0.0011	-0.0004	-0.0004
	Y0	0.2259	0.9327	0.4467	0.2259	0.2259	-0.0006	0.0070	0.0030	-0.0007	-0.0007
	MD	-0.4523	-0.9064	-0.8087	-0.4522	-0.4522	0.0000	-0.0068	-0.0019	0.0004	0.0004

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	-0.0126	0.0008	-0.0043	-0.0043	0.0032	-0.0004	0.0005	0.0002
	Y0	0.1482	0.0260	-0.0356	-0.0358	0.021	0.0086	-0.0006	-0.0004
	MD	-0.1608	-0.0252	0.0313	0.0315	-0.0178	-0.009	0.0011	0.0005
n = 5,000	Y1	-0.0121	-0.0016	-0.0054	-0.0054	0.0069	-0.0024	-0.0003	-0.0005
	Y0	0.1532	0.0284	-0.0342	-0.0344	0.0473	0.019	0.0001	0.0003
	MD	-0.1653	-0.0300	0.0288	0.0290	-0.0404	-0.0214	-0.0004	-0.0007
n = 2,000	Y1	-0.0127	0.0019	-0.0043	-0.0043	0.013	-0.0028	-0.0011	-0.0003
	Y0	0.1558	0.0283	-0.0362	-0.0364	0.0885	0.034	0.0001	-0.0004
	MD	-0.1685	-0.0264	0.0319	0.0321	-0.0755	-0.0369	-0.0012	0.0001

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	Y1	0.0031	0.0002	0.0001	0.0001	0.0029	0.0002	0.0003	0.0003	0.0070	-0.0002
	Y0	0.0186	-0.0004	-0.0003	-0.0003	0.006	-0.0004	0.0016	0.0016	-0.0099	-0.0002
	MD	-0.0156	0.0005	0.0004	0.0004	-0.003	0.0005	-0.0013	-0.0013	0.0169	0.0000
n = 5,000	Y1	0.0065	-0.0005	-0.0004	-0.0004	0.0065	-0.0005	-0.0002	-0.0002	0.0144	-0.0014
	Y0	0.0398	0.0002	0.0002	0.0002	0.0238	0.0003	0.0018	0.0018	-0.0053	0.0007
	MD	-0.0333	-0.0007	-0.0006	-0.0006	-0.0173	-0.0007	-0.002	-0.002	0.0197	-0.0020
n = 2,000	Y1	0.0113	-0.0003	0.0001	0.0001	0.0118	-0.0003	0.0001	0.0001	0.0347	-0.0023
	Y0	0.0663	-0.0005	-0.001	-0.001	0.0721	-0.0004	0.0022	0.0022	-0.0436	0.0002
	MD	-0.055	0.0002	0.0011	0.0011	-0.0603	0.0001	-0.0021	-0.0021	0.0783	-0.0025

Table 3.23: **Variance** of the estimators where $\theta_1 = \theta_2 = 2$ and the proportion of complete cases is 0.85.

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0021	0.0094	0.0075	0.0021	0.0021	0.0019	0.0040	0.0039	0.0019	0.0019
	Y0	0.0021	0.0384	0.0066	0.0022	0.0022	0.0018	0.0133	0.0035	0.0019	0.0019
	MD	0.0078	0.0327	0.0205	0.0079	0.0079	0.0069	0.0139	0.0118	0.0069	0.0069
n = 5,000	Y1	0.0041	0.0195	0.0149	0.0043	0.0043	0.0035	0.0081	0.0076	0.0036	0.0036
	Y0	0.0043	0.0743	0.0127	0.0044	0.0044	0.0037	0.027	0.0069	0.0037	0.0037
	MD	0.0156	0.0639	0.0403	0.0159	0.0159	0.0132	0.0269	0.0236	0.0133	0.0133
n = 2,000	Y1	0.0102	0.0484	0.0356	0.0105	0.0105	0.0088	0.0214	0.0177	0.0089	0.0089
	Y0	0.0107	0.1995	0.0333	0.0111	0.0111	0.0088	0.0705	0.0172	0.0089	0.0089
	MD	0.0388	0.1741	0.0988	0.0395	0.0395	0.0323	0.0698	0.0565	0.0325	0.0325

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.004	0.0039	0.0019	0.0019	0.0047	0.0051	0.0040	0.0020
	Y0	0.0125	0.0034	0.0019	0.0019	0.0921	0.0159	0.0038	0.0023
	MD	0.0132	0.0117	0.007	0.007	0.0904	0.0319	0.0146	0.0077
n = 5,000	Y1	0.0081	0.0076	0.0036	0.0036	0.0095	0.0101	0.0081	0.0040
	Y0	0.0253	0.0067	0.0037	0.0037	0.1516	0.0261	0.0074	0.0046
	MD	0.0256	0.0233	0.0134	0.0134	0.1524	0.0569	0.0292	0.0154
n = 2,000	Y1	0.0214	0.0177	0.0089	0.0089	0.0261	0.0240	0.0210	0.0099
	Y0	0.0658	0.0166	0.009	0.009	0.3122	0.0533	0.0193	0.0114
	MD	0.0661	0.0557	0.0327	0.0327	0.3215	0.1238	0.0761	0.0385

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	Y1	0.0051	0.0020	0.0019	0.0019	0.0048	0.0020	0.0018	0.0018	0.0053	0.0020
	Y0	0.0534	0.0023	0.0022	0.0022	0.0605	0.0023	0.0019	0.0019	0.2079	0.0024
	MD	0.0469	0.0077	0.0072	0.0072	0.0573	0.0077	0.0066	0.0066	0.1792	0.0077
n = 5,000	Y1	0.0103	0.0040	0.0038	0.0038	0.0098	0.0040	0.0036	0.0036	0.011	0.004
	Y0	0.0835	0.0046	0.0044	0.0044	0.1344	0.0046	0.0039	0.0039	0.3712	0.0047
	MD	0.0738	0.0154	0.0145	0.0145	0.1266	0.0154	0.0135	0.0135	0.3181	0.0155
n = 2,000	Y1	0.0280	0.0099	0.0092	0.0092	0.0266	0.0099	0.0089	0.0089	0.0316	0.01
	Y0	0.1728	0.0115	0.0107	0.0107	0.6379	0.0114	0.0099	0.0099	1.2752	0.0123
	MD	0.1475	0.0385	0.0356	0.0355	0.5982	0.0385	0.0337	0.0337	1.1101	0.0395

Table 3.24: MSE of the estimators where $\theta_1 = \theta_2 = 2$ and the proportion of complete cases is 0.85.

