5-2022

Statistical methods for assessing drug interactions and identifying effect modifiers using observational data.

Qian Xu
University of Louisville

STATISTICAL METHODS FOR ASSESSING DRUG INTERACTIONS AND IDENTIFYING EFFECT MODIFIERS USING OBSERVATIONAL DATA

By

Qian Xu
B.S. in Mathematics, China Agricultural University, 2013
M.S. in Statistics, The George Washington University, 2015

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

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

May 2022

STATISTICAL METHODS FOR ASSESSING DRUG INTERACTIONS AND IDENTIFYING EFFECT MODIFIERS USING OBSERVATIONAL DATA

By

Qian Xu
B.S. in Mathematics, China Agricultural University, 2013
M.S. in Statistics, The George Washington University, 2015

A Dissertation Approved on

April 22, 2022

by the following Dissertation Committee:

Maiying Kong, Ph.D., Committee Chair

Karunarathna B. Kulasekera, Ph.D.

Jiapeng Huang, M.D., Ph.D.

Demetra Antimisiaris, Pharm.D.

Jeremy Gaskins, Ph.D.

Qi Zheng, Ph.D.

DEDICATION

To my parents for their care, guidance, encouragement, advice and support.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my esteemed advisor Dr. Maiying Kong for her unparalleled guidance, insightful inspiration, and unwavering support for my research over the last four years. It has been a great privilege and joy to study under her supervision. Furthermore, it is my honor to benefit from her personality and diligence, which I will treasure my whole life.

I would also like to extend my extreme gratitude to my dissertation committee. My sincere and hearty gratitude goes to Dr. Qi Zheng for his ingenious suggestions and invaluable contributions to my research projects, to Dr. Jiapeng Huang for providing me with precious opportunities to work on COVID-19 clinical research and hand hygiene clinical trial with his strong research teams, to Dr. Demetra Antimisiaris for her valuable guidance on drug interactions, and to Dr. Karunarathna B. Kulasekera and Dr. Jeremy Gaskins for their consistent support and time to serve on my dissertation committee.

My warm gratitude also goes to the Department of Bioinformatics and Biostatistics for providing me with a great environment to do research, thank all the faculty members for enriching my knowledge in biostatistics and related fields.

My acknowledgements also go to my classmates and friends. Their support and generosity move me a lot. Last but not least, I have to thank my family members who have been supporting me on every step on the road of my life.

ABSTRACT

STATISTICAL METHODS FOR ASSESSING DRUG INTERACTIONS AND IDENTIFYING EFFECT MODIFIERS USING OBSERVATIONAL DATA

Qian Xu

April 22, 2022

This dissertation consists of three projects related to causal inference based on observational data.

In the first project, we propose a double robust to identify the effect modifiers and estimate optimal treatment. Observational studies differ from experimental studies in that assignment of subjects to treatments is not randomized but rather occurs due to natural mechanisms, which are usually hidden from the researchers. Many statistical methods to identify the treatment effect and select the optimal personalized treatment for experimental studies may not be suitable for observational studies any more. In this project, we propose a flexible outcome model to select the optimal personalized treatment which is suitable for experimental studies as well as observational studies. In the proposed model, the control group response profile is captured by a non-parametric function, and treatment heterogeneity is captured by the interaction term between treatment and a linear combination of covariates. L1 penalty and A-learning method are proposed to select the important variables in the interaction terms, thus the effect modifiers can be obtained and the optimal treatment can be determined. The proposed approach is quite flexible and has a doubly

robust nature in terms of that the estimated individual treatment effect is consistent if either the control group response profile or the propensity score model is correctly specified.

In the second project, we propose a statistical method for assessing drug interactions with binary treatments. With advances in medicine, many drugs and treatments become available. On the one hand, polydrug use (*i.e.*, using more than one drug at a time) has been used to treat patients with multiple morbid conditions, and polydrug use may cause severe side effects. On the other hand, combination treatments have been successfully developed to treat severe diseases such as cancer and chronic diseases. Observational data, such as electronic health record data, may provide useful information for assessing drug interactions. In this project we propose using marginal structural models to assess the average treatment effect and causal interaction of two drugs by controlling confounding variables. The causal effect and the interaction of two drugs are assessed using the weighted likelihood approach, with weights being the inverse probability of the treatment assigned. Simulation studies were conducted to examine the performance of the proposed method, which showed that the proposed method was able to estimate the causal parameters consistently. Case studies were conducted to examine the joint effect of metformin and glyburide use on reducing the hospital readmission for type 2 diabetic patients, and to examine the joint effect of antecedent statins and opioids use on the immune and inflammatory biomarkers for COVID-19 hospitalized patients.

In the third project, we propose a statistical methods for assessing treatment interactions where treatment could be measured in a continuous scale such as different dose levels or intensity of treatment. Combination treatment has been often used to treat certain disease such as cancer or alcohol use disorders (AUD). For example, medication and psychotherapy could be applied together to treat patients with AUD. Observational data from electronic health records or claims data are examples of such

data resources which could be used to examine treatment effects and treatment interactions. In the second project, we proposed the generalized MSMs and provide the procedures for estimating ATE and treatment interactions using observational data, where the confounding variables are controlled via the IPTW method. Nevertheless, this method presents the MSMs and algorithms for estimating ATE and treatment interaction when two treatments are used together and each drug has only two levels (present or not), which is unsuitable for the situation when each treatment has multiple levels or in continuous scale. In this project, we propose the marginal structural semiparametric model (MSSM) to estimate ATE and treatment interactions, where the generalized propensity score (GPS) method and spline functions are applied, and each treatment either includes multiple levels or in continuous scale. The statistical method developed here can be used to investigate ATE and treatment interaction on treatment effect, as well as on adverse event, depending on the outcome of interest.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Doubly robust methods for identifying effect modifiers and estimating optimal treatment based on observational data

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

In this project, we propose a flexible outcome model, where the control group response profile is captured by a non-parametric function, and treatment heterogeneity is captured by the interaction term between treatment and a linear combination of covariates. L1 penalty (Zou and Hastie, 2005a) and A-learning method (Schulte

et al., 2014) are proposed to select the important variables in the interaction terms, thus the effect modifiers can be obtained and the optimal treatment can be determined. The proposed approach is quite flexible, and it can be applied to data from either randomized trials or observational studies.

1.2 Statistical methods for assessing drug interactions using observational data with binary treatments

With advances in medicine, many drugs have become available to treat patients. Polydrug use may cause adverse side effects, and has increasingly caused concerns (Tramontina et al., 2018). The data depositories from routine clinical practice provide great opportunity to study the treatment effect of combination therapy and the adverse effect of polydrug use. Observational data from electronic health records or claims data are examples of such data resources which could be used to examine treatment efficacy or adverse outcomes due to combination therapy or polydrug use. In either type of data sets, confounding variables, which are causally related to both treatment selection and outcome variable, must be controlled to obtain unbiased estimates for treatment effect and drug interaction.

As multiple drug use increases, drug interaction has been recognized as an important problem recently, yet the rigorous statistical methods assessing drug interaction based on clinical observational data are lacking. The assessment of drug interaction based on observational data requires thoughtful consideration of confounding factors such as patient-specific characteristics and comorbidity. The most critical issue using clinical observational data to assess drug interactions is to control the confounding factors, which impact both treatment selection and outcome variables (Hernan and Robins, 2020). Propensity score based methods such as the inverse probability of treatment weighting (IPTW) and doubly robust methods have been

applied to estimate causal effects (Rosenbaum and Rubin (1983); Yan et al. (2019)). In particular, IPTW method along with marginal structural models (MSMs) have been applied to estimate causal parameters such as average treatment effect (Cole and Hernán (2008); Robins et al. (2000)).

In this project, we propose using MSMs in the framework of generalized linear models to assess the interaction of two drugs. The statistical method we develop here can be used to investigate drug interaction on treatment effect, as well as drug interaction on adverse event, which depends on the outcome of interest.

1.3 Statistical methods for assessing drug interactions using observational data with treatment in continuous scale

It is common that several drugs (treatments) are prescribed to a patient by health providers, particularly, when the patient has multiple morbid conditions. Preventing drug interaction plays a vital role in maximizing patient benefit from polydrug use. Observational data from electronic health records or claims data are examples of such data resources which could be used to examine drug interactions.

However, one of the greatest challenges using clinical observational data to assess drug interactions is to control the confounding factors, which impact both treatment selection and outcome variables (Hernan and Robins, 2020). Despite propensity score based methods are popular in observational studies, a main practical difficulty of these methods is that the propensity score must be estimated. Previous studies have revealed that slight misspecification of the propensity score model can result in substantial bias of estimated treatment effects (e.g. Kang and Schafer (2007) and Smith and Todd (2005)). Evidence suggests that covariate balancing propensity score (CBPS) method is more reliable, which is robust to mild misspecification of the parametric propensity score model (Imai and Ratkovic, 2014).

In previous project, we proposed generalized MSMs and provide the procedures for estimating ATE and drug interaction using observational data. Nevertheless, this method presents the MSMs and algorithms for estimating ATE and drug interaction when two drugs are used together and each drug has only two levels (present or not), which is unsuitable for the situation when each drug has multiple levels or in continuous scale.

In this project, we propose a marginal structural semiparametric model (MSSM) to estimate ATE and treatment interaction with treatment in multiple levels or in continuous scale. The generalized propensity score (Hirano and Imbens, 2004) enables us to investigate the drugs with multiple levels or in continuous scale. Once the generalized propensity scores are obtained, the MSSM based on the weighted sample can be applied to estimate the interaction. Thus, the proposed MSSM along with IPTW method provides rigorous statistical method for assessing treatment interactions when treatment is measured either in multi-levels or in continuous scale.

CHAPTER 2

DOUBLY ROBUST METHODS FOR IDENTIFYING EFFECT MODIFIERS AND ESTIMATING OPTIMAL TREATMENT BASED ON OBSERVATIONAL DATA

2.1 Introduction

As advancing in medicine, many drugs have been developed to treat patients. It is often that a treatment may only work for patients with certain characteristics (Dijkman et al., 2009). Therefore, studying the treatment heterogeneity and selecting optimal treatment have drawn much attention in the literature (Murphy, 2003; Zhang et al., 2012; Lu et al., 2013; Tian et al., 2014; Foster et al., 2015, 2016; Lipkovich et al., 2017). Many statistical methods were developed to identify the treatment effect and select the optimal personalized treatment for experimental studies. For example, Tian et al. (2014) considered a simple method for estimating interactions between a treatment and a large number of covariates based on experimental data.

Observational studies differ from experimental studies in that assignment of subjects to treatments is not randomized but rather occurs due to natural mechanisms, which are usually hidden from the researchers. Although observational studies such as electronic health records and registry data often provide rich data resources under routine clinical settings, many statistical methods developed for experimental studies may not be suitable for observational studies anymore, because the relation-

ship between treatment and outcomes in observational studies is often confounded. Appropriate statistical methods for assessing treatment heterogeneity and selecting optimal treatments need to handle the possible confounding.

In the literature on optimal treatment, Laber and Zhao (2015) developed methods for optimal treatment selection based on decision trees (Breiman, 2001), which can facilitate communication with health care providers. In contrast, regression-based approaches (see, e.g., Qian and Murphy (2011); Brinkley et al. (2010)) typically face the choice between constructing parsimonious models leading to more interpretable decision rules but subject to model misspecification, or more complex models that may avoid misspecification but result in "unintelligible" treatment rules. Tian et al. (2014) proposed such a regression-based model, where the interaction term of treatment and covariates are included in the model to capture the contrasts between groups, and further provide guiding decisions about future subjects. However, questions have been raised about the assumption of correctly specification of propensity score model. A major problem with this method is the estimates would be biased if the propensity score model is misspecified.

In this chapter, we propose a flexible outcome model, where the control group response profile is captured by a non-parametric function, and treatment heterogeneity is captured by the interaction term between treatment and a linear combination of covariates. L1 penalty (Zou and Hastie, 2005a) and A-learning method (Schulte et al., 2014) are proposed to select the important variables in the interaction terms, thus the effect modifiers can be obtained and the optimal treatment can be determined. The most important contribution of the proposed method is double robust to model misspecification for consistent estimation for optimal treatment regime. A-learning requires only the part of the outcome regression including contrasts among treatments and propensity scores, which is more robustness, and the form of the decision rules defining the optimal regime is easy interpretable. Besides, the proposed method

can handle the high dimensional data by L1 penalty. Our simulation studies showed that the estimates are still consistent if either the propensity score model is correctly specified or the control group response profile is correctly specified. The proposed approach is quite flexible, and it can be applied to data from either randomized trials or observational studies.

The structure of the remainder of this chapter is as follows: in Section 2.2, we describe the context and propose a method using L1 penalty (Zou and Hastie, 2005a) and A-learning method (Schulte et al., 2014) to estimate the contrast function and identify the optimal treatment for a given patient. In Section 2.3, we describe the simulation study we executed to examine the performance of the proposed method. In Section 2.4, we applied the proposed method to study which group of patients more likely benefit from statin use in control of inflammation for COVID-19 patients based on an observational study. Finally, in Section 2.5, we provide further discussion of the results and general conclusions.

2.2 Proposed doubly robust method for estimating optimal treatment

Let $(Y, T, \boldsymbol{X})$ indicate the triplet for the p-dimensional baseline covariates, treatment received, and outcome variable. Without loss of generality, we consider the case with two possible treatment choices, that is, $T \in \{0, 1\}$ with 0 and 1 denoting receiving placebo(control) and getting treated, respectively. Let $\{(Y_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$ denote a random sample consisting of n i.i.d replicates of $(Y, T, \boldsymbol{X})$. The sample could result from either a randomized experimental design or an observational study.

The following model (2.1) has often been used (see, e.g., Lu et al., 2013; Fu et al., 2016; Lipkovich et al., 2017) to identify the subgroup which may benefit from the treatment and select the optimal treatment regime:

$$E(Y|\boldsymbol{X}, T) = h^*(\boldsymbol{X}) + g^*(\boldsymbol{X})T. \tag{2.1}$$

The interaction term $g^*(\boldsymbol{X})T$ plays a vital role in quantifying the benefits that a patient receives from the treatment and hence identifying the optimal treatment. To further elaborate, we use the concept of potential outcomes (Rubin, 1974). Let $Y^{(0)}$ and $Y^{(1)}$ denote respectively, the potential outcomes for a subject with covariates $\boldsymbol{X}$ receiving control and treatment. Here we assume that exchangeability and consistency hold (Hernan and Robins, 2020), where (i) exchangeability: $Y^{(t)} \perp T|\boldsymbol{X}$ for $t = 0, 1$, i.e., given $\boldsymbol{X}$, the potential outcome is independent of the treatment received; and (ii) consistency: $Y = TY^{(1)} + (1 - T)Y^{(0)}$, i.e., the observed outcome equals the potential outcome corresponding to the treatment the subject receives. Under the two conditions and (2.1), we observe that

$$h^*(\boldsymbol{X}) = E(Y|\boldsymbol{X}, T = 0) = E(Y^{(0)}|\boldsymbol{X}, T = 0) = E(Y^{(0)}|\boldsymbol{X}) \quad (2.2)$$

$$h^*(\boldsymbol{X}) + g^*(\boldsymbol{X}) = E(Y|\boldsymbol{X}, T = 1) = E(Y^{(1)}|\boldsymbol{X}, T = 1) = E(Y^{(1)}|\boldsymbol{X}) \quad (2.3)$$

where the first equalities in both (2.2) and (2.3) follow from (2.1), the second equalities follow from the consistency condition, and the last equalities follow from the exchangeability condition. Subsequently, $E(Y^{(1)} - Y^{(0)}|\boldsymbol{X}) = g^*(\boldsymbol{X})$. Therefore, $g^*(\boldsymbol{X})$ in (2.1) is a contrast function, capturing the change of response brought by treatment for a subject with covariates $\boldsymbol{X}$, whereas the function $h^*(\boldsymbol{X})$ is a baseline function, reflecting the response profile if the subject receives no treatment.

In practice, the primary interest is to identify the subjects who may have a beneficial treatment effect, that is, to pinpoint $\{\boldsymbol{X} : E[Y^{(1)} - Y^{(0)}|\boldsymbol{X}] = g^*(\boldsymbol{X}) > 0\}$. Thus, we focus on estimating the contrast function $g^*(\boldsymbol{X})$ instead of $h^*(\boldsymbol{X})$, and subsequently finding out the subjects who may have a beneficial treatment effect.

To facilitate a clinical decision, we would like the $g^*(\boldsymbol{X})$ function to be simple and to capture the main variables for decision making on which treatment works better for a person with covariate $\boldsymbol{X}$. The function $h^*(\boldsymbol{X})$ captures the response

profile under control, which is not of direct interest and is considered as a nuisance function. However, a good estimation of $h^*(\boldsymbol{X})$ can facilitate an accurate estimation of the contrast function $g^*(\boldsymbol{X})$. In the following, we develop a doubly robust estimator for $g^*(\boldsymbol{X})$ in that $g^*(\boldsymbol{X})$ can be estimated consistently if either propensity score model $\pi^*(\boldsymbol{X}) = P(T = 1|\boldsymbol{X})$ or the outcome model $h^*(\boldsymbol{X})$ is correctly specified.

2.2.1 Doubly robust method

Let assume that $g^*(\boldsymbol{X})$ is a linear combination of $\boldsymbol{X}$, that is,

$$g^*(\boldsymbol{X}) = g(\boldsymbol{X}; \boldsymbol{\beta}^*) = \beta_0^* + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_1^* = \widetilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\beta}^*,$$

where $\widetilde{\boldsymbol{X}} = (1, \boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\beta}^* = (\beta_0^*, \boldsymbol{\beta}_1^{*T})^{\mathrm{T}}$ with $\beta_0^* \in \mathbb{R}$ and $\boldsymbol{\beta}_1^* \in \mathbb{R}^p$. Lu et al. (2013) and Lipkovich et al. (2017) proposed to estimate $g^*(\boldsymbol{X}, \boldsymbol{\beta}^*)$ by minimizing the following squared-loss function with respect to $\boldsymbol{\beta}$

$$L(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i - h_1^*(\boldsymbol{X}_i) - \widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}\big(T_i - \pi^*(\boldsymbol{X}_i)\big)\Big\}^2, \tag{2.4}$$

where $h_1^*(\boldsymbol{X}) = E(Y|\boldsymbol{X})$. Consequently, an estimating equation for $\boldsymbol{\beta}^*$ can be obtained by setting the first derivative of the objective function (2.4) to zero:

$$\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{X}}_i\{T_i - \pi^*(\boldsymbol{X}_i)\}\Big\{Y_i - h_1^*(\boldsymbol{X}_i) - \widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}\big(T_i - \pi^*(\boldsymbol{X}_i)\big)\Big\} = 0. \tag{2.5}$$

However, as $\pi^*(\boldsymbol{X})$ is involved in both of the two terms $T_i - \pi^*(\boldsymbol{X}_i)$ and $Y_i - h_1^*(\boldsymbol{X}_i) - \widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}(T_i - \pi^*(\boldsymbol{X}_i))$ in the product in Equation (2.5), the expectation of the left hand side of Equation (2.5) has a value zero only if $\pi^*(\boldsymbol{X})$ is correctly specified. Therefore, the solution to (2.5) may not provide consistent estimates for $\boldsymbol{\beta}^*$ if $\pi^*(\boldsymbol{X})$ is misspecified.

We note that

$$
\begin{aligned}
h_1^*(\boldsymbol{X}) &= E(Y|\boldsymbol{X}) = E((1-T)Y^{(0)} + TY^{(1)}|\boldsymbol{X}) \quad \text{(by consistency)} \\
&= E(Y^{(0)}|\boldsymbol{X})(1-\pi^*(\boldsymbol{X})) + E(Y^{(1)}|\boldsymbol{X})\pi^*(\boldsymbol{X}) \quad \text{(by exchangeability)} \\
&= h^*(\boldsymbol{X})(1-\pi^*(\boldsymbol{X})) + (h^*(\boldsymbol{X}) + g^*(\boldsymbol{X}))\pi^*(\boldsymbol{X}) \quad \text{(by consistency and (2.1))} \\
&= h^*(\boldsymbol{X}) + g^*(\boldsymbol{X})\pi^*(\boldsymbol{X}).
\end{aligned}
$$

Therefore, Equation (2.5) is essentially equivalent to

$$
\frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i\{T_i - \pi^*(\boldsymbol{X}_i)\}\Big\{Y_i - h^*(\boldsymbol{X}_i) - T_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}\Big\} = 0. \tag{2.6}
$$

In Equation (2.6), the propensity score $\pi^*(\boldsymbol{X})$ is only included in the term $T_i - \pi^*(\boldsymbol{X}_i)$, whereas $Y_i - h^*(\boldsymbol{X}_i) - T_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}$ is solely related to the outcome model. In the following Proposition 2.2.1, we show that $E[\widetilde{\boldsymbol{X}}_i\{T_i - \pi(\boldsymbol{X}_i)\}\{Y_i - h(\boldsymbol{X}_i) - T_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}^*\}] = 0$ if either the propensity score is correctly specified (*i.e.*, $\pi(\cdot) = \pi^*(\cdot)$) or the response profile for subjects under control is correctly specified (*i.e.*, $h(\cdot) = h^*(\cdot)$).

Proposition 2.2.1. *Assuming that $g(\boldsymbol{X};\beta) = \widetilde{\boldsymbol{X}}^{\mathrm{T}}\beta$, if either the propensity score $\pi^*(\boldsymbol{X})$ or the response profile $h^*(\boldsymbol{X})$ is correctly specified, then*

$$
E\left[\widetilde{\boldsymbol{X}}_i\{T_i - \pi(\boldsymbol{X}_i)\}\Big\{Y_i - h(\boldsymbol{X}_i) - T_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}^*\Big\}\right] = 0
$$

Proposition 2.2.1 implies that that Equation (2.6) is a valid estimating equation if either $h^*(\boldsymbol{X})$ or $\pi^*(\boldsymbol{X})$ is correctly specified. We denote its solution by $\hat{\boldsymbol{\beta}}_{DR}$. Equation (2.6) has been referred as A-learning in the literature (Schulte et al., 2014). The proof of Proposition 2.2.1 is provided in the Appendix.

As $h^*(\boldsymbol{X})$ and $\pi^*(\boldsymbol{X})$ are typically unknown in practice, they are often replaced by their estimates $\hat{h}(\boldsymbol{X})$ and $\hat{\pi}(\boldsymbol{X})$. Lu et al. (2013) proposed to estimate the nuisance function $h_1^*(\boldsymbol{X})$ via a parametric model, such as a constant or a linear

combination of $\boldsymbol{X}$, which can be adopted to estimate $h^*(\boldsymbol{X})$ as well. $\pi^*(\boldsymbol{X})$, also known as a propensity score function, is often estimated via the logistic regression in an observational study, and is often assumed to be a known constant (say 0.5) for a randomized study. With $h^*(\boldsymbol{X})$ and $\pi^*(\boldsymbol{X})$ replaced by $\hat{h}(\boldsymbol{X})$ and $\hat{\pi}(\boldsymbol{X})$ in Equation (2.6), simple algebra yields that

$$\hat{\boldsymbol{\beta}}_{DR} = \left(\sum_{i=1}^{n} \{T_i - \hat{\pi}(\boldsymbol{X}_i)\} T_i \widetilde{\boldsymbol{X}}_i \widetilde{\boldsymbol{X}}_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{n} \{T_i - \hat{\pi}(\boldsymbol{X}_i)\} \widetilde{\boldsymbol{X}}_i \{Y_i - \hat{h}(\boldsymbol{X}_i)\} \right). \quad (2.7)$$

Next we show that $\hat{\boldsymbol{\beta}}_{DR}$ is consistent, if either the propensity score model $\pi^*(\boldsymbol{X})$ or the response function under control $h^*(\boldsymbol{X})$ can be consistently estimated. Therefore, the solution to Equation (2.6) is doubly robust (Koch et al., 2018; Yan et al., 2019).

Theorem 2.2.1. *Under Conditions (C1)–(C3) in the Appendix if* $\min\{\|\hat{\pi} - \pi^*\|_\infty, \|\hat{h} - h^*\|_\infty\} = o_p(1)$*, then* $\hat{\boldsymbol{\beta}}_{DR} \to_p \boldsymbol{\beta}$*.*

The proof of Theorem 2.2.1 is provided in the appendix. Theorem 2.2.1 indicates that our proposed doubly robust estimator $\hat{\boldsymbol{\beta}}_{DR}$ achieves the estimation consistency. Next, we establish its asymptotic distribution.

Define $\boldsymbol{B}(\pi) = E\left[\{1 - \pi(\boldsymbol{X})\} T \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{X}}^{\mathrm{T}}\right]$. Denote $\tilde{\pi}$ and $\tilde{h}$ as the functions to which $\hat{\pi}$ and $\hat{h}$ converge in probability respectively. $\tilde{\pi} \neq \pi^*$ (or $\tilde{h} \neq h^*$) implies that π^* (or h^*) is not consistently estimated.

Theorem 2.2.2. *Under Conditions (C1)–(C3) in the Appendix, suppose* $\|\hat{\pi} - \tilde{\pi}\|_\infty = o_p(n^{-\alpha_1})$ *and* $\|\hat{h} - \tilde{h}\|_\infty = o_p(n^{-\alpha_2})$ *for some* $\tilde{\pi} \in \Pi, \tilde{h} \in \mathcal{H}$*, and* $\alpha_1 + \alpha_2 > 1/2$*.*

(i) If $\tilde{\pi} = \pi^*$ *and Condition (C4):* $E[\{\pi^*(\boldsymbol{X}) - \hat{\pi}(\boldsymbol{X})\} \widetilde{\boldsymbol{X}} \{h^*(\boldsymbol{X}) - \tilde{h}(\boldsymbol{X})\}] = \mathbb{E}_n \psi_{\tilde{h}}(T_i, \boldsymbol{X}_i) + o_p(n^{-1/2})$*,* $E[\psi_{\tilde{h}}(T, \boldsymbol{X})] = 0$*, and* $E[\|\psi_{\tilde{h}}(T, \boldsymbol{X})\|^2] < \infty$ *holds, then*

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{DR} - \boldsymbol{\beta}^*) \to_d N(0, \boldsymbol{B}^{-1}(\pi^*) \boldsymbol{\Sigma}(\tilde{h}) \boldsymbol{B}^{-1}(\pi^*)),$$

where

$$\begin{aligned}\boldsymbol{\Sigma}(\tilde{h}) =& E\left[\left(\{T-\pi^*(\boldsymbol{X})\}\widetilde{\boldsymbol{X}}\{h^*(\boldsymbol{X})+\epsilon-\tilde{h}(\boldsymbol{X})\}+\psi_{\tilde{h}}(T,\boldsymbol{X})\right)\right.\\ &\left.\times\left(\{T-\pi^*(\boldsymbol{X})\}\widetilde{\boldsymbol{X}}\{h^*(\boldsymbol{X})+\epsilon-\tilde{h}(\boldsymbol{X})\}+\psi_{\tilde{h}}(T,\boldsymbol{X})\right)^{\mathrm{T}}\right].\end{aligned}$$

(ii) If $\tilde{h}=h^*$ *and Condition (C5):* $E[\{T-\tilde{\pi}(\boldsymbol{X})\}\widetilde{\boldsymbol{X}}\{h^*(\boldsymbol{X})-\hat{h}(\boldsymbol{X})\}]=\mathbb{E}_n\phi_{\tilde{\pi}}(Y_i,$ $T_i,\boldsymbol{X}_i)+o_p(n^{-1/2})$, $E[\phi_{\tilde{\pi}}(Y,T,\boldsymbol{X})]=0$, *and* $E[\|\phi_{\tilde{\pi}}(Y,T,\boldsymbol{X})\|^2]<\infty$ *holds, then*

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{DR}-\boldsymbol{\beta}^*)\to_d N(0,\boldsymbol{B}^{-1}(\tilde{\pi})\boldsymbol{\Sigma}(\tilde{\pi})\boldsymbol{B}^{-1}(\tilde{\pi})),$$

where

$$\begin{aligned}\boldsymbol{\Sigma}(\tilde{\pi}) =& E\left[\left(\{T-\tilde{\pi}(\boldsymbol{X})\}\widetilde{\boldsymbol{X}}\epsilon+\phi_{\tilde{\pi}}(Y,T,\boldsymbol{X})\right)\right.\\ &\left.\times\left(\{T-\tilde{\pi}(\boldsymbol{X})\}\widetilde{\boldsymbol{X}}\epsilon+\phi_{\tilde{\pi}}(Y,T,\boldsymbol{X})\right)^{\mathrm{T}}\right].\end{aligned}$$

We discuss the feasibility of conditions in Theorem 2.2.2 in the following two remarks.

Remark 2.2.1. *The conditions that* $\|\hat{\pi}-\tilde{\pi}\|_\infty=o_p(n^{-\alpha_1})$ *and* $\|\hat{h}-\tilde{h}\|_\infty=o_p(n^{-\alpha_2})$ *with* $\alpha_1+\alpha_2>1/2$ *are often satisfied by different approaches. For example, if parametric models are used to obtain* $\hat{\pi}$ *and* $\hat{h}$, $\|\hat{\pi}-\tilde{\pi}\|_\infty=o_p(n^{-1/2})$ *and* $\|\hat{h}-\tilde{h}\|_\infty=o_p(n^{-1/2})$ *under Conditions (C1) and (C3). If multivariate kernel methods are used to acquire both* $\hat{\pi}$ *and* $\hat{h}$, $\|\hat{\pi}-\tilde{\pi}\|_\infty=o_p(n^{-2/(p+4)}\log n)$ *and* $\|\hat{h}-\tilde{h}\|_\infty=o_p(n^{-2/(p+4)}\log n)$ *(see, e.g., Yang and Tschernig, 1999) and the condition holds if* $p<4$.

Remark 2.2.2. *Conditions (C4) and (C5) are also satisfied by many commonly used estimation methods, including the maximum likelihood estimation, the estimating equation approaches, nonparametric kernel regression, and smoothing splines regression. Conditions (C4) and (C5) are inspired by the condition 2.6' (ii) in Chen et al.*

(2003).

Theorem 2.2.2 indicates that the asymptotic variance of $\hat{\boldsymbol{\beta}}_{DR}$ depends on both $\hat{\pi}$ and $\hat{h}$. It is worthwhile to mention that the asymptotic covariance matrix in Theorem 1 of Lu et al. (2013) is $\boldsymbol{B}^{-1}(\pi^*)\boldsymbol{\Sigma}(\tilde{h})\boldsymbol{B}^{-1}(\pi^*)$ in Theorem 2.2.2 (i) with $\tilde{\pi} = \pi^*$. If both π^* and h^* are consistently estimated, we obtain the following corollary.

Corollary 2.2.1. *Under Conditions (C1)–(C3) in the Appendix, suppose* $\|\hat{\pi}-\pi^*\|_\infty = o_p(n^{-\alpha_1})$ *and* $\|\hat{h} - h^*\|_\infty = o_p(n^{-\alpha_2})$*, and* $\alpha_1 + \alpha_2 > 1/2$*, then*

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{DR} - \boldsymbol{\beta}^*) \to_d N(0, \boldsymbol{B}^{-1}(\pi^*)E\left[(T - \pi^*(\boldsymbol{X}))^2\epsilon^2\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{T}}\right]\boldsymbol{B}^{-1}(\pi^*)).$$

From Theorem 2.2.1, the estimated parameters $\boldsymbol{\beta}$ for the contrast function $g^*(\boldsymbol{X})$ based on the estimating equation (2.6) has doubly robust properties. Although estimating the mean function $h^*(\boldsymbol{X})$ is not of primary interest, the correct specification of $h^*(\boldsymbol{X})$ can improve the efficiency of the estimator $\hat{\boldsymbol{\beta}}_{DR}$ and guarantee a consistent estimator for $\boldsymbol{\beta}$. In the literature, simple function forms such as constant or linear combination of covariates have been proposed to estimate $h^*(\boldsymbol{X})$ or $h_1^*(\boldsymbol{X})$ (Lu et al., 2013). However, such specification could be far from true model, which could result in biased estimates if the propensity score model is not correctly specified. In this article, we propose using a more flexible ensemble method to estimate the response profile $h^*(\boldsymbol{X})$. The ensemble method uses several commonly used methods and selects the optimal one (i.e., having the smallest predicted mean squared errors (MSE)) based on the 10-fold cross-validation method, such as multiple linear regression model, multivariate adaptive regression spline (MARS), and the gradient boosting method (GBM). MARS (Friedman, 1991) has been used to alleviate the bias problem and achieve the ideal performance in causal inferences (Foster et al., 2015; Powers et al., 2018). MARS is a non-parametric regression technique which uses hinge functions as the basis functions, and MARS can be seen as an extension

of linear models that automatically models nonlinearities and interactions between variables (Friedman, 1991). The maximum degree of interactions of predictors and the number of terms retained in the final model are two important tuning parameters in MARS model, and we performed a grid search and 10-fold cross-validation method to select the two optimal tuning parameters. MARS technique has been implemented in the R-package "earth" (Milborrow, 2019). GBM is a tree based approach and iteratively adds basis functions in a greedy fashion so that each additional basis function further reduces the selected loss function (Friedman, 2001). GBM and multiple linear model are very commonly used to estimate the response profile. The proposed ensemble method uses MARS, GBM and linear regression model and selects the one which provides the smallest MSE based on 10-fold cross-validation method.

We propose the following algorithm to estimate the parameter β in the contrast function $g(\boldsymbol{X};\beta)$:

(i) Estimate $\pi^*(\boldsymbol{X})$: we propose to estimate $\pi^*(\boldsymbol{X})$ via the logistic regression regardless the data is from an observational study or a randomized trial. The resulting estimate is denoted as $\hat{\pi}(\boldsymbol{X})$. Note that for a randomized study a known constant, say 0.5, has been used as $\pi^*(\boldsymbol{X})$.

(ii) Estimate $E(Y^{(0)}|\boldsymbol{X}) = h^*(\boldsymbol{X})$ nonparametrically: we fit the model $h^*(\boldsymbol{X})$ using the proposed ensemble method and the observations from control subjects and obtain a predicted model $\hat{h}(\boldsymbol{X})$. We predict $\hat{h}(\boldsymbol{X}_i), i = 1, \cdots, n$, the potential outcome under control for all subjects regardless of the treatment assignment.

(iii) Estimate the parameters $\boldsymbol{\beta}^*$ in the contrast function $g(X;\beta^*)$ by the following expression:

$$\hat{\beta}_{DR} = \left(\sum_{i=1}^{n} \{T_i - \hat{\pi}(\boldsymbol{X}_i)\} T_i \widetilde{\boldsymbol{X}}_i \widetilde{\boldsymbol{X}}_i^{\mathrm{T}} \right)^{-1} \left(\sum_{i=1}^{n} \{T_i - \hat{\pi}(\boldsymbol{X}_i)\} \{Y_i - \hat{h}(\boldsymbol{X}_i)\} \widetilde{\boldsymbol{X}}_i \right)$$

(iv) Identify patients who benefit from treatment and identify the effect modifiers which impact the treatment heterogeneity. On the one hand, we can identify the patients who benefit from the treatment as those who have covariate $\boldsymbol{X}$ such that $g(\boldsymbol{X};\hat{\beta}) > 0$, or $g(\boldsymbol{X};\hat{\beta})$ is great than a clinically meaningful threshold. On the other hand, we can also identify the variables which are significant in the contrast function $g(\boldsymbol{X};\hat{\beta})$. These variables impact treatment heterogeneity and are often referred as effect modifiers.

2.2.2 Doubly robust method with variable selection

In clinical research and practice, it is important to know the markers or variables which impact the treatment heterogeneity (i.e., effect modifiers). In this section, we incorporate the variable selection in proposed doubly robust method by using the adaptive lasso method, which is an extension of traditional lasso proposed by Tibshirani (1996) with additional coefficient specific weights on the penalty term (Zou, 2006). Under certain conditions, Zou (2006) demonstrated that adaptive lasso estimators have oracle properties, which refer to consistent variable selection. We note

that the estimating equation (2.6) can be written in the following form:

$$
\begin{aligned}
&\sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i\{T_i - \pi^*(\boldsymbol{X}_i)\}\Big\{Y_i - h^*(\boldsymbol{X}_i) - T_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}\Big\} \\
=&\sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i\{T_i - \pi^*(\boldsymbol{X}_i)\}\Big\{Y_i - h^*(\boldsymbol{X}_i)\Big\} - \sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i\{T_i - \pi^*(\boldsymbol{X}_i)\}T_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta} \\
=&\sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i\{T_i - \pi^*(\boldsymbol{X}_i)\}\Big\{Y_i - h^*(\boldsymbol{X}_i)\Big\} \\
&\qquad - \sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i\sqrt{[T_i - \pi^*(\boldsymbol{X}_i)]T_i}\sqrt{[T_i - \pi^*(\boldsymbol{X}_i)]T_i}\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta} \\
\approx&\sum_{i=1}^{n} \widetilde{\boldsymbol{X}}_i\sqrt{[T_i - \pi^*(\boldsymbol{X}_i)]T_i + (1-T_i)\delta}\left(\frac{\{T_i - \pi^*(\boldsymbol{X}_i)\}\Big\{Y_i - h^*(\boldsymbol{X}_i)\Big\}}{\sqrt{[T_i - \pi^*(\boldsymbol{X}_i)]T_i + (1-T_i)\delta}}\right. \\
&\qquad \left. - \sqrt{[T_i - \pi^*(\boldsymbol{X}_i)]T_i + (1-T_i)\delta}\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}\right),
\end{aligned}
$$

which is the derivative of the objective function

$$
\sum_{i=1}^{n}\left(\frac{\{T_i - \pi^*(\boldsymbol{X}_i)\}\Big\{Y_i - h^*(\boldsymbol{X}_i)\Big\}}{\sqrt{[T_i - \pi^*(\boldsymbol{X}_i)]T_i + (1-T_i)\delta}} - \sqrt{[T_i - \pi^*(\boldsymbol{X}_i)]T_i + (1-T_i)\delta}\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2. \quad (2.8)
$$

Here δ is a small value, e.g., 10^{-5}.

The objective function (2.8) can be simplified as:

$$
\sum_{i=1}^{n}\left(Y_i^{New} - \boldsymbol{X}_i^{New\mathrm{T}}\boldsymbol{\beta}\right)^2, \quad (2.9)
$$

where $Y_i^{New} = \frac{\{T_i - \pi^*(\boldsymbol{X}_i)\}\Big\{Y_i - h^*(\boldsymbol{X}_i)\Big\}}{\sqrt{[T_i - \pi^*(\boldsymbol{X}_i)]T_i + (1-T_i)\delta}}$, and $\boldsymbol{X}_i^{New} = \sqrt{[T_i - \pi^*(\boldsymbol{X}_i)]T_i + (1-T_i)\delta}\widetilde{\boldsymbol{X}}_i$. Hence, we perform variable selection with adaptive lasso by adding the L1 penalty with specific weights for the objective function (2.9), and the adaptive lasso estimators are:

$$
\hat{\beta}_{(adLasso)} = \mathrm{argmin}_{\beta}\sum_{i=1}^{n}\left(Y_i^{New} - \boldsymbol{X}_i^{New\mathrm{T}}\boldsymbol{\beta}\right)^2 + \lambda\sum_{j=1}^{p}\hat{w}_i|\beta_j|,
$$

where $\hat{w}_i = \frac{1}{|\hat{\beta}_i(OLS)|}$, and $\hat{\beta}_i(OLS)$ is estimated via ordinary least squares regression fitted by Y^{New} and $\boldsymbol{X}^{New}$. The tuning parameter λ is selected by minimizing Bayesian Information Criterion (BIC). Note that the intercept term in the contrast function is not penalized.

2.3 Simulation studies

2.3.1 Simulation settings

As a proof-of-concept, we designed a simulation study to implement the proposed method. The design of the simulation study were to a large extent influenced by the simulation study conducted by Lu et al. (2013), which is for randomized controlled trials (RCTs). Our approach is applicable to both RCTs and observational studies.

Our aims in the simulation studies were to demonstrate that our proposed method was doubly robust in estimating the contrast function under a variety of underlying data generating conditions. The outcome was designed such that a higher value indicated a more desirable response. As in Lu et al. (2013), we used 10 covariates. We carried out simulation study for both experimental and observational data, and while Lu et al. (2013) considered only experimental data in their simulation studies.

In all scenarios, the dimension of covariates was set to 10 as in Lu et al. (2013) and $\boldsymbol{X} \sim N(\mathbf{0}_{10}, \boldsymbol{\Sigma})$, where $\mathbf{0}_d$ indicated the zero vector of length d and $\boldsymbol{\Sigma}$ had a AR(1) structure with correlation coefficient being 0.3. We considered the outcome model $Y = h(\boldsymbol{X}; \gamma) + Tg(\boldsymbol{X}; \beta) + \epsilon$, where the contrast function $g(X; \boldsymbol{\beta}) = \widetilde{X}^{\mathrm{T}}\boldsymbol{\beta}$ with $\boldsymbol{\beta} = (1, 1, \mathbf{0}_7^{\mathrm{T}}, -1, 1.2)^{\mathrm{T}}$, and $\epsilon \sim N(0, 1)$. The treatment assignment T either was a Bernoulli trial with probability 0.5 in RCTs or followed a conditional logit model $\mathrm{logit}(Pr[T = 1|\boldsymbol{X}]) = \mathrm{logit}(\pi^*(\boldsymbol{X})) = \tilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\phi}$, with $\boldsymbol{\phi} = (1, 2, -1, \mathbf{0}_6^{\mathrm{T}}, 1.5, -0.5)^{\mathrm{T}}$.

The $h()$ function controls the complexity of the relationship between the outcome Y and the covariates $\boldsymbol{X}$ under no treatment. We consider three scenarios of $h(\cdot)$ as follows:

(S1) Linear model: $h(\boldsymbol{X};) = 1 + \boldsymbol{X}_1^{\mathrm{T}}$, where $_1 = (1, -1, 0.5, \mathbf{0}_7^{\mathrm{T}})^{\mathrm{T}}$.

(S2) Quadratic model: $h(\boldsymbol{X};) = 1 + (\boldsymbol{X}_1^{\mathrm{T}})(\boldsymbol{X}_2^{\mathrm{T}})$, where $_1$ is the same as in Scenario S1 and $_2 = (0.4, 0.3, 0.8, 0.1, \mathbf{0}_6^{\mathrm{T}})^{\mathrm{T}}$.

(S3) Exponential model: $h(\boldsymbol{X};) = 1 + (\boldsymbol{X}_1^{\mathrm{T}})^2/4 + \exp(\boldsymbol{X}_2^{\mathrm{T}})/4$, where $_1$ and $_2$ are the same as in Scenario S2.

For each scenario, we consider two sample sizes $n = 500$ and 2000, and two types of treatment models, RCTs with probability 0.5 and the observational study with the aforementioned conditional logit model.

For each generated data set of sample size n, we used the proposed doubly robust method to estimate the contrast function, where the propensity score was estimated using the logistic regression and $h(\boldsymbol{X};\gamma)$ was estimated using the linear function and the ensemble method, respectively. These results were presented in Tables 2.1-2.3 under the respective columns "drLM" and "drEns" on the rows with row title "Est. PS". To compare, we also estimated the contrast function $g(\boldsymbol{X};\beta)$ using the constant propensity score, say $\pi(\boldsymbol{X};\phi) = 0.5$, which was the true propensity scores for RCTs but a mis-specified propensity score model for observational studies (see the row blocks "PS=0.5" in Tables 2.1-2.3). In addition, we applied the linear model of the form $E(Y) = \widetilde{\boldsymbol{X}}^{\mathrm{T}}\gamma + T\widetilde{\boldsymbol{X}}^{\mathrm{T}}\beta$ to estimate the contrast function, which was presented under the column "LM" in Tables 2.1-2.2. As a comparison, we also applied the 2-stage approach proposed by Lu et al. (2013) using Equation (2.5) to estimate $h_1(\boldsymbol{X};\gamma)$ (instead of $h(\boldsymbol{X};\gamma)$) first then estimate the contrast function. The results for estimating $h_1(\boldsymbol{X};\gamma)$ by using the linear function and using the ensemble method were presented in Tables 2.1-2.3 under the column "2sLM" and "2sEns" respectively.

For the proposed doubly robust approach, we also examined the performance when variable selection was applied (see Tables 2.1-2.3 with columns "drLMla", "drEnsla"). Under each simulation scenario and for each sample size, we generated 1000 simulated datasets. For each dataset, we applied the 7 methods to estimate the contrast function with either estimated propensity score or constant propensity scores. The performance of these methods were summarized by the prediction accuracy in estimating the parameters β in the contrast function and the contrast function itself. The prediction accuracy in estimating the parameters β were summarized by the mean squares for errors (MSE) (i.e.,$||\hat{\beta} - \beta||^2$) (see Table 2.1).The precision in estimating the contrast function was summarized by the percent correct decision (PCD, $\frac{1}{n}\sum_{i=1}^{n} I[sign(g(\boldsymbol{X}_i;\hat{\beta})) = sign(g(\boldsymbol{X}_i;\beta))]$) (see Table 2.2). We constructed Y such that larger values were more desirable; thus, $g(\boldsymbol{X}_i;\hat{\beta}) > 0$ implied that subject i should be prescribed the treatment, while $g(\boldsymbol{X}_i;\hat{\beta}) < 0$ implied that subject i should not be prescribed the treatment. PCD measured how well the predicted decision matched the known best decision. We also examined the performance for variable selection based on the proposed doubly robust approach. We summarized the number of correctly dropped covariates (Corr0), and the number of incorrectly dropped covariates (Incorr0) in Table 2.3.

2.3.2 Simulation results

The simulation scenarios varied in terms of sample size (n=500 and 2000), nature of the -functions (i.e., linear, quadratic, and exponential models), and experimental design (i.e., RCTs versus observational studies). The simulation results for different metrics were summarized in Tables 2.1-2.3. From Table 2.1 and based on the MSE under different models, we concluded that (i) the MSE resulted from doubly robust ensemble method (drEns) method had the smallest MSE than all other

methods in almost all scenarios; (ii) the proposed doubly robust ensemble method with variable selection using adaptive lasso (drEnsla) further reduced the MSE than drEns in all scenarios; (iii) drEns performed much better than all the other methods when treatment assignment dependent on covariates (i.e., observational study) while the propensity score was set as a constant 0.5, that was the case both propensity score model and the response model were mis-specified. It was clear that using the non-parametric ensemble method in the doubly robust approach can alleviate the impact due to the misspecification of the propensity score model. (iv) Linear model (LM) model itself or LM in the 2 stage approach (2sLM) or doubly robust approach (drLM) could result in larger MSE when $h(\boldsymbol{X};\gamma)$ was not linear. In summary, it was clearly that drEns was the winner among all the mentioned methods here.

Table 2.2 presented the predicted correct decision (PCD) for the contrast function. From Table 2.2, it is clear that (i) PCD was the highest for doubly robust ensemble method with variable selection method (drEnsla) for almost all scenarios, followed by doubly robust ensemble method (drEns); (ii) the largest PCD improvement for drEns versus 2sEns was from observational study and PS was set as 0.5: the improvement was from 85.9% to 91.5% (n=500) and 87.9% to 95.6% (n=2000) when $h(\boldsymbol{X};\gamma)$ was quadratic, and the improvement was from 88.9% to 90.0% (n=500) and 94.0% to 95.5% (n=2000) when $h(\boldsymbol{X};\gamma)$ was exponential function. Overall, the proposed drEns and drEnsla method led, a great majority of the time, to the correct decision regarding whether a particular subject should receive treatment or not.

Table 2.3 showed the number of the correctly dropped covariates among the 7 covariates with zero coefficients in the contrast function (Correct 0), and the number of the incorrectly dropped covariates among the 3 covariates with non-zero coefficients and intercept in the contrast function (Incorrect 0). When $h(\boldsymbol{X};\gamma)$ was linear function, both methods dropped almost all 7 irrelevant covariates (i.e., Correct0 $>$ 6.9) and picked up the relevant covariates (Incorrect0=0). On the other hand, the

number of incorrectly dropped covariates was 0 in almost all scenarios (Incorr0 in Table 2.3), based on our proposed doubly robust methods (i.e., drLMla, drEnsla), indicating that our proposed method selects the important variables very well.

Table 2.1: Mean squared errors for $\boldsymbol{\beta}$ based on 1000 simulated datasets under each scenario.

			LM	2sLM	2sEns	drLM	drEns	drLMla	drEnla
Sample size n=500									
RCT	Est.PS	Linear	0.106	0.111	0.118	0.109	0.110	0.040	0.040
		Quad	0.353	0.319	0.235	0.325	0.152	0.208	0.062
		Exp	0.189	0.184	0.248	0.184	0.154	0.088	0.063
	PS=.5	Linear	0.106	0.108	0.117	0.106	0.108	0.039	0.040
		Quad	0.353	0.342	0.246	0.353	0.153	0.243	0.064
		Exp	0.189	0.187	0.261	0.189	0.156	0.094	0.065
Obs.study	Est.PS	Linear	0.135	0.384	0.475	0.279	0.281	0.113	0.120
		Quad	2.031	0.724	0.461	0.770	0.386	0.768	0.236
		Exp	1.085	0.471	0.457	0.423	0.380	0.295	0.208
	PS=.5	Linear	0.135	0.356	1.183	0.135	0.147	0.066	0.078
		Quad	2.031	1.735	1.203	2.031	0.505	2.051	0.438
		Exp	1.085	0.872	1.381	1.085	0.743	1.102	0.682
Sample size n=2000									
RCT	Est.PS	Linear	0.026	0.026	0.027	0.026	0.026	0.009	0.009
		Quad	0.087	0.078	0.035	0.078	0.029	0.044	0.010
		Exp	0.044	0.042	0.041	0.042	0.029	0.018	0.010
	PS=.5	Linear	0.026	0.026	0.027	0.026	0.026	0.009	0.009
		Quad	0.087	0.087	0.036	0.087	0.029	0.053	0.010
		Exp	0.044	0.044	0.042	0.044	0.029	0.019	0.010
Obs.study	Est.PS	Linear	0.031	0.085	0.085	0.063	0.063	0.024	0.024
		Quad	1.687	0.169	0.091	0.184	0.072	0.253	0.029
		Exp	0.926	0.104	0.088	0.095	0.072	0.045	0.029
	PS=.5	Linear	0.031	0.241	0.828	0.031	0.032	0.014	0.015
		Quad	1.687	1.509	0.854	1.687	0.135	1.724	0.110
		Exp	0.926	0.727	0.987	0.926	0.172	0.915	0.155

Table 2.2: Percent correct decision (PCD) based on 1000 simulated datasets under each scenario.

			LM	2sLM	2sEns	drLM	drEns	drLMla	drEnla
Sample size n=500									
RCT	Est.PS	Linear	0.961	0.960	0.960	0.961	0.961	0.977	0.977
		Quad	0.930	0.932	0.949	0.933	0.953	0.948	0.971
		Exp	0.949	0.949	0.950	0.949	0.953	0.967	0.971
	PS=.5	Linear	0.961	0.961	0.960	0.961	0.961	0.977	0.977
		Quad	0.930	0.930	0.949	0.930	0.953	0.943	0.971
		Exp	0.949	0.948	0.950	0.949	0.953	0.965	0.971
Obs.study	Est.PS	Linear	0.956	0.928	0.923	0.937	0.937	0.962	0.961
		Quad	0.802	0.903	0.922	0.900	0.927	0.899	0.947
		Exp	0.882	0.920	0.922	0.924	0.928	0.942	0.950
	PS=.5	Linear	0.956	0.920	0.911	0.956	0.954	0.969	0.967
		Quad	0.802	0.790	0.859	0.802	0.915	0.801	0.923
		Exp	0.882	0.885	0.889	0.882	0.900	0.882	0.915
Sample size n=2000									
RCT	Est.PS	Linear	0.980	0.980	0.980	0.980	0.980	0.989	0.989
		Quad	0.965	0.966	0.978	0.967	0.979	0.976	0.988
		Exp	0.974	0.975	0.978	0.975	0.979	0.984	0.988
	PS=.5	Linear	0.980	0.980	0.980	0.980	0.980	0.989	0.989
		Quad	0.965	0.965	0.978	0.965	0.979	0.974	0.988
		Exp	0.974	0.974	0.978	0.974	0.979	0.984	0.988
Obs.study	Est.PS	Linear	0.979	0.966	0.967	0.970	0.970	0.982	0.982
		Quad	0.810	0.951	0.965	0.949	0.968	0.945	0.980
		Exp	0.890	0.961	0.966	0.963	0.968	0.976	0.980
	PS=.5	Linear	0.979	0.931	0.951	0.979	0.978	0.985	0.985
		Quad	0.810	0.800	0.879	0.810	0.956	0.809	0.961
		Exp	0.890	0.897	0.940	0.890	0.955	0.891	0.958

2.4 Case study

To demonstrate the use of our proposed method, we applied it to study the effect of antecedent statins use in a large database of hospitalized coronavirus disease 2019 (COVID-19) patients, which was established by the University of Louisville Center of Excellence for Research in Infectious Disease (CERID). Statins were a class of drugs that lower the level of cholesterol in the blood by reducing the production of cholesterol by the liver, and statins also had pleiotropic effects (non-lipid, often beneficial effects). Cholesterol was critical to the normal function of every cell in the body, and it also contributed to the development of atherosclerosis. The impact of statins

on COVID-19 severity and recovery was important given their high prevalence of use among individuals at risk for severe COVID-19 (Daniels et al., 2021). We used this COVID-19 dataset to study the impact of statins on the immune and inflammatory biomarker. The patients were formed into two groups according to their antecedent statins use: control group (i.e., no statin used, n= 815), and statin use group (n= 658). The target immune and inflammatory biomarker was a special type of white blood cells: lymphocyte percentage, that were involved in the fight against infection. Lymphocytes percentage has a normal range 20% to 40% of circulating white blood cells (McPherson and Pincus, 2021). Recent study has showed lymphopenia, which refers to a reduced level of lymphocyte, can indicate certain viral diseases, including COVID-19 (Tavakolpour et al., 2020), and a meta analysis demonstrated that lymphopenia on admission was related with poor outcome in COVID-19 patients (Huang and Pranata, 2020). Race, gender, age, BMI, and different comorbidity conditions (see Table 2.4) were possible confounding variables because they are associated with treatment choices as well outcome risks.

We obtained the estimated interaction parameters of covariates and statins using the proposed double robust method, and the results were presented in Table 2.5. The proposed drEnsla method identified several covariates which impacted the optimal treatment selection (see Table 2.5): gender, BMI, race, pulmonary comorbidity, neoplastic comorbidity, cerebrovascular accident (CVA) comorbidity, diabetes, and thrombosis. For example, a regression coefficient of -1.258 for gender (female as reference level) indicated that the statins lowered 1.258% of lymphocyte percentage for males more than females; a regression coefficient of 1.390 for BMI indicated that higher BMI led to higher statins effect; and the regression coefficients for the race (hispanic, black, other versus white) -3.087, -2.388, and -5.471, respectively, indicated that statins had larger treatment effect for the White COVID-19 patients.

It was also clear that statins had larger treatment effect on lymphocyte percentage in the COVID-19 patients with pulmonary, neoplastic, CVA, and diabetes than in the COVID-19 patients without these comorbidities (regression coefficients are 1.107, 4.765, 1.124, and 1.432, respectively), and statins had smaller treatment effect on lymphocyte percentage in patients with thrombosis than patients without thrombosis (regression coefficient: -3.854).

Based on the estimates of contrast function, we constructed the optimal treatment based on the patient characteristics, the optimal treatment for the patients with contrast function greater than 0 was the statins treatment, and the optimal treatment for the patients with contrast function less than or equal to 0 should be assigned to the control. We further presented the distribution of probability of obtaining Lymphocyte (%) larger than the given value for treatment group, the control group, and the optimal treatment, respectively (see Figure 2.1). From Figure 2.1, it is clear that the patients in the optimal group (green dot line) would have a higher probability of obtaining a normal lymphocyte percentage (20% to 40%), compared to the patients in either treatment group or control group. We also selected two pairs of patients with similar conditions as examples (Table 2.5). In each pair both patients had the similar characteristics thus have similar contrast functions, which was greater than 0 in the first pair, and less than 0 in the second pair. Based on the optimal treatment rule, the patients in the first pair should be recommended to be treated. Indeed, we saw that the patient from the first pair in the treated group (patient P1t in Table 2.5) had a normal lymphocyte percentage 26.7%, but patient P1c from the control group had a poor lymphocyte percentage 16.0%. The patients in the second pair shouldn't be recommended to be treated with statins. Indeed, we saw that the patient P2c from the control group had a better lymphocyte percentage 16.4% than the patient P2t's lymphocyte percentage 5.0% from the treated group. The observed matched pairs clearly showed the selected optimal treatment benefits the patients with similar

characteristics.

This result could reflect the impacts of the effect modifiers for statins use on innate immune function. The mechanism how these covariates modify the statin effect may be further investigated, which is beyond the scope of this work.

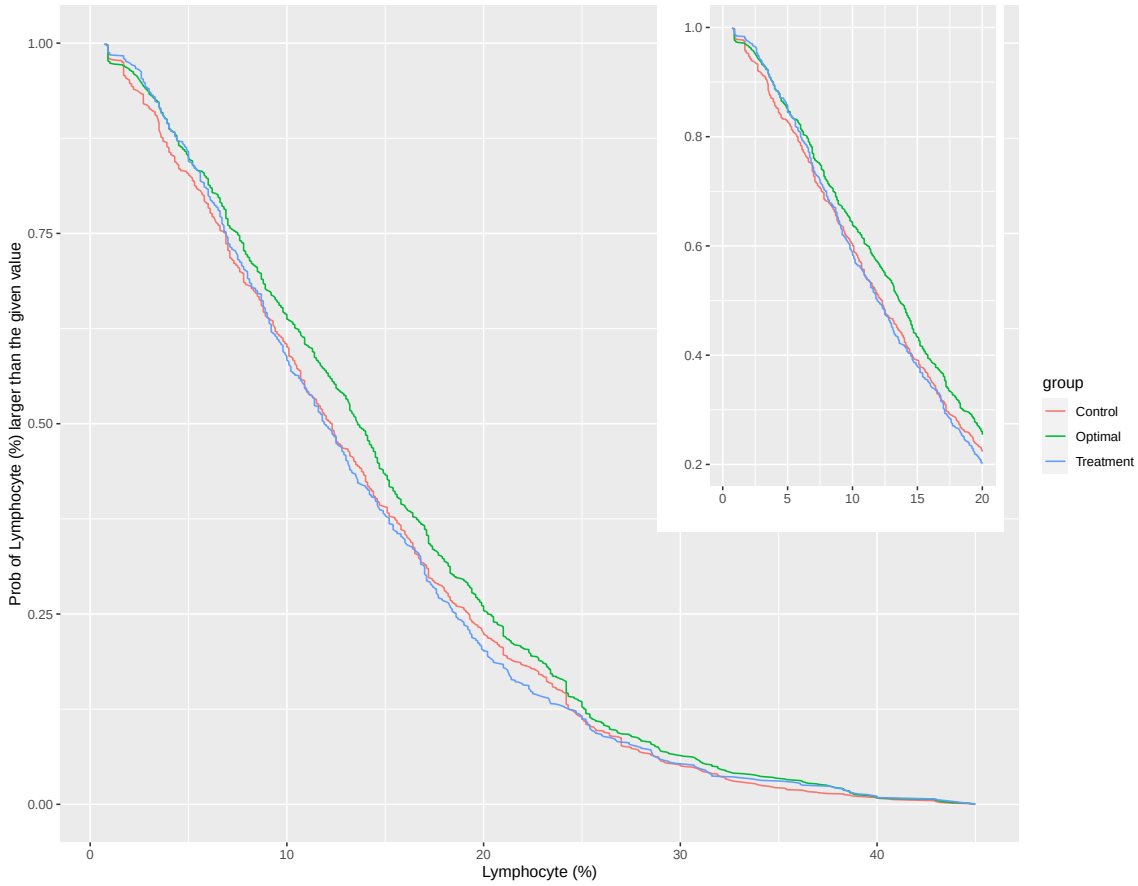

Figure 2.1: The probability of obtaining Lymphocyte (%) larger than the given value, where the blue, red, and green curves represent the treatment group; the control group; and the optimal group based on the proposed method, respectively.

2.5 Conclusion and discussion

This manuscript presents a method for identifying covariates that interact with a binary treatment to affect the outcome, hence characterizing subgroups of a study

population that have differing average treatment effects. The method is capable of handling a large number of covariates, due to the complexity control exerted by way of penalized regression. Also, the method is capable of handling both experimental data and observational data.

This work does have some limitations. So far, it applies only to point-in-time treatments. Extending this work to dynamic treatment regimens would be beneficial, particularly given that individualized treatment regimes are receiving heightened attention in personalized medicine. Also, the method should be studied for its effectiveness with binary or categorical outcomes, as the work to date has focused on continuous outcome measures. The characterization of the subgroups resulting from the selected covariates is not automatic, but a separate step following estimating the control group response profile. Nevertheless, this approach appears flexible and powerful approach for selecting the variables which impact the treatment heterogeneity, for deciding which patients would benefit more from the treatment.

Table 2.3: Variable selection results based on 1000 simulated datasets under each scenario. Correct0 # is the average of number of covariates correctly estimated as 0 in the contrast function (True value is 7); Incorrect0 # is the average number of covariates incorrectly estimated as 0 in the contrast function.

			Correct0 #		Incorrect0 #	
			drLMla	drEnsla	drLMla	drEnsla
Sample size n=500						
RCT	Est.PS	Linear	6.942	6.941	0	0
		Quad	6.210	6.872	0	0
		Exp	6.691	6.840	0	0
	PS=.5	Linear	6.940	6.940	0	0
		Quad	5.942	6.858	0.002	0
		Exp	6.641	6.821	0	0
Obs.study	Est.PS	Linear	6.976	6.969	0.013	0.017
		Quad	6.585	6.849	0.372	0.059
		Exp	6.834	6.855	0.095	0.046
	PS=.5	Linear	6.821	6.792	0	0.001
		Quad	4.838	6.079	0.493	0.042
		Exp	4.773	5.176	0.248	0.048
Sample size n=2000						
RCT	Est.PS	Linear	6.977	6.977	0	0
		Quad	6.409	6.979	0	0
		Exp	6.843	6.965	0	0
	PS=.5	Linear	6.973	6.973	0	0
		Quad	6.299	6.983	0	0
		Exp	6.840	6.968	0	0
Obs.study	Est.PS	Linear	6.991	6.992	0	0
		Quad	6.436	6.972	0.155	0
		Exp	6.799	6.955	0	0
	PS=.5	Linear	6.910	6.910	0	0
		Quad	4.633	6.254	0.171	0.002
		Exp	4.771	5.975	0.007	0.001

Table 2.4: Baseline characteristics of hospitalized COVID-19 patients, statins case study.

	Control group (N=815)	Statins use group (N=658)
Age (Mean±SD)	59 ± 17.2	68.3 ± 11.68
BMI	32 ± 9.7	32.4 ± 8.28
Race		
Hispanic	91 (11.17%)	27 (4.10%)
Non-Hispanic black	205 (25.15%)	186 (28.27%)
Non-Hispanic white	468 (57.42%)	425 (64.59%)
Non-Hispanic others	51 (6.26%)	20 (3.04%)
Gender		
Male	397 (48.71%)	345 (52.43%)
Female	418 (51.29%)	313 (47.57%)
Pulmonary count		
0	616 (75.58%)	457 (69.45%)
1	176 (21.60%)	181 (27.51%)
2	22 (2.70%)	20 (3.04%)
3	1 (0.12%)	0 (0.00%)
Cad count		
0	326 (40.00%)	44 (6.69%)
1	258 (31.66%)	125 (19.00%)
2	155 (19.02%)	257 (39.06%)
3	47 (5.77%)	123 (18.69%)
4-6	29 (3.56%)	109 (16.56%)
Liver count		
0	782 (95.95%)	629 (95.59%))
1-2	33 (4.05%)	29 (4.41%)
Neoplastic	61 (7.48%)	70 (10.64%)
CVA	58 (7.12%)	114 (17.33%)
Renal disease	111 (13.62%)	189 (28.72%)
Diabetes	214 (26.26%)	380 (57.75%)
CVD Thrombosis	41 (5.03%)	41 (6.23%)

Table 2.5: The estimated coefficients for the interaction between the covariates and statins use based on the hospitalized COVID-19 patients' cohort, where the response variable was lymphocyte percentage, and 2 pairs of patients were shown as examples.

	Est. of drEnsla	Patient Pair 1				Patient Pair 2			
	Coeff. (SE)	P1c: No statins use		P1t: statins use		P2c: No statins use		P2t: statins use	
Intercept	0.533	Ori. Value	Std. Value	Ori. Value	Std. Value	Ori. Value	Std. Value	Ori. Value	Std. Value
Age	0	92	1.809	88	1.558	67	0.241	66	0.178
Gender Male (Ref: Female)	-1.258	0	0	0	0	1	1	1	1
BMI	1.390	29.8	-0.255	30.4	-0.178	26.7	-0.592	27.4	-0.513
Race Hispanic (Ref: White)	-3.087	0	0	0	0	0	0	0	0
Race black (Ref: White)	-2.388	0	0	0	0	0	0	0	0
Race other (Ref: White)	-5.471	0	0	0	0	0	0	0	0
Pulmonary count	1.107	0	-0.572	1	1.347	0	-0.572	0	-0.572
Cad count	0	1	-0.414	4	1.846	5	2.599	3	1.092
Liver count	0	0	-0.205	0	-0.205	0	-0.205	0	-0.205
Neoplastic	4.765	0	0	0	0	0	0	0	0
CVA	1.124	0	0	0	0	0	0	0	0
Renal disease	0	1	1	1	1	0	0	0	0
Diabetes	1.432	1	1	1	1	0	0	0	0
Thrombosis	-3.854	0	0	0	0	1	1	1	1
Contrast Function $g(\boldsymbol{X})$			>0		>0		<0		<0
Lymphocyte percentage		16.0		26.7		16.4		5.0	

CHAPTER 3

STATISTICAL METHODS FOR ASSESSING DRUG INTERACTIONS USING OBSERVATIONAL DATA WITH BINARY TREATMENTS

3.1 Introduction

With advances in medicine, many drugs have become available to treat patients. Drug interactions could be concerns when multiple drugs are applied to the same patients. For example, several drugs are prescribed to a patient by health providers when the patient has multiple morbid conditions. Polydrug use is referred as using more than one drug or treatment during the same time period to treat different underlying morbid conditions (Vogt-Ferrier, 2011; Naveiro-Rilo et al., 2014; Tramontina et al., 2018). Polydrug use may cause adverse side effects, and has increasingly caused concerns (Tramontina et al., 2018). On the other hand, the combination of multiple treatments are developed to treat the same condition, such as cancer (Mokhtari et al., 2017). Combination therapy may be synergistic, thus more effective than each single drug (*i.e.*, monotherapy) (Lee et al., 2007). In developing combination therapy for a certain disease, drug interaction often refers to the situation where the effect of two drugs is more or less than the predicted additive effect. When the effect of combination of two drugs is more than their predicted additive effect, the two drugs are said to be synergistic (Kong and Lee, 2006; Lee et al., 2007). When the effect of combination of two drugs is less than their predicted additive effect, the two

drugs are said to be antagonistic (Kong and Lee, 2006; Lee et al., 2007). However, drug interaction can also refer to adverse effect when polydrug is used. When the outcome of interest is treatment efficacy, the synergistic treatment effect is preferred. In the case that the outcome is the adverse side effect, the synergistic effect would amplify the adverse side effect, and should be avoided.

Although the topic on assessing drug interaction is not new, many statistical models have been developed for assessing drug interaction using *in-vitro* cell culture data (Kong and Lee, 2006; Lee et al., 2007) and *in-vivo* human tumor xenograft data (Fang et al., 2004). These models have the form of $EY = h(d_1 + \rho d_2 + \tau d_1 d_2)$, where Y is the outcome variable, h is the dose-response curve for drug 1, d_1 and d_2 are respectively the dose level for drug 1 and drug 2, ρ captures the relative potency of drug 2 versus drug 1, and τ captures the interaction of two drugs. $\tau > 0$ indicates that the combination of two drugs is synergistic, and $\tau < 0$ indicates that the combination of the two drugs is antagonistic. These approaches on assessing drug interactions are consistent with isobologram and combination index (Kong and Lee, 2006; Lee et al., 2007; Zhao and Yang, 2014; Noguchi et al., 2019). However, these methods do not control confounding variables, and they are not applicable to observational data from clinical settings. Drug-drug interaction and drug-herb interaction in clinical studies have drawn much attention (Malone et al., 2004; Baxter and Preston, 2010; Mokhtari et al., 2017). However, much of these drug interactions in clinical settings were discovered through clinical experience. As multiple drug use increases, drug interaction has been recognized as an important problem recently, yet the rigorous statistical methods assessing drug interaction based on clinical observational data are lacking. The data depositories from routine clinical practice provide great opportunity to study the treatment effect of combination therapy or the adverse effect of polydrug use. Observational data from electronic health records or claims data are examples of such data resources which can be used to examine treatment efficacy due

to combination therapy or adverse outcomes due to polydrug use.