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0534	0.0103	0.1424	0.0534	0.0534	0.0019	0.0040	0.0039	0.0019	0.0019
	Y0	0.0533	0.9063	0.2045	0.0534	0.0534	0.0018	0.0133	0.0035	0.0019	0.0019
	MD	0.2127	0.8461	0.6802	0.2127	0.2127	0.0069	0.0139	0.0118	0.0069	0.0069
n = 5,000	Y1	0.0558	0.0204	0.1500	0.0559	0.0559	0.0035	0.0081	0.0076	0.0036	0.0036
	Y0	0.0557	0.9570	0.2139	0.0558	0.0558	0.0037	0.0270	0.0069	0.0037	0.0037
	MD	0.2217	0.8918	0.7063	0.2220	0.2220	0.0132	0.0269	0.0236	0.0133	0.0133
n = 2,000	Y1	0.0615	0.0491	0.1667	0.0617	0.0617	0.0088	0.0214	0.0177	0.0089	0.0089
	Y0	0.0617	1.0694	0.2328	0.0621	0.0621	0.0088	0.0705	0.0172	0.0089	0.0089
	MD	0.2434	0.9957	0.7528	0.244	0.244	0.0323	0.0698	0.0565	0.0325	0.0325

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0042	0.0039	0.0019	0.0019	0.0047	0.0051	0.0040	0.0020
	Y0	0.0344	0.004	0.0032	0.0032	0.0925	0.0159	0.0038	0.0023
	MD	0.039	0.0123	0.008	0.008	0.0907	0.0320	0.0146	0.0077
n = 5,000	Y1	0.0082	0.0076	0.0036	0.0036	0.0095	0.0101	0.0081	0.0040
	Y0	0.0487	0.0075	0.0049	0.0049	0.1538	0.0265	0.0074	0.0046
	MD	0.0529	0.0242	0.0142	0.0142	0.1540	0.0574	0.0292	0.0154
n = 2,000	Y1	0.0216	0.0177	0.0089	0.0089	0.0263	0.0240	0.0210	0.0099
	Y0	0.0900	0.0174	0.0104	0.0104	0.3200	0.0545	0.0193	0.0114
	MD	0.0945	0.0564	0.0337	0.0338	0.3272	0.1252	0.0761	0.0385

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	Y1	0.0051	0.002	0.0019	0.0019	0.0048	0.002	0.0018	0.0018	0.0053	0.002
	Y0	0.0537	0.0023	0.0022	0.0022	0.0605	0.0023	0.0019	0.0019	0.208	0.0024
	MD	0.0472	0.0077	0.0072	0.0072	0.0573	0.0077	0.0066	0.0066	0.1795	0.0077
n = 5,000	Y1	0.0103	0.0040	0.0038	0.0038	0.0098	0.0040	0.0036	0.0036	0.0112	0.004
	Y0	0.0851	0.0046	0.0044	0.0044	0.1350	0.0046	0.0039	0.0039	0.3712	0.0047
	MD	0.0749	0.0154	0.0145	0.0145	0.1269	0.0154	0.0135	0.0135	0.3184	0.0155
n = 2,000	Y1	0.0281	0.0099	0.0092	0.0092	0.0267	0.0099	0.0089	0.0089	0.0328	0.01
	Y0	0.1772	0.0115	0.0107	0.0107	0.6431	0.0114	0.0099	0.0099	1.277	0.0123
	MD	0.1505	0.0385	0.0356	0.0355	0.6018	0.0385	0.0337	0.0337	1.1161	0.0395

Table 3.25: Bias of the estimators where $\theta_1 = \theta_2 = 0.5$ and the proportion of complete cases is 0.7

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.2014	0.1917	0.2971	0.2014	0.2014	0.0002	0.0003	0.0002	0.0001	0.0001
	Y0	0.4029	1.4157	0.6916	0.4028	0.4028	0.0007	-0.0029	-0.0005	0.0006	0.0006
	MD	-0.2015	-1.2240	-0.3945	-0.2014	-0.2014	-0.0005	0.0032	0.0007	-0.0004	-0.0004
n = 5,000	Y1	0.2012	0.1915	0.2959	0.2012	0.2012	0.0017	0.0014	0.0014	0.0016	0.0016
	Y0	0.4017	1.4022	0.6875	0.4017	0.4017	0.0024	0.0076	0.0047	0.0024	0.0024
	MD	-0.2005	-1.2107	-0.3916	-0.2005	-0.2005	-0.0006	-0.0062	-0.0033	-0.0009	-0.0009
n = 2,000	Y1	0.2006	0.1905	0.2919	0.2005	0.2005	-0.0006	-0.0012	-0.0015	-0.0007	-0.0007
	Y0	0.4016	1.4140	0.6955	0.4017	0.4017	-0.0001	0.0071	0.0033	-0.0002	-0.0002
	MD	-0.2011	-1.2236	-0.4036	-0.2012	-0.2012	-0.0005	-0.0082	-0.0049	-0.0005	-0.0005

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0141	-0.0019	0.0128	0.0128	-0.0011	0.0022	0.0000	0.0001
	Y0	0.21	0.0318	-0.0549	-0.0554	0.0649	0.0250	-0.0002	0.0001
	MD	-0.1959	-0.0337	0.0676	0.0682	-0.0660	-0.0228	0.0003	0.0001
n = 5,000	Y1	0.0153	-0.0008	0.0141	0.0142	-0.0029	0.0029	-0.0008	-0.0004
	Y0	0.2193	0.0368	-0.0525	-0.053	0.1034	0.0385	-0.0028	-0.0015
	MD	-0.204	-0.0376	0.0667	0.0672	-0.1063	-0.0356	0.0021	0.0011
n = 2,000	Y1	0.0132	-0.0039	0.0118	0.0119	-0.0038	0.0086	0.0012	0.0005
	Y0	0.2239	0.0361	-0.0551	-0.0556	0.2368	0.0863	0.0042	0.0011
	MD	-0.2107	-0.04	0.0669	0.0675	-0.2406	-0.0776	-0.0031	-0.0006

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	Y1	-0.0005	0.0001	0.0001	0.0001	0.0003	0.0001	-0.0378	-0.0247	0.0034	0.0005
	Y0	0.0480	0.0015	0.0021	0.0017	-0.5896	0.0015	0.2408	0.1705	-0.0270	0.0016
	MD	-0.0485	-0.0013	-0.0019	-0.0016	0.5899	-0.0013	-0.2786	-0.1952	0.0304	-0.0012
n = 5,000	Y1	-0.0016	-0.0004	0.0000	0.0000	-0.0015	-0.0004	-0.0376	-0.0251	0.0074	0.0003
	Y0	0.0756	0.0008	0.0026	0.0020	-0.5415	0.0008	0.2430	0.1707	-0.0722	0.0021
	MD	-0.0772	-0.0012	-0.0026	-0.0020	0.5399	-0.0012	-0.2805	-0.1957	0.0796	-0.0019
n = 2,000	Y1	-0.0010	0.0005	-0.0007	-0.0007	-0.0025	0.0005	-0.0400	-0.0254	0.0167	0.0021
	Y0	0.1365	0.0060	0.0067	0.0053	-0.4607	0.0060	0.2382	0.1731	-0.1107	0.0103
	MD	-0.1376	-0.0055	-0.0074	-0.0060	0.4583	-0.0055	-0.2782	-0.1984	0.1274	-0.0081

Table 3.26: Bias of the estimators where $\theta_1 = \theta_2 = 0.5$ and the proportion of complete cases is 0.85

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.1132	0.1060	0.1840	0.1132	0.1132	0.0002	0.0004	0.0003	0.0003	0.0003
	Y0	0.2266	0.9382	0.4473	0.2267	0.2267	0.0013	0.0009	0.0015	0.0012	0.0012
	MD	-0.1134	-0.8323	-0.2633	-0.1134	-0.1135	-0.0011	-0.0005	-0.0012	-0.0009	-0.0009
n = 5,000	Y1	0.1128	0.1055	0.1823	0.1128	0.1128	0.0006	0.0005	0.0007	0.0005	0.0005
	Y0	0.2259	0.9360	0.4467	0.2259	0.2259	-0.0008	0.0058	0.0023	-0.0007	-0.0008
	MD	-0.1132	-0.8305	-0.2644	-0.1131	-0.1131	0.0013	-0.0052	-0.0016	0.0012	0.0012
n = 2,000	Y1	0.1130	0.1059	0.1831	0.1127	0.1127	-0.001	-0.0007	0.0004	-0.001	-0.001
	Y0	0.2274	0.9351	0.4481	0.2271	0.2271	-0.0005	0.005	0.002	-0.0008	-0.0008
	MD	-0.1144	-0.8292	-0.2650	-0.1144	-0.1144	-0.0005	-0.0057	-0.0016	-0.0002	-0.0002

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0053	-0.0004	0.0048	0.0048	-0.0005	0.0005	-0.0001	-0.0001
	Y0	0.1476	0.0265	-0.0344	-0.0346	0.0252	0.0103	0.0002	0.0002
	MD	-0.1423	-0.0269	0.0392	0.0395	-0.0257	-0.0098	-0.0003	-0.0003
n = 5,000	Y1	0.0055	0	0.005	0.005	-0.0015	0.0000	-0.0006	-0.0005
	Y0	0.1531	0.0273	-0.0364	-0.0366	0.0442	0.0175	-0.0012	-0.0007
	MD	-0.1476	-0.0273	0.0413	0.0416	-0.0456	-0.0174	0.0006	0.0002
n = 2,000	Y1	0.0044	-0.0004	0.0034	0.0034	-0.0021	0.0020	-0.0008	-0.0006
	Y0	0.1537	0.0273	-0.0361	-0.0364	0.0859	0.0334	-0.0005	-0.0001
	MD	-0.1493	-0.0276	0.0396	0.0398	-0.0881	-0.0314	-0.0003	-0.0006

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	Y1	-0.0003	-0.0001	-0.0002	-0.0002	0.0001	-0.0001	-0.0235	-0.0138	0.0013	-0.0001
	Y0	0.0206	0.0006	0.0006	0.0005	-0.4486	0.0006	0.1596	0.1138	-0.0162	0.0009
	MD	-0.0209	-0.0007	-0.0008	-0.0007	0.4487	-0.0007	-0.1832	-0.1276	0.0175	-0.0011
n = 5,000	Y1	-0.0012	-0.0005	-0.0005	-0.0005	-0.0008	-0.0005	-0.0236	-0.0141	0.0021	-0.0005
	Y0	0.0375	-0.0001	0.0003	0.0001	-0.4276	-0.0001	0.1601	0.1136	-0.0321	0.0006
	MD	-0.0387	-0.0004	-0.0008	-0.0006	0.4267	-0.0004	-0.1838	-0.1277	0.0342	-0.0012
n = 2,000	Y1	-0.0012	-0.0005	-0.0005	-0.0005	-0.0015	-0.0006	-0.0237	-0.0142	0.0062	-0.0007
	Y0	0.0375	-0.0001	0.0003	0.0001	-0.3802	0.0011	0.1628	0.1154	-0.0953	0.0034
	MD	-0.0387	-0.0004	-0.0008	-0.0006	0.3788	-0.0018	-0.1865	-0.1296	0.1015	-0.0041

Table 3.27: Variance of the estimators where $\theta_1 = \theta_2 = 0.5$ and the proportion of complete cases is 0.85

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0007	0.0009	0.0021	0.0008	0.0008	0.0006	0.0006	0.0011	0.0006	0.0006
	Y0	0.0021	0.0387	0.0066	0.0022	0.0022	0.0018	0.013	0.0033	0.0018	0.0018
	MD	0.0011	0.0384	0.0054	0.0012	0.0012	0.001	0.0123	0.0022	0.001	0.001
n = 5,000	Y1	0.0014	0.0018	0.0041	0.0016	0.0016	0.0012	0.0014	0.0022	0.0013	0.0013
	Y0	0.0043	0.0758	0.0129	0.0044	0.0044	0.0035	0.0281	0.0069	0.0035	0.0035
	MD	0.0021	0.0757	0.0108	0.0024	0.0024	0.0019	0.0268	0.0048	0.002	0.002
n = 2,000	Y1	0.0036	0.0047	0.0108	0.0040	0.0039	0.0031	0.0034	0.0056	0.0032	0.0032
	Y0	0.0110	0.1910	0.0322	0.0114	0.0114	0.0091	0.0704	0.0174	0.0092	0.0092
	MD	0.0054	0.1922	0.0281	0.0060	0.0060	0.0047	0.0673	0.0123	0.0049	0.0049

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0006	0.0011	0.0006	0.0006	0.0008	0.0014	0.0008	0.0007
	Y0	0.0123	0.0032	0.0018	0.0018	0.0961	0.0164	0.0037	0.0022
	MD	0.0116	0.0022	0.001	0.001	0.0906	0.0122	0.0020	0.0013
n = 5,000	Y1	0.0014	0.0023	0.0013	0.0013	0.0017	0.0028	0.0016	0.0014
	Y0	0.0267	0.0068	0.0036	0.0036	0.1544	0.0263	0.0076	0.0045
	MD	0.0254	0.0046	0.002	0.002	0.1445	0.0192	0.0041	0.0026
n = 2,000	Y1	0.0034	0.0056	0.0032	0.0032	0.0043	0.0071	0.0041	0.0034
	Y0	0.066	0.0168	0.0094	0.0094	0.3112	0.0528	0.0189	0.0112
	MD	0.0631	0.0118	0.0051	0.0051	0.2893	0.0373	0.0101	0.0064

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	Y1	0.0008	0.0007	0.0006	0.0006	0.0008	0.0007	0.0008	0.0006	0.0007	0.0007
	Y0	0.0546	0.0023	0.0021	0.0020	0.0714	0.0023	0.0038	0.0018	0.2414	0.0023
	MD	0.0527	0.0014	0.0013	0.0013	0.0757	0.0014	0.0023	0.0011	0.2450	0.0014
n = 5,000	Y1	0.0016	0.0014	0.0013	0.0013	0.0017	0.0014	0.0017	0.0013	0.0014	0.0014
	Y0	0.0853	0.0046	0.0041	0.0041	0.2146	0.0046	0.0085	0.0040	0.4385	0.0049
	MD	0.0827	0.0027	0.0025	0.0025	0.2258	0.0027	0.0052	0.0025	0.4461	0.0030
n = 2,000	Y1	0.0016	0.0014	0.0013	0.0013	0.0042	0.0034	0.0043	0.0032	0.0038	0.0034
	Y0	0.0853	0.0046	0.0041	0.0041	0.5786	0.0115	0.0196	0.0093	1.2566	0.0118
	MD	0.0827	0.0027	0.0025	0.0025	0.6094	0.0066	0.0116	0.0055	1.2797	0.0071

Table 3.28: MSE of the estimators where $\theta_1 = \theta_2 = 0.5$ and the proportion of complete cases is 0.85