Recently several data mining algorithms were proposed to identify drug interaction quantitatively using clinical data (Tatonetti et al., 2012; VanderWeele and Knol, 2014; VanderWeele, 2015), a common used multivariate regression model has the form of $g(EY) = X^T\gamma + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2$, where g is a known link function, and X is a vector of covariates. However, this type of model adjusts the confounding variables in regression, and assumes that the treatment effects given X are the same for different values of covariates. The parameters $(\tau_1, \tau_2, \tau_{12})$ capture the average treatment effects (ATEs) and drug interactions only under certain link function g. The generalized marginal structural models (MSMs) directly model the relationship between the treatment (or combination treatment) and its potential outcome. The generalized MSMs can be written as $g(EY^{(d_1,d_2)}) = \tau_0 + \tau_1 d_1 + \tau_2 d_2 + \tau_{12} d_1 d_2$, where $Y^{(d_1,d_2)}$ describes the potential outcome if a patient had received treatment (d_1, d_2) (Rubin, 2005). Note that not everyone receives the treatment (d_1, d_2). Under certain assumptions, the inverse probability of treatment weighting (IPTW) method weights each subject with the inverse of probability of the treatment received, and the resulting weighted sample for patients receiving treatment (d_1, d_2) has a similar distribution as the entire study cohort. That is, in the weighted sample, the distribution of each covariate across different treatment groups are similar, the parameters in the MSMs thus are obtained from the weighted sample and have causal interpretation (Robins et al., 2000; Cole and Hernán, 2008; Hernan and Robins, 2020). The IPTW are often obtained from generalized propensity score (GPS) models (Imbens, 2000; Yan et al., 2019). Literature has indicated that the variances of the causal parameters from IPTW method are optimal if propensity score models include confounding variables (i.e., variables associated with both outcome and treatment selection) and predictor variables (i.e., variables associated with outcome but not treatment selection) (Brookhart et al., 2006; Craycroft et al., 2020). In Section 3.2, we propose a gener-

alized MSM to assess ATEs and drug interactions, where the confounding variables and predictor variables are obtained from fitting a multivariate regression model using the elastic net variable selection method (Zou and Hastie, 2005b), and the selected variables are included in the propensity score estimation. The proposed statistical methods in this article can be used to investigate drug interaction on treatment efficacy as well as drug interaction on adverse event, which depends on the outcome of interest. In Section 3.3, extensive simulation studies are conducted to examine the performance of the proposed method. In Section 3.4, one case study is conducted to examine the joint effect of metformin and glyburide in reducing hospital readmission in type 2 diabetic patients, and another case study is conducted to examine the joint effect of statins and opioids on the immune and inflammatory biomarkers in hospitalized COVID-19 patients. Section 3.5 is devoted to conclusions and discussions.

The major contributions of this article include: (1) we establish a connection between the MSMs (suitable for clinical data) and the drug interaction model originated for pre-clinical data; (2) we investigate the impact of variable selection on estimating ATEs and drug interactions; (3) we compare the performance of the proposed methods with traditional multivariate regression models in estimating ATEs and drug interactions; and (4) we illustrate the application of the proposed methods in estimating ATEs and drug interactions based on clinical data sets.

3.2 Statistical method to assess drug interactions

3.2.1 The proposed models

Let us consider evaluating drug interactions of two drugs (say drug 1 and drug 2). Let us assume that we have observed quadruplets (D_1, D_2, X, Y) for each subject. D_1 and D_2 denote the treatment received, and (D_1, D_2) takes values in $\mathcal{D}$, where

$\mathcal{D} = \{(0, 0), (1, 0), (0, 1), \text{and } (1, 1)\}$. Specifically,

$$(D_1, D_2) = \begin{cases} (0,0), & \text{if the patient neither receives drug 1 nor drug 2;} \\ (1,0), & \text{if the patient receives drug 1 but not drug 2;} \\ (0,1), & \text{if the patient receives drug 2 but not drug 1;} \\ (1,1), & \text{if the patient receives both drug 1 and drug 2.} \end{cases}$$

Y denotes the outcome variables from an exponential family, which include Gaussian distribution for continuous outcome and binomial distribution for binary outcome. X denotes the confounding variables which impact both treatment choice and outcome variable. One example for confounding variable is comorbid conditions, which impact the patient's treatment choices as well as his or her health outcome. The outcome variable Y could be continuous (e.g., the reading value for a biomarker) or binary (e.g., death, hospital readmission). In the literature, the additive effect of drug 1 and drug 2 was obtained from two dose-response curves when drug 1 and drug 2 are applied alone (Kong and Lee, 2006; Lee et al., 2007). If the treatment effect of the combination treatment of drug 1 and drug 2 is more than the additive effect, the combination treatment is said to be synergistic; and alternatively, if the treatment effect of the combination treatment of drug 1 and drug 2 is less than the additive effect, the combination treatment is said to be antagonistic. The model in the form of $h(d_1 + \rho d_2 + \tau d_1 d_2)$ has been used to capture drug interaction in *in-vitro* study (Kong and Lee, 2006; Lee et al., 2007). The generalized MSM proposed in this section is on the line with the form of $E(Y) = h(d_1 + \rho d_2 + \tau d_1 d_2)$ but for potential outcome. Potential outcomes can help understand what the ATEs and drug interaction parameters represent (Rubin, 2005). In the case of two drugs, there are four potential outcomes (say, $Y^{(0,0)}, Y^{(1,0)}, Y^{(0,1)}$, and $Y^{(1,1)}$) for each patient. The potential outcome $Y^{(d_1,d_2)}$ for a patient would be the outcome when the patient had

received treatment combination (d_1, d_2), where d_1 and d_2 take values 0 or 1. In practice, only one potential outcome is observed, say Y, which is the potential outcome $Y^{(d_1,d_2)}$ corresponding to the treatment the patient received, say $(D_1, D_2)= (d_1, d_2)$. That is, $Y = \sum_{(d_1,d_2)\in\mathcal{D}} I_{\{(D_1,D_2)=(d_1,d_2)\}} Y^{(d_1,d_2)}$. Here $I_{\{(D_1,D_2)=(d_1,d_2)\}}$ is an indicator function and takes value 1 if $(D_1, D_2) = (d_1, d_2)$ and zero otherwise. We assume the outcome variable follows an exponential family distribution, and we propose the following generalized MSM to assess ATEs and drug interaction:

$$g\left(E\{Y^{(d_1,d_2)}\}\right) = \tau_0 + \tau_1 d_1 + \tau_2 d_2 + \tau_{12} d_1 d_2. \tag{3.1}$$

Here g is a known monotonic link function. When the outcome is continuous, we may take g as the identity link function, and model (3.1) becomes

$$E\{Y^{(d_1,d_2)}\} = \tau_0 + \tau_1 d_1 + \tau_2 d_2 + \tau_{12} d_1 d_2. \tag{3.2}$$

When the outcome is binary, we may take g as the logit link function, and model (3.1) becomes

$$logit\left(E\{Y^{(d_1,d_2)}\}\right) = \tau_0 + \tau_1 d_1 + \tau_2 d_2 + \tau_{12} d_1 d_2. \tag{3.3}$$

When outcome is the count data, the log-link function may be applied. The generalized MSM (3.1) and its special cases (3.2) and (3.3) are distinctive to the traditional marginal regression model $g\left(E\{Y|d_1, d_2\}\right) = \tilde{\tau}_0 + \tilde{\tau}_1 d_1 + \tilde{\tau}_2 d_2 + \tilde{\tau}_{12} d_1 d_2$ and the traditional multivariate regression model $g\left(E\{Y|X, d_1, d_2\}\right) = X^T\gamma + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2$ in that (i) the generalized MSMs are structural and they model the relationship between potential outcome and treatment (d_1, d_2) as if everyone had received the treatment (d_1, d_2), thus the parameters in the MSM have causal interpretation and capture the ATEs and drug interactions; (ii) the multivariate regression model could capture

the ATEs and drug interactions if the multivariate regression model has an identity link function; (iii) when the multivariate regression model has an logit link function, the multivariate model can not be used to capture the ATEs and drug interactions; (iv) the traditional marginal regression model does not capture the ATEs and drug interactions. Heuristic proofs for these statements are provided in the Appendix. The proposed generalized MSM (3.1) is quite similar to the response surface models for assessing drug interactions proposed by Kong and Lee (2006) and Lee et al. (2007). However, the models proposed in (Kong and Lee, 2006; Lee et al., 2007) are used to assess drug interactions based on *in-vitro* cell culture data, where there are no confounding variables involved. Here we extend these models to assess drug interactions based on clinical observational data, where confounding variables often exist and thus must be controlled for any valid inferences for causal relationship between combination treatment and outcome variable.

Let us first consider that the link function g as the identity link function with the resulting model (3.2). The ATE due to drug 1 in the absence of drug 2 is defined as $E\left(Y^{(1,0)}\right) - E\left(Y^{(0,0)}\right)$, which is captured by τ_1 based on model (3.2). The ATE due to drug 2 is defined as $E\left(Y^{(0,1)}\right) - E\left(Y^{(0,0)}\right)$, which is captured by τ_2. We can also see that the ATE due to drug 1 in the presence of drug 2 is $E\left(Y^{(1,1)}\right) - E\left(Y^{(0,1)}\right)$, which equals $\tau_1 + \tau_{12}$. That is, when $\tau_{12} \neq 0$, the ATE due to drug 1 varies depending on whether drug 2 is administered or not. When $\tau_{12} > 0$, the ATE due to drug 1 in the presence of drug 2 is larger than that in the absence of drug 2, indicating that drug 1 and drug 2 are synergistic. Vise versa, when $\tau_{12} < 0$, the ATE due to drug 1 in the presence of drug 2 is less than that in the absence of drug 2, indicating that drug 1 and drug 2 are antagonistic. This definition is consistent with the definition of drug interaction on the additive scale (VanderWeele, 2015), which states that the sum of ATE of each individual treatment is different from the ATE when the two

drugs are applied together. That is,

$$\big(E(Y^{(1,0)}) - E(Y^{(0,0)})\big) + \big(E(Y^{(0,1)}) - E(Y^{(0,0)})\big) \neq E(Y^{(1,1)}) - E(Y^{(0,0)}). \quad (3.4)$$

Note that $E(Y^{(1,0)}) - E(Y^{(0,0)}) = \tau_1$ and $E(Y^{(0,1)}) - E(Y^{(0,0)}) = \tau_2$, while $E(Y^{(1,1)}) - E(Y^{(0,0)}) = \tau_1 + \tau_2 + \tau_{12}$. Equation (3.4) above is equivalent to $\tau_{12} \neq 0$, thus the parameter τ_{12} captures drug interaction on the additive scale.

When the outcome is binary, the logistic MSM (3.3) is often used. We illustrate that the causal parameter τ_{12} captures drug interaction in the odds ratio scale (VanderWeele, 2015). That is, τ_{12} captures whether the product of causal odds ratios of each individual treatment is different from the odds ratio when two treatments are applied together. That is,

$$OR^{(1,0) \text{ vs } (0,0)} \times OR^{(0,1) \text{ vs } (0,0)} \neq OR^{(1,1) \text{ vs } (0,0)}. \quad (3.5)$$

Equation (3.5) is equivalent to $\tau_{12} \neq 0$, and τ_{12} captures drug interaction in the odds ratio scale (VanderWeele, 2015).

When risk ratio is the quantity of interest, the following MSM with log-link function could be applied:

$$log(Pr(Y^{(d_1,d_2)} = 1)) = \tau_0 + \tau_1 d_1 + \tau_2 d_2 + \tau_{12} d_1 d_2.$$

The interaction on the multiplication scale is defined as that the product of causal risk ratios of individual treatment is different from the causal risk ratio when two treatments are applied together (VanderWeele, 2015). That is, $\frac{Pr(Y^{(1,0)}=1)}{Pr(Y^{(0,0)}=1)} \times \frac{Pr(Y^{(0,1)}=1)}{Pr(Y^{(0,0)}=1)} \neq \frac{Pr(Y^{(1,1)}=1)}{Pr(Y^{(0,0)}=1)}$, which is equivalent to $\tau_{12} \neq 0$. Therefore, τ_{12} captures drug interaction in the multiplication scale (VanderWeele, 2015).

In summary, the parameter τ_{12} in the generalized MSM (3.1) captures drug

interaction. $\tau_{12} > 0$ implies that the effect of the combination treatments is more than expected, indicating a synergistic effect; $\tau_{12} < 0$ implies that the effect of the combination treatments is less than expected, indicating an antagonistic effect.

3.2.2 Estimation of the parameters in MSMs

Note that the aforementioned generalized MSMs are models for potential outcome $Y^{(d_1,d_2)}$ when the subject had received combination treatment (d_1, d_2). However the potential outcome $Y^{(d_1,d_2)}$ is not observed if the subject does not receive the combination treatment (d_1, d_2). Although the parameters $\tau = (\tau_0, \tau_1, \tau_2, \tau_{12})$ in the generalized MSM (3.1) have causal interpretation, an appropriate estimating method must control for confounding variables. Following the literature in the causal inference, we apply the IPTW method to estimate the causal parameters τ in the generalized MSMs (Robins et al., 2000; Cole and Hernán, 2008). The IPTW method essentially creates a weighted sample where the distributions of each confounding variable across different treatment groups are similar, thus removing the confounding effect between the treatment assignment and the outcome variable. The consistency of the estimates for τ provided by the MSMs with IPTW holds under the following assumptions (Rosenbaum and Rubin, 1983; Imbens, 2000; Gelman and Hill, 2006; Cole and Hernán, 2008; Greenland and Mansournia, 2015; Hernan and Robins, 2020): (i) Weak ignorability (*i.e.*, weak unconfoundness): $Y^{(d_1,d_2)} \perp (D_1, D_2)|X$ for each pair $(d_1, d_2) \in \mathcal{D}$. That is, given X, the potential outcome is independent of the treatment received; (ii) Positivity: $Pr(D_1 = d_1, D_2 = d_2|X) > 0$ for all $(d_1, d_2) \in \mathcal{D}$ and X; (iii) Consistency: $Y = \sum_{(d_1,d_2)\in\mathcal{D}} I_{\{(D_1,D_2)=(d_1,d_2)\}} Y^{(d_1,d_2)}$, that is, the observed outcome is the same as the potential outcome corresponding the treatment received; and (iv) Correctly specification of propensity score model. The first three assumptions are key in the causal inference literature. The fourth assumption is specifically required for the IPTW method (Robins et al., 2000; Kang and Schafer, 2007; Cole and Hernán, 2008). Vi-

olation of any one of the assumptions may result in biased estimates for ATEs and drug interactions when our proposed IPTW method is applied.

Let us denote $(d_{1i}, d_{2i}, x_i, y_i)$ as the observed quadruplets for i^{th} subject ($i = 1, 2, ..., n$). The IPTW weight for i^{th} subject is:

$$w_i = \frac{1}{Pr(D_1 = d_{1i}, D_2 = d_{2i}|X = x_i)}.$$

Thus, we form a weighted sample where i^{th} subject has w_i copies of $(d_{1i}, d_{2i}, x_i, y_i)$ instead of 1 copy. Let us denote the distribution of X in the weighted sample as f^*, and in the original sample as f. The weight for an observation with covariate x and treatment (d_1, d_2) is defined as $w(d_1, d_2|x) = \frac{1}{Pr(D_1=d_1, D_2=d_2|x)} \triangleq \frac{1}{f(d_1,d_2|x)}$. Then the conditional distribution of X given $(D_1, D_2) = (d_1, d_2)$ in the weighted sample is:

$$f^*(x|d_1, d_2) \triangleq w(d_1, d_2|x) f(x|d_1, d_2) = \frac{1}{f(d_1, d_2|x)} \frac{f(x, d_1, d_2)}{f(d_1, d_2)} = \frac{f(x)}{f(d_1, d_2)} \propto f(x).$$

Thus, in the weighted sample, the distribution of X in each treatment group is proportional to the marginal distribution of X in the population where the original sample comes from. That is, X is not associated with treatment assignment anymore in the weighted sample. In practice, the probability of treatment assignment, say $Pr(D_1 = d_1, D_2 = d_2|X = x)$, needs to be estimated. In the literature, parametric methods (e.g., multinomial regression (Imbens, 2000), covariate balance propensity score (CBPS) method (Imai and Ratkovic, 2014)) and non-parametric method (e.g., generalized boosting method (GBM) (McCaffrey et al., 2013)) have been proposed to estimate the GPSs (Yan et al., 2019). Recent studies indicate that a more efficient estimate for ATEs can be obtained by including only the confounding variables and predictor variables in the GPS estimation (Brookhart et al., 2006; Craycroft et al., 2020). Note that the confounding variables and predictor variables are associated with the outcome variable, we propose applying the LASSO method (i.e., using L_1

penalty) (Zou and Hastie, 2005b) to the multivariate regression model to select the variables which are associated with the outcome variable. We assume that the multivariate regression model includes treatment D_1, D_2, D_1D_2, and all covariates. We apply the LASSO method to the multivariate regression model to select the covariates associated with the outcome, and the selected covariates are included in the GPS model. GPSs are also called balance score (Rosenbaum and Rubin, 1983). A metric to evaluate whether a covariate is balanced is the standardized mean difference (SMD) among all groups. The SMD between two treatment groups, say between treatment (d_1, d_2) and (d_1', d_2'), is defined as:

$$SMD_{\{(d_1,d_2) \text{ vs } (d_1',d_2')\}} = \frac{\bar{X}_{(d_1,d_2)} - \bar{X}_{(d_1',d_2')}}{\sqrt{\left(S^2_{(d_1,d_2)} + S^2_{(d_1',d_2')}\right)/2}},$$

where $\bar{X}_{(d_1,d_2)}$ and $\bar{X}_{(d_1',d_2')}$ are, respectively, the sample means of the covariate in the (d_1, d_2) group and (d_1', d_2') group. $S^2_{(d_1,d_2)}$ is the standard deviation of the covariate based on the observations in the (d_1, d_2) group. The SMDs in the weighted sample are defined similarly but with $\bar{X}_{(d_1,d_2)}$ being the weighted sample mean in the (d_1, d_2) group. The summarized metric of covariate balance across all groups for a covariate is obtained by the mean of SMDs over all pairs of different treatment groups. A SMD greater than 0.1 is considered a sign of imbalance of the covariate (Zhang et al., 2019). SMD allows researchers to quantitatively compare balance in measured baseline covariates between two groups in the IPTW weighted sample (Austin and Stuart, 2015). However, SMD also can be used to assess balance of confounding variables in the original sample, which helps determine whether or not weighting is needed to correct for confounding variables. Further SMD can be used to assess the balance of confounding variables under different GPS models. The GPS model should be correctly specified and result in balanced confounding variables. In this article, we apply the multinomial logistic regression to estimate GPSs, which is easy to imple-

ment and fast to compute. However, in case that the confounding variables in the weighted sample are not balanced, more advanced models such as CBPS model (Imai and Ratkovic, 2014) and GBM (McCaffrey et al., 2013) can be applied to estimate the GPSs. The balance of the confounding variables in the weighted sample should be assessed and achieved.

To obtain the GPSs, say $Pr(D_1 = d_{1i}, D_2 = d_{2i}|X = x_i)$ for i^{th} subject ($i = 1, ..., n$), we apply the multinomial logistic regression model which has the following form:

$$log\frac{Pr(D_1 = d_1, D_2 = d_2|X = x)}{Pr(D_1 = 0, D_2 = 0|X = x)} = (1, x^T)\beta^{(d_1,d_2)} = \beta_0^{(d_1,d_2)} + x^T\beta_1^{(d_1,d_2)}.$$

Here $\beta^{(d_1,d_2)} = (\beta_0^{(d_1,d_2)}, \beta_1^{(d_1,d_2)T})^T$ for $(d_1, d_2) \in \mathcal{D}$, and $\beta^{(0,0)}$ takes a vector of zero values. The GPS can be obtained as:

$$Pr(D_1 = d_1, D_2 = d_2|X = x) = \frac{\exp\left(\beta_0^{(d_1,d_2)} + x^T\beta_1^{(d_1,d_2)}\right)}{\sum_{(d_1',d_2')\in\mathcal{D}} \exp\left(\beta_0^{(d_1',d_2')} + x^T\beta_1^{(d_1',d_2')}\right)}. \tag{3.6}$$

The parameters $\beta^{(d_1,d_2)}$ with $(d_1, d_2) \in \mathcal{D} - (0, 0)$ can be estimated from the maximum likelihood (ML) method using the observed treatment assignments (d_{1i}, d_{2i}) and confounding variables x_i $(i = 1, ..., n)$. Once the parameters $\beta^{(d_1,d_2)}$ are estimated, the weight for i^{th} subject is obtained as $\hat{w}_i = \frac{1}{\hat{Pr}(D_1=d_{1i},D_2=d_{2i}|X=x_i)}$, where the $\hat{Pr}(D_1 = d_{1i}, D_2 = d_{2i}|X = x_i)$ is obtained by replacing $\beta_0^{(d_1,d_2)}$ and $\beta_1^{(d_1,d_2)}$ by their ML estimates in equation (3.6).

Under the four assumptions (i.e., weak ignorability, positivity, consistency, and correct specification of GPS model), there is no confounding anymore in the weighted sample. The parameters $\tau = (\tau_0, \tau_1, \tau_2, \tau_{12})$ in the generalized MSM (3.1) can be

obtained by maximizing the weighted log-likelihood function. That is,

$$\hat{\tau} = (\hat{\tau}_0, \hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_{12})^T = \arg\max \sum_{i=1}^{n} \hat{w}_i l(\tau; d_{1i}, d_{2i}, y_i).$$

Here $l(\tau; d_{1i}, d_{2i}, y_i)$ is the log-likelihood function for i^{th} observation. When the outcome is continuous and the identity link function is used, the log-likelihood function for i^{th} observation is

$$\begin{aligned} l(\tau; d_{1i}, d_{2i}, y_i) &= log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \tau_0 - \tau_1 d_{1i} - \tau_2 d_{2i} - \tau_{12} d_{1i} d_{2i})^2}{2\sigma^2}}) \\ &= -\frac{(y_i - \tau_0 - \tau_1 d_{1i} - \tau_2 d_{2i} - \tau_{12} d_{1i} d_{2i})^2}{2\sigma^2} - \frac{1}{2} log(2\pi\sigma^2). \end{aligned}$$

When the outcome is binary and the logit link function is used, the log-likelihood function for i^{th} observation is

$$l(\tau; d_{1i}, d_{2i}, y_i) = y_i(\tau_0 + \tau_1 d_{1i} + \tau_2 d_{2i} + \tau_{12} d_{1i} d_{2i}) - log(1 + e^{\tau_0 + \tau_1 d_{1i} + \tau_2 d_{2i} + \tau_{12} d_{1i} d_{2i}}).$$

The similar work (Robins et al., 2000) on MSMs indicates that the weighted ML results in consistent estimator for the causal parameter for τ under the four assumptions for causal inference. The weighted ML estimate for τ can be obtained by using the R package *survey* (Lumley, 2004), where the weights are obtained by the GPS model which achieves the balance of confounding variables. Although a robust variance estimator for $\hat{\tau}$ can be obtained from the *survey* package, it does not incorporate the uncertainty in estimating the GPSs in the IPTW method. Instead, we use the bootstrap sampling techniques to estimate the variance of $\hat{\tau}$. That is, we obtain B (say, 100) bootstrap samples from the original sample. For the b^{th} bootstrap sample ($b = 1, \cdots, B$), we repeat the same estimating process as outlined to obtain an estimate $\hat{\tau}^{(b)}$ for τ. $\widehat{\text{Var}}(\hat{\tau})$, the estimate of the variance of $\hat{\tau}$, is obtained as the variance of the B bootstrap estimates $\hat{\tau}^{(b)}$ ($b = 1, \cdots, B$) (Mooney et al., 1993). In the following

section, extensive simulation studies were carried out to examine the performance of the proposed methods in estimating ATEs and drug interactions.

3.3 Simulation studies

In this section, we carried out extensive simulation studies to examine the performance of the proposed method in estimating ATEs and drug interactions using the generalized MSMs with IPTW method.

3.3.1 Design of simulation studies

In our simulation studies, we examined the performance of our proposed method under three different sample sizes (say n=500, 1000, and 5000) for continuous responses as well as binary responses (say Y). For continuous responses Y, we considered the following two regression models, the first one assumed homogeneous treatment effects (i.e., the conditional treatment effect given X was the same over different covariates X):

$$Y = \tilde{\mathbf{X}}^T\gamma + \tau_1^* D_1 + \tau_2^* D_2 + \tau_{12}^* D_1 D_2 + \epsilon; \tag{3.7}$$

and the second one assumed heterogeneous treatment effects (i.e., the conditional treatment effect given X varied over different covariates X):

$$Y = \tilde{\mathbf{X}}^T\gamma + \tau_1^* D_1 + \tau_2^* D_2 + \tau_{12}^* D_1 D_2 + \delta_1 X_1^2 D_1 + \delta_2 X_2^2 D_2 + \delta_3 X_3 D_1 D_2 + \epsilon. \tag{3.8}$$

Here we set $\tau_1^* = \tau_2^* = 1$, and τ_{12}^* was taking values from -1 to 1 by a step of 0.5, $\delta_1 = \delta_2 = \delta_3 = 1$. $\tilde{\mathbf{X}} = (1, \mathbf{X}^T)^T$, where $\mathbf{X}$ is a vector of p covariates ($p = 10$) with each covariate being independently normally distributed with mean zero and variance 1. ϵ was a random error with normal distribution of mean 0 and variance 0.5^2. γ was set as $(1, 2, -2, 2, -2, 2, \mathbf{0}_5)^T$ so that the outcome models only depended on the first 5 covariates. Here $\mathbf{0}_q$ represented a vector of zero with q components.

For binary responses Y, we considered the following two logistic regression models, the first one assumed homogeneous treatment effects:

$$logitPr(Y = 1|X, D_1, D_2) = \tilde{\mathbf{X}}^T\gamma + \tau_1^* D_1 + \tau_2^* D_2 + \tau_{12}^* D_1 D_2, \tag{3.9}$$

and the second one assumed heterogeneous treatment effects:

$$\begin{aligned} logitPr(Y = 1|X, D_1, D_2) =& \tilde{\mathbf{X}}^T\gamma + \tau_1^* D_1 + \tau_2^* D_2 + \tau_{12}^* D_1 D_2 \\ &+ \delta_1 X_1^2 D_1 + \delta_2 X_2^2 D_2 + \delta_3 X_3 D_1 D_2. \end{aligned} \tag{3.10}$$

Here $\tau_1^* = \tau_2^* = 2$, and τ_{12}^* was taking values from -1 to 1 by a step of 0.5, $\delta_1 = \delta_2 = \delta_3 = 1$, $\tilde{\mathbf{X}} = (1, \mathbf{X}^T)^T$ were set the same as those for the continuous outcomes, and γ was set as $(0.1, 1, 1, 1, 1, -1, \mathbf{0}_5)^T$.

Given $\mathbf{X}$, the treatment assignment (say, (D_1, D_2)) was generated using the multinomial distribution with the following probabilities:

$$Pr((D_1, D_2) = (d_1, d_2)) = \frac{\exp\left(\beta_0^{(d_1,d_2)} + \mathbf{X}^T\beta_1^{(d_1,d_2)}\right)}{\sum_{(d_1',d_2')\in\mathcal{D}} \exp\left(\beta_0^{(d_1',d_2')} + \mathbf{X}^T\beta_1^{(d_1',d_2')}\right)}, \tag{3.11}$$

where $\beta_0^{(0,0)} = \beta_0^{(1,0)} = \beta_0^{(0,1)} = \beta_0^{(1,1)} = 0$, $\beta_1^{(0,0)} = \mathbf{0}_{10}^T$, $\beta_1^{(1,0)} = 0.2\times(1, -1, -1, \mathbf{0}_2, 0.5, -1, \mathbf{0}_3)^T$, $\beta_1^{(0,1)} = 0.2\times(0.5, -1.5, 0.5, \mathbf{0}_2, 1.5, -1, \mathbf{0}_3)^T$, and $\beta_1^{(1,1)} = 0.2\times(1, -1, 0.5, \mathbf{0}_2, 1, 0.5, \mathbf{0}_3)^T$. Thus, in our simulation setting, the first three covariates were confounding variables, which were related to both the treatment assignment and the outcome; X_4 and X_5 were predictor variables, which were only related to the outcome but not the treatment assignment; X_6 and X_7 were instrumental variables, which were only related to the treatment assignment; and all the other three variables X_8, X_9 and X_{10} were spurious variables which were neither related to the treatment assignment nor to the outcome. Based on the generated covariates and treatments, we generated the con-

tinuous responses using models (3.7) and (3.8), respectively. We also generated the binary responses using models (3.9) and (3.10), respectively.

For each outcome model (four models in total), each τ_{12}^* (five specifications), and each sample size (n=500, 1000, and 5000), we generated 1000 simulated data sets. The data generating and estimating procedures were carried out in the following steps.

Step 1: A data set was generated by (i) first generating n observations for the p covariates, say $\mathbf{X}_i$ $(i = 1, \ldots, n)$, where $\mathbf{X}_i \sim MVN(0, I^2)$; (ii) generating treatment assignment (D_{i1}, D_{i2}) based on a multinomial distribution with probability of treatment assignment model (3.11) and the covariate $\mathbf{X}_i$ $(i = 1, \ldots, n)$; (iii) generating the outcome Y_i based on the outcome model along with the covariate $\mathbf{X}_i$ generated in (i) and treatment assignment (D_{i1}, D_{i2}) $(i = 1, \ldots, n)$ generated in (ii).

Step 2: For each simulated data set, the four potential outcomes $\{Y_i^{(0,0)}, Y_i^{(1,0)}, Y_i^{(0,1)}, Y_i^{(1,1)}\}$ corresponding to the covariate $\mathbf{X}_i$ were generated using the outcome model $(i = 1, \ldots, n)$. For continuous outcome, the true sample ATEs for drug 1 used alone versus control, for drug 2 used alone versus control and, and drug interaction were, respectively, obtained as

$$ATE_{(1,0) \text{ vs. } (0,0)} = \tfrac{1}{n} \sum_{i=1}^{n} (Y_i^{(1,0)} - Y_i^{(0,0)}),$$
$$ATE_{(0,1) \text{ vs. } (0,0)} = \tfrac{1}{n} \sum_{i=1}^{n} (Y_i^{(0,1)} - Y_i^{(0,0)}), \text{ and}$$
$$ATE_{(1,1) \text{ vs. } (0,1)} - ATE_{(1,0) \text{ vs. } (0,0)} = \tfrac{1}{n} \sum_{i=1}^{n} \big(Y_i^{(1,1)} - Y_i^{(0,1)} - (Y_i^{(1,0)} - Y_i^{(0,0)})\big).$$

For binary outcome, the true sample ATEs and drug interaction in terms of odds ratio were obtained as

$$ATE_{(1,0) \text{ vs. } (0,0)} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^{(1,0)}}{1 - \frac{1}{n}\sum_{i=1}^{n} Y_i^{(1,0)}} \Big/ \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^{(0,0)}}{1 - \frac{1}{n}\sum_{i=1}^{n} Y_i^{(0,0)}},$$
$$ATE_{(0,1) \text{ vs. } (0,0)} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^{(0,1)}}{1 - \frac{1}{n}\sum_{i=1}^{n} Y_i^{(0,1)}} \Big/ \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^{(0,0)}}{1 - \frac{1}{n}\sum_{i=1}^{n} Y_i^{(0,0)}}, \text{ and}$$

$$\frac{ATE_{(1,1)\text{ vs. }(0,1)}}{ATE_{(1,0)\text{ vs. }(0,0)}} = \left(\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^{(1,1)}}{1-\frac{1}{n}\sum_{i=1}^{n} Y_i^{(1,1)}} \Big/ \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^{(0,1)}}{1-\frac{1}{n}\sum_{i=1}^{n} Y_i^{(0,1)}} \right) \Big/ \left(\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^{(1,0)}}{1-\frac{1}{n}\sum_{i=1}^{n} Y_i^{(1,0)}} \Big/ \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i^{(0,0)}}{1-\frac{1}{n}\sum_{i=1}^{n} Y_i^{(0,0)}} \right).$$

Step 3: For each dataset, we estimated the parameters $\tau = (\tau_0, \tau_1, \tau_2, \tau_{12})$ in the generalized MSM without using the IPTW and with using the IPTW, and we also estimated τ using multivariate regression model as specified in model (3.7) for continuous outcome and model (3.9) for binary outcome. In the IPTW approach, the following five sets of covariates were used to estimate GPSs and weights: (i) all available covariates, i.e. $X_1 \sim X_{10}$; (ii) true confounding variables only, i.e. $X_1 \sim X_3$; (iii) confounding variables and instrumental variables, i.e. $X_1 \sim X_3$ and $X_6 \sim X_7$; (iv) confounding variables and predictors, i.e. $X_1 \sim X_5$; and (v) the set of covariates selected based the LASSO method.

Step 4: Bootstrap resampling method was used to estimate the variance for each estimated causal parameter, say τ_1, τ_2, and τ_{12}.

Step 5: Steps 1-4 were repeated for 1000 times. The true ATEs and drug interaction were obtained as the mean of the 1000 true sample ATEs and drug interactions calculated in Step 2. For continuous outcome, the true τ_1, τ_2 and τ_{12} were obtained,respectively, as the mean of 1000 true sample $ATE_{(1,0)\text{ vs. }(0,0)}$, $ATE_{(0,1)\text{ vs. }(0,0)}$, and $ATE_{(1,1)\text{ vs. }(0,1)} - ATE_{(1,0)\text{ vs. }(0,0)}$. For binary outcome, the true τ_1, τ_2 and τ_{12} were obtained, respectively, as the mean of 1000 $log(ATE_{(1,0)\text{ vs. }(0,0)})$, $log(ATE_{(0,1)\text{ vs. }(0,0)})$, and $log(ATE_{(1,1)\text{ vs. }(0,1)} - ATE_{(1,0)\text{ vs. }(0,0)})$.