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0135	0.0121	0.0359	0.0136	0.0136	0.0006	0.0006	0.0011	0.0006	0.0006
	Y0	0.0535	0.9190	0.2067	0.0536	0.0536	0.0018	0.013	0.0033	0.0018	0.0018
	MD	0.0139	0.7311	0.0747	0.0141	0.0141	0.001	0.0123	0.0022	0.001	0.001
n = 5,000	Y1	0.0141	0.0129	0.0373	0.0143	0.0143	0.0012	0.0014	0.0022	0.0013	0.0013
	Y0	0.0553	0.9519	0.2125	0.0555	0.0555	0.0035	0.0281	0.007	0.0035	0.0035
	MD	0.0149	0.7655	0.0807	0.0152	0.0152	0.0019	0.0268	0.0048	0.002	0.002
n = 2,000	Y1	0.0164	0.0159	0.0443	0.0167	0.0166	0.0031	0.0034	0.0056	0.0032	0.0032
	Y0	0.0628	1.0654	0.2330	0.0630	0.0630	0.0091	0.0704	0.0174	0.0092	0.0092
	MD	0.0185	0.8798	0.0984	0.0191	0.0191	0.0047	0.0674	0.0123	0.0049	0.0049

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0006	0.0011	0.0006	0.0006	0.0008	0.0014	0.0008	0.0007
	Y0	0.0123	0.0032	0.0018	0.0018	0.0967	0.0165	0.0037	0.0022
	MD	0.0318	0.0029	0.0026	0.0026	0.0912	0.0123	0.0020	0.0013
n = 5,000	Y1	0.0014	0.0023	0.0013	0.0013	0.0017	0.0028	0.0016	0.0014
	Y0	0.0501	0.0075	0.0049	0.0049	0.1564	0.0267	0.0076	0.0045
	MD	0.0472	0.0053	0.0037	0.0038	0.1466	0.0195	0.0041	0.0026
n = 2,000	Y1	0.0034	0.0056	0.0032	0.0032	0.0071	0.0041	0.0034	0.0181
	Y0	0.0896	0.0175	0.0107	0.0107	0.0528	0.0189	0.0112	0.0092
	MD	0.0853	0.0126	0.0066	0.0066	0.2893	0.0373	0.0101	0.0064

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	Y1	0.0008	0.0007	0.0006	0.0006	0.0008	0.0007	0.0014	0.0008	0.0007	0.0007
	Y0	0.0550	0.0023	0.0021	0.0020	0.2726	0.0023	0.0292	0.0148	0.2416	0.0023
	MD	0.0531	0.0014	0.0013	0.0013	0.2770	0.0014	0.0358	0.0174	0.2453	0.0014
n = 5,000	Y1	0.0016	0.0014	0.0013	0.0013	0.0017	0.0014	0.0023	0.0015	0.0015	0.0014
	Y0	0.0867	0.0046	0.0041	0.0041	0.3974	0.0046	0.0341	0.0169	0.4395	0.0049
	MD	0.0842	0.0027	0.0025	0.0025	0.4079	0.0027	0.0390	0.0188	0.4473	0.0030
n = 2,000	Y1	0.0016	0.0014	0.0013	0.0013	0.0042	0.0034	0.0043	0.0032	0.0038	0.0034
	Y0	0.0853	0.0046	0.0041	0.0041	0.5786	0.0115	0.0196	0.0093	1.2655	0.0118
	MD	0.0827	0.0027	0.0025	0.0025	0.6094	0.0066	0.0116	0.0055	1.2899	0.0071

Table 3.29: Variance of the estimators where $\theta_1 = \theta_2 = 0.5$ and the proportion of complete cases is 0.7

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0008	0.0012	0.0034	0.0009	0.0009	0.0006	0.0006	0.001	0.0006	0.0006
	Y0	0.0027	0.1039	0.0151	0.0029	0.0029	0.002	0.014	0.0036	0.002	0.002
	MD	0.0014	0.1045	0.0141	0.0017	0.0017	0.0011	0.0132	0.0024	0.0011	0.0011
n = 5,000	Y1	0.0017	0.0024	0.0068	0.0020	0.0019	0.0013	0.0014	0.0023	0.0013	0.0013
	Y0	0.0055	0.2085	0.0292	0.0058	0.0058	0.0036	0.0271	0.0069	0.0037	0.0037
	MD	0.0028	0.2100	0.0267	0.0034	0.0034	0.002	0.0253	0.0047	0.0021	0.0021
n = 2,000	Y1	0.0043	0.0060	0.0163	0.0048	0.0048	0.0032	0.0033	0.0056	0.0032	0.0032
	Y0	0.0136	0.5344	0.0721	0.0145	0.0145	0.0093	0.071	0.0173	0.0093	0.0093
	MD	0.0069	0.5367	0.0668	0.0084	0.0084	0.0054	0.0675	0.0124	0.0056	0.0056

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0006	0.001	0.0006	0.0006	0.0015	0.0023	0.0015	0.0008
	Y0	0.0132	0.0034	0.002	0.002	0.2105	0.0334	0.0119	0.0036
	MD	0.0124	0.0023	0.0012	0.0012	0.1921	0.0237	0.0063	0.0023
n = 5,000	Y1	0.0014	0.0023	0.0013	0.0013	0.0032	0.0046	0.0032	0.0017
	Y0	0.0263	0.0067	0.0038	0.0038	0.3167	0.0485	0.0233	0.0068
	MD	0.0244	0.0045	0.0022	0.0022	0.2838	0.0324	0.0123	0.0043
n = 2,000	Y1	0.0034	0.0055	0.0032	0.0032	0.0082	0.0108	0.0080	0.0042
	Y0	0.0672	0.0167	0.0096	0.0096	0.5357	0.0802	0.0580	0.0164
	MD	0.0634	0.0118	0.0059	0.0059	0.4663	0.0497	0.0302	0.0100

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	Y1	0.0014	0.0008	0.0007	0.0007	0.0015	0.0008	0.0016	0.0006	0.0008	0.0008
	Y0	0.0962	0.0041	0.0027	0.0026	0.3234	0.0041	0.0149	0.0027	0.8468	0.0047
	MD	0.0968	0.0026	0.0022	0.0021	0.3500	0.0026	0.0091	0.0020	0.8608	0.0033
n = 5,000	Y1	0.0029	0.0017	0.0014	0.0014	0.0032	0.0017	0.0034	0.0013	0.0018	0.0017
	Y0	0.1359	0.0074	0.0050	0.0050	0.5761	0.0074	0.0280	0.0042	1.2465	0.0075
	MD	0.1409	0.0047	0.0039	0.0038	0.6323	0.0047	0.0160	0.0029	1.2731	0.0050
n = 2,000	Y1	0.0072	0.0042	0.0036	0.0036	0.0081	0.0042	0.0085	0.0033	0.0046	0.0042
	Y0	0.2346	0.0175	0.0117	0.0117	2.1213	0.0175	0.0680	0.0111	3.1547	0.0183
	MD	0.2551	0.0107	0.0089	0.0088	2.2921	0.0107	0.0392	0.0079	3.2202	0.012

Table 3.30: MSE of the estimators where $\theta_1 = \theta_2 = 0.5$ and the proportion of complete cases is 0.7

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0414	0.0379	0.0916	0.0415	0.0415	0.0019	0.0040	0.0040	0.0019	0.0019
	Y0	0.1650	2.1081	0.4934	0.1652	0.1652	0.0019	0.0130	0.0035	0.0020	0.0020
	MD	0.0420	1.6028	0.1698	0.0423	0.0423	0.0069	0.0136	0.0118	0.0069	0.0069
n = 5,000	Y1	0.0141	0.0129	0.0373	0.0143	0.0143	0.0036	0.0083	0.0078	0.0037	0.0037
	Y0	0.0553	0.9519	0.2125	0.0555	0.0555	0.0038	0.0271	0.0069	0.0039	0.0039
	MD	0.0149	0.7655	0.0807	0.0152	0.0152	0.0133	0.0268	0.0237	0.0134	0.0134
n = 2,000	Y1	0.0164	0.0159	0.0443	0.0167	0.0166	0.0091	0.0216	0.0185	0.0092	0.0092
	Y0	0.0628	1.0654	0.2330	0.0630	0.0630	0.0092	0.0703	0.0172	0.0092	0.0092
	MD	0.0185	0.8798	0.0984	0.0191	0.0191	0.0324	0.0694	0.0567	0.0326	0.0326

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0008	0.001	0.0008	0.0008	0.0015	0.0023	0.0015	0.0008
	Y0	0.0574	0.0044	0.0051	0.0051	0.2147	0.0341	0.0119	0.0036
	MD	0.0508	0.0034	0.0058	0.0059	0.1965	0.0242	0.0063	0.0023
n = 5,000	Y1	0.0016	0.0023	0.0015	0.0015	0.0032	0.0046	0.0032	0.0017
	Y0	0.0744	0.008	0.0065	0.0066	0.3274	0.0499	0.0233	0.0068
	MD	0.066	0.0059	0.0066	0.0067	0.2951	0.0337	0.0123	0.0043
n = 2,000	Y1	0.0035	0.0056	0.0034	0.0034	0.0083	0.0109	0.0080	0.0042
	Y0	0.1173	0.018	0.0126	0.0127	0.5917	0.0877	0.0581	0.0164
	MD	0.1078	0.0134	0.0104	0.0105	0.5243	0.0557	0.0302	0.0100

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	Y1	0.0014	0.0008	0.0007	0.0007	0.0015	0.0008	0.0031	0.0012	0.0009	0.0008
	Y0	0.0985	0.0041	0.0027	0.0026	0.6711	0.0041	0.0729	0.0318	0.8475	0.0047
	MD	0.0991	0.0026	0.0022	0.0021	0.6980	0.0026	0.0867	0.0401	0.8616	0.0033
n = 5,000	Y1	0.0029	0.0017	0.0014	0.0014	0.0032	0.0017	0.0048	0.0019	0.0019	0.0017
	Y0	0.0074	0.0050	0.0050	0.1469	0.8693	0.0074	0.0870	0.0334	1.2515	0.0075
	MD	0.1469	0.0047	0.0040	0.0038	0.9238	0.0047	0.0947	0.0412	1.2793	0.0050
n = 2,000	Y1	0.0072	0.0042	0.0036	0.0036	0.0081	0.0042	0.0101	0.0039	0.0049	0.0042
	Y0	0.2532	0.0175	0.0118	0.0117	2.3336	0.0175	0.1248	0.0411	3.1666	0.0184
	MD	0.2740	0.0107	0.0090	0.0088	2.5021	0.0107	0.1166	0.0473	3.2361	0.0121

Table 3.31: MSE of the estimators where $\theta_1 = \theta_2 = 0.5$ and the proportion of complete cases is 0.7

Missing data methods		Complete Cases analysis					Within MI				
Causal estimators		g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW	g-formula	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0414	0.0379	0.0916	0.0415	0.0415	0.0019	0.0040	0.0040	0.0019	0.0019
	Y0	0.1650	2.1081	0.4934	0.1652	0.1652	0.0019	0.0130	0.0035	0.0020	0.0020
	MD	0.0420	1.6028	0.1698	0.0423	0.0423	0.0069	0.0136	0.0118	0.0069	0.0069
n = 5,000	Y1	0.0141	0.0129	0.0373	0.0143	0.0143	0.0036	0.0083	0.0078	0.0037	0.0037
	Y0	0.0553	0.9519	0.2125	0.0555	0.0555	0.0038	0.0271	0.0069	0.0039	0.0039
	MD	0.0149	0.7655	0.0807	0.0152	0.0152	0.0133	0.0268	0.0237	0.0134	0.0134
n = 2,000	Y1	0.0164	0.0159	0.0443	0.0167	0.0166	0.0091	0.0216	0.0185	0.0092	0.0092
	Y0	0.0628	1.0654	0.2330	0.0630	0.0630	0.0092	0.0703	0.0172	0.0092	0.0092
	MD	0.0185	0.8798	0.0984	0.0191	0.0191	0.0324	0.0694	0.0567	0.0326	0.0326

Missing data methods		Across MI				IPCCW			
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW
n = 10,000	Y1	0.0008	0.001	0.0008	0.0008	0.0015	0.0023	0.0015	0.0008
	Y0	0.0574	0.0044	0.0051	0.0051	0.2147	0.0341	0.0119	0.0036
	MD	0.0508	0.0034	0.0058	0.0059	0.1965	0.0242	0.0063	0.0023
n = 5,000	Y1	0.0016	0.0023	0.0015	0.0015	0.0032	0.0046	0.0032	0.0017
	Y0	0.0744	0.008	0.0065	0.0066	0.3274	0.0499	0.0233	0.0068
	MD	0.066	0.0059	0.0066	0.0067	0.2951	0.0337	0.0123	0.0043
n = 2,000	Y1	0.0035	0.0056	0.0034	0.0034	0.0083	0.0109	0.0080	0.0042
	Y0	0.1173	0.018	0.0126	0.0127	0.5917	0.0877	0.0581	0.0164
	MD	0.1078	0.0134	0.0104	0.0105	0.5243	0.0557	0.0302	0.0100

Missing data methods		AIPCCW - Model				AIPCCW - MC				Restricted AIPCCW	
Causal estimators		IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	N_IPTW	AIPTW	N_AIPTW	IPTW	AIPTW
n = 10,000	Y1	0.0014	0.0008	0.0007	0.0007	0.0015	0.0008	0.0031	0.0012	0.0009	0.0008
	Y0	0.0985	0.0041	0.0027	0.0026	0.6711	0.0041	0.0729	0.0318	0.8475	0.0047
	MD	0.0991	0.0026	0.0022	0.0021	0.6980	0.0026	0.0867	0.0401	0.8616	0.0033
n = 5,000	Y1	0.0029	0.0017	0.0014	0.0014	0.0032	0.0017	0.0048	0.0019	0.0019	0.0017
	Y0	0.0074	0.0050	0.0050	0.1469	0.8693	0.0074	0.0870	0.0334	1.2515	0.0075
	MD	0.1469	0.0047	0.0040	0.0038	0.9238	0.0047	0.0947	0.0412	1.2793	0.0050
n = 2,000	Y1	0.0072	0.0042	0.0036	0.0036	0.0081	0.0042	0.0101	0.0039	0.0049	0.0042
	Y0	0.2532	0.0175	0.0118	0.0117	2.3336	0.0175	0.1248	0.0411	3.1666	0.0184
	MD	0.2740	0.0107	0.0090	0.0088	2.5021	0.0107	0.1166	0.0473	3.2361	0.0121

Table 3.32: Bias, variance, and MSE of AIPCCW-MC estimators with more flexible imputation mode where $\theta_1 = \theta_2 = 0.5$ and the proportion of complete cases is 0.30