Step 6: The mean of the 1000 estimates (Est), mean squared errors (MSE), the mean of 1000 standard errors (SE) based on the bootstrap method in Step 4, the empirical standard deviation for the 1000 estimates (E.SD) for each causal parameter, and the true coverage rate (CR) of the 95% confidence intervals for the true causal parameters $(\tau_1, \tau_2, \tau_{12})$ from Step 5 were obtained.

Step 7: Repeat Steps 1-6 for each of the 60 combined settings (four outcome models,

five specifications for τ_{12}^*, and three sample sizes.)

The boxplots of the quantities based on the 1000 simulated data with heterogeneous treatment effects for the three sample sizes and different τ_{12}^* were shown in Figure 3.1 for continuous outcomes and Figure 3.2 for binary outcomes, and the boxplots of these quantities based on the 1000 simulated data sets with homogeneous treatment effects were shown in Figure A1.1 for continuous outcomes and Figure A1.2 for binary outcomes in the Appendix. The summarized metrics, including Est, MSE, SE, E.SD, and the CR for the true causal parameter were presented in Table 3.1 for continuous outcomes and Table 3.2 for binary outcomes for $\tau_{12}^* = 0$ and the sample size $n = 1000$. The summarized simulation results with sample size $n = 500$ and $n = 5000$ were reported in Tables A1.1 $\sim$ A1.2 in the Appendix for continuous outcomes, and in Tables A1.3 $\sim$ A1.4 in the Appendix for binary outcomes.

We also assessed the balance of the covariates in our simulation study. It has been well known that the propensity scores are used to balance covariates among different treatment groups (Rosenbaum and Rubin, 1983). For each simulated data set, we calculated the propensity scores by using the multinomial regression models and different sets of covariates. We calculated the average of SMDs for each covariate to assess the balance of covariate among the four treatment groups. Figure A1.3 in the Appendix showed the boxplots of the average of SMDs among all pairs of the four treatment groups for each covariate without and with IPTW for the 1000 simulated data sets with sample size $n = 5000$, where we presented three sets of covariates in the GPS models: (i) all covariates; (ii) confounding variables and predictors; and (iii) the selected variables via LASSO. A dashed horizontal line at the height 0.1 indicated the threshold on whether the covariate was balanced or not. From Figure A1.3, we can see that (i) the confounding variables (i.e., $X_1 \sim X_3$) and instrumental variables (i.e., X_6 and X_7) were not balanced in the original samples, and the predictors (i.e., X_4 and X_5) and the spurious variables (i.e., $X_8 \sim X_{10}$) were less unbalanced in the original sample;

(ii) the GPS model included all covariates balanced all covariates; (iii) the GPS model included confounding variables and predictors only balanced these variables; (iv) the GPS model included the selected variables via LASSO balanced confounding variables and predictors. All GPS models resulted in balanced confounding variables but not necessary instrumental variables. To estimate ATEs and drug interaction appropriately, the confounding variables were important to be controlled. It was clear that different GPS models balanced all the confounding variables across the four groups.

3.3.2 Simulation results

Based on the simulation results in Figures 3.1 $\sim$ 3.2 and Tables 3.1 $\sim$ 3.2 in the current article and Figures A1.1 $\sim$ A1.2 and Tables A1.1 $\sim$ A1.4 in the Appendix, we concluded that (i) the estimated ATEs based on the proposed method for drug 1 were close to the true ATEs (see the first row from the third to the seventh boxplots versus the first boxplot for each τ_{12}^* in Figures 3.1 $\sim$ 3.2, S1 $\sim$ S2), while the estimated ATEs without IPTW (see the first row the second boxplot for each τ_{12}^*) were far from the true ATEs; (ii) the estimated ATEs based on the proposed method for drug 2 were also close to the true ATEs for drug 2 (see the second row from the third to the seventh boxplot versus the first boxplot for each τ_{12}^* in Figures 3.1 $\sim$ 3.2, S1 $\sim$ S2), while the estimated ATEs without IPTW for drug 2 were far from the true ATEs (see the second row the second boxplot versus the first boxplot for each τ_{12}^*; (iii) the estimated drug interactions (say $\hat{\tau}_{12}$) based on the MSMs with IPTW were also close to the true drug interactions, while the estimated drug interactions based on MSM without IPTW were far from the true drug interactions (see the third

row in Figures 3.1 $\sim$ 3.2, S1 $\sim$ S2); (iv) as the sample size increased (n=500, 1000, 5000 in the first, second, and third column, respectively, in Figures 3.1 $\sim$ 3.2, S1 $\sim$ S2), the estimated ATEs and drug interactions based on the proposed methods became more accurate and variations became smaller; (v) although the IPTW-based estimates for ATEs and drug interactions using different sets of covariates were close to the true values, the variations of the estimated ATEs and drug interactions with GPS including confounding variables and predictors were the smallest one, followed by the GPS model with covariates selected by LASSO (see MSE and E.SD in Tables 3.1 $\sim$ 3.2, A1.1 $\sim$ A1.4 in the Appendix); (vi) the SEs based on the bootstrap method were close to the E.SD for each causal parameter, and the true CRs were close to the nominal rate of 0.95 (Tables 3.1 $\sim$ 3.2, A1.1 $\sim$ A1.4), indicating that the bootstrap method performed well in estimating the variances of the causal parameters; (vii) the multivariate regression models could provide unbiased estimates for ATEs and drug interactions when the underlying link function was the identity link function, and the multivariate regression models may not provide unbiased estimates for ATEs and drug interactions when the underlying link function was the logit link function (Tables 3.1 $\sim$ 3.2, A1.1 $\sim$ A1.4). We concluded that the proposed method performed well in estimating ATEs and drug interactions in these simulation studies.

We would like to clarify that the first boxplot in each block in Figures 3.1 $\sim$ 3.2, A1.1 $\sim$ A1.2 was the boxplot of the true sample ATEs and drug interactions based on the potential outcomes obtained from the underlying outcome models in Step 2 in the simulation algorithm. The four potential outcomes for each patient in a simulated data set were obtained from the underlying models with the generated covariates and four different treatment combinations. The true ATEs and drug interaction reported in Tables 3.1 $\sim$ 3.2 in this article and Tables A1.1 $\sim$ A1.4 in the Appendix were the mean of the 1000 true sample ATEs and drug interactions. The true sample ATEs and drug interactions were not estimates from a sample but were obtained with a

known underlying outcome model with plugging in the generated covariates and each one of the four treatment combinations. While the other quantities were obtained from certain estimation procedures as outlined in the simulation study. This may explain why the variations for the true sample ATEs were smaller than the other quantities in Figures 3.1 ∼3.2, A1.1 ∼ A1.2 and Tables 3.1 ∼ 3.2, A1.1 ∼ A1.4.

3.4 Case studies

3.4.1 Case study 1: glyburide/metformin on readmission rates for diabetes

Recent evidence suggests that the management of hyperglycemia in the hospitalized patient has critical impact on the clinical outcomes for both morbidity and mortality (Strack et al., 2014). Literature has shown that the combination of glyburide and metformin significantly reduced fasting plasma glucose and 2 hour postprandial glucose values compared with either monotherapy (Garber et al., 2003). This was a randomized, three arm parallel group, double blinded trial, of 486 participants, with inadequate glycemic control that had attempted first line diet and exercise alone. The use of first-line combination treatment of glyburide and metformin versus glyburide or metformin monotherapy was studied. The combination treatment provided superior glycemic control over component monotherapy, allowing more patients in this study, to achieve American Diabetes Association (ADA) treatment goals with lower component doses in drug-naive patients with type 2 diabetes (Garber et al., 2003). In our case study, we studied the joint effect of metformin and glyburide on readmission rates in individuals admitted to a hospital obtained from electronic health records data. The data set consisted of patients who were hospitalized with diabetic encounter and had recorded diabetes medications. The data set was created from the health fact database which included patients' comprehensive clinical records from 130 US hospitals for years 1999-2008, and the data set was available

in the University of California Irvine (UCI) machine learning repository database. The details of the data set can be obtained from Strack et al. (2014). We obtained data for a total of 1233 patients who had type 2 diabetes with HbA1c test result greater than 8% and did not have any other diabetes medications other than glyburide and metformin. Per American Diabetic Association guidelines, HbA1c levels 7% are the recommended target, thus the patients in our case study would be considered patients with inadequate glucose control, unless they are older frail adults. Among the patients, 738 patients did not use any diabetes medication (say, control group), 150 patients used glyburide only, 198 patients used metformin only, and 147 patients used both glyburide and metformin (see Table 3.3A). The readmission rate without adjusting any confounding variables were 43.8%, 40.7%, 30.8%, and 32.7%, respectively, for control group, glyburide only group, metformin only group, and the combination of glyburide and metformin group (Table 3.3A). The covariates include race, gender, age, weight, the comorbidity conditions prior encounter hospital admission such as circulatory diseases, diabetes, digestive diseases, genitourinary diseases, musculoskeletal disease, respiratory diseases, injury, neoplasms, and other diseases. The SMD for the original samples were larger than 0.1, indicating unbalance of the covariates in the original sample (Figure A1.4 in the Appendix). Thus, the IPTW method, which adjust confounding variables, should be applied to estimate the ATEs and drug interaction.

We applied LASSO method to the outcome model, which included all the covariates, treatment indicator variables and the interaction of the two treatments. The selected covariates were race, gender, age, and the comorbidity conditions, which were included in the GPS model. We first applied the multinomial regression model to estimate the GPSs, however the balance of the selected covariates among different treatment groups were not achieved. Instead, we applied the CBPS method (Imai and Ratkovic, 2014) to estimate the GPS and balance the selected variables. The

balances of covariates in terms of SMD for this case study on original observed sample and the IPTW weighted sample were presented in Figure A1.4 in the Appendix. It clearly showed that the SMD for the covariates across the four different groups were small in the weighted sample, indicating balance of the covariates and comorbid conditions. The histograms of the GPSs for the four groups (see Figure A1.5 in the Appendix) indicated that there were not zero or one probabilities, which implied that the positivity assumption was not violated. The MSM with logit-link function was used to assess treatment effect and drug interaction. The estimated readmission rates in the weighted samples were 44.0%, 32.2%, 33.2%, and 31.8%, respectively, for control group, glyburide only group, metformin only group, and the combination of glyburide and metformin group (see the column under "Readmission with IPTW (LASSO)" in Table 3.3A). The resulting parameters for MSM without IPTW and with IPTW were reported in Table 1B. Due to the presence of confounding variables, the inference for ATEs and drug interaction should be made based on the estimates applying MSM with IPTW (LASSO). Based on Table 3.3B, both glyburide and metformin used alone significantly reduced the readmission rate, with OR as $e^{-0.503} = 0.604$ (p=0.019) for glyburide used alone, and $e^{-0.458} = 0.633$ (p=0.015) for metformin used alone. Their combination also significantly reduced the readmission rate with an OR=$e^{-0.524} = 0.592$ (p=0.012). However, the interaction parameter was not significantly different from zero (p=0.193), indicating that the two drugs did not have interacting effect on hospital readmission based on this case study.

The model proposed can potentially detect a causal relationship between the medications and outcome of re-hospitalization, if confounding variables were controlled. The covariates included in this analysis sought to address variables that make a patient more vulnerable to hospitalization when added glucose lowering treatments are assigned, such as age and disease burden. This approach helps diminish the likelihood of including re-hospitalization for hypoglycemia which is an problematic reason

for re-hospitalization in older and frail patients (e.g. older age and higher diagnosis burden). In this example, the medications available in the data base excluded medications other than diabetes medications. Clinically, there are potentially other confounding variables, such as medications that increase glucose levels and impair diabetic control, such as prednisone, which could probably explain the inability to detect a significant interaction effect of metformin and glyburide on re-hospitalization. Another confounder and limitation to the data set is that how the patient actually takes a medication and manages diabetes at home is very difficult to account for. Skipping doses, in addition to poor eating choices or lack of exercise, can negate the impact of medications in any individual. Lastly, although re-hospitalization is the ultimate outcome that most health systems seek to avoid, the breadth of reasons a patient could be re-hospitalized for is vast. In this analysis, all re-hospitalizations were considered the outcome of interest. Re-hospitalizations due to poor glucose control such as urinary frequency, dizziness, fainting, blurry vision and others, were not identifiable.

3.4.2 Case study 2: Effect of antecedent statins and opioids use on inflammatory biomarkers in hospitalized COVID-19 patients

Statins are a class of drugs that lower the level of cholesterol in the blood by reducing the production of cholesterol by the liver, and statins also have pleiotropic effects (non-lipid, often beneficial effects). Cholesterol is critical to the normal function of every cell in the body, and it also contributes to the development of atherosclerosis. The impact of statins on coronavirus disease 2019 (COVID-19) severity and recovery is important given their high prevalence of use among individuals at risk for severe COVID-19 (Daniels et al., 2021). Statins have a pleiotropic effect in addition to cholesterol lowering mechanisms. This pleiotropic effect is independent of statin effects on cholesterol. Statins inhibit production of proinflammatory cytokines, reactive oxygen

species and diminish platelet reactivity (Oesterle et al., 2017). In other words, statins appear to diminish the harms of inflammation and clotting that persons infected with COVID-19 experience, which may lead to lower morbidity and mortality. Some studies indicate that some persons receiving statins during COVID-19 infection have lower morbidity and mortality compared to some not receiving statins (De Spiegeleer et al., 2020). Countering the theory that statins diminish COVID-19 infection morbidity and mortality is that statins may up-regulate the ACE2 receptor activity in the lungs, which is how SARS-Co-V2 enters cells, implying that statins may increase infection risk (Tikoo et al., 2015).

Opioids are a class of drugs that include the illegal drug heroin, synthetic opioids such as fentanyl, and pain relievers available legally by prescriptions (https://www.drugabuse.gov/drug-topics/opioids). There is increasing recognition that persons engaging in chronic opioid misuse have higher rates of viral and other infections. Opioid use can make people more vulnerable to infection via suppression of immune surveillance, in contrast to statin use which is theorized to increase viral load by easing COVID-19's ability to infect. Opioid misuse is associated with altered inflammatory response. Opioid use is associated with elevated risk of infection due to immune suppression, and paradoxically immune suppression may play a protective role during COVID-19 infection due to mitigated inflammatory response (Ataei et al., 2020). Louisville Kentucky is an epidemic area for opioid crisis. Statins and opoids are therefore theorized to impact COVID-19 outcomes in unconfirmed and variable ways. One approach is to study their impact on biomarkers of inflammation and infection. We used a large database of hospitalized COVID-19 patients established by the University of Louisville Center of Excellence for Research in Infectious Disease (CERID) to study the impact of opioids and statins on the immune and inflammatory biomarkers among patients hospitalized for COVID-19. The data used in this study consisted of 685 adult inpatients hospitalized with COVID-19 at nine different hospitals within

the Louisville metropolitan area from March 9, 2020 to June 20, 2020. To assess the effect of antecedent statins and opioids use on the immune and inflammatory biomarkers, the patients were formed into four treatment groups according to their antecedent statins and opioids use: control group (*i.e.*, neither statin nor opioids used, n=402), statins only group (n=214), opioids only group (n=34), and statin plus opioids group (n=35). The target immune and inflammatory biomarkers of interest regarding acute COVID-19 infection outcomes, the outcome variables, included laboratory Ct values (*i.e.*, the cycle threshold value in RT-PCR tests for the coronavirus, the smaller the Ct value, the more severe the infection), neutrophils percentage (a special type of white blood cells that are involved in the fight against infection), lymphocytes percentage (white blood cells that are one of the body's main types of immune cells), activated partial thromboplastin time (aPTT) (characterizing coagulation of the blood), and procalcitonin (a higher level of procalcitonin indicates a response to a pro-inflammatory stimulus). Several chronic co-morbidities and factors are associated with baseline elevated pro-inflammatory states and baseline elevated inflammatory bio-markers (de Lucena et al., 2020). Race, gender, age, weight, body mass index (BMI), and different comorbidity conditions are possible confounding variables because they are associated with treatment choices as well outcome risks.

Generalized propensity scores were obtained from multinomial regression models and were employed to balance the covariates in each group. Figure A1.6 in the Appendix presented the SMDs for the original sample and the weighted sample. It showed that the SMDs in the original sample were much larger than the threshold value 0.1, and the SMDs in the weighted sample were smaller or close to the threshold value 0.1, indicating that covariates were balanced among the four groups in the weighted sample. We then applied IPTW to estimate the ATEs and drug interactions on the different outcomes. Table 3.4 presented the group means of the outcome variables based on the original observed sample (Table 3.4A) and the weighted sam-

ples (Table 3.4B). We reported the estimated ATE and drug interaction parameters using MSM without IPTW as well as with IPTW (see Table 3.5). Due to the unbalanced covariates in the original sample, the statistical inference for ATEs and drug interactions should be based on the weighted sample, *i.e.*, the estimates from MSMs with IPTW. The variances for the estimated parameters in the MSMs were obtained from the bootstrap method. Since there were five biomarkers in this investigation, Bonferroni correction were applied to adjust the p-values for statistical inferences. Based on Table 3.5, statins significantly increased neutrophils percentage and aPTT, and statins significantly decreased lymphocytes percentage. Opioids used alone did not change the aforementioned biomarkers significantly (although the p-value for the procalcitonin level was 0.046, it became insignificant after Bonferroni corrections). In particular, statins used alone increased aPTT significantly, and opioids used alone did not change aPTT significantly. However, the effect of statins on aPTT in the presence of opioids use was significantly decreased comparing with the effect of statins on aPTT in the absence of opioids by a magnitude of -9.5 unit (adjusted p=0.004×5=0.02). That is, statins and opioids did interact on aPTT significantly. The multiple pathways associated with immune response and sequent inflammatory cascade may be reflected in this analysis. For example, an increase in aPTT in statin use alone may indicate the impact of statin use on higher viral load and intensity of infection, which in turn is associated with higher inflammatory levels. This finding supports the increased ACE2 receptor up-regulation hypothesis with statin use. And, the presence of opioids mitigating aPTT levels in statin users supports the hypothesis that opioids may mitigate inflammatory response. This result could also reflect unknown impacts of opioids and statins on innate immune function and inflammation. The mechanism of the interaction of statins and opioids may be further investigated, which is beyond the scope of this work.

We noted that the histogram of GPSs for this case study (Figure A1.7 in the

Appendix) did indicate that some propensity scores were close to zero and some were close to one, which implied that the positivity assumption might be violated. We performed sensitivity analysis by weight trimming (Lee et al., 2011) to the range of $(\frac{1}{0.95}, \frac{1}{0.05})$. The parameters in the generalized MSM with the trimmed weights were presented in Table A1.5 in the Appendix. The results were similar to those in Table 3.5. However, the violation of the positivity assumption was a concern, and a confirmatory conclusion requires a further study.

3.5 Conclusion and discussion

In this article, we propose the generalized MSMs and provide the procedures for estimating ATEs and drug interactions using observational data, where the confounding variables are controlled via the IPTW method. This proposed method, paired with strong clinical modeling and the appropriate data set, provide a novel approach to studying medication signals embedded in complex clinical scenarios. Our extensive simulation studies illustrate that the proposed models and estimating algorithms provide consistent estimates for the causal parameters in the MSMs and capture the true ATEs and drug interactions under the assumptions of weak ignorability, positivity, consistency and correct specification of GPS. However, in practice, there could be many factors that can confound the translation from the medication list to outcomes. If there are unmeasured confounders, the underlying assumptions for the proposed method do not hold, thus the proposed method may not be suitable to assess ATEs and drug interactions. However, the focus of this article is the analytic methodology of drug-drug interaction under these common assumptions. The cases presented illustrate drug-drug interactions that may be detected from observational data for potential prospective clinical study. The two case studies illustrate the method described in this paper, with acknowledgement that the case studies are not based on comprehensive clinical models aimed at controlling all confounding variables. The

case studies illustrate the MSM modeling and identify drug-drug interactions associated with outcomes of consequence, providing a potential starting point for clinical studies. Presentation of these case studies illustrates the potential utility of the model in mining observational retrospective data to identify potential drug-drug synergies for future clinical study.

The positivity assumption for the GPS approach is important, and a violation of the positivity assumption may result in biased estimates for ATEs and drug interactions. The simulation studies in Section 3.3 represent the situations where the positivity assumption holds. The histograms of the GPSs for patients assigned to each group for a sample of size 5000 based on the simulation settings for the GPS model in Section 3.3 were shown in Figure A1.8 in the Appendix, which clearly indicated that the probability assigned to each group was positive (i.e., the positivity assumption held). We also carried out additional simulation studies by letting $\beta_1^{(1,0)} = (1.5, 1, -1, \mathbf{0}_2, 1.5, 1, \mathbf{0}_3)^T$, $\beta_1^{(0,1)} = (0.5, -1.5, 0.5, \mathbf{0}_2, 1.5, -1, \mathbf{0}_3)^T$, and $\beta_1^{(1,1)} = (1, -1, 0.5, \mathbf{0}_2, 1, 1.5, \mathbf{0}_3)^T$ in the GPS model (3.11) to simulate the cases where the positivity assumption was violated (Lee et al., 2007). The histograms of the GPSs for patients assigned to each group for a sample of size 5000 were shown in Figure A1.9 in the Appendix, which clearly indicated the violation of the positivity assumption. The boxplots for the simulation results for binary outcome model (3.10) with independent covariates were presented in Figure A1.10, indicating that the MSMs model based on IPTW method can result in biased estimates for ATEs and drug interactions when the positivity assumption is violated. We also allowed that $\mathbf{X}_i$ followed a multivariate normal with mean 0 and variance matrix as Σ, where $\Sigma = (1-\rho)I_{p\times p} + \rho J \times J^T$ with J being a vector of p ones. We set $\rho = 0.2$ for moderately correlated case and $\rho = 0.5$ for strongly correlated case. The simulation results (not shown) indicated that the correlated covariates may impact the variance of the estimated ATEs and drug interactions, and the biases of the estimated ATEs

and drug interactions were largely impacted by the GPS specifications.

Once the generalized propensity scores are obtained, one can use the Horvitz-Thompson survey sampling weighted estimator (Horvitz and Thompson, 1952) to estimate the mean of the potential outcome, further estimate ATEs and drug interactions. Never the less, the Horvitz-Thompson method neither directly estimates ATEs and drug interactions nor estimates their variances. On the contrary, the generalized MSMs based on the weighted sample directly estimate ATEs and drug interactions and their variances. Thus, the generalized MSMs along with IPTW method provide rigorous statistical method for assessing drug interactions. With robust interdisciplinary collaboration and clinical modeling, the impact of drug interactions detected using this method, may identify significant leads for biomedical clinical study.

This article presents the MSMs and algorithms for estimating ATEs and drug interactions when two drugs are used together and each drug has only two levels (present or not). However, the doses of the medicine sometimes vary among patients and may be adjusted based on patients' conditions, where the treatments could be continuous variables. The proposed approach could be extended to the situation when each drug has multiple levels or in continuous scale using the GPS (Hirano and Imbens, 2004), which will be investigated in our future research.

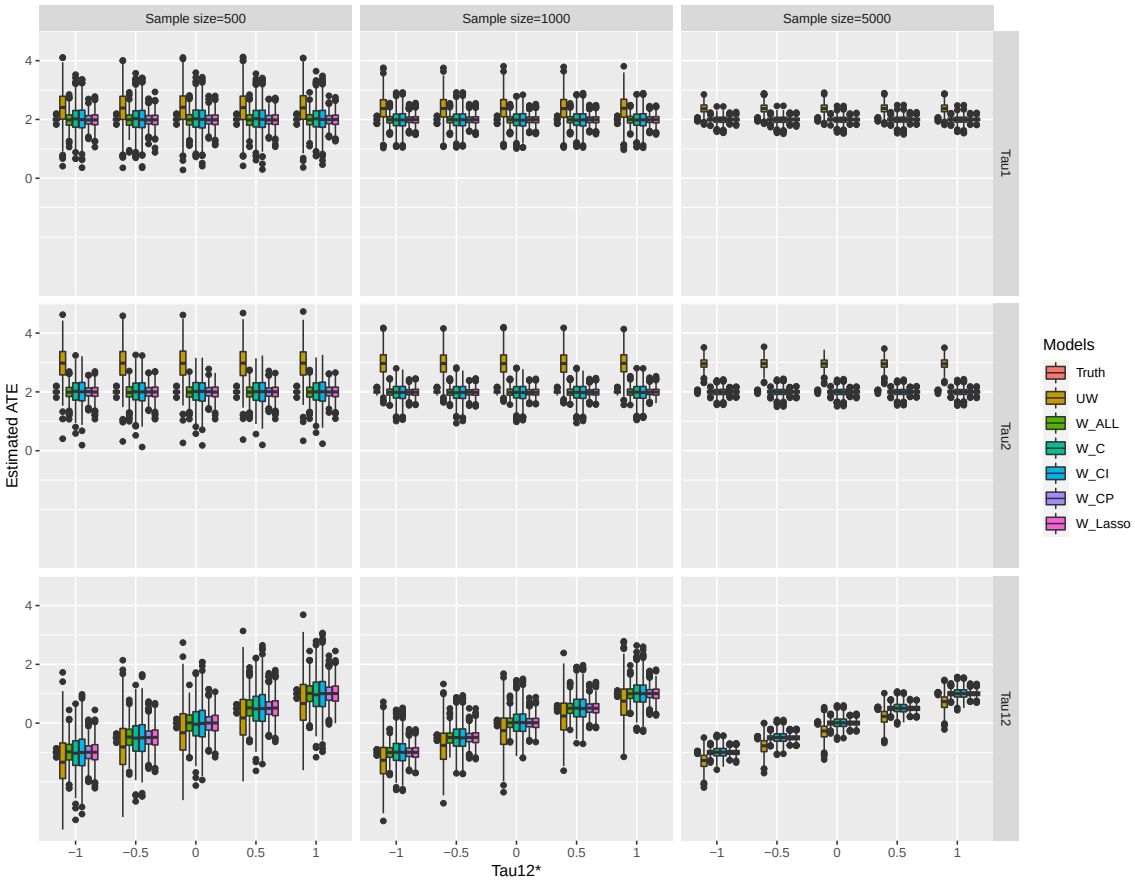

Figure 3.1: Boxplots of 1000 estimated ATEs and drug interactions for continuous outcome with heterogeneous treatment effects and different sample sizes. The first row and the second row, respectively, showed the estimated ATEs for drug 1 (*i.e.* $\hat{\tau}_1$) and drug 2 (*i.e.* $\hat{\tau}_2$), with different specification of τ_{12}^*. The third row showed the estimated τ_{12} to capture drug interaction. In each block (*i.e.* for a fixed τ_{12}^*), the first boxplot was the true ATE or drug interaction, the second boxplot was estimates according to MSMs without IPTW, and the third to the seventh boxplots were estimates according to MSMs with IPTW being obtained from the following five different sets of covariates: (i) all covariates, (ii) confounding variables only, (iii) confounding variables and instrumental variables, (iv) confounding variables and predictors, and (v) covariates selected by LASSO.

Table 3.1: Summarized metrics for the estimated ATEs and drug interactions based on 1000 simulated data sets for continuous outcomes with fixed $\tau_{12}^* = 0$ and sample size $n = 1000$.

		UW	W-All	W-C	W-CI	W-CP	W-LASSO	Reg. model
Homogeneous treatment effect for continuous outcome (Model (3.7)): The true ATEs and drug interaction were (1.000, 1.000, 0.000).								
τ_1	Est.	1.370	0.991	0.983	0.983	0.989	0.989	0.998
	MSE	0.309	0.018	0.079	0.083	0.012	0.012	0.002
	SE	0.400	0.163	0.278	0.289	0.128	0.156	0.045
	E.SD	0.414	0.134	0.281	0.288	0.109	0.110	0.047
	CR	0.825	0.969	0.936	0.939	0.960	0.985	0.950
τ_2	Est.	1.960	0.991	0.985	0.981	0.993	0.993	0.999
	MSE	1.088	0.016	0.076	0.080	0.010	0.010	0.002
	SE	0.400	0.161	0.276	0.289	0.120	0.151	0.045
	E.SD	0.408	0.127	0.276	0.282	0.0980	0.099	0.047
	CR	0.328	0.979	0.944	0.947	0.968	0.990	0.936
τ_{12}	Est.	-0.352	0.009	0.024	0.024	0.013	0.014	0.004
	MSE	0.467	0.029	0.157	0.161	0.021	0.021	0.004
	SE	0.568	0.225	0.392	0.408	0.173	0.213	0.064
	E.SD	0.586	0.171	0.396	0.401	0.144	0.145	0.067
	CR	0.890	0.982	0.946	0.954	0.976	0.988	0.940
Heterogeneous treatment effect for continuous outcome (Model (3.8)): The true ATEs and drug interaction were (2.001, 1.999, 0.001).								
τ_1	Est.	2.370	1.99	1.982	1.982	1.989	1.989	2.014
	MSE	0.316	0.024	0.086	0.091	0.017	0.022	0.010
	SE	0.413	0.187	0.292	0.302	0.155	0.183	0.107
	E.SD	0.429	0.162	0.297	0.305	0.138	0.154	0.107
	CR	0.836	0.958	0.939	0.932	0.964	0.964	0.944
τ_2	Est.	2.961	1.988	1.982	1.979	1.990	1.991	1.943
	MSE	1.099	0.020	0.080	0.083	0.014	0.018	0.012
	SE	0.415	0.180	0.288	0.301	0.145	0.176	0.107
	E.SD	0.421	0.149	0.288	0.293	0.126	0.140	0.105
	CR	0.360	0.975	0.956	0.956	0.974	0.981	0.924
τ_{12}	Est.	-0.266	0.010	0.025	0.024	0.015	0.011	0.071
	MSE	0.480	0.057	0.181	0.186	0.046	0.053	0.044
	SE	0.619	0.282	0.424	0.440	0.238	0.277	0.199
	E.SD	0.639	0.240	0.425	0.432	0.215	0.233	0.200
	CR	0.919	0.983	0.950	0.958	0.966	0.977	0.934

Table 3.2: Summarized metrics for the estimated ATEs and drug interactions based on 1000 simulated data sets for binary outcomes with fixed $\tau_{12}^* = 0$ and sample size $n = 1000$.

		UW	W-All	W-C	W-CI	W-CP	W-LASSO	Reg. model
Homogeneous treatment effect for binary outcome (Model (3.9)): The true ATEs and drug interaction were (1.155, 1.160, 0.135).								
τ_1	Est.	1.046	1.156	1.163	1.162	1.157	1.154	2.055
	MSE	0.055	0.034	0.038	0.039	0.032	0.033	0.900
	SE	0.196	0.170	0.179	0.183	0.164	0.170	0.297
	E.SD	0.196	0.168	0.181	0.185	0.163	0.167	0.289
	CR	0.906	0.939	0.942	0.937	0.947	0.939	0.134
τ_2	Est.	1.104	1.163	1.165	1.163	1.165	1.163	2.062
	MSE	0.048	0.032	0.037	0.038	0.030	0.032	0.900
	SE	0.196	0.172	0.180	0.185	0.165	0.172	0.298
	E.SD	0.198	0.166	0.178	0.182	0.161	0.165	0.284
	CR	0.932	0.954	0.937	0.947	0.944	0.953	0.132
τ_{12}	Est.	0.414	0.168	0.164	0.166	0.165	0.169	0.022
	MSE	0.226	0.129	0.133	0.138	0.122	0.129	0.249
	SE	0.365	0.340	0.345	0.353	0.328	0.339	0.478
	E.SD	0.363	0.332	0.340	0.346	0.322	0.332	0.468
	CR	0.873	0.947	0.941	0.950	0.941	0.949	0.948
Heterogeneous treatment effect for binary outcome (Model (3.10)): The true ATEs and drug interaction are (1.635, 1.638, -0.278).								
τ_1	Est.	1.502	1.638	1.645	1.644	1.639	1.637	2.606
	MSE	0.067	0.040	0.044	0.045	0.038	0.039	1.038
	SE	0.213	0.193	0.201	0.204	0.187	0.192	0.313
	E.SD	0.208	0.183	0.195	0.197	0.179	0.182	0.299
	CR	0.907	0.954	0.954	0.950	0.953	0.955	0.110
τ_2	Est.	1.626	1.649	1.649	1.649	1.648	1.648	2.657
	MSE	0.052	0.042	0.046	0.048	0.040	0.041	1.141
	SE	0.218	0.199	0.205	0.210	0.191	0.199	0.321
	E.SD	0.212	0.191	0.199	0.205	0.185	0.190	0.311
	CR	0.952	0.956	0.956	0.957	0.957	0.954	0.104
τ_{12}	Est.	0.040	-0.236	-0.244	-0.239	-0.244	-0.237	-0.663
	MSE	0.330	0.218	0.216	0.224	0.203	0.215	0.477
	SE	0.593	0.581	0.581	0.590	0.568	0.580	0.707
	E.SD	0.450	0.435	0.435	0.443	0.420	0.431	0.554
	CR	0.901	0.945	0.946	0.947	0.950	0.943	0.914

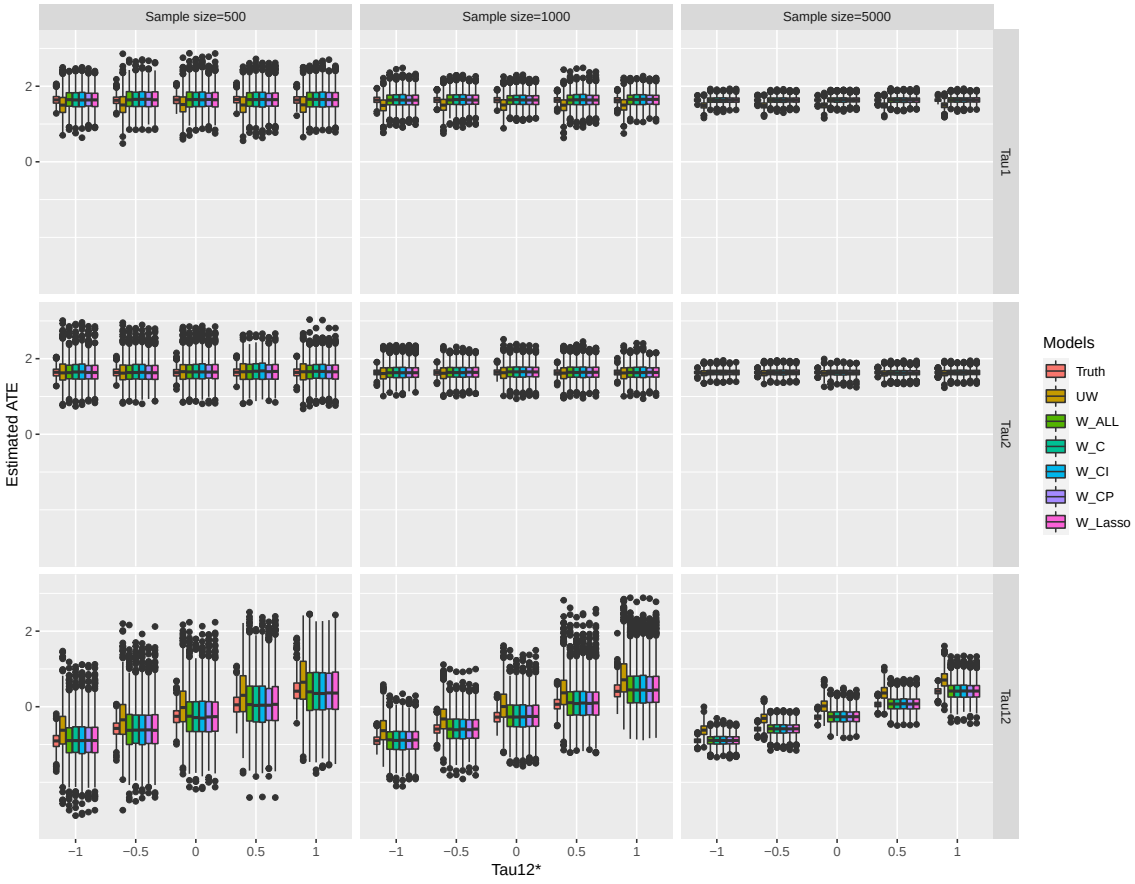

Figure 3.2: Boxplots of 1000 estimated ATEs and drug interactions for binary outcome with heterogeneous treatment effects and different sample sizes. The first row and the second row, respectively, showed the estimated ATEs (OR in log scale) for drug 1 (*i.e.* $\hat{\tau}_1$) and drug 2 (*i.e.* $\hat{\tau}_2$), with different specification of τ_{12}^*. The third row showed the estimated τ_{12} to capture drug interaction. In each block (*i.e.* for a fixed τ_{12}^*), the first boxplot was the true ATE or drug interaction, the second boxplot was estimates according to MSMs without IPTW, and and the third to the seventh boxplots were estimates according to MSMs with IPTW being obtained from the following five different sets of covariates: (i) all covariates, (ii) confounding variables only, (iii) confounding variables and instrumental variables, (iv) confounding variables and predictors, and (v) covariates selected by LASSO.