		AIPW_IPW	N_AIPW_IPW	AIPW_AIPW	N_AIPW_AIPW
Bias					
n=10000	Y1	0.0029	0.0001	0.0175	0.0034
	Y0	-0.0742	0.0001	-0.0031	0.0487
	MD	0.0771	0.0000	0.0206	-0.0453
n=5000	Y1	0.0010	-0.0004	0.0177	0.0031
	Y0	-0.0398	-0.0014	0.0070	0.0521
	MD	0.0408	0.0010	0.0106	-0.0490
n=2000	Y1	0.0000	0.0005	0.0151	0.0027
	Y0	-0.0264	0.0013	0.2085	0.0776
	MD	0.0265	-0.0009	-0.1934	-0.0749
Variance					
n=10000	Y1	0.0015	0.0008	0.0015	0.0006
	Y0	0.4284	0.0038	0.0150	0.0031
	MD	0.4570	0.0024	0.0094	0.0024
n=5000	Y1	0.0032	0.0017	0.0033	0.0013
	Y0	0.7187	0.0069	0.0299	0.0054
	MD	0.7782	0.0044	0.0189	0.0042
n=2000	Y1	0.0081	0.0042	0.0082	0.0033
	Y0	3.8938	0.0167	103.9473	0.1636
	MD	4.0835	0.0103	103.9990	0.1619
MSE					
n=10000	Y1	0.0015	0.0008	0.0018	0.0006
	Y0	0.4339	0.0038	0.0151	0.0054
	MD	0.4630	0.0024	0.0099	0.0045
n=5000	Y1	0.0032	0.0017	0.0036	0.0013
	Y0	0.7203	0.0069	0.0300	0.0081
	MD	0.7799	0.0044	0.0190	0.0066
n=2000	Y1	0.0081	0.0042	0.0084	0.0033
	Y0	3.8945	0.0167	103.9907	0.1696
	MD	4.0842	0.0103	104.0363	0.1675

Table 3.33: Bias, variance, and MSE of AIPCCW-MC estimators with more flexible imputation mode where $\theta_1 = \theta_2 = 0.5$ and the proportion of complete cases is 0.15

		Bias			
		AIPW_IPW	N_AIPW_IPW	AIPW_AIPW	N_AIPW_AIPW
n=10000	Y1	0.0012	-0.0001	0.0060	0.0011
	Y0	-0.0654	0.0002	-0.0064	0.0295
	MD	0.0665	-0.0003	0.0124	-0.0284
n=5000	Y1	0.0002	-0.0005	0.0058	0.0008
	Y0	-0.0493	-0.0007	-0.0035	0.0303
	MD	0.0495	0.0002	0.0093	-0.0295
n=2000	Y1	-0.0005	-0.0006	0.0056	0.0006
	Y0	-0.0214	-0.0001	0.0136	0.0380
	MD	0.0209	-0.0005	-0.0080	-0.0374
		Variance			
n=10000	Y1	0.0008	0.0007	0.0008	0.0006
	Y0	0.0745	0.0023	0.0041	0.0020
	MD	0.0789	0.0013	0.0024	0.0012
n=5000	Y1	0.0017	0.0014	0.0017	0.0013
	Y0	0.2298	0.0045	0.0092	0.0042
	MD	0.2412	0.0026	0.0057	0.0026
n=2000	Y1	0.0042	0.0034	0.0042	0.0032
	Y0	0.7265	0.0112	0.0645	0.0152
	MD	0.7593	0.0064	0.0572	0.0113
		MSE			
n=10000	Y1	0.0008	0.0007	0.0008	0.0006
	Y0	0.0788	0.0023	0.0042	0.0028
	MD	0.0833	0.0013	0.0026	0.0020
n=5000	Y1	0.0017	0.0014	0.0017	0.0013
	Y0	0.2323	0.0045	0.0092	0.0051
	MD	0.2437	0.0026	0.0058	0.0035
n=2000	Y1	0.0042	0.0034	0.0042	0.0032
	Y0	0.7269	0.0112	0.0647	0.0166
	MD	0.7597	0.0064	0.0572	0.0127

Bibliography

[1] Issa J Dahabreh et al. "Heterogeneity of Treatment Effects". In: *Methods in Comparative Effectiveness Research* (2017), pp. 227–271.

[2] Issa J Dahabreh, Rodney Hayward, and David M Kent. "Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence". In: *International Journal of Epidemiology* 45.6 (2016), pp. 2184–2193.

[3] David M Kent, Ewout Steyerberg, and David van Klaveren. "Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects". In: *BMJ* (Dec. 2018), k4245. DOI: 10.1136/bmj.k4245. URL: https://doi.org/10.1136%2Fbmj.k4245.

[4] David M. Kent et al. "The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement". In: *Annals of Internal Medicine* 172.1 (Nov. 2019), p. 35. DOI: 10.7326/m18-3667. URL: https://doi.org/10.7326%2Fm18-3667.

[5] David M. Kent et al. "The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement: Explanation and Elaboration". In: *Annals of Internal Medicine* 172.1 (Nov. 2019), W1. DOI: 10.7326/m18-3668. URL: https://doi.org/10.7326%2Fm18-3668.

[6] David van Klaveren et al. "Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting". In: *Journal of Clinical Epidemiology* 114 (Oct. 2019), pp. 72–83. DOI: 10.1016/j.jclinepi.2019.05.029. URL: https://doi.org/10.1016%5C%2Fj.jclinepi.2019.05.029.

[7] J. F. Burke et al. "Using Internally Developed Risk Models to Assess Heterogeneity in Treatment Effects in Clinical Trials". In: *Circulation: Cardiovascular Quality and Outcomes* 7.1 (Jan. 2014), pp. 163–169. DOI: 10.1161/circoutcomes.113.000497. URL: https://doi.org/10.1161%2Fcircoutcomes.113.000497.

[8] Donald B Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of Educational Psychology* 66.5 (1974), p. 688.

[9] James M Robins and Sander Greenland. "Causal inference without counterfactuals: comment". In: *Journal of the American Statistical Association* 95.450 (2000), pp. 431–435.

[10] Miguel A Hernán and James M Robins. *Causal inference (forthcoming)*. Boca Raton, FL: Chapman & Hall/CRC, 2020.

[11] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. "Subgroup identification from randomized clinical trial data". In: *Statistics in medicine* 30.24 (2011), pp. 2867–2880.

[12] Lu Tian et al. "A simple method for estimating interactions between a treatment and a large number of covariates". In: *Journal of the American Statistical Association* 109.508 (2014), pp. 1517–1532.

[13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[14] for the ALLHAT Research Group et al. " Rationale and Design for the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) ". In: *American Journal of Hypertension* 9.4 (Apr. 1996), pp. 342–360. eprint: http://oup.prod.sis.lan/ajh/article-pdf/9/4/342/7449697/9-4-342.pdf. URL: https://doi.org/10.1016/0895-7061(96)00037-4.

[15] The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. "Major Outcomes in High-Risk Hypertensive Patients Randomized to Angiotensin-Converting Enzyme Inhibitor or Calcium Channel Blocker vs DiureticThe Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT)". In: *JAMA* 288.23 (Dec. 2002), pp. 2981–2997. ISSN: 0098-7484. DOI: 10.1001/jama.288.23.2981. eprint: https://jamanetwork.com/journals/jama/articlepdf/195626/joc21962.pdf. URL: https://doi.org/10.1001/jama.288.23.2981.