Table 3.3: Summarized information for the study of glyburide and metformin on hospitalized diabetic patients.

A: Sample size and estimated readmission rate			
	Sample Size	Observed Readmission	Estimated Readmission with IPTW (LASSO)
Control	738	43.8%	44.0%
Gly. only	150	40.7%	32.2%
Met. only	198	30.8%	33.2%
Gly.+Met.	147	32.7%	31.8%
Overall	1233	40.0%	35.3%

B: Estimated ATE and drug interaction in logit (Est) and odds ratio (OR) scale						
	MSM without IPTW			MSM with IPTW(LASSO)		
Parameter	Est (SE)	OR	P-Value	Est (SE)	OR	P-Value
τ_g	-0.127 (0.169)	0.881	0.454	-0.503 (0.189)	0.604	0.008
τ_m	-0.558 (0.173)	0.572	0.001	-0.458 (0.196)	0.633	0.020
τ_{gm}	0.212 (0.289)	1.237	0.464	0.437 (0.317)	1.549	0.168
Total	-0.473 (0.197)	0.623	0.018	-0.524 (0.233)	0.592	0.025

Note: τ_g, Glyburide vs Control; τ_m, Metformin vs Control; τ_{gm}, Drug interaction; Total, Gly.+Met. vs Control .

Abbreviations: Gly., Glyburide; Met., Metformin.

Table 3.4: Summarized statistics for the antecedent opioid use and statins use for hospitalized COVID-19 patients

A: Summarized mean and standard deviation based on the observed sample				
	Control	**Statins Only**	**Opioids Only**	**Statins+Opioids**
Sample Size(N=685)	402	214	34	35
Ct	28.1±5.59	26.1±5.78	26.0±6.13	26.1±4.79
Neutropct	71.1±14.03	74.4±14.3	69.3±12.94	69.4±14.15
Lymphopct	19.3±11.39	15.5±10.38	19.6±9.61	18.7±11.35
aPTT	30.7±7.02	37.0±27.08	33.0±6.35	29.3±4.09
Procalcitonin	2.0±14.49	3.2±28.99	0.5±0.89	1.3±3.59
B: Summarized mean and standard deviation based on the weighted sample with IPTW-All				
	Control	**Statins Only**	**Opioids Only**	**Statins+Opioids**
Ct	27.3±5.76	26.6±5.62	25.5±6.09	27.1±4.14
Neutropct	71.1±13.41	75.0±13.08	71.8±13.57	71.6±12.86
Lymphopct	18.8±10.75	15.7±9.95	17.8±9.4	18.2±10.79
aPTT	30.6±6.39	35.4±20.92	33.8±6.91	29.0±3.73
Procalcitonin	1.8±12.2	3.3±30.69	0.5±0.82	1.2±3.54

Table 3.5: The estimated ATEs and drug interactions for opioids and statins use

	Biomarkers/	MSM without IPTW			MSM with IPTW-All		
	Parameters	Est	SE	P-Value	Est	SE	P-Value*
Ct	τ_s	-2.031	0.529	<0.001	-0.771	0.698	0.270
	τ_o	-2.136	1.286	0.111	-1.853	1.545	0.231
	τ_{so}	2.123	1.713	0.229	2.412	2.022	0.234
	$\tau_s + \tau_o + \tau_{so}$	-2.043	1.222	0.109	-0.213	1.210	0.860
Neutro.	τ_s	3.362	1.209	0.006	3.837	1.295	0.003
	τ_o	-1.758	3.038	0.569	0.719	3.923	0.855
	τ_{so}	-3.307	4.229	0.441	-4.077	4.793	0.396
	$\tau_s + \tau_o + \tau_{so}$	-1.703	3.026	0.578	0.480	2.961	0.871
Lymph.	τ_s	-3.746	0.929	<0.001	-3.090	1.016	0.002
	τ_o	0.341	2.189	0.878	-1.029	2.646	0.698
	τ_{so}	2.825	3.290	0.398	3.535	3.718	0.342
	$\tau_s + \tau_o + \tau_{so}$	-0.580	2.513	0.819	-0.584	2.744	0.832
aPTT	τ_s	6.302	2.546	0.015	4.758	1.791	0.008
	τ_o	2.286	1.808	0.225	3.155	2.373	0.185
	τ_{so}	-10.047	3.404	0.009	-9.526	3.241	0.004
	$\tau_s + \tau_o + \tau_{so}$	-1.459	1.180	0.234	-1.612	1.192	0.178
Procal.	τ_s	1.245	2.212	0.574	1.498	2.425	0.537
	τ_o	-1.487	0.822	0.083	-1.294	0.642	0.045
	τ_{so}	-0.433	2.201	0.845	-0.840	2.386	0.725
	$\tau_s + \tau_o + \tau_{so}$	-0.676	1.078	0.536	-0.636	1.018	0.533

* The p-values were obtained from the Wald tests without Bonferroni correction. The Bonferroni corrected p-values would be the minimal value of 1 and 5 times of the p-values reported.

CHAPTER 4

STATISTICAL METHODS FOR ASSESSING DRUG INTERACTIONS USING OBSERVATIONAL DATA WITH TREATMENT IN CONTINUOUS SCALE

4.1 Introduction

It is common that several treatments are applied to treat patients with the same condition. For example, patients with alcohol use disorder often receive both medical treatment and psychotherapy. It is also common that multiple drugs are prescribed to a patient by health providers to treat different morbid conditions. Using more than one drug or treatment during the same time period to treat different underlying morbid conditions is referred to as polydrug use (Vogt-Ferrier, 2011; Naveiro-Rilo et al., 2014). Polydrug use could enhance the desired treatment effects. However, polydrug use may also cause severe adverse side effects, and polydrug use has increasingly caused concerns (Tramontina et al., 2018).

In developing combination therapy for certain disease, treatment interaction often refers to the situation where the effect of two treatments is more or less than the predicted additive effect. When the effect of combination of two treatments is more than their predicted additive effect, the two treatments are said to be synergistic (Kong and Lee, 2006, 2008). However, treatment interaction can also refer to adverse effect when polydrug is used. When the outcome of interest is treatment efficacy, the synergistic treatment effect is preferred. In the case that the outcome is the adverse

side effect, the synergistic effect would amplify the adverse side effect, and should be avoided. Previous study suggests that healthcare providers use various database and Web resources depending on the potential drug interaction evidence they are seeking (Grizzle et al., 2019). Observational data from electronic health records or claims data are examples of such data resources which could be used to examine treatment interactions.

Recently, a considerable literature has centered around the theme of assessing treatment interaction using the existing data depositories from routine clinical practice. Many statistical models have been developed for assessing drug interaction using *in-vitro* cell culture data (Kong and Lee, 2006, 2008) and *in-vivo* human tumor xenograft data (Fang et al., 2004). Drug-drug interaction and drug-herb interaction in clinical studies have drawn much attention (Vogt-Ferrier, 2011; Baxter and Preston, 2010; VanderWeele, 2015). Much research discovered drug interactions clinically through experience or experiments derived by dose-response curves for each pair of drugs (Tatonetti et al., 2012; Li et al., 2015; Malone et al., 2004). Data mining algorithms were also proposed to identify drug interaction quantitatively using clinical data (Noguchi et al., 2019). However, one of the greatest challenges using clinical observational data to assess drug interactions is to control the confounding factors, which impact both treatment selection and outcome variables (Hernan and Robins, 2020). Propensity score based methods such as the inverse probability of treatment weighting (IPTW) and doubly robust methods have been applied to estimate causal effects (Rosenbaum and Rubin, 1983; Yan et al., 2019). In particular, the marginal structural models (MSMs) along with IPTW method have been applied to estimate causal parameters such as average treatment effect (ATE) (Cole and Hernán, 2008; Robins et al., 2000). Despite propensity score based methods are popular in observational studies, a main practical difficulty of these methods is that the propensity score must be estimated. Previous studies have revealed that slight misspecification

of the propensity score model can result in substantial bias of estimated treatment effects (e.g. Kang and Schafer (2007) and Smith and Todd (2005)). Evidence suggests that covariate balancing propensity score (CBPS) method is more reliable, which is robust to mild misspecification of the parametric propensity score model (Imai and Ratkovic, 2014).

In previous chapter, we proposed generalized MSMs and provide the procedures for estimating ATE and drug interaction using observational data, where the confounding variables are controlled via the IPTW method, which is presented in Chapter 3. Nevertheless, this method presents the MSMs and algorithms for estimating ATE and drug interaction when two drugs are used together and each drug has only two levels (present or not), which is unsuitable for the situation when each drug has multiple levels or in continuous scale. The generalized propensity score (Hirano and Imbens, 2004) enables us to investigate the drug interactions with each drug in multiple levels or in continuous scale. Covariate balancing generalized propensity score (CBGPS) methodology, which minimizes the association between covariates and the continuous treatments, is a robust approach to model misspecification by directly optimizing sample covariate balance among different groups (Fong et al., 2018).

In this project, we propose a marginal structural semiparametric model (MSSM) to estimate ATE and treatment interaction with treatment in multiple levels or in continuous scale, which is presented in Section 4.2. Once the generalized propensity scores are obtained, the MSSM based on the weighted sample can be applied directly to estimate the causal drug interaction. Thus, the MSSM along with IPTW method could provide rigorous statistical method for assessing treatment interaction.

The statistical method we develop here can be used to investigate treatment interaction on treatment effect, as well as drug interaction on adverse event, which depends on the outcome of interest and the investigated treatments/drugs.

4.2 Statistical method to assess treatment interactions

Let us consider evaluating treatment interactions of two treatments (say treatment 1 and treatment 2). Let us assume that we have observed quadruplets (D_1, D_2, X, Y) for each subject. D_k is the dose level of the kth treatment in the set of $\mathcal{D}_k$ $(k = 1, 2)$. Y denotes the outcome variable. X denotes the baseline covariates which include all potential confounding variables. Let denote $Y^{(d_1,d_2)}$ as the potential outcome when the two treatment levels being d_1 and d_2, respectively. Note that there is a potential outcome for each possible combination treatment (d_1, d_2), thus there are many potential outcomes for each subject. However, there is only one observed outcome for each subject. The following statistical method is developed under the same assumptions as in the literature for causal inferences. These assumptions include consistency and no unmeasured confounding. The consistency assumes that the observed outcome Y is the same as the potential outcome $Y^{(d_1,d_2)}$ corresponding to the observed treatment combination, that is $Y = \sum_{d_1 \in \mathcal{D}_1, d_2 \in \mathcal{D}_2} Y^{(d_1,d_2)} I_{(D_1,D_2)=(d_1,d_2)}$. The assumption for "No unmeasured confounding" assumes that $Y^{(d_1,d_2)}$ is independent to the treatment assignment $\boldsymbol{D} = (D_1, D_2)^T$ given the covariates X. The goal of the current study is to access treatment effect for each individual treatment as well as to examine whether the combination of the treatment is synergistic. We assume that the potential outcome $Y^{(d_1,d_2)}$ follows an exponential family with mean $\mu(d_1, d_2)$ and variance function as $v(\mu)$. That is

$$g(E\{Y^{(d_1,d_2)}\}) = g(\mu(d_1, d_2)) = f(d_1, d_2). \tag{4.1}$$

Equation (4.1) is an extension of generalized MSMs model, and we use $f(d_1, d_2)$ denotes the response profile for (D_1, D_2) at the values (d_1, d_2). Here g is a known monotonic link function. When the outcome is continuous, we may take g as the

identity link function, and model (4.1) becomes

$$E\{Y^{(d_1,d_2)}\} = f(d_1, d_2). \tag{4.2}$$

When the outcome is binary, we may take g as the logit link function, and model (4.1) becomes

$$logit\left(E\{Y^{(d_1,d_2)}\}\right) = f(d_1, d_2). \tag{4.3}$$

Assuming D_1 is a multi-level treatment with K levels (i.e., $\mathcal{D}_1 = \{0, 1, \cdots, K-1\}$), and D_2 is a continuous treatment with support interval $\mathcal{D}_2$. The term $f(d_1, d_2)$ in Equation (4.1) can be written as:

$$f(d_1, d_2) = \tau_0 + f_0(d_2) + \sum_{k \in \mathcal{D}_1 - \{0\}} \mathbb{1}_{d_1=k}\{\tau_k + f_k(d_2)\}, \tag{4.4}$$

where $f_k(0) = 0$ for $k \in \mathcal{D}_1 = \{0, 1, \cdots, K-1\}$, and $d_2 \in \mathcal{D}_2$.

Here τ_0 and $\tau_0 + \tau_k$ are the mean response when treatment 1 is used alone at level 0 and k, respectively, in the absence of treatment 2. $\tau_0 + f_0(d_2)$ and $\tau_0 + f_0(d_2) + \tau_k + f_k(d_2)$ are the mean response when treatment 1 is used at level 0 and k, respectively, in the presence of treatment 2 at dose d_2. Note that not all potential outcomes can be observed, only the potential outcome corresponding the treatment received is observed. Literature has indicated that the inverse probability of treatment weighted (IPTW) sample could remove the confounding and the causal parameters in Equation (4.4) can be obtained using the IPTW sample.

Let us first consider that the link function g as the identity link function with the resulting model (4.2). The ATE due to treatment 1 at the level k in the absence of treatment 2 is defined as $E\left(Y^{(k,0)}\right) - E\left(Y^{(0,0)}\right)$, which is captured by τ_k based on model (4.2) and Equation (4.4). The ATE due to treatment 2 is defined as $E\left(Y^{(0,d_2)}\right) - E\left(Y^{(0,0)}\right)$, which is captured by $f_0(d_2)$. We can also see that the ATE due

to treatment 1 at the level k in the presence of treatment 2 is $E\left(Y^{(k,d_2)}\right) - E\left(Y^{(0,d_2)}\right)$, which equals $\tau_k + f_k(d_2)$. That is, when $f_k(d_2) \neq 0$, the ATE due to treatment 1 at the level k varies depending on whether treatment 2 is administered or not. When $f_k(d_2) > 0$, the ATE due to treatment 1 at the level k in the presence of treatment 2 is larger than that in the absence of treatment 2, indicating that treatment 1 at the level k and treatment 2 are synergistic. Vise versa, when $f_k(d_2) < 0$, the ATE due to treatment 1 at the level k in the presence of treatment 2 is less than that in the absence of treatment 2, indicating that treatment 1 at the level k and treatment 2 are antagonistic. This definition is consistent with the definition of drug interaction on the additive scale (VanderWeele, 2015), which states that the sum of ATE of each individual treatment at the level k or the dosage d_2 is different from the ATE when the two drugs are applied together. That is,

$$\left(E(Y^{(k,0)}) - E(Y^{(0,0)})\right) + \left(E(Y^{(0,d_2)}) - E(Y^{(0,0)})\right) \neq E(Y^{(k,d_2)}) - E(Y^{(0,0)}). \quad (4.5)$$

Note that $E(Y^{(k,0)}) - E(Y^{(0,0)}) = \tau_k$ and $E(Y^{(0,d_2)}) - E(Y^{(0,0)}) = f_0(d_2)$, while $E(Y^{(k,d_2)}) - E(Y^{(0,0)}) = \tau_k + f_0(d_2) + f_k(d_2)$. Equation (4.5) above is equivalent to $f_k(d_2) \neq 0$, thus the function $f_k(d_2)$ captures drug interaction on the additive scale.

When the outcome is binary, the logistic MSSM (4.3) is often used. We illustrate that the function $f_k(d_2)$ captures drug interaction in the odds ratio scale (VanderWeele, 2015). That is, $f_k(d_2)$ captures whether the product of causal odds ratios of each individual treatment at the level k or the dosage d_2 is different from the odds ratio when two treatments are applied together. That is,

$$OR^{(k,0) \text{ vs } (0,0)} \times OR^{(0,d_2) \text{ vs } (0,0)} \neq OR^{(k,d_2) \text{ vs } (0,0)}. \quad (4.6)$$

Equation (4.6) is equivalent to $f_k(d_2) \neq 0$, and $f_k(d_2)$ captures drug interaction in the odds ratio scale (VanderWeele, 2015).

When risk ratio is the quantity of interest, the following MSSM with log-link function could be applied:

$$log(Pr(Y^{(d_1,d_2)} = 1)) = f(d_1, d_2).$$

The interaction on the multiplication scale is defined as that the product of causal risk ratios of individual treatment at the level k or the dosage d_2 is different from the causal risk ratio when two treatments are applied together (VanderWeele, 2015). That is, $\frac{Pr(Y^{(k,0)}=1)}{Pr(Y^{(0,0)}=1)} \times \frac{Pr(Y^{(0,d_2)}=1)}{Pr(Y^{(0,0)}=1)} \neq \frac{Pr(Y^{(k,d_2)}=1)}{Pr(Y^{(0,0)}=1)}$, which is equivalent to $f_k(d_2) \neq 0$. Therefore, $f_k(d_2)$ captures drug interaction in the multiplication scale (VanderWeele, 2015).

In summary, the function $f_k(d_2)$ in the MSSM (4.1) with the specification of $f(d_1, d_2)$ in Equation (4.4) captures drug interaction. $f_k(d_2) > 0$ implies that the effect of the combination treatments is more than expected, indicating a synergistic effect; $f_k(d_2) < 0$ implies that the effect of the combination treatments is less than expected, indicating an antagonistic effect. Compared to the fixed causal parameter τ_{12} in the generalized MSM (3.1), which captures two binary drugs' interaction in previous Chapter, the function $f_k(d_2)$ is more flexible to describe the phenomena when synergy and antagonism are interspersed in the different regions of drug combinations of treatment 1 at the level k and treatment 2 with the dosage d_2. Hence, the investigation of the interaction of drug combinations of treatment 1 at the level k and treatment 2 with the dosage d_2 is essentially to test if the function $f_k(d_2)$ is 0. That is, we test the following null hypothesis H_0 against the alternative H_a:

$$H_0 : f_k(d_2) = 0, \text{ for all } k = 0, 1, \cdots, K-1 \text{ and all } d_2 \in \mathcal{D}_2$$

$$H_1 : f_k(d_2) \neq 0, \text{ for at least one } k \in \{1, \cdots, K-1\} \text{ and some } d_2 \in \mathcal{D}_2$$

Although a number of statistical methods have been applied to estimate the functions $f_k(d_2)$, we propose to estimate the function $f_k(d_2)$ with the orthogonal spline basis functions (Perperoglou et al., 2019). That is, the function $f_k(d_2)$ can be written as a combination of the orthogonal spline functions, i.e. $f_k(d_2) = B(d_2)\boldsymbol{\gamma}^{(k)} = \sum_{j=1}^{J} \gamma_j^{(k)} B_j(d_2)$. Because the basis spline functions are orthogonal, the two hypothesis tests for H_0 against H_1 are equivalent

$$H_0 : \gamma_j^{(k)} = 0 \text{ for all } k \in \{1, \cdots, K-1\} \text{ and all } j \in \{1, \cdots, J\}$$

$$H_1 : \gamma_j^{(k)} \neq 0 \text{ for at least one } k \in \{1, \cdots, K-1\} \text{ and one } j \in \{1, \cdots, J\}.$$

Without loss of generality, let assume the observed sample as $(d_{1i}, d_{2i}, x_i, y_i)$ for i^{th} subject $(i = 1, \cdots, n)$, and $d_{1i} \in \{0, 1, \cdots, K-1\}$. $f_k(d_2) = \sum_{j=1}^{J} \gamma_j^{(k)} B_j(d_2)$ for $k = 1, \cdots, K-1$ could be written as:

$$\begin{pmatrix} f_k(d_{21}) \\ f_k(d_{22}) \\ \cdots \\ f_k(d_{2i}) \\ \cdots \\ f_k(d_{2n}) \end{pmatrix} = \begin{pmatrix} B_1(d_{21}) & B_2(d_{21}) & \cdots & B_J(d_{21}) \\ B_1(d_{22}) & B_2(d_{22}) & \cdots & B_J(d_{22}) \\ \cdots & \cdots & \cdots & \cdot \\ B_1(d_{2i}) & B_2(d_{2i}) & \cdots & B_J(d_{2i}) \\ \cdots & \cdots & \cdots & \cdot \\ B_1(d_{2n}) & B_2(d_{2n}) & \cdots & B_J(d_{2n}) \end{pmatrix} \begin{pmatrix} \gamma_1^k \\ \gamma_2^k \\ \cdots \\ \gamma_j^k \\ \cdots \\ \gamma_J^k \end{pmatrix}.$$

Similarly, $f_0(d_2)$ in Equation (4.4) could also be expressed as the linear combination of $B_j(d_2)$ for $j = 1, \cdots, J$, i.e. $f_0(d_2) = B(d_2)\boldsymbol{\beta} = \sum_{j=1}^{J} \beta_j B_j(d_2)$. Hence, Equation (4.4) could be written as:

$$f(d_1, d_2) = \tau_0 + \sum_{k \in \mathcal{D}_1 - \{0\}} \mathbb{1}_{d_1 = k}\{\tau_k\} + \sum_{j=1}^{J} \beta_j B_j(d_2) + \sum_{k \in \mathcal{D}_1 - \{0\}} \mathbb{1}_{d_1 = k}\{\sum_{j=1}^{J} \gamma_j^{(k)} B_j(d_2)\}, \tag{4.7}$$

where $f_k(0) = 0$ for $k \in \mathcal{D}_1 = \{0, 1, \cdots, K-1\}$. The linear equality constraint can often be removed by reformulating the problem: one of unknown parameters can be expressed as a function of the others.

Note that the aforementioned the extension of MSSM are models for potential outcome $Y^{(d_1,d_2)}$ when the subject had received the drug combinations of treatment 1 at the level $d_1 = k$ and treatment 2 with the dosage d_2. However the potential outcome $Y^{(d_1,d_2)}$ is not observed if the subject does not receive the drug combinations of treatment 1 at the level k and treatment 2 with the dosage d_2. Although the function $\boldsymbol{\gamma}^{(k)}$ in $f_k(d_2)$ in the MSSM (4.1) with specification of $f(d_1, d_2)$ as (4.7) has causal interpretation, an appropriate estimating method must control for confounding variables. Following the literature in the causal inference, we apply the IPTW method to estimate the function $\boldsymbol{\gamma}^{(k)}$ in the MSSM (4.1) with specification of $f(d_1, d_2)$ as (4.7) (Robins et al., 2000; Cole and Hernán, 2008). The IPTW method essentially creates a weighted sample where the distributions of each confounding variable across different treatment groups are similar, thus removing the confounding effect between the treatment assignment and the outcome variable. The consistency of the estimates for $\boldsymbol{\gamma}^{(k)}$ provided by the MSSM with IPTW holds under the following assumptions (Rosenbaum and Rubin, 1983; Imbens, 2000; Gelman and Hill, 2006; Cole and Hernán, 2008; Greenland and Mansournia, 2015; Hernan and Robins, 2020): (i) Weak ignorability (*i.e.*, weak unconfoundness): $Y^{(d_1,d_2)} \perp (D_1, D_2)|X$ for each pair $(d_1, d_2) \in \mathcal{D}$. That is, given X, the potential outcome is independent of the treatment received; (ii) Positivity: $Pr(D_1 = d_1, D_2 = d_2|X) > 0$ for all $(d_1, d_2) \in \mathcal{D}$ and X; (iii) Consistency: $Y = \sum_{(d_1,d_2)\in\mathcal{D}} I_{\{(D_1,D_2)=(d_1,d_2)\}} Y^{(d_1,d_2)}$, that is, the observed outcome is the same as the potential outcome corresponding the treatment received; and (iv) Correctly specification of propensity score model. The first three assumptions are key in the causal inference literature. The fourth assumption is specifically required for the IPTW method (Robins et al., 2000; Kang and Schafer, 2007; Cole and Hernán,

2008). Violation of any one of the assumptions may result in biased estimates for ATEs and drug interactions when our proposed IPTW method is applied.

The IPTW weight for i^{th} subject is:

$$w_i = \frac{1}{Pr(D_1 = d_{1i}, D_2 = d_{2i}|X = x_i)}.$$

Thus, we form a weighted sample where i^{th} subject has w_i copies of $(d_{1i}, d_{2i}, x_i, y_i)$ instead of 1 copy. In the weighted sample, X is not associated with treatment assignment anymore in the weighted sample. In practice, the probability of treatment assignment, say $Pr(D_1 = d_1, D_2 = d_2|X = x)$, needs to be estimated. The generalized propensity score (Hirano and Imbens, 2004) enables us to investigate the drugs with multiple levels or in continuous scale. In the literature, parametric methods (e.g., multinomial regression (Imbens, 2000), covariate balance propensity score (CBPS) method (Imai and Ratkovic, 2014)) and non-parametric method (e.g., generalized boosting method (GBM) (McCaffrey et al., 2013)) have been proposed to estimate the GPSs (Yan et al., 2019). In particular, the weight for the interaction effects of both treatment D_1 and D_2 are the product of two weights (Kang et al., 2014). The weight for subject i can be written as:

$$w_i = \frac{1}{Pr(D_1 = d_{1i}, D_2 = d_{2i}|X = x_i)} = \frac{1}{e_1(D_{1i}|\boldsymbol{X}_i)} \times \frac{1}{e_2(D_{2i}|D_{1i}, \boldsymbol{X}_i)},$$

where $e_1(D_{1i}|\boldsymbol{X}_i)$ is the probability of treatment 1 assignment, and $e_2(D_{2i}|D_{1i}, \boldsymbol{X}_i)$ is the probability of treatment 2 assignment conditional on treatment 1.

Under the four assumptions (i.e., weak ignorability, positivity, consistency, and correct specification of GPS model), there is no confounding anymore in the weighted sample. The parameter $\boldsymbol{\gamma}^{(k)}$ in $f_k(d_2)$ in the MSSM (4.1) with specification of $f(d_1, d_2)$ as (4.7) can be obtained by maximizing the following objective function, which is the

weighted log-likelihood function. That is,

$$\hat{\theta}^{(k)} = (\hat{\tau}_0, \hat{\tau}_k, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}^{(k)})^T = \arg\max \sum_{i=1}^{n} \hat{w}_i l(\theta^{(k)}; d_{1i}, d_{2i}, y_i).$$

Here $l(\theta^{(k)}; d_{1i}, d_{2i}, y_i)$ is the log-likelihood function for i^{th} observation, and λ is a tunning parameter. When the outcome is continuous and the identity link function is used, the log-likelihood function for i^{th} observation is

$$\begin{aligned} l(\theta^{(k)}; d_{1i} = k, d_{2i}, y_i) &= log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \tau_0 - \tau_k - \sum_{j=1}^{J} \beta_j B_j(d_2) - \sum_{j=1}^{J} \gamma_j^{(k)} B_j(d_2))^2}{2\sigma^2}}) \\ &= -\frac{(y_i - \tau_0 - \tau_k - \sum_{j=1}^{J} \beta_j B_j(d_2) - \sum_{j=1}^{J} \gamma_j^{(k)} B_j(d_2))^2}{2\sigma^2} - \frac{1}{2} log(2\pi\sigma^2). \end{aligned}$$

When the outcome is binary and the logit link function is used, the log-likelihood function for i^{th} observation is

$$\begin{aligned} l(\theta^{(k)}; d_{1i}, d_{2i}, y_i) = y_i(\tau_0 + \tau_k + \sum_{j=1}^{J} \beta_j B_j(d_2) + \sum_{j=1}^{J} \gamma_j^{(k)} B_j(d_2)) \\ - log(1 + e^{\tau_0 + \tau_k + \sum_{j=1}^{J} \beta_j B_j(d_2) + \sum_{j=1}^{J} \gamma_j^{(k)} B_j(d_2)}). \end{aligned}$$

The similar work (Robins et al., 2000) on MSMs indicates that the weighted ML results in consistent estimator for the causal parameter for $\theta^{(k)}$ under the four assumptions for causal inference. The weighted ML estimate for $\theta^{(k)}$ can be obtained by using the R package *survey*(Lumley, 2004), where the weights are obtained by the GPS model which achieves the balance of confounding variables. Although a robust variance estimator for $\hat{\theta}^{(k)}$ can be obtained from the *survey* package, it does not incorporate the uncertainty in estimating the GPSs in the IPTW method. Instead, we use the bootstrap sampling techniques to estimate the variance of $\hat{\theta}^{(k)}$. That is, we obtain B (say, 100) bootstrap samples from the original sample. For the b^{th} bootstrap sample ($b = 1, \cdots, B$), we repeat the same estimating process as outlined to obtain

an estimate $\hat{\theta}^{(kb)}$ for $\theta^{(k)}$. $\widehat{\mathrm{Var}}(\hat{\theta}^{(k)})$, the estimate of the variance of $\hat{\theta}^{(k)}$, is obtained as the variance of the B bootstrap estimates $\hat{\theta}^{(kb)}$ $(b = 1, \cdots, B)$ (Mooney et al., 1993).

4.3 Case study

To demonstrate the use of our proposed method, we applied it to investigate the treatment interaction between medication and psychology counseling on the remission for patients with alcohol use disorder in the Kentucky Medicaid database.

Alcohol abuse and dependence is a serious threats throughout United States (Smothers et al., 2004). According to the CDC, Kentucky is the state with the third-highest problem with binge drinking, alcohol use disorder is a critical problem to the health and well-being of Kentucky residents. Alcoholism treatment and alcohol rehab are quite important to help people overcome addiction and regain full control over their lives.

Patients with alcohol abuse and dependence quite often received drug therapy. Naltrexone and acamprosate have well established efficacy and are first-line treatments (Crowley, 2015). For patients who do not respond to naltrexone or acamprosate, disulfiram or topiramate will be considered (Crowley, 2015). On the other hand, patients with alcohol abuse and dependence usually have psychological issues, such as delusions and irrational thinking, anxiety and paranoia attention, memory and cognitive problems, aggression, anger and irritability and so on. Psychologists should play an increasing role in assessing and treating addictive (Miller and Brown, 1997) and psychology counseling is most used treatment for the patients with alcohol abuse and dependence.

We used this alcohol use disorder in the Kentucky Medicaid dataset to study the joint impact of drug therapy (such as naltrexone, disulfiram, acamprosate, and topiramate) and psychology counseling, which the patients received within one year window after the diagnosis of alcohol use disorder, on the remission of the alcohol use

disorder or not. Here psychology counseling is treatment 1, which is a binary variable (level 0: without psychology counseling; level 1: with psychology counseling), and total drug supply days is treatment 2, which is a continuous variable. The patients were formed into four groups according to their treatment: control group (i.e., no any treatment used, n= 2543), psychology counseling only group (n= 1414), drug therapy only group (n= 956), and both treatments group (n= 173). Race, gender, age, urban/rural, and different comorbidity conditions (see Table 4.1) were possible confounding variables because they are associated with treatment choices as well outcome risks.

Table 4.1: Baseline characteristics of patients with alcohol use disorder.