[16] The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. "Major Cardiovascular Events in Hypertensive Patients Randomized to Doxazosin vs ChlorthalidoneThe Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT)". In: *JAMA* 283.15 (Apr. 2000), pp. 1967–1975. ISSN: 0098-7484. DOI: 10.1001/jama.283.15.1967. eprint: https://jamanetwork.com/journals/jama/articlepdf/192604/joc00401.pdf. URL: https://doi.org/10.1001/jama.283.15.1967.

[17] Barry R Davis et al. "Relationship of antihypertensive treatment regimens and change in blood pressure to risk for heart failure in hypertensive patients randomly assigned to doxazosin or chlorthalidone: further analyses from the Antihypertensive and Lipid-Lowering treatment to prevent Heart Attack Trial". In: *Annals of internal medicine* 137.5_Part_1 (2002), pp. 313–320.

[18] Jason Nelson et al. "Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials". In: *International Journal of Epidemiology* 45.6 (July 2016), pp. 2075–2088. DOI: 10.1093/ije/dyw118. eprint: http://oup.prod.sis.lan/ije/article-pdf/45/6/2075/24170598/dyw118.pdf. URL: https://doi.org/10.1093/ije/dyw118.

[19] Tyler J VanderWeele et al. "Selecting Optimal Subgroups for Treatment Using Many Covariates". en. In: *Epidemiology* 30.3 (May 2019), pp. 334–341.

[20] Stephen R. Cole and Elizabeth A. Stuart. "Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial". In: *American Journal of Epidemiology* 172.1 (June 2010), pp. 107–115. ISSN: 0002-9262. DOI: 10.1093/aje/kwq084. eprint: http://oup.prod.sis.lan/aje/article-pdf/172/1/107/187351/kwq084.pdf. URL: https://doi.org/10.1093/aje/kwq084.

[21] "The use of propensity scores to assess the generalizability of results from randomized trials". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 174.2 (2011), pp. 369–386. ISSN: 09641998, 1467985X. URL: http://www.jstor.org/stable/23014404.

[22] Jin-Liern Hong et al. “Generalizing Randomized Clinical Trial Results: Implementation and Challenges Related to Missing Data in the Target Population”. In: *American Journal of Epidemiology* 187.4 (Aug. 2017), pp. 817–827. ISSN: 0002-9262. DOI: 10.1093/aje/kwx287. eprint: http://oup.prod.sis.lan/aje/article-pdf/187/4/817/25082017/kwx287.pdf. URL: https://doi.org/10.1093/aje/kwx287.

[23] Issa J. Dahabreh et al. “Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals”. In: *Biometrics* 0.ja (2018). DOI: 10.1111/biom.13009. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13009. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13009.

[24] Issa J Dahabreh et al. “Study Designs for Extending Causal Inferences from a Randomized Trial to a Target Population”. In: *American Journal of Epidemiology* (Dec. 2020). kwaa270. ISSN: 0002-9262. DOI: 10.1093/aje/kwaa270. eprint: https://academic.oup.com/aje/advance-article-pdf/doi/10.1093/aje/kwaa270/34923675/kwaa270.pdf. URL: https://doi.org/10.1093/aje/kwaa270.

[25] Issa J. Dahabreh et al. “Extending inferences from a randomized trial to a new target population”. In: *Statistics in Medicine* 39.14 (2020), pp. 1999–2014. DOI: https://doi.org/10.1002/sim.8426. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8426. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8426.

[26] Daniel Westreich et al. “Transportability of Trial Results Using Inverse Odds of Sampling Weights”. In: *American Journal of Epidemiology* 186.8 (May 2017), pp. 1010–1014. eprint: http://oup.prod.sis.lan/aje/article-pdf/186/8/1010/24330634/kwx164.pdf.

[27] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. “Estimation of Regression Coefficients When Some Regressors Are Not Always Observed”. In: *Journal of the American Statistical Association* 89.427 (1994), pp. 846–866. ISSN: 01621459. URL: http://www.jstor.org/stable/2290910.

[28] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[29] BaoLuo Sun and Eric J. Tchetgen Tchetgen. "On Inverse Probability Weighting for Nonmonotone Missing at Random Data". In: *Journal of the American Statistical Association* 113.521 (2018), pp. 369–379. DOI: 10.1080/01621459.2016.1256814. eprint: https://doi.org/10.1080/01621459.2016.1256814. URL: https://doi.org/10.1080/01621459.2016.1256814.

[30] Shaun Seaman et al. "What Is Meant by "Missing at Random"?" In: *Statist. Sci.* 28.2 (May 2013), pp. 257–268. DOI: 10.1214/13-STS415. URL: https://doi.org/10.1214/13-STS415.

[31] Fabrizia Mealli and Donald B. Rubin. "Clarifying missing at random and related definitions, and implications when coupled with exchangeability". In: *Biometrika* 102.4 (Sept. 2015), pp. 995–1000. ISSN: 0006-3444. DOI: 10.1093/biomet/asv035. eprint: http://oup.prod.sis.lan/biomet/article-pdf/102/4/995/649481/asv035.pdf. URL: https://doi.org/10.1093/biomet/asv035.

[32] Alan E. Gelfand, Adrian F. M. Smith, and Tai-Ming Lee. "Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling". In: *Journal of the American Statistical Association* 87.418 (1992), pp. 523–532. ISSN: 01621459. URL: http://www.jstor.org/stable/2290286.

[33] "Coronary artery surgery study (CASS): A randomized trial of coronary artery bypass surgery: Comparability of entry characteristics and survival in randomized patients and nonrandomized patients meeting randomization criteria". In: *Journal of the American College of Cardiology* 3.1 (1984), pp. 114–128. ISSN: 0735-1097. DOI: https://doi.org/10.1016/S0735-1097(84)80437-4. URL: http://www.sciencedirect.com/science/article/pii/S0735109784804374.

[34] "Coronary artery surgery study (CASS): a randomized trial of coronary artery bypass surgery. Survival data". en. In: *Circulation* 68.5 (Nov. 1983), pp. 939–950.

[35] CLAUDIA SCHMOOR, MANFRED OLSCHEWSKI, and MARTIN SCHUMACHER. "RANDOMIZED AND NON-RANDOMIZED PATIENTS IN CLINICAL TRIALS: EXPERIENCES WITH COMPREHENSIVE COHORT STUDIES". In: *Statistics in*

Medicine 15.3 (1996), pp. 263–271. DOI: 10.1002/(SICI)1097-0258(19960215)15:3<263::AID-SIM165>3.0.CO;2-K.

[36] Neil J Perkins et al. "Principled Approaches to Missing Data in Epidemiologic Studies". In: *American Journal of Epidemiology* 187.3 (Nov. 2017), pp. 568–575. ISSN: 0002-9262. DOI: 10.1093/aje/kwx348. eprint: http://oup.prod.sis.lan/aje/article-pdf/187/3/568/24134556/kwx348.pdf. URL: https://doi.org/10.1093/aje/kwx348.

[37] Ofer Harel et al. "Multiple Imputation for Incomplete Data in Epidemiologic Studies". In: *American Journal of Epidemiology* 187.3 (Nov. 2017), pp. 576–584. ISSN: 0002-9262. DOI: 10.1093/aje/kwx349. eprint: http://oup.prod.sis.lan/aje/article-pdf/187/3/576/24134569/kwx349.pdf. URL: https://doi.org/10.1093/aje/kwx349.

[38] BaoLuo Sun et al. "Inverse-Probability-Weighted Estimation for Monotone and Non-monotone Missing Data". In: *American Journal of Epidemiology* 187.3 (Nov. 2017), pp. 585–591. ISSN: 0002-9262. DOI: 10.1093/aje/kwx350. eprint: http://oup.prod.sis.lan/aje/article-pdf/187/3/585/24134640/kwx350.pdf. URL: https://doi.org/10.1093/aje/kwx350.

[39] Lingling Li et al. "On weighting approaches for missing data". en. In: *Stat. Methods Med. Res.* 22.1 (Feb. 2013), pp. 14–30.

[40] Robin Mitra and Jerome P Reiter. "A comparison of two methods of estimating propensity scores after multiple imputation". In: *Statistical Methods in Medical Research* 25.1 (2016). PMID: 22687877, pp. 188–204. DOI: 10.1177/0962280212445945. eprint: https://doi.org/10.1177/0962280212445945. URL: https://doi.org/10.1177/0962280212445945.

[41] Clémence Leyrat et al. "Propensity score analysis with partially observed covariates: How should multiple imputation be used?" In: *Statistical Methods in Medical Research* 28.1 (2019). PMID: 28573919, pp. 3–19. DOI: 10.1177/0962280217713032. eprint: https://doi.org/10.1177/0962280217713032. URL: https://doi.org/10.1177/0962280217713032.

[42] BBL Penning de Vries and RHH Groenwold. "Comments on propensity score matching following multiple imputation". In: *Statistical Methods in Medical Research* 25.6 (2016). PMID: 27852808, pp. 3066–3068. DOI: 10.1177/0962280216674296. eprint: https://doi.org/10.1177/0962280216674296. URL: https://doi.org/10.1177/0962280216674296.

[43] Katherine Evans, Isabel Fulcher, and Eric J Tchetgen Tchetgen. "A coherent likelihood parametrization for doubly robust estimation of a causal effect with missing confounders". In: *arXiv preprint arXiv:2007.10393* (2020).

[44] James Robins. "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7.9 (1986), pp. 1393–1512. ISSN: 0270-0255. DOI: https://doi.org/10.1016/0270-0255(86)90088-6. URL: http://www.sciencedirect.com/science/article/pii/0270025586900886.

[45] James M. Robins. "Association, Causation, and Marginal Structural Models". In: *Synthese* 121.1/2 (1999), pp. 151–179. ISSN: 00397857, 15730964. URL: http://www.jstor.org/stable/20118224.

[46] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. DOI: 10.1017/CBO9780511802256.

[47] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models". In: *Journal of the American Statistical Association* 94.448 (1999), 1096–1120 (with Rejoinder, 1135–1146). DOI: 10.1080/01621459.1999.10473862. eprint: https://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1999.10473862. URL: https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1999.10473862.

[48] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models". In: *Journal of the American Statistical Association* 94.448 (1999), 1096–1120 (with Rejoinder, 1135–1146). DOI: 10.1080/01621459.1999.10473862. eprint: https://amstat.tandfonline.

com/doi/pdf/10.1080/01621459.1999.10473862. URL: https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1999.10473862.

[49] Iris Eekhout et al. "Missing data: a systematic review of how they are reported and handled". In: *Epidemiology* 23.5 (2012), pp. 729–732.

[50] Roderick J A Little and Donald B Rubin. *Statistical Analysis with Missing Data*. en. John Wiley & Sons, Aug. 2014.

[51] Shaun R Seaman and Ian R White. "Review of inverse probability weighting for dealing with missing data". In: *Statistical Methods in Medical Research* 22.3 (2013). PMID: 21220355, pp. 278–295. DOI: 10.1177/0962280210395740. eprint: https://doi.org/10.1177/0962280210395740. URL: https://doi.org/10.1177/0962280210395740.

[52] PAUL R. ROSENBAUM and DONALD B. RUBIN. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (Apr. 1983), pp. 41–55. ISSN: 0006-3444. DOI: 10.1093/biomet/70.1.41. eprint: https://academic.oup.com/biomet/article-pdf/70/1/41/662954/70-1-41.pdf. URL: https://doi.org/10.1093/biomet/70.1.41.

[53] Rohit Bhattacharya et al. "Identification In Missing Data Models Represented By Directed Acyclic Graphs". In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. Ed. by Ryan P. Adams and Vibhav Gogate. Vol. 115. Proceedings of Machine Learning Research. Tel Aviv, Israel: PMLR, July 2020, pp. 1149–1158. URL: http://proceedings.mlr.press/v115/bhattacharya20b.html.

[54] Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. "Full Law Identification in Graphical Models of Missing Data: Completeness Results". In: *Proceedings of machine learning research* 119 (July 2020), pp. 7153–7163. ISSN: 2640-3498. URL: https://europepmc.org/articles/PMC7716645.

[55] Bas BL Penning de Vries and Rolf HH Groenwold. "A comparison of two approaches to implementing propensity score methods following multiple imputation". In: *Epidemiology, Biostatistics and Public Health* 14.4 (2017).

[56] Emily Granger, Jamie C. Sergeant, and Mark Lunt. "Avoiding pitfalls when combining multiple imputation and propensity scores". In: *Statistics in Medicine* 38.26 (2019), pp. 5120–5132. DOI: 10.1002/sim.8355. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8355. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8355.

[57] Geert Molenberghs et al. *Handbook of missing data methodology*. CRC Press, 2014.

[58] E.J. Williamson, A. Forbes, and R. Wolfe. "Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder". In: *Statistics in Medicine* 31.30 (2012), pp. 4382–4400. DOI: 10.1002/sim.5643. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.5643. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5643.

[59] Barry C Arnold and S James Press. "Compatible conditional distributions". In: *Journal of the American Statistical Association* 84.405 (1989), pp. 152–156.

[60] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

[61] JINGCHEN LIU et al. "On the stationary distribution of iterative imputations". In: *Biometrika* 101.1 (2014), pp. 155–173. ISSN: 00063444. URL: http://www.jstor.org/stable/43305601.

[62] Ying Huang and Xiao-Hua Zhou. "Identification of the optimal treatment regimen in the presence of missing covariates". In: *Statistics in Medicine* 39.4 (2020), pp. 353–368. DOI: 10.1002/sim.8407. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8407. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8407.

[63] Victor Chernozhukov et al. "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1 (Jan. 2018), pp. C1–C68. ISSN: 1368-4221. DOI: 10.1111/ectj.12097. eprint: https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf. URL: https://doi.org/10.1111/ectj.12097.

[64] Vira Semenova and Victor Chernozhukov. "Debiased machine learning of conditional average treatment effects and other causal functions". In: *The Econometrics Journal* (Aug. 2020). utaa027. ISSN: 1368-4221. DOI: 10.1093/ectj/utaa027. eprint: https://academic.oup.com/ectj/advance-article-pdf/doi/10.1093/ectj/utaa027/33800361/utaa027.pdf. URL: https://doi.org/10.1093/ectj/utaa027.

[65] Kenneth Joseph Arrow. *Social Choice and Individual Values*. 1951.

[66] R. Duncan Luce. "On the possible psychophysical laws." In: *Psychological Review* 66.2 (1959), pp. 81–95.

[67] Judea Pearl. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (Dec. 1995), pp. 669–688. ISSN: 0006-3444. DOI: 10.1093/biomet/82.4.669. eprint: https://academic.oup.com/biomet/article-pdf/82/4/669/698263/82-4-669.pdf. URL: https://doi.org/10.1093/biomet/82.4.669.