	Control group (N=2543)	Psychology counseling only group (N=1414)	Drug therapy only group (n=956)	Both treatments group (n=173)
Age (Mean±SD)	35.7±12.55	35.4±12.85	37.7±11.51	34.5±9.53
Gender				
Male	1060 (41.68%)	661 (46.75%)	331 (34.62%)	59 (34.1%)
Female	1483 (58.32%)	753 (53.25%)	625 (65.38%)	114 (65.9%)
Race				
Hispanic	16 (0.63%)	9 (0.64%)	10 (1.05%)	0 (0.00%)
Non-Hispanic black	296 (11.64%)	233 (16.48%)	59 (6.17%)	15 (8.67%)
Non-Hispanic white	1846 (72.59%)	944 (66.76%)	741 (77.51%)	127 (73.41%)
Non-Hispanic others	29 (1.14%)	20 (1.41%)	8 (0.84%)	3 (1.73%)
Non-Hispanic missing	356 (14%)	208 (14.71%)	138 (14.44%)	28 (16.18%)
Myocardial infarction	35 (1.38%	31 (2.19%)	28 (2.93%)	2 (1.16%)
Congestive heart failure	62 (2.44%)	45 (3.18%)	37 (3.87%)	1 (0.58%)
Peripheral vascular disease	58 (2.28%)	35 (2.48%)	45 (4.71%)	8 (4.62%)
Cerebrovascular disease	55 (2.16%)	28 (1.98%)	45 (4.71%)	3 (1.73%)
Dementia	5 (0.2%)	5 (0.35%)	0 (0.00%)	0 (0.00%)
Chronic pulmonary disease	696 (27.37%)	386 (27.3%)	327 (34.21%)	48 (27.75%)
Rheumatic disease	43 (1.69%)	22 (1.56%)	28 (2.93%)	3 (1.73%)
Peptic ulcer disease	25 (0.98%)	19 (1.34%)	18 (1.88%)	3 (1.73%)
Mild liver disease	180 (7.08%)	107 (7.57%)	93 (9.73%)	28 (16.18%)
Diabetes	245 (9.63%)	136 (9.62%)	148 (15.48%)	17 (9.83%)
Hemiplegia or paraplegia	16 (0.63%)	10 (0.71%)	8 (0.84%)	1 (0.58%)
Renal disease	71 (2.79%)	38 (2.69%)	35 (3.66%)	2 (1.16%)
Cancer	31 (1.22%)	26 (1.84%)	19 (1.99%)	0 (0.00%)
Moderate/severe liver disease	9 (0.35%)	6 (0.42%)	4 (0.42%)	0 (0.00%)
Metastatic solid tumour	13 (0.51%)	3 (0.21%)	6 (0.63%)	0 (0.00%)
Opioid use disorder	606 (23.83%)	366 (25.88%)	189 (19.77%)	67 (38.73%)
Mental disorder including anxiety/depression	1788 (70.31%)	967 (68.39%)	721 (75.42%)	121 (69.94%)
Alcohol history	4 (0.16%)	3 (0.21%)	5 (0.52%)	0 (0.00%)
Tobacco use	1776 (69.84%)	1066 (75.39%)	642 (67.15%)	127 (73.41%)
Remission	156 (6.13%)	109 (7.71%)	94 (9.83%)	32 (18.5%)

Generalized propensity scores obtained from the product of logistic regression model for psychology counseling and continuous propensity score model for drug therapy were employed to balance the covariates in each group. We applied proposed MSSM model to estimate the ATEs and drug interactions on the remission. We estimated the functions $f_0(d_2)$ and $f_1(d_2)$ by the orthogonal polynomials with the degree 3. That is, $f_0(d_2) = B(d_2)\boldsymbol{\beta} = \sum_{j=1}^{3} \beta_j B_j(d_2)$, and $f_1(d_2) = B(d_2)\boldsymbol{\gamma}^{(1)} = \sum_{j=1}^{3} \gamma_j^{(1)} B_j(d_2)$.

Table 4.2: The interaction parameter estimates for psychology counseling and drug therapy

Parameters	Est	SE	P-Value*
τ_0	-2.644	0.069	<0.001
τ_1	0.370	0.117	0.002
β_1	9.737	2.144	<0.001
β_2	-1.526	4.165	0.714
β_3	-15.205	6.533	0.020
$\gamma_1^{(1)}$	-2.318	7.709	0.764
$\gamma_2^{(1)}$	-15.553	18.490	0.400
$\gamma_3^{(1)}$	-13.316	27.941	0.634

* The p-values were obtained from the Wald tests.

We reported the estimated drug interaction parameters $\gamma_j^{(1)}$, $(j = 1, 2, 3)$ using proposed MSSM model (see Table 4.2). Based on Table 4.2, all $\gamma_j^{(1)}$, $(j = 1, 2, 3)$ are not significant. That is, psychology counseling and drug therapy did not interact on the remission significantly.

4.4 Conclusion and discussion

In this project, we developed a MSSM model to assess drug interactions when one drug has multi-levels ans another drug in continuous scale. There are more work to be done for this research project. First, we will work out the details about the asymptotic normality and theoretic derivation. For the simulation study, we will design and explore a wide spectrum of settings to examine the performance of the proposed method.

REFERENCES

Ataei, M., Shirazi, F. M., Lamarine, R. J., Nakhaee, S., and Mehrpour, O. (2020). A double-edged sword of using opioids and covid-19: a toxicological view. *Substance Abuse Treatment, Prevention, and Policy*, 15(1):1–4.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (iptw) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, 34(28):3661–3679.

Baxter, K. and Preston, C. L. (2010). *Stockley's Drug Interactions*, volume 495. Pharmaceutical Press London.

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

Brinkley, J., Tsiatis, A., and Anstrom, K. J. (2010). A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics*, 66(2):512–522.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156.

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664.

Craycroft, J. A., Huang, J., and Kong, M. (2020). Propensity score specification for optimal estimation of average treatment effect with binary response. *Statistical Methods in Medical Research*, 29(12):3623–3640.

Crowley, P. (2015). Long-term drug treatment of patients with alcohol dependence. *Australian prescriber*, 38(2):41.

Daniels, L. B., Ren, J., Kumar, K., Bui, Q. M., Zhang, J., Zhang, X., Sawan, M. A., Eisen, H., Longhurst, C. A., and Messer, K. (2021). Relation of prior statin and anti-hypertensive use to severity of disease among patients hospitalized with covid-19: Findings from the american heart association's covid-19 cardiovascular disease registry. *PloS One*, 16(7):e0254635.

de Lucena, T. M. C., da Silva Santos, A. F., de Lima, B. R., de Albuquerque Borborema, M. E., and de Azevêdo Silva, J. (2020). Mechanism of inflammatory response in associated comorbidities in covid-19. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):597–600.

De Spiegeleer, A., Bronselaer, A., Teo, J. T., Byttebier, G., De Tré, G., Belmans, L., Dobson, R., Wynendaele, E., Van De Wiele, C., Vandaele, F., et al. (2020). The effects of arbs, aceis, and statins on clinical outcomes of covid-19 infection among nursing home residents. *Journal of the American Medical Directors Association*, 21(7):909–914.

Dijkman, B., Kooistra, B., and Bhandari, M. (2009). How to work with a subgroup analysis. *Canadian Journal of Surgery*, 52(6):515.

Fang, H.-B., Tian, G.-L., and Tan, M. (2004). Hierarchical models for tumor xenograft experiments in drug development. *Journal of Biopharmaceutical Statistics*, 14(4):931–945.

Fong, C., Hazlett, C., and Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177.

Foster, J. C., Nan, B., Shen, L., Kaciroti, N., and Taylor, J. M. (2016). Permutation testing for treatment–covariate interactions and subgroup identification. *Statistics in Biosciences*, 8(1):77–98.

Foster, J. C., Taylor, J. M., Kaciroti, N., and Nan, B. (2015). Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. *Biostatistics*, 16(2):368–382.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

Fu, H., Zhou, J., and Faries, D. E. (2016). Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. *Statistics in Medicine*, 35(19):3285–3302.

Garber, A. J., Donovan Jr, D. S., Dandona, P., Bruce, S., and Park, J.-S. (2003). Efficacy of glyburide/metformin tablets compared with initial monotherapy in type 2 diabetes. *The Journal of Clinical Endocrinology & Metabolism*, 88(8):3598–3604.

Gelman, A. and Hill, J. (2006). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge university press.

Greenland, S. and Mansournia, M. A. (2015). Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *European Journal of Epidemiology*, 30(10):1101–1110.

Grizzle, A. J., Horn, J., Collins, C., Schneider, J., Malone, D. C., Stottlemyer, B., and Boyce, R. D. (2019). Identifying common methods used by drug interaction experts for finding evidence about potential drug-drug interactions: web-based survey. *Journal of Medical Internet Research*, 21(1):e11182.

Hernan, M. and Robins, J. (2020). Causal inference: What if. boca raton: Chapman & hill/crc.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 226164:73–84.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Huang, I. and Pranata, R. (2020). Lymphopenia in severe coronavirus disease-2019 (covid-19): systematic review and meta-analysis. *Journal of Intensive Care*, 8(1):1–10.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Kang, J., Su, X., Liu, L., and Daviglus, M. L. (2014). Causal inference of interaction effects with inverse propensity weighting, g-computation and tree-based standardization. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 7(5):323–336.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, pages 523–539.

Koch, B., Vock, D. M., and Wolfson, J. (2018). Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics*, 74(1):8–17.

Kong, M. and Lee, J. J. (2006). A generalized response surface model with varying relative potency for assessing drug interaction. *Biometrics*, 62(4):986–995.

Kong, M. and Lee, J. J. (2008). A semiparametric response surface model for assessing drug interaction. *Biometrics*, 64(2):396–405.

Kwemou, M. (2016). Non-asymptotic oracle inequalities for the Lasso and group Lasso in high dimensional logistic model. *ESAIM: Probability and Statistics*, 20:309–331.

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

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS One*, 6(3):e18174.

Lee, J. J., Kong, M., Ayers, G. D., and Lotan, R. (2007). Interaction index and different methods for determining drug interaction in combination therapy. *Journal of Biopharmaceutical statistics*, 17(3):461–480.

Li, P., Huang, C., Fu, Y., Wang, J., Wu, Z., Ru, J., Zheng, C., Guo, Z., Chen, X., Zhou, W., et al. (2015). Large-scale exploration and analysis of drug combinations. *Bioinformatics*, 31(12):2007–2016.

Lipkovich, I., Dmitrienko, A., and B D'Agostino Sr, R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36(1):136–196.

Lu, W., Zhang, H. H., and Zeng, D. (2013). Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, 22(5):493–504.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19.

Malone, D. C., Armstrong, E. P., Abarca, J., Grizzle, A. J., Hansten, P. D., Van Bergen, R. C., Duncan-Edgar, B. S., Solomon, S. L., and Lipton, R. B. (2004). Identification of serious drug–drug interactions: results of the partnership to prevent drug–drug interactions. *Journal of the American Pharmacists Association*, 44(2):142–151.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414.

McPherson, R. A. and Pincus, M. R. (2021). *Henry's Clinical Diagnosis and Management by Laboratory Methods E-book*. Elsevier Health Sciences.

Milborrow, S. (2019). Earth: Multivariate adaptive regression splines.

Miller, W. R. and Brown, S. A. (1997). Why psychologists should treat alcohol and drug problems. *American psychologist*, 52(12):1269.

Mokhtari, R. B., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B., and Yeger, H. (2017). Combination therapy in combating cancer. *Oncotarget*, 8(23):38022.

Mooney, C. Z., Mooney, C. F., Mooney, C. L., Duval, R. D., and Duvall, R. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Number 95. sage.

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

Naveiro-Rilo, J. C., Diez-Juárez, D., Flores-Zurutuza, M. L., Pérez, A., et al. (2014). Quality of life in the elderly on polymedication and with multiple morbidities. *Revista Espanola de Geriatria y Gerontologia*, 49(4):158–164.

Noguchi, Y., Tachi, T., and Teramachi, H. (2019). Review of statistical methodologies for detecting drug–drug interactions using spontaneous reporting systems. *Frontiers in Pharmacology*, 10:1319.

Oesterle, A., Laufs, U., and Liao, J. K. (2017). Pleiotropic effects of statins on the cardiovascular system. *Circulation Research*, 120(1):229–243.

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019). A review of spline function procedures in r. *BMC Medical Research Methodology*, 19(1):1–16.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37(11):1767–1787.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):1180.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

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

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

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a Review Journal of the Institute of Mathematical Statistics*, 29(4):640.

Smith, J. A. and Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1-2):305–353.

Smothers, B. A., Yahr, H. T., and Ruhl, C. E. (2004). Detection of alcohol use disorders in general hospital admissions in the united states. *Archives of internal medicine*, 164(7):749–756.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., and Clore, J. N. (2014). Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014.

Tatonetti, N. P., Patrick, P. Y., Daneshjou, R., and Altman, R. B. (2012). Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125):125ra31–125ra31.

Tavakolpour, S., Rakhshandehroo, T., Wei, E. X., and Rashidian, M. (2020). Lymphopenia during the covid-19 infection: What it shows and what can be learned. *Immunology Letters*, 225:31.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tikoo, K., Patel, G., Kumar, S., Karpe, P. A., Sanghavi, M., Malek, V., and Srinivasan, K. (2015). Tissue specific up regulation of ace2 in rabbit model of atherosclerosis by atorvastatin: role of epigenetic histone modifications. *Biochemical Pharmacology*, 93(3):343–351.

Tramontina, M. Y., Ferreira, M. B., Castro, M. S. d., and Heineck, I. (2018). Comorbidities, potentially dangerous and low therapeutic index medications: factors linked to emergency visits. *Ciencia & Saude Coletiva*, 23:1471–1482.

Van Der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Science & Business Media.

VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.

VanderWeele, T. J. and Knol, M. J. (2014). A tutorial on interaction. *Epidemiologic Methods*, 3(1):33–72.

Vogt-Ferrier, N. (2011). Older patients, multiple comorbidities, polymedication... should we treat everything? *European Geriatric Medicine*, 2(1):48–51.

Yan, X., Abdia, Y., Datta, S., Kulasekera, K., Ugiliweneza, B., Boakye, M., and Kong, M. (2019). Estimation of average treatment effects among multiple treatment groups by using an ensemble approach. *Statistics in Medicine*, 38(15):2828–2846.

Yang, L. and Tschernig, R. (1999). Multivariate bandwidth selection for local linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):793–815.

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

Zhang, Z., Kim, H. J., Lonjon, G., Zhu, Y., et al. (2019). Balance diagnostics after propensity score matching. *Annals of Translational Medicine*, 7(1).

Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105(1):397–411.

Zhao, W. and Yang, H. (2014). *Statistical Methods in Drug Combination Studies*, volume 69. CRC Press.

Zheng, Q., Peng, L., and He, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *Annals of Statistics*, 43(5):2225.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005a). Addendum: regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(5):768–768.

Zou, H. and Hastie, T. (2005b). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

APPENDIX

Appendix 1: Technical Conditions

We begin with introducing some necessary notations. For any square matrix $\boldsymbol{A}$, we use $\sigma_{\min}(\boldsymbol{A})$ and $\sigma_{\max}(\boldsymbol{A})$ to denote its minimum and maximum eigenvalues, respectively. We adopt the following empirical process notations as follows: for a generic variable Z and function f, $\mathbb{G}_n(f) = \mathbb{G}_n(f(Z_i)) := n^{-1/2}\sum_{i=1}^n (f(Z_i) - E[f(Z_i)])$ and $\mathbb{E}_n f(Z_i) := n^{-1}\sum_{i=1}^n f(Z_i)$. Let $\|A\|_q$ denote the L_q norm of A, where A can be a vector, matrix, or function. In particular, $\|A\|$ denote the L_2 norm of A,

(C1) $\boldsymbol{X}$ has a compact support $\mathcal{X}$ in $\mathbb{R}^p$, $E(\epsilon) = 0$ and $E(\epsilon^2) < \infty$.

(C2) There exists some $\nu > 0$ such that $\nu \leq \inf_{\boldsymbol{x}\in\mathcal{X}} \pi^*(\boldsymbol{x}) \leq \sup_{\boldsymbol{x}\in\mathcal{X}} \pi^*(\boldsymbol{x}) \leq 1 - \nu$.

(C3) $\hat{\pi}(\cdot) \in \Pi$ and $\hat{h}(\cdot) \in \mathcal{H}$, where Π and $\mathcal{H}$ are two functional space including π^* and h^*, respectively. $\nu \leq \inf_{\pi\in\Pi,\boldsymbol{x}\in\mathcal{X}} \pi(\boldsymbol{x}) \leq \sup_{\pi\in\Pi,\boldsymbol{x}\in\mathcal{X}} \pi(\boldsymbol{x}) \leq 1 - \nu$ and $\sup_{h\in\mathcal{H}} \|h\|_\infty < \infty$. Moreover, the class $\mathcal{F} = \{(T - \pi(\boldsymbol{X}))\widetilde{\boldsymbol{X}}\{h^*(\boldsymbol{X}) - h(\boldsymbol{X}) + \epsilon\} : \pi \in \Pi, h \in \mathcal{H}\}$ is Donsker, where we refer the definition of a Donsker class by Van Der Vaart and Wellner (1996).

Remark 4.1.1. *Condition (C1) is commonly adopted in the literature (see, e.g., Zhao and Li, 2012; Zheng et al., 2015; Kwemou, 2016). It reflects the data standardization at the pre-processing stage. Condition (C2) is often satisfied in practice, for example, if π^* follows a logistic model, then π^* is bounded away from 0 and 1, under Condition (C1). By Condition (C2), we obtain that $\tau < \sigma_{\min}(E(\pi^*(\boldsymbol{X})\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{T}})) \leq \sigma_{\max}(E(\pi^*(\boldsymbol{X})\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{T}})) < \tau^{-1}$ for some constant $\tau > 0$. The boundedness condition*

for Π *and* $\mathcal{H}$ *in Condition (C3) is usually satisfied given that* $\mathcal{X}$ *is compact. The condition that* $\mathcal{F}$ *is a Donsker class in Condition (C3) is often met. If* Π *and* $\mathcal{H}$ *are collections of parametric models indexed by parameters in compact subset of* $\mathbb{R}^p$*, that is,* $\Pi = \{\rho(\boldsymbol{x};\theta) : \theta \in \Theta\}$ *and* $\mathcal{H} = \{\kappa(\boldsymbol{x};\gamma) : \gamma \in \Gamma\}$ *for some fixed function* ρ *and* κ*, where* Θ *and* Γ *are some compact subsets of* $\mathbb{R}^p$*, then* $\mathcal{F}$ *is Donsker by Lemma 2.6.15 and Theorem 2.5.2 in Van Der Vaart and Wellner (1996). For example,* ρ *is the logit function and* κ *is a linear function.* Π *and* $\mathcal{H}$ *are also allowed to be some nonparametric classes (Van Der Vaart and Wellner, 1996). We also refer to Chen et al. (2003) for a detailed discussion.*

Appendix 2: Proof of proposition and theorems

This section includes the proof of proposition and theorems in Chapter 2.

Proof of Proposition 2.2.1: If the propensity score $\pi^*(\boldsymbol{X})$ is correctly specified and $h^*(\boldsymbol{X})$ is specified as $h'(\boldsymbol{X})$, then

$$\begin{aligned}
&E\left\{\widetilde{\boldsymbol{X}}\{Y - h'(\boldsymbol{X}) - T\widetilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\beta}\}\{T - \pi^*(X)\}\right\} \\
=&E\left\{E_{\boldsymbol{X}}\left[\widetilde{\boldsymbol{X}}\{h^*(\boldsymbol{X}) + \epsilon - h'(\boldsymbol{X})\}\{T - \pi^*(\boldsymbol{X})\}\Big|\boldsymbol{X}\right]\right\} \\
=&E\left\{\widetilde{\boldsymbol{X}}\{h^*(\boldsymbol{X}) - h'(\boldsymbol{X})\}E_{\boldsymbol{X}}\{T - \pi^*(\boldsymbol{X})|\boldsymbol{X}\}\right\} = 0.
\end{aligned}$$

If the response profile $h^*(\boldsymbol{X})$ is correctly specified but the propensity score $\pi^*(\boldsymbol{X})$ is specified as $\pi'(\boldsymbol{X})$, then

$$E\left\{\widetilde{\boldsymbol{X}}\{Y - h^*(\boldsymbol{X}) - T\widetilde{\boldsymbol{X}}^{\mathrm{T}}\boldsymbol{\beta}\}\{T - \pi'(\boldsymbol{X})\}\right\} = E\left\{\widetilde{\boldsymbol{X}}\epsilon\{T - \pi'(\boldsymbol{X})\}\right\} = 0.$$

Therefore, we proved that the estimating equation holds when either the propensity score $\pi^*(\boldsymbol{X})$ or the response profile $h^*(\boldsymbol{X})$ is correctly specified. □

Proof of Theorem 2.2.1: We note that

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{DR} &= \left(\mathbb{E}_n\left[\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right]\right)^{-1}\mathbb{E}_n\left[\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{Y_i - \hat{h}(\boldsymbol{X}_i)\}\right] \\
&= \left(\mathbb{E}_n\left[\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right]\right)^{-1} \\
&\qquad \times \mathbb{E}_n\left[\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) + T_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon_i - \hat{h}(\boldsymbol{X}_i)\}\right] \\
&= \boldsymbol{\beta}^* + \left(\mathbb{E}_n\left[\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right]\right)^{-1} \\
&\qquad \times \mathbb{E}_n\left[\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) + \epsilon_i - \hat{h}(\boldsymbol{X}_i)\}\right]
\end{aligned}$$

First, we show that for any $\pi(\cdot)$ such that $\nu \leq \pi(\boldsymbol{X}) \leq 1-\nu$ for all $\boldsymbol{X}$, $\sigma_{\max}((\mathbb{E}_n[\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}])^{-1}) \leq (\nu\tau)^{-1}$ in probability, where ν is defined in Condition (C2) and τ is defined in Remark 4.1.1 presented later in this section.

By the law of large number and the fact $T^2 = T$,

$$\mathbb{E}_n\left[(T_i - (1-\nu))T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right] \to_p E\left((T-(1-\nu))T\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{T}}\right) = \nu E\left(\pi^*(\boldsymbol{X})\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{T}}\right) \tag{4.8}$$

Since $\nu \leq \pi(\boldsymbol{X}) \leq 1-\nu$, $(T-\pi(\boldsymbol{X}))T \geq (T-(1-\nu))T$. Then uniformly over all nonzero vector $\boldsymbol{u} \in \mathbb{R}^{p+1}$, $(T_i - \pi(\boldsymbol{X}_i))T_i\boldsymbol{u}^{\mathrm{T}}\widetilde{X}_i\widetilde{X}_i^{\mathrm{T}}\boldsymbol{u} \geq (T_i-(1-\nu))T_i\boldsymbol{u}^{\mathrm{T}}\widetilde{X}_i\widetilde{X}_i^{\mathrm{T}}\boldsymbol{u}$ and $\mathbb{E}_n[(T_i - \pi(\boldsymbol{X}_i))T_i\boldsymbol{u}^{\mathrm{T}}\widetilde{X}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{u}] \geq \mathbb{E}_n[(T_i-(1-\nu))T_i\boldsymbol{u}^{\mathrm{T}}\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\boldsymbol{u}] \geq \nu\tau\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}$ in probability, where the last inequality follows from (4.8) and Condition (C2). Therefore, we obtain that in probability

$$\begin{aligned}
&\sigma_{\min}\left(\mathbb{E}_n\left[(T_i - \hat{\pi}(\boldsymbol{X}_i))T_i\widetilde{X}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right]\right) \geq \nu\tau \quad \text{and} \\
&\sigma_{\max}\left(\left(\mathbb{E}_n\left[(T_i - \hat{\pi}(\boldsymbol{X}_i))T_i\widetilde{X}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right]\right)^{-1}\right) \leq (\nu\tau)^{-1}.
\end{aligned} \tag{4.9}$$

Next, we evaluate $n^{-1}\sum_{i=1}^{n}\{T_i-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{Y_i-\hat{h}(\boldsymbol{X}_i)\}$. We obtain that

$$
\begin{aligned}
&\left\|n^{-1}\sum_{i=1}^{n}\{T_i-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)+\epsilon_i-\hat{h}(\boldsymbol{X}_i)\}\right\| \\
\leq&\left\|n^{-1}\sum_{i=1}^{n}\{T_i-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\epsilon_i\right\|+\left\|n^{-1}\sum_{i=1}^{n}\{T_i-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)-\hat{h}(\boldsymbol{X}_i)\}\right\| \\
\leq&\left\|n^{-1}\sum_{i=1}^{n}T_i\widetilde{\boldsymbol{X}}_i\epsilon_i\right\|+\sup_{\pi\in\Pi}\left\|n^{-1}\sum_{i=1}^{n}\pi(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i\epsilon_i\right\| \\
&+\sup_{h\in\mathcal{H}}\left\|n^{-1}\sum_{i=1}^{n}\{T_i-\pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)-h(\boldsymbol{X}_i)\}\right\| \\
&+\|\hat{\pi}-\pi^*\|_\infty\|h^*-\hat{h}\|_\infty n^{-1}\sum_{i=1}^{n}\left\|\widetilde{\boldsymbol{X}}_i\right\| \\
=:&I_1+I_2+I_3+I_4
\end{aligned}
$$

where the first inequality is trivial and the second inequality follows from that $\hat{h}\in\mathcal{H}$ and $\hat{\pi}\in\Pi$ in Condition (C3).

We deal with the four terms separately. By the law of large number, $I_1=o_p(1)$. Since $E(\pi(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i\epsilon_i)=0$ for all $\pi(\cdot)\in\Pi$,

$$
I_2\leq\sup_{\pi\in\Pi}n^{-1/2}\left\|\mathbb{G}_n\left[\pi(\boldsymbol{X}_i)\widetilde{\boldsymbol{X}}_i\epsilon_i\right]\right\|=o_p(1),
$$

following from that $\{\pi(\boldsymbol{X})\widetilde{\boldsymbol{X}}\epsilon,\pi\in\Pi\}$ is Glivenko–Cantelli from Condition (C3). Since $E(\{T_i-\pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)-h(\boldsymbol{X}_i)\})=0$ for all $h\in\mathcal{H}$,

$$
I_3\leq\sup_{h\in\mathcal{H}}n^{-1/2}\left\|\mathbb{G}_n\left[\{T_i-\pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)-h(\boldsymbol{X}_i)\}\right]\right\|=o_p(1),
$$

following from that $\{\{T-\pi^*(\boldsymbol{X})\}\widetilde{\boldsymbol{X}}\{h^*(\boldsymbol{X}_i)-h(\boldsymbol{X}_i)\},h\in\mathcal{H}\}$ is Glivenko–Cantelli from Condition (C3). For the term I_4, by the law of large number and Condition (C3) $I_4=o_p(1)$, if $\min\{\|\hat{h}-h^*\|_\infty,\|\hat{\pi}-\pi^*\|_\infty\}=o_p(1)$.

Combining I_1 – I_4 together yields that $n^{-1}\sum_{i=1}^{n}\{T_i-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{Y_i-\hat{h}(\boldsymbol{X}_i)\}=$

$o_p(1)$. This and (4.9) imply that $\hat{\boldsymbol{\beta}}_{DR} \to_p \boldsymbol{\beta}^*$. This completes the proof of Theorem 2.2.1. □

Proof of Theorem 2.2.2: From the proof of Theorem 2.2.1, we have

$$\begin{aligned} & n^{1/2}(\hat{\boldsymbol{\beta}}_{DR} - \boldsymbol{\beta}^*) \\ = & \left(\mathbb{E}_n\left[\{T_i - \tilde{\pi}(\boldsymbol{X}_i)\}T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right] + \mathbb{E}_n\left[\{\tilde{\pi}_i - \hat{\pi}(\boldsymbol{X}_i)\}T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right]\right)^{-1} \\ & \times n^{-1/2}\sum_{i=1}^{n}\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) + \epsilon_i - \hat{h}(\boldsymbol{X}_i)\}. \end{aligned}$$

By the law of large number, $\mathbb{E}_n[\{T_i - \tilde{\pi}(\boldsymbol{X}_i)\}T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}] \to_p E[T - \tilde{\pi}(\boldsymbol{X})\}T\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{T}}]$. By Condition (C3),

$$\left\|\mathbb{E}_n\left[\{\tilde{\pi}(\boldsymbol{X}_i) - \hat{\pi}(\boldsymbol{X}_i)\}T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right]\right\| \leq \|\tilde{\pi} - \hat{\pi}\|_\infty \mathbb{E}_n\left[\left\|T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right\|\right] = o_p(1).$$

By Continuous mapping theorem,

$$\left(\mathbb{E}_n\left[\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}T_i\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^{\mathrm{T}}\right]\right)^{-1} \to_p \left(E\left[\{T - \tilde{\pi}(\boldsymbol{X})\}T\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{X}}^{\mathrm{T}}\right]\right)^{-1} = \boldsymbol{B}^{-1}(\tilde{\pi}). \tag{4.10}$$

By simple algebra, we obtain that

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\{T_i-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)+\epsilon_i-\hat{h}(\boldsymbol{X}_i)\} \\
=& n^{-1/2}\sum_{i=1}^{n}\{T_i-\tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\epsilon_i+n^{-1/2}\sum_{i=1}^{n}\{\tilde{\pi}(\boldsymbol{X}_i)-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\epsilon_i \\
& +n^{-1/2}\sum_{i=1}^{n}\{T_i-\tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)-\tilde{h}(\boldsymbol{X}_i)\} \\
& +n^{-1/2}\sum_{i=1}^{n}\{\tilde{\pi}(\boldsymbol{X}_i)-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)-\tilde{h}(\boldsymbol{X}_i)\} \\
& +n^{-1/2}\sum_{i=1}^{n}\{T_i-\tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{\tilde{h}(\boldsymbol{X}_i)-\hat{h}(\boldsymbol{X}_i)\} \\
& +n^{-1/2}\sum_{i=1}^{n}\{\tilde{\pi}(\boldsymbol{X}_i)-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{\tilde{h}(\boldsymbol{X}_i)-\hat{h}(\boldsymbol{X}_i)\} \\
=& n^{-1/2}\sum_{i=1}^{n}\{T_i-\tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\epsilon_i \\
& +n^{-1/2}\sum_{i=1}^{n}\{T_i-\tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)-\tilde{h}(\boldsymbol{X}_i)\} \\
& +n^{-1/2}\sum_{i=1}^{n}\{\tilde{\pi}(\boldsymbol{X}_i)-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)-\tilde{h}(\boldsymbol{X}_i)\} \\
& +n^{-1/2}\sum_{i=1}^{n}\{T_i-\tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{\tilde{h}(\boldsymbol{X}_i)-\hat{h}(\boldsymbol{X}_i)\}+o_p(1),
\end{aligned}
$$

where the last equality follows from Lemma 4.1.1 presented later in this section.

(i): if $\tilde{\pi}=\pi^*$, then by Lemma 4.1.1

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\{T_i-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)+\epsilon_i-\hat{h}(\boldsymbol{X}_i)\} \\
=& n^{-1/2}\sum_{i=1}^{n}\{T_i-\pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\left\{h^*(\boldsymbol{X}_i)+\epsilon_i-\tilde{h}(\boldsymbol{X}_i)\right\} \\
& +n^{-1/2}\sum_{i=1}^{n}\{\pi^*(\boldsymbol{X}_i)-\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i)-\tilde{h}(\boldsymbol{X}_i)\}+o_p(1).
\end{aligned}
$$

Following the same arguments used for (i) of Lemma 4.1.1, we can show that $\mathbb{G}_n[\{\pi^*(\boldsymbol{X}_i)-$

$\hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) - \tilde{h}(\boldsymbol{X}_i)\}] = o_p(1)$. Thus,

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) + \epsilon_i - \hat{h}(\boldsymbol{X}_i)\} \\
=& n^{-1/2}\sum_{i=1}^{n}\{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) + \epsilon_i - \tilde{h}(\boldsymbol{X}_i)\} \\
& \quad + \mathbb{G}_n\left[\{\pi^*(\boldsymbol{X}_i) - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) - \tilde{h}(\boldsymbol{X}_i)\}\right] \\
& \quad + n^{1/2}E\left[\{\pi^*(\boldsymbol{X}) - \hat{\pi}(\boldsymbol{X})\}\widetilde{\boldsymbol{X}}\{h^*(\boldsymbol{X}) - \tilde{h}(\boldsymbol{X})\}\right] + o_p(1) \\
=& n^{-1/2}\sum_{i=1}^{n}\{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) - \tilde{h}(\boldsymbol{X}_i) + \epsilon_i\} \\
& \quad + n^{1/2}E\left[\{\pi^*(\boldsymbol{X}) - \hat{\pi}(\boldsymbol{X})\}\widetilde{\boldsymbol{X}}\{h^*(\boldsymbol{X}) - \tilde{h}(\boldsymbol{X})\}\right] + o_p(1) \\
=& n^{-1/2}\sum_{i=1}^{n}\{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) + \epsilon_i - \tilde{h}(\boldsymbol{X}_i)\} + n^{-1/2}\sum_{i=1}^{n}\phi_{\tilde{h}}(T_i, \boldsymbol{X}_i) + o_p(1) \\
\to_d & N(0, \boldsymbol{\Sigma}(\tilde{h})),
\end{aligned}
$$

by the central limit theorem. This and (4.10) together imply that

$$
n^{1/2}(\hat{\boldsymbol{\beta}}_{DR} - \boldsymbol{\beta}^*) \to_d N(0, \boldsymbol{B}^{-1}(\pi^*)\boldsymbol{\Sigma}(\tilde{h})\boldsymbol{B}^{-1}(\pi^*)).
$$

(ii): if $\tilde{h} = h^*$, we obtain that

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) + \epsilon_i - \hat{h}(\boldsymbol{X}_i)\} \\
=& n^{-1/2}\sum_{i=1}^{n}\{T_i - \tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\epsilon_i + n^{-1/2}\sum_{i=1}^{n}\{T_i - \tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) - \hat{h}(\boldsymbol{X}_i)\} + o_p(1).
\end{aligned}
$$

Similarly, using the same arguments for (i) of Lemma 4.1.1, we show that $\mathbb{G}_n[\{T_i -$

$\tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) - \hat{h}(\boldsymbol{X}_i)\}] = o_p(1)$. Therefore,

$$
\begin{aligned}
& n^{-1/2}\sum_{i=1}^{n}\{T_i - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{h^*(\boldsymbol{X}_i) + \epsilon_i - \hat{h}(\boldsymbol{X}_i)\} \\
=& n^{-1/2}\sum_{i=1}^{n}\{T_i - \tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\epsilon_i + n^{1/2}E\left[\{T - \tilde{\pi}(\boldsymbol{X})\}\widetilde{\boldsymbol{X}}\{h^*(\boldsymbol{X}) - \hat{h}(\boldsymbol{X})\}\right] + o_p(1) \\
=& n^{-1/2}\sum_{i=1}^{n}\{T_i - \tilde{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\epsilon_i + n^{-1/2}\sum_{i=1}^{n}\phi_{\tilde{\pi}}(Y_i, T_i, \boldsymbol{X}_i) + o_p(1) \\
\to_d & N(0, \boldsymbol{\Sigma}(\tilde{\pi})),
\end{aligned}
$$

by the central limit theorem. Then by (4.10), we obtain that

$$n^{1/2}(\boldsymbol{\beta}_{DR} - \boldsymbol{\beta}^*) \to_d N(0, \boldsymbol{B}^{-1}(\tilde{\pi})\boldsymbol{\Sigma}(\tilde{\pi})\boldsymbol{B}^{-1}(\tilde{\pi}))$$

This completes the proof of Theorem 2.2.2. □

For a generic function $g(\cdot)$. let $\mathcal{H}_g(\eta) := \{h(\cdot) \in \mathcal{H} : \|h - g\|_\infty \leq \eta\}$. Likewise, $\Pi_g(\eta) := \{h(\cdot) \in \Pi : \|\pi - g\|_\infty \leq \eta\}$.

Lemma 4.1.1. *Under Conditions (C1)–(C3), if* $\|\hat{\pi} - \tilde{\pi}\|_\infty = o_p(n^{-\alpha_1})$, $\|\hat{h} - \tilde{h}\|_\infty = o_p(n^{-\alpha_2})$, $\mathcal{H}_{\tilde{h}}(1) \subset \mathcal{H}$, *and* $\alpha_1+\alpha_2 > 1/2$, *then* **(i)** $n^{-1/2}\sum_{i=1}^{n}\{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{\tilde{h}(\boldsymbol{X}_i) - \hat{h}(\boldsymbol{X}_i)\} = o_p(1)$; **(ii)** $n^{-1/2}\sum_{i=1}^{n}\{\tilde{\pi}(\boldsymbol{X}_i) - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{\tilde{h}(\boldsymbol{X}_i) - \hat{h}(\boldsymbol{X}_i)\} = o_p(1)$; **(iii)** $n^{-1/2}\sum_{i=1}^{n}\{\tilde{\pi}(\boldsymbol{X}_i) - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\epsilon_i = o_p(1)$.

Proof: **(i)** $\|\hat{h} - \tilde{h}\|_\infty = o_p(n^{-\alpha_2})$ implies that given any $\eta, \delta > 0$, there exists N, such that for all $n > N$, $P(\|\hat{h} - \tilde{h}\|_\infty > \eta) < \delta$.

Thus, we restrict our attention on the functional space $\mathcal{H}_{\tilde{h}}(\eta) := \{h(\cdot) : \|h - \tilde{h}\|_\infty \leq \eta\}$. Without loss of generality, we assume $\mathcal{H}_{\tilde{h}}(\eta) \subset \mathcal{H}$. Noting that for any

$h \in \mathcal{H}_{\tilde{h}}(\eta)$, $E[\{T - \pi^*(\boldsymbol{X})\}\widetilde{\boldsymbol{X}}\{\tilde{h}(\boldsymbol{X}) - h(\boldsymbol{X})\}] = 0$,

$$\begin{aligned}
&\left| n^{-1/2} \sum_{i=1}^{n} \{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{\tilde{h}(\boldsymbol{X}_i) - \hat{h}(\boldsymbol{X}_i)\} \right| \\
\leq &\sup_{h \in \mathcal{H}_{\tilde{h}}(\eta)} \eta \left| \mathbb{G}_n \left[\{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i \eta^{-1}\{\tilde{h}(\boldsymbol{X}_i) - h(\boldsymbol{X}_i)\} \right] \right| \\
\leq &\sup_{h \in \mathcal{H}_{\tilde{h}}(1)} \eta \left| \mathbb{G}_n \left[\{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{\tilde{h}(\boldsymbol{X}_i) - h(\boldsymbol{X}_i)\} \right] \right| \leq \eta M_\delta,
\end{aligned}$$

for some constant M_δ, with probability at least $1 - \delta$ uniformly over n. The last inequality follows from that under Condition (C3), $\{\mathbb{G}_n[\{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\eta^{-1}\{\tilde{h}(\boldsymbol{X}_i) - h(\boldsymbol{X}_i)\}, h \in \mathcal{H}_{\tilde{h}}(1)\}$ converges weakly to a Gaussian process indexed by h and subsequently, the sequence $\sup_{h \in \mathcal{H}_{\tilde{h}}(1)} \left| \mathbb{G}_n \left[\{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\eta^{-1}\{\tilde{h}(\boldsymbol{X}_i) - h(\boldsymbol{X}_i)\} \right] \right|$ is tight. Therefore, for all $n > N$

$$P\left(\left| n^{-1/2} \sum_{i=1}^{n} \{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{\tilde{h}(\boldsymbol{X}_i) - \hat{h}(\boldsymbol{X}_i)\} \right| > \eta M_\delta \right) < 2\delta.$$

This implies that $n^{-1/2} \sum_{i=1}^{n} \{T_i - \pi^*(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{\tilde{h}(\boldsymbol{X}_i) - \hat{h}(\boldsymbol{X}_i)\} = o_p(1)$.

(ii) By the boundedness of $\boldsymbol{X}$,

$$\begin{aligned}
&\left\| n^{-1/2} \sum_{i=1}^{n} \{\tilde{\pi}(\boldsymbol{X}_i) - \hat{\pi}(\boldsymbol{X}_i)\}\widetilde{\boldsymbol{X}}_i\{\tilde{h}(\boldsymbol{X}_i) - \hat{h}(\boldsymbol{X}_i)\} \right\| \\
\leq &n^{-1/2} \|\hat{\pi} - \tilde{\pi}\|_\infty \|\hat{h} - \tilde{h}\|_\infty \sum_{i=1}^{n} \|\widetilde{\boldsymbol{X}}_i\| = o_p(n^{-(1/2+\alpha_1+\alpha_2)}) \sum_{i=1}^{n} \|\widetilde{\boldsymbol{X}}_i\| = o_p(1).
\end{aligned}$$

(iii) The proof follows from the same arguments as used for **(i)** with $\mathcal{H}_{\tilde{h}}(\eta)$ replaced by $\Pi_{\tilde{\pi}}(\eta)$.

Thus, the proof of Lemma 4.1.1 is completed. □

Appendix 3: Heuristic arguments for the relationship between multivariate regression models and MSMs

First let us assume that the multivariate regression model with the identity link function (i.e., Model (3.7)) holds, we then have

$$\begin{aligned}
E(Y^{(d_1,d_2)}) &= E_X\{E(Y^{(d_1,d_2)}|X)\} \\
&= E_X\{E(Y^{(d_1,d_2)}|X,(D_1,D_2)=(d_1,d_2))\} \quad \text{by the weak ignorability assumption (i)} \\
&= E_X\{E(Y|X,(D_1,D_2)=(d_1,d_2))\} \quad \text{by the consistency assumption (iii)} \\
&= E\{\tilde{\mathbf{X}}^T\gamma + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2\} \quad \text{by the model (3.7)} \\
&= E\{\tilde{\mathbf{X}}^T\gamma\} + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2.
\end{aligned}$$

Taking $\tau_0 = E\{X\gamma\}$, the correct specification of the multivariate regression model (3.7) implies the MSM (3.2). Thus, the multiple variate regression model (3.7) is able to capture the causal parameters as specified in the MSM (3.2).

Let us assume that the multivariate outcome model (3.8) is the underlying true outcome model, we then have

$$\begin{aligned}
E(Y^{(d_1,d_2)}) &= E_X\{E(Y|X,(D_1,D_2)=(d_1,d_2))\} \\
&= E_X\{\tilde{\mathbf{X}}^T\gamma + \tau_1^* D_1 + \tau_2^* D_2 + \tau_{12}^* D_1 D_2 + \delta_1 X_1^2 D_1 + \delta_2 X_2^2 D_2 \\
&\qquad + \delta_3 X_3 D_1 D_2 | X,(D_1,D_2)=(d_1,d_2)\} \\
&= E_X\{\tilde{\mathbf{X}}^T\gamma + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2 + \delta_1 X_1^2 d_1 + \delta_2 X_2^2 d_2 + \delta_3 X_3 d_1 d_2\} \\
&= E_X\{\tilde{\mathbf{X}}^T\gamma\} + (\tau_1^* + \delta_1 E X_1^2) d_1 + (\tau_2^* + \delta_2 E X_2^2) d_2 + (\tau_{12}^* + \delta_3 E X_3) d_1 d_2.
\end{aligned}$$

Thus, fitting a model of form (3.7) would result in regression coefficients for d_1, d_2, and $d_1 d_2$ as $\tau_1^* + \delta_1 E X_1^2$, $\tau_2^* + \delta_2 E X_2^2$), and $\tau_{12}^* + \delta_3 E X_3$, respectively. These regression coefficients do capture the ATEs and drug interaction specified in the MSM (3.2).

Let us assume that the multivariate model (3.7) holds. Fitting a marginal regression model with treatment indicator variables only, while ignoring the confounding variables, would result in biased estimates for ATEs. This can be illustrated by

$$\begin{aligned} E(Y|(D_1, D_2) = (d_1, d_2)) &= E\{X\gamma + \tau_1^* D_1 + \tau_2^* D_2 + \tau_{12}^* D_1 D_2 | (D_1, D_2) = (d_1, d_2)\} \\ &= E\{X\gamma | (D_1, D_2) = (d_1, d_2)\} + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2. \end{aligned}$$

For the observational data, the confounding variables are associated with treatment selection, the first term is a function of d_1 and d_2. Thus the parameters in the marginal model for d_1, d_2, $d_1 d_2$ are not τ_1^*, τ_2^* and τ_{12}^* any more, and the parameters in the marginal model do not have causal interpretation.

When the multivariate logistic regression model (3.9) is applied, even though the treatment effects in logit-scale given X are the same across different X, the ATEs in logit-scale are different from those conditional treatment effects. To examine this,

$$\begin{aligned} E(Y^{(d_1,d_2)}) &= E_X\{E(Y|X, (D_1, D_2) = (d_1, d_2))\} \\ &= E_X\left\{ \frac{exp\big(\tilde{\mathbf{X}}^T\gamma + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2\big)}{1 + exp\big(\tilde{\mathbf{X}}^T\gamma + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2\big)} \right\} \\ &= \left\{ \frac{exp\big(E_X(\tilde{\mathbf{X}}^T\gamma) + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2\big)}{1 + exp\big(E_X(\tilde{\mathbf{X}}^T\gamma) + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2\big)} \right\} \\ &\quad \times \left\{ 1 + \frac{1 - exp\big(E_X(\tilde{\mathbf{X}}^T\gamma) + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2\big)}{\big(1 + exp\big(E_X(\tilde{\mathbf{X}}^T\gamma) + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2\big)\big)^2} Var(\tilde{\mathbf{X}}^T\gamma)) \right\} \end{aligned}$$

The last equation is derived from the second order Taylor expansion at $E_X(\tilde{\mathbf{X}}^T\gamma) + \tau_1^* d_1 + \tau_2^* d_2 + \tau_{12}^* d_1 d_2$. It is clear that the regression coefficients for d_1, d_2, and $d_1 d_2$ in the multivariate logistic regression model do not correspond to the the regression coefficients in the MSM (3.3). Thus, even for the homogeneous treatment effect model

(3.9), the regression coefficients for treatments do not have causal interpretation unless $(1-\tilde{p})(1-2\tilde{p})=0$, where $\tilde{p}=\frac{exp\Big(E_X(\tilde{\mathbf{X}}^T\gamma)+\tau_1^*d_1+\tau_2^*d_2+\tau_{12}^*d_1d_2\Big)}{1+exp\Big(E_X(\tilde{\mathbf{X}}^T\gamma)+\tau_1^*d_1+\tau_2^*d_2+\tau_{12}^*d_1d_2\Big)}$.

Appendix 4: Supplementary table and figures

This section includes the additional figures and tables in Chapters 2-4.

Table A1.1: Summarized statistics for the estimated ATEs and drug interactions based on 1000 simulated datasets for continuous outcomes with fixed $\tau_{12}^* = 0$ and sample size $n = 500$.

		UW	W-All	W-C	W-CI	W-CP	W-Lasso	Reg. model
Homogeneous treatment effect for continuous outcome (Model (3.7)): The true ATEs and drug interaction were (1.000, 1.000, 0.000).								
τ_1	Est.	1.405	0.992	1.025	1.022	0.996	0.997	1.000
	MSE	0.493	0.051	0.164	0.177	0.030	0.030	0.004
	SE	0.562	0.300	0.402	0.423	0.220	0.287	0.064
	E.SD	0.574	0.226	0.405	0.420	0.175	0.174	0.064
	CR	0.880	0.982	0.937	0.941	0.973	0.994	0.951
τ_2	Est.	1.982	0.997	1.018	1.011	1.009	1.008	1.002
	MSE	1.296	0.047	0.160	0.179	0.027	0.027	0.004
	SE	0.566	0.300	0.399	0.423	0.213	0.285	0.065
	E.SD	0.576	0.217	0.399	0.423	0.163	0.165	0.065
	CR	0.587	0.983	0.951	0.951	0.976	0.996	0.941
τ_{12}	Est.	-0.398	-0.001	-0.021	-0.015	-0.006	-0.006	-0.002
	MSE	0.859	0.089	0.336	0.369	0.054	0.054	0.008
	SE	0.799	0.420	0.563	0.594	0.305	0.399	0.091
	E.SD	0.838	0.299	0.580	0.608	0.232	0.232	0.091
	CR	0.912	0.985	0.947	0.951	0.980	0.997	0.949
Heterogeneous treatment effect for continuous outcome (Model (3.8)): The true ATEs and drug interaction were (1.999, 2.000, -0.001).								
τ_1	Est.	2.401	1.988	2.022	2.017	1.993	1.990	2.010
	MSE	0.503	0.065	0.174	0.189	0.041	0.057	0.019
	SE	0.580	0.327	0.421	0.442	0.253	0.318	0.152
	E.SD	0.592	0.259	0.421	0.438	0.210	0.244	0.151
	CR	0.882	0.976	0.944	0.950	0.972	0.975	0.944
τ_2	Est.	2.984	1.995	2.016	2.008	2.007	1.998	1.948
	MSE	1.314	0.058	0.170	0.190	0.038	0.051	0.022
	SE	0.587	0.320	0.415	0.439	0.242	0.311	0.152
	E.SD	0.594	0.250	0.415	0.438	0.206	0.237	0.154
	CR	0.602	0.976	0.944	0.957	0.962	0.981	0.929
τ_{12}	Est.	-0.304	0.008	-0.012	-0.006	0.004	0.007	0.072
	MSE	0.915	0.148	0.373	0.412	0.105	0.135	0.081
	SE	0.871	0.490	0.607	0.640	0.386	0.477	0.281
	E.SD	0.909	0.387	0.612	0.643	0.327	0.370	0.280
	CR	0.921	0.977	0.948	0.950	0.975	0.977	0.942

Table A1.2: Summarized statistics for the estimated ATEs and drug interactions based on 1000 simulated data sets for continuous outcomes with fixed $\tau_{12}^* = 0$ and sample size $n = 5000$.

		UW	W-All	W-C	W-CI	W-CP	W-Lasso	Reg. model
Homogeneous treatment effect for binary outcome (Model (3.7)): The true ATEs and drug interaction were (1.000, 1.000, 0.000).								
τ_1	Est.	1.374	1.001	0.992	0.993	1.000	1.000	1.000
	MSE	0.168	0.003	0.014	0.015	0.002	0.002	0.000
	SE	0.178	0.054	0.121	0.124	0.045	0.049	0.020
	E.SD	0.168	0.051	0.118	0.121	0.042	0.042	0.020
	CR	0.452	0.952	0.949	0.957	0.963	0.976	0.940
τ_2	Est.	1.953	1.000	0.995	0.995	1.000	1.000	1.000
	MSE	0.939	0.002	0.014	0.015	0.002	0.002	0.000
	SE	0.179	0.051	0.120	0.124	0.041	0.045	0.020
	E.SD	0.174	0.048	0.119	0.123	0.039	0.039	0.021
	CR	0.001	0.957	0.955	0.949	0.951	0.968	0.935
τ_{12}	Est.	-0.361	0.000	0.013	0.012	0.001	0.001	0.001
	MSE	0.191	0.004	0.028	0.029	0.003	0.003	0.001
	SE	0.253	0.071	0.170	0.174	0.060	0.064	0.029
	E.SD	0.247	0.067	0.167	0.170	0.057	0.057	0.030
	CR	0.694	0.966	0.949	0.955	0.952	0.966	0.937
Heterogeneous treatment effect for continuous outcome (Model (3.8)): The true ATEs and drug interaction were (1.999, 2.001, 0.000).								
τ_1	Est.	2.371	2.000	1.991	1.992	1.999	2.000	2.016
	MSE	0.168	0.004	0.015	0.016	0.003	0.003	0.002
	SE	0.184	0.067	0.127	0.131	0.059	0.066	0.047
	E.SD	0.172	0.064	0.124	0.126	0.057	0.061	0.047
	CR	0.477	0.949	0.953	0.962	0.952	0.962	0.924
τ_2	Est.	2.958	2.002	1.997	1.997	2.002	2.002	1.952
	MSE	0.947	0.003	0.015	0.016	0.002	0.003	0.004
	SE	0.185	0.063	0.125	0.129	0.055	0.062	0.047
	E.SD	0.179	0.061	0.124	0.130	0.053	0.058	0.047
	CR	0.002	0.952	0.951	0.944	0.955	0.957	0.814
τ_{12}	Est.	-0.276	0.002	0.014	0.014	0.002	0.003	0.067
	MSE	0.147	0.009	0.032	0.034	0.008	0.009	0.012
	SE	0.276	0.102	0.184	0.189	0.093	0.101	0.089
	E.SD	0.267	0.097	0.180	0.184	0.090	0.094	0.087
	CR	0.836	0.950	0.952	0.951	0.952	0.951	0.889

Table A1.3: Summarized statistics for the estimated ATEs and drug interactions based on 1000 simulated data sets for binary outcomes with fixed $\tau_{12}^* = 0$ and sample size $n = 500$.

		UW	W-All	W-C	W-CI	W-CP	W-Lasso	Reg. model
Homogeneous treatment effect for binary outcome (Model (3.9)): The true ATEs and drug interaction were (1.160, 1.156, 0.134).								
τ_1	Est.	1.052	1.159	1.161	1.160	1.160	1.158	2.111
	MSE	0.098	0.071	0.077	0.079	0.066	0.070	1.106
	SE	0.281	0.256	0.259	0.267	0.240	0.254	0.458
	E.SD	0.275	0.249	0.257	0.263	0.238	0.246	0.437
	CR	0.940	0.953	0.950	0.954	0.943	0.955	0.474
τ_2	Est.	1.136	1.179	1.189	1.190	1.179	1.178	2.137
	MSE	0.095	0.073	0.081	0.084	0.068	0.071	1.171
	SE	0.284	0.258	0.261	0.270	0.241	0.257	0.458
	E.SD	0.288	0.252	0.264	0.269	0.242	0.249	0.449
	CR	0.948	0.944	0.941	0.949	0.935	0.950	0.434
τ_{12}	Est.	0.393	0.165	0.159	0.161	0.163	0.167	0.012
	MSE	0.361	0.259	0.268	0.277	0.242	0.255	0.526
	SE	0.809	0.796	0.790	0.804	0.770	0.793	1.001
	E.SD	0.515	0.485	0.495	0.503	0.467	0.480	0.696
	CR	0.936	0.959	0.957	0.957	0.957	0.964	0.964
Heterogeneous treatment effect for binary outcome (Model (3.10)): The true ATEs and drug interaction were (1.639, 1.632, -0.259).								
τ_1	Est.	1.520	1.656	1.655	1.655	1.652	1.655	2.688
	MSE	0.118	0.092	0.094	0.098	0.083	0.089	1.332
	SE	0.308	0.290	0.292	0.299	0.274	0.288	0.494
	E.SD	0.302	0.285	0.285	0.294	0.269	0.280	0.468
	CR	0.923	0.947	0.949	0.945	0.95	0.953	0.446
τ_2	Est.	1.659	1.666	1.674	1.676	1.663	1.665	2.737
	MSE	0.114	0.095	0.102	0.106	0.088	0.093	1.464
	SE	0.320	0.302	0.302	0.312	0.283	0.300	0.508
	E.SD	0.316	0.290	0.297	0.304	0.276	0.285	0.483
	CR	0.942	0.945	0.951	0.945	0.944	0.947	0.428
τ_{12}	Est.	0.070	-0.179	-0.196	-0.190	-0.186	-0.181	-0.626
	MSE	1.306	1.200	1.187	1.215	1.156	1.181	1.623
	SE	2.056	2.071	2.066	2.075	2.055	2.069	2.255
	E.SD	1.067	1.065	1.060	1.071	1.045	1.056	1.197
	CR	0.963	0.966	0.968	0.970	0.971	0.969	0.953

Table A1.4: Summarized statistics for the estimated ATEs and drug interactions based on 1000 simulated datasets for binary outcomes with fixed $\tau_{12}^{*} = 0$ and sample size $n = 5000$.

		UW	W-All	W-C	W-CI	W-CP	W-Lasso	Reg. model
Homogeneous treatment effect for binary outcome (Model (3.9)): The true ATEs and drug interaction were (1.155, 1.157, 0.131).								
τ_1	Est.	1.040	1.157	1.154	1.155	1.156	1.157	2.009
	MSE	0.022	0.006	0.007	0.008	0.006	0.006	0.746
	SE	0.087	0.073	0.079	0.080	0.072	0.073	0.124
	E.SD	0.089	0.076	0.082	0.083	0.075	0.076	0.129
	CR	0.734	0.945	0.938	0.938	0.936	0.945	0.000
τ_2	Est.	1.103	1.159	1.158	1.159	1.159	1.160	2.017
	MSE	0.012	0.006	0.007	0.008	0.006	0.006	0.755
	SE	0.087	0.074	0.079	0.081	0.072	0.074	0.125
	E.SD	0.087	0.073	0.080	0.081	0.072	0.073	0.124
	CR	0.894	0.946	0.945	0.942	0.942	0.948	0.000
τ_{12}	Est.	0.381	0.130	0.132	0.131	0.131	0.130	-0.012
	MSE	0.092	0.023	0.026	0.026	0.022	0.023	0.063
	SE	0.158	0.142	0.148	0.151	0.139	0.142	0.200
	E.SD	0.158	0.140	0.149	0.151	0.138	0.140	0.198
	CR	0.651	0.949	0.946	0.946	0.949	0.950	0.901
Heterogeneous treatment effect for binary outcome (Model (3.10)): The true ATEs and drug interaction were (1.634, 1.636, -0.277).								
τ_1	Est.	1.494	1.637	1.635	1.636	1.636	1.637	2.538
	MSE	0.029	0.008	0.008	0.009	0.007	0.008	0.835
	SE	0.094	0.083	0.088	0.089	0.081	0.083	0.130
	E.SD	0.094	0.083	0.089	0.091	0.082	0.083	0.128
	CR	0.685	0.947	0.937	0.940	0.942	0.947	0.000
τ_2	Est.	1.614	1.633	1.632	1.632	1.633	1.633	2.578
	MSE	0.011	0.009	0.009	0.010	0.008	0.009	0.907
	SE	0.097	0.085	0.090	0.092	0.083	0.085	0.133
	E.SD	0.099	0.088	0.092	0.094	0.086	0.088	0.136
	CR	0.935	0.947	0.943	0.938	0.944	0.949	0.000
τ_{12}	Est.	0.019	-0.265	-0.263	-0.264	-0.264	-0.264	-0.650
	MSE	0.132	0.041	0.042	0.045	0.038	0.040	0.201
	SE	0.194	0.182	0.186	0.189	0.179	0.182	0.235
	E.SD	0.195	0.187	0.190	0.196	0.181	0.186	0.237
	CR	0.666	0.938	0.941	0.937	0.943	0.939	0.643

Table A1.5: The estimated ATEs and drug interactions for opioids and statins use based on COVID-19 data set and weights trimmed to the range of $(\frac{1}{0.95}, \frac{1}{0.05})$.

	Biomarkers/ Parameters	MSM with IPTW-All Est	SE	P-Value*
Ct	τ_s	-0.809	0.844	0.338
	τ_o	0.687	1.798	0.703
	τ_{so}	-0.768	1.996	0.701
	$\tau_s + \tau_o + \tau_{so}$	-0.890	1.259	0.481
Neutro.	τ_s	3.841	1.181	0.001
	τ_o	-3.609	3.465	0.299
	τ_{so}	1.413	4.646	0.761
	$\tau_s + \tau_o + \tau_{so}$	1.646	3.080	0.594
Lymph.	τ_s	-3.122	1.035	0.003
	τ_o	2.454	2.658	0.357
	τ_{so}	-2.326	3.638	0.523
	$\tau_s + \tau_o + \tau_{so}$	-2.994	2.333	0.201
aPTT	τ_s	4.766	1.670	0.005
	τ_o	-0.071	0.850	0.934
	τ_{so}	-6.536	2.312	0.005
	$\tau_s + \tau_o + \tau_{so}$	-1.840	1.766	0.299
Procal.	τ_s	1.427	2.612	0.585
	τ_o	-1.443	0.730	0.049
	τ_{so}	-0.084	2.746	0.976
	$\tau_s + \tau_o + \tau_{so}$	-0.099	1.474	0.947

* The p-values were obtained from the Wald tests without Bonferroni correction. The Bonferroni corrected p-values would be the minimal value of 1 and 5 times of the p-values reported.

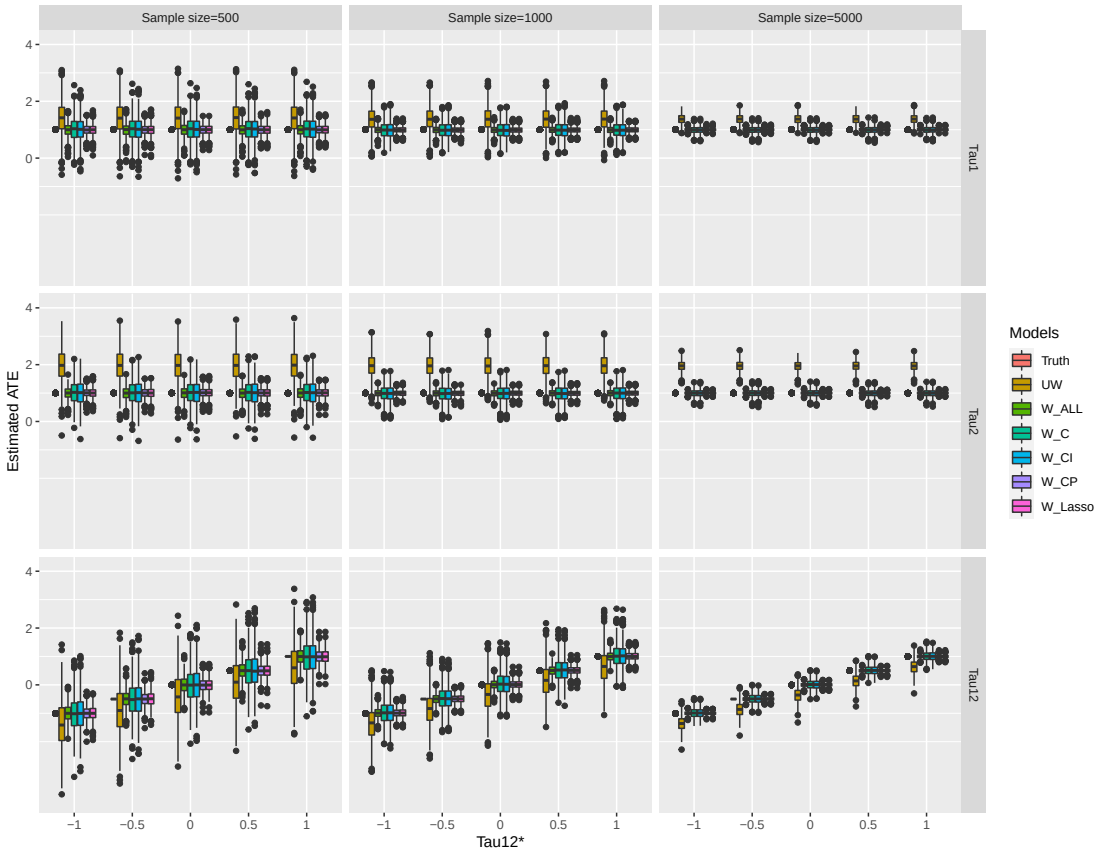

Figure A1.1: Boxplots of 1000 estimated ATEs and drug interactions for continuous outcome with homogeneous treatment effects and different sample sizes. The first row and the second row, respectively, showed the estimated ATEs for drug 1 (*i.e.* $\hat{\tau}_1$) and drug 2 (*i.e.* $\hat{\tau}_2$), with different specification of τ_{12}^*. The third row showed the estimated τ_{12} to capture drug interaction. In each block (*i.e.* for a fixed τ_{12}^*), the first boxplot was the true ATE or drug interaction parameter, the second boxplot was estimates according to MSMs without IPTW, and the third to the seventh boxplots were estimates according to MSMs with IPTW with weights being estimated from the five different sets of covariates: (i) all covariates, (ii) confounders only, (iii) confounders and instrumental variables, (iv) confounders and predictors, and (v) covariates selected by Lasso.

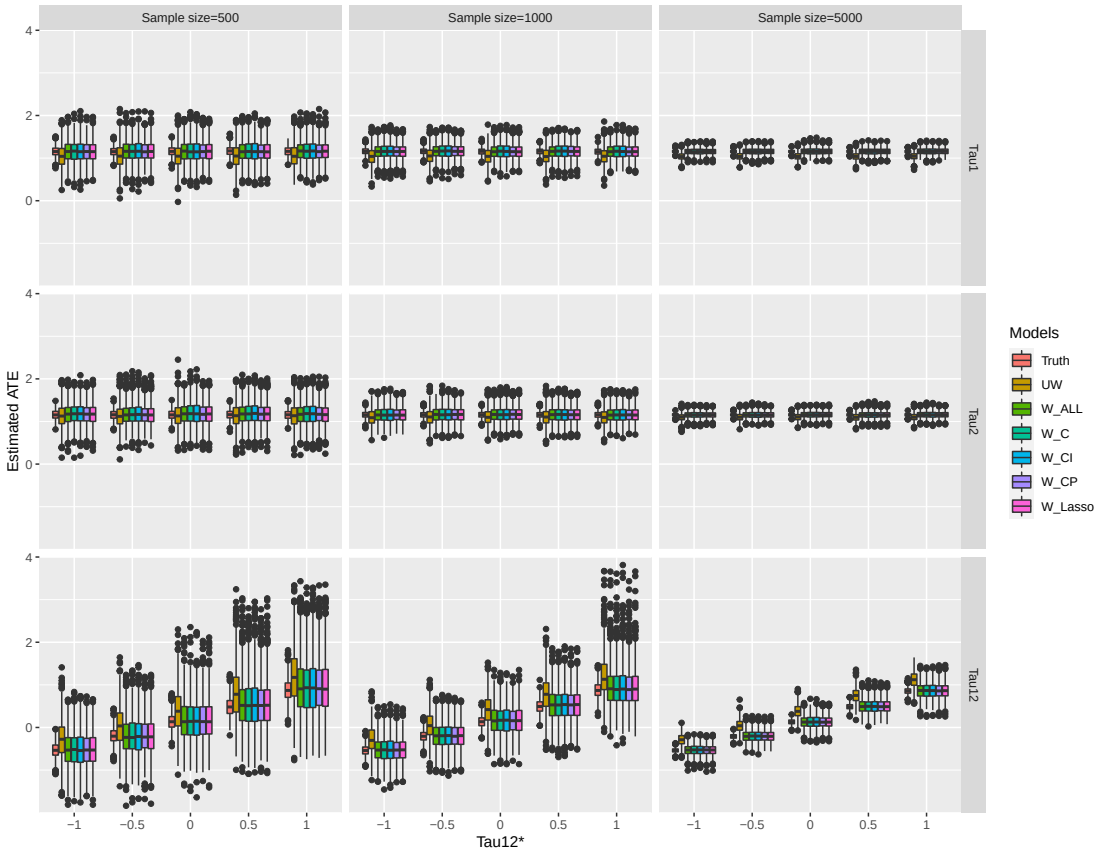

Figure A1.2: Boxplots of 1000 estimated ATEs and drug interactions for binary outcome with homogeneous treatment effects and different sample sizes. The first row and the second row, respectively, showed the estimated ATEs (odds ratio) in log scale for drug 1 (*i.e.* $\hat{\tau}_1$) and drug 2 (*i.e.* $\hat{\tau}_2$), with different specification of τ_{12}^*. The third row showed the estimated τ_{12} to capture drug interaction. In each block (*i.e.* for a fixed τ_{12}^*), the first boxplot was the true ATE or drug interaction parameter in log scale, the second boxplot was estimates according to MSMs without IPTW, and and the third to the seventh boxplots were estimates according to MSMs with IPTW with weights being estimated from the five different sets of covariates: (i) all covariates, (ii) confounders only, (iii) confounders and instrumental variables, (iv) confounders and predictors, and (v) covariates selected by Lasso.

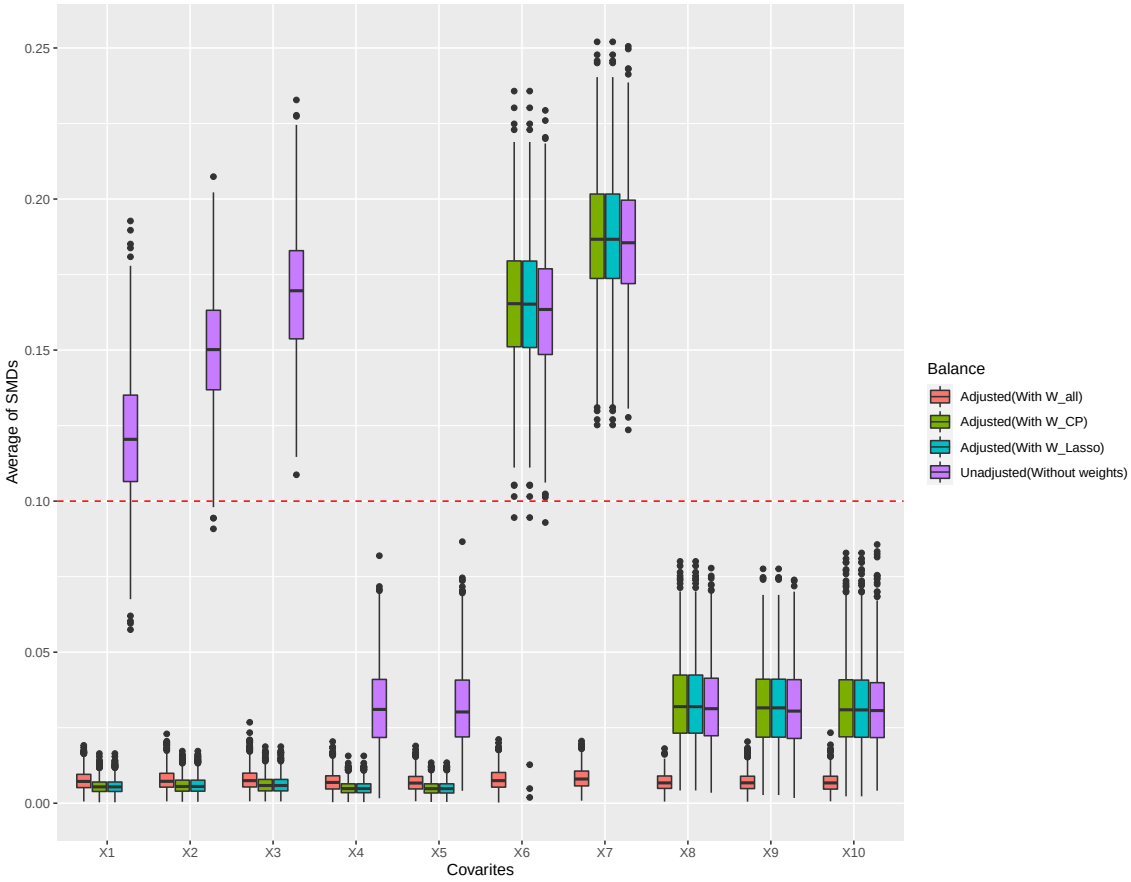

Figure A1.3: Boxplots of the average of standard mean difference (SMD) of each variable among all pairs of the four treatment groups based on 1000 simulated datasets (n=5000). The intercept of horizon line is 0.1.

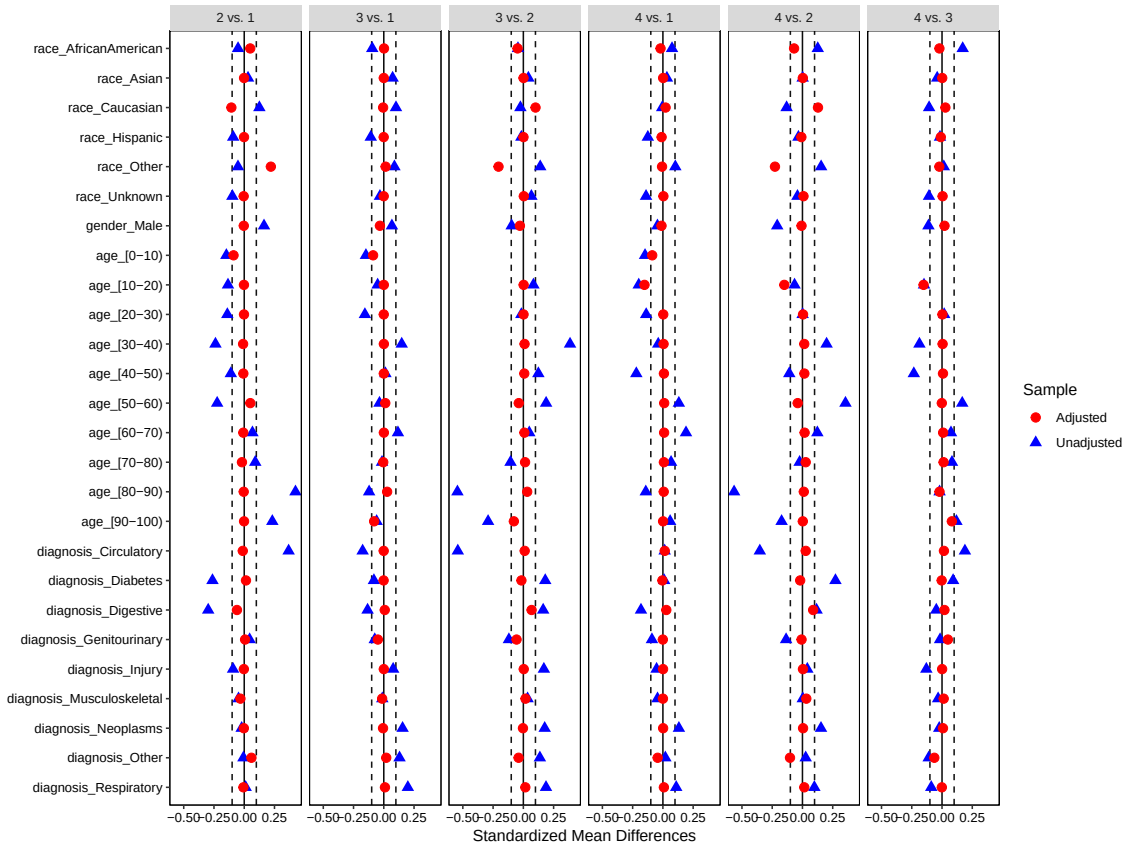

Figure A1.4: Covariates balance diagnose for the study of glyburide and metformin on diabetic patients, where 1 denotes control group, 2 denotes glyburide group, 3 denotes metformin group, and 4 denotes the combination of glyburide and metformin.

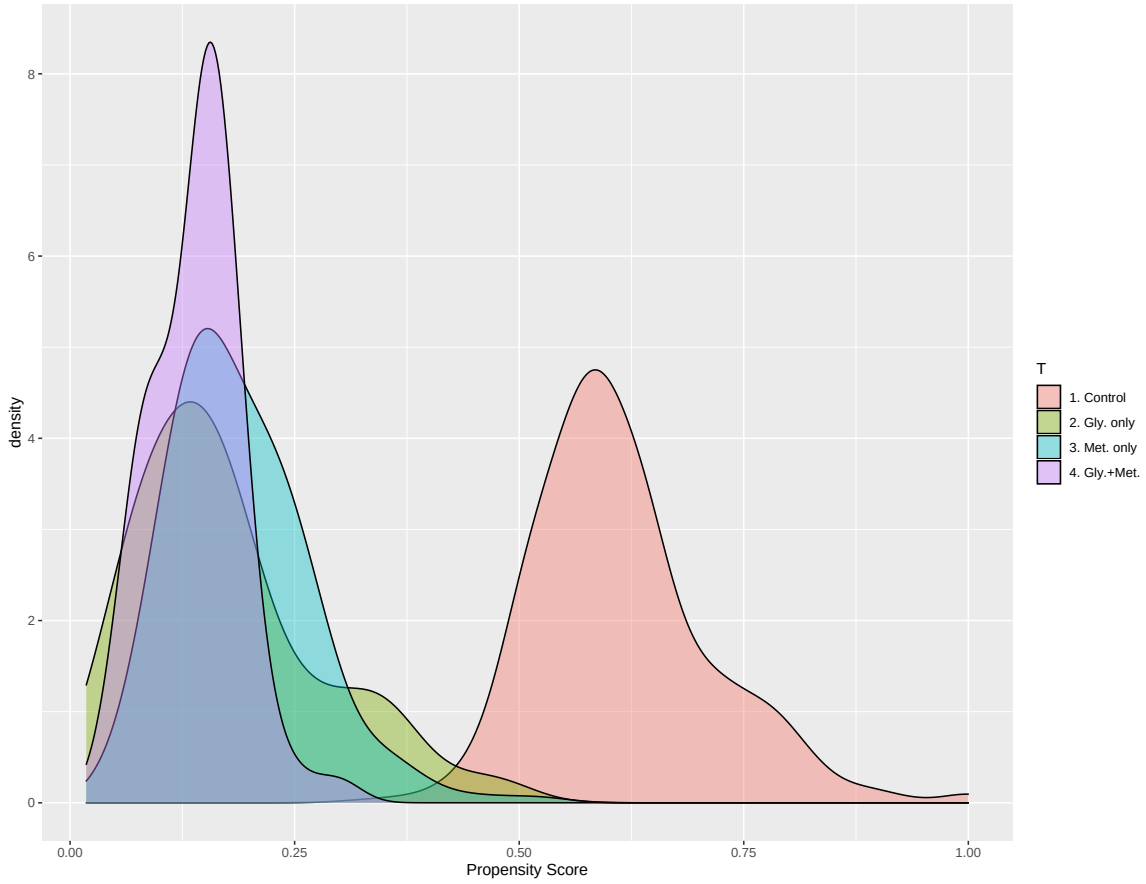

Figure A1.5: The histogram of generalized propensity scores for the study of glyburide and metformin on diabetic patients

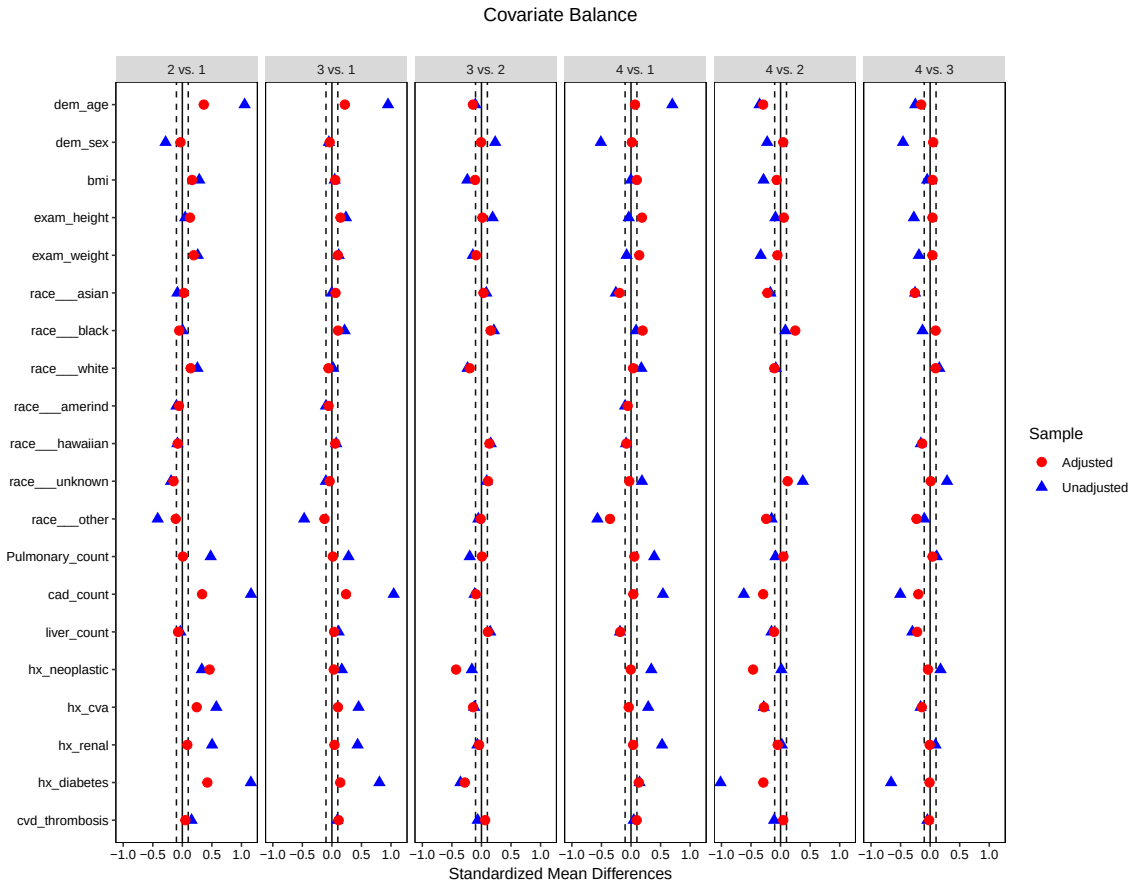

Figure A1.6: Covariates balance diagnose for the study of antecedent statin and opioid use for hospitalized COVID-19 patients, where 1 denotes control group, 2 denotes statins group, 3 denotes opioids group, and 4 denotes the combination of statins and opioids.

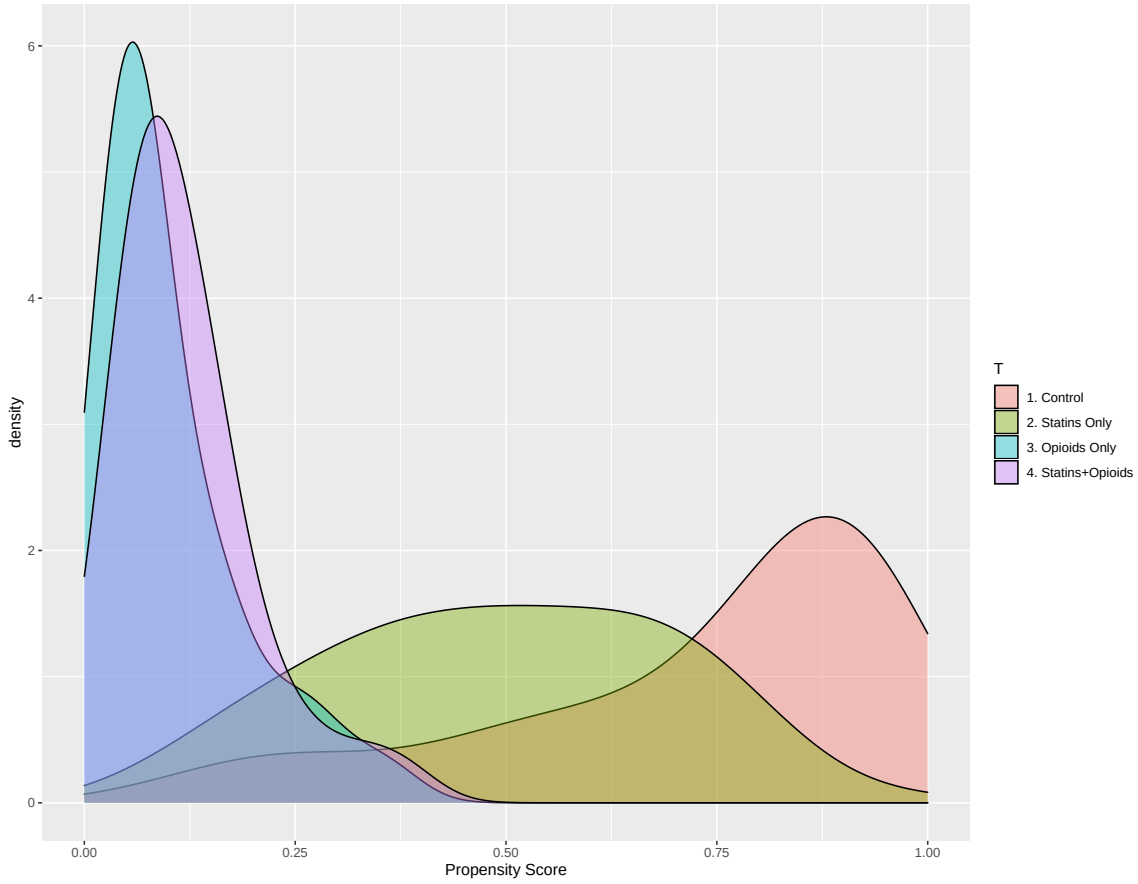

Figure A1.7: The histogram of generalized propensity scores for the study of antecedent statin and opioid use for hospitalized COVID-19 patients.

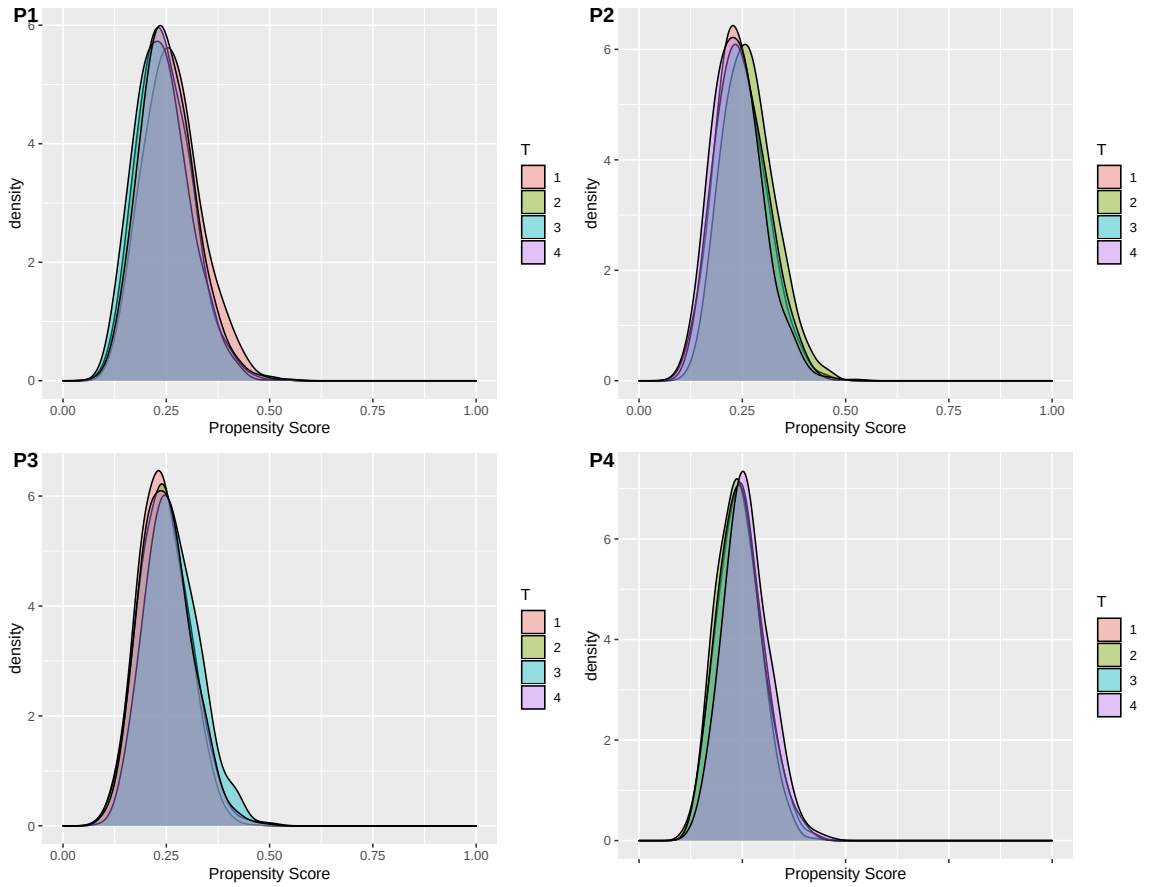

Figure A1.8: The histogram of generalized propensity scores of a sample (n=5000) with good overlap based on the generalized GPS model (11) with specified coefficients in the simulation setting in Section 3.3.1.

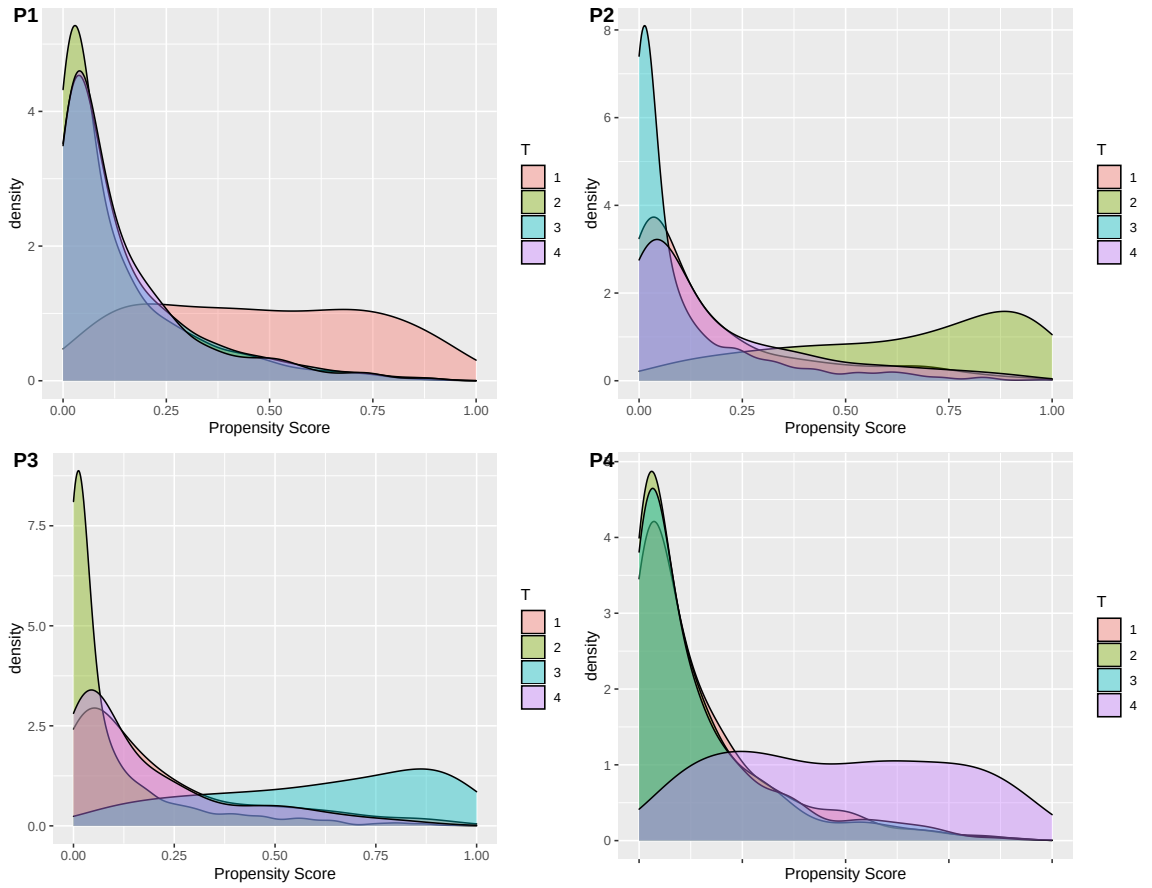

Figure A1.9: The histogram of generalized propensity scores of a sample (n=5000) with bad overlap based on the generalized GPS model (11) with specified coefficients in the discussion in Section 3.5.

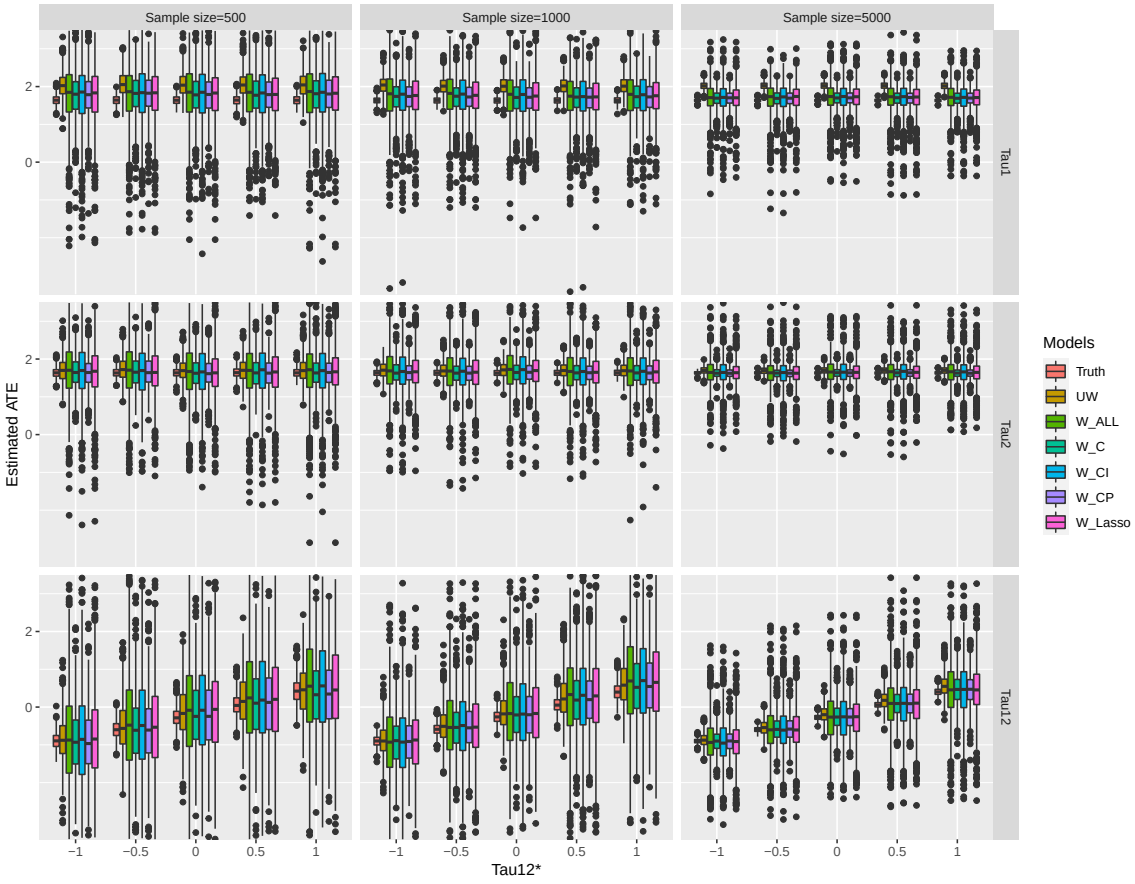

Figure A1.10: The boxplots of the simulation results for the binary heterogenous treatment effect outcome model with bad overlap for independent covariates.

CURRICULUM VITA

NAME **Qian Xu**

EDUCATION

Ph.D. in Biostatistics, *University of Louisville*, Louisville KY. 2018-2022

M.S. in Statistics, *The George Washington University*, Washington DC, 2013-2015

B.S. in Mathematics, *China Agricultural University*, Beijing, 2009-2013

EMPLOYMENT

Research Assistant, Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY, 2018-2022.

Statistician, Amarex Clinical Research LLC, Germantown, MD, 2015-2018

PUBLICATIONS

Qian Xu, Demetra Antimisiaris, Maiying Kong. (2022+). **Statistical Methods for Assessing Drug Interactions Using Observational Data**. *Journal of Applied Statistics* (Revised, Under Review).

Qian Xu, John Craycroft, Qi Zheng, Demetra Antimisiaris, Maiying Kong. (2022+). **Doubly Robust Approach for Identifying Effect Modifiers and Estimating Optimal Treatment Based on Observational Data.** (To be submitted).

Lynn Roser, Harideep Samanapally, Tshura Ali, Qian Xu, Vidyulata Salunkhe, Fnu

Deepti, Trevor McGuffin, Stephen Furmanek, Alex Glynn, Julio Ramirez, Christopher Jones, Forest W Arnold, Sean Clifford, Siddharth Pahwa, Maiying Kong, Jiapeng Huang. (2022+). **Comparing the effect of in-hospital cardiovascular events on clinical outcomes among adults hospitalized with SARS-CoV-2 Community-Acquired Pneumonia at two different waves of COVID-19**. (To be submitted).

Stephen Furmanek, Vidyulata Salunkhe, Siddharth Pahwa, Harideep Samanapally, Pavani Nathala, Qian Xu, Tshura Ali, Fnu Deepti, Alex Glynn, Trevor McGuffin, Derek Titus, Ian Farah, Christopher Jones, Julio Ramirez, Sean Clifford, Forest Arnold, Maiying Kong, Lynn Roser, Jiapeng Huang. (2022+). **Echocardiographic Features and Troponin Levels Are Associated with Survival Time in Hospitalized COVID-19 Patients with Cardiovascular Events**. *Journal of Cardiothoracic and Vascular Anesthesia* (Under Review).

Marc Fromer, Qian Xu, Tyler Mouw, Maiying Kong, John Myers, Yana Feygin, Indranil Ghosh, Robert Martin II, Kelly McMasters, Prejesh Philips, Charles Scoggins, C Tyler Ellis, Michael Egger. (2022+). **Diagnosis of Pancreatic Cancer After Cholecystectomy in the Elderly.** *Journal of the American College of Surgeons* (Under Review).

Michael Egger, Yana Feygin, Anton Khodakov, C. Tyler Ellis, Indranil Ghosh, Maiying Kong, Qian Xu, Vivek Sharma, Rebecca Redman, Adam Rojan, John Myers, Kelly McMasters. (2022+). **Utilization of second and third line chemotherapy regimens at endof life in aggressive stage IV gastrointestinal cancers**. *Cancer Medicine* (Under Review).

Grace Kuravackel, Emily Roth, Heidi Cooleycook, Qian Xu, Maiying Kong. (2022+).

Problem Behavior in Autism Spectrum Disorder and its Effect on Family. *Journal of Family Issues* (Under Review).

Qian Xu, Yang Liu, Darius Cepulis, Ann Jerde, Rachel Sheppard, Wesley Reichle, LaShawn Scott, Leah Oppy, Gina Stevenson, Sarah Bishop, Sean Clifford, Peng Liu, Maiying Kong, Jiapeng Huang. (2022). **Hand Hygiene Behaviors Monitored by an Electronic System in the Intensive Care Unit-A Prospective Observational Study**. *American Journal of Infection Control* (In Press).

Pavani Nathala, Vidyulata Salunkhe, Harideep Samanapally, Qian Xu, Stephen Furmanek, Omar Fahmy, Fnu Deepti, Alex Glynn, Trevor McGuffin, Dylan Goldsmith, Jessica Petrey, Tshura Ali, Derek Titus, Ruth Carrico, Julio Ramirez, Demetra Antimisiaris, Sean Clifford, Siddharth Pahwa, Lynn Roser, Maiying Kong, Jiapeng Huang. (2022). **Electrocardiographic Changes and Their Outcome Correlations in Hospitalized COVID-19 Patients**. *Journal of Cardiothoracic and Vascular Anesthesia* (In Press).

Rachel Speer, Idoia Meaza, Jennifer Toyoda, Yuan Lu, Qian Xu, Ronald Walter, Maiying Kong, Haiyan Lu, J. Calvin Kouokam, John Pierce Wise. (2022). **Particulate hexavalent chromium alters microRNAs in human lung cells that target key carcinogenic pathways**. *Toxicology and Applied Pharmacology* 438, 115890.

Jamie Young, Matthew Cave, Qian Xu, Maiying Kong, Jianxiang Xu, Qian Lin, Yi Tan, Lu Cai. (2022). **Whole Life Exposure to Low Dose Cadmium Alters Diet-Induced NAFLD**. *Toxicology and Applied Pharmacology* 436, 115855.

Qian Xu, Yang Liu, Darius Cepulis, Ann Jerde, Rachel A. Sheppard, Kaitlin Tretter, Leah Oppy, Gina Stevenson, Sarah Bishop, Sean P. Clifford MD, Peng Liu, Maiying

Kong, Jiapeng Huang. (2021). **Implementing an Electronic Hand Hygiene System Improved Compliance in the Intensive Care Unit.** *American Journal of Infection Control.* 49(12), P1535-1542.

Qian Xu, Harideep Samanapally, Pavani Nathala, Vidyulata Salunkhe, Stephen Furmanek, Meredith N. Cahill, Trevor McGuffin, Tahboub Mohammad, Bradford Marsili, Jessica Petrey, Ruth Carrico, Julio Ramirez, Ozan Akca, Sean P. Clifford, Siddharth Pahwa, Lynn Roser, Maiying Kong, Jiapeng Huang. (2021). **Outcomes and Risk Factors for Cardiovascular Events in Hospitalized COVID-19.** *Journal of Cardiothoracic and Vascular Anesthesia.* 35(12), P3581-3593.

Bin Zhou, Adrienne Gentry, Qian Xu, Jamie L. Young, Xiaofang Yan, Kelly Pagidas, Yu Yang, Walter H. Watson, Maiying Kong, Lu Cai, and Jonathan H. Freedman. (2021). **Effects of Cadmium and High-Fat Diet on Essential Metal Concentration in the Mouse Testis**. *Toxicology Reports.* 8, P718–723.

Kaicheng Song, Qian Xu, Heidi M. Koenig, Sean P. Clifford, Maiying Kong, Mark S. Slaughter, Yuguang Huang, Jiapeng Huang. (2020). **Validation of a Novel NeurOsTM Cerebral Oximetry Monitor Against the INVOSTM Monitor During Cardiac Surgery**. *Journal of Cardiothoracic and Vascular Anesthesia.* 35(7), P2009-2018.

John Mazzocco, Rekha Jagadapillai, Evelyne Gozal, Maiying Kong, Qian Xu, Gregory Barnes, Jonathan Freedman. (2020). **Disruption of Essential Metal Homeostasis in the Brain by Cadmium and High-Fat Diet**. *Toxicology Reports.* 7, P1164-1169.

Kaicheng Song, Qian Xu, Victoria H. Knott, Chen B. Zhao, Maiying Kong, Mark S.

Slaughter, Yuguang Huang, Jiapeng Huang. (2020). **Liposomal Bupivacaine Based Erector Spinae Block for Cardiac Surgery**. *Journal of Cardiothoracic and Vascular Anesthesia*. 35(5), P1555-1559.

Qian Xu, Maiying Kong, Jiapeng Huang. (2019). **Letter to Editor in Response to "Regional Cerebral Oxygen Saturation and Mortality in Patients with Left Ventricular Assist Device"**. *Journal of Cardiothoracic and Vascular Anesthesia*. 33(7), P2107-2108.

Xiaotong Qin, Liming Yang, Qian Xu, Yannan Chao. (2013). **Least square based minimum class variance twin support vector machines**. *Journal of Computational Information System*. 17, P6929-6936.

PRESENTATIONS

Statistical Methods for Assessing Drug Interactions Using Observational Data, *Kentucky Chapter of the American Statistical Association 2021 Spring Meeting*, Apr 2021.

Statistical Methods for Assessing Drug Interactions Using Observational Data, *2021 Joint Statistical Meetings*, Aug 2021.

Doubly Robust Approach for Identifying Effect Modifiers and Estimating Optimal Treatment Based on Observational Data, *2021 Southern Regional Council on Statistics Summer Research Conference*, Jekyll Island, Georgia, Oct 2021.

Doubly Robust Approach for Identifying Effect Modifiers and Estimating Optimal Treatment Based on Observational Data, *Eastern North American Region - ENAR 2022 Spring Meeting*, Houston, Texas, Mar 2022.

HONORS AND AWARDS

University Fellowship, *Graduate School, University of Louisville*, 2018-2020

Graduate Research Assistantship, *Department of Bioinformatics and Biostatistics, University of Louisville*, 2020-2022

School of Public Health and Information Sciences Scholarship, *School of Public Health and Information Sciences, University of Louisville*, Sep 2021

Harshbarger Travel Award, *Southern Regional Council on Statistics*, Oct 2021

Winter & Spring 2022 GSC Travel Award, *Graduate School, University of Louisville*, Feb 2022

Winter & Spring 2022 GSC Travel Award, *Graduate School, University of Louisville*, Mar 2022

Ph.D. Dean's Award, *School of Public Health and Information Sciences, University of Louisville*, May 2022

Graduate Dean's Citation, *Graduate School, University of Louisville*, May 2022

CERTIFICATES

SAS Certified Base Programmer for SAS 9, *SAS*. Aug 2014

SAS Certified Advanced Programmer for SAS 9, *SAS*. Feb 2015

SAS Certified Clinical Trials Programmer Using SAS 9, *SAS*. May 2020