TOWARDS DEVELOPMENT OF BEST PRACTICE METHODS OF CAUSAL INFERENCE TO ASSESS TREATMENT SELECTION BIOMARKERS FROM NON-RANDOMIZED DATA

by

HULYA KOCYIGIT

(Under the Direction of Kevin K. Dobbin)

ABSTRACT

In this dissertation, we present three novel contributions, providing a new methodology, examining the proposed method's performances, and extensive the study in literature. The first paper of this dissertation focuses on statistical methods for developing biomarkers that provide integration of reliable indicators of effectiveness for guiding adjuvant chemotherapy treatment selection for cases utilizing the tumor's biological makeup. When we directly attempt to evaluate a biomarker’s performance without considering the influence of covariates on treatment assignment, the result can lead to inaccurate evaluation of biomarker performance. To minimize the influence of covariates on treatment, outcome, or both, that can produce bias, we have employed various causal inference methods in a lung cancer dataset. Chapter 3 aims to present the general framework for the treatment selection process in literature, consisting of the intersection of machine learning, causal inference, and biomarkers. We use parametric, and machine learning techniques to estimate propensity scores and then apply pair matching techniques that rely on these scores to adjust the existence of extraneous factors. Different associations between treatment or outcome and covariates are studied and assessed in terms of results in outcome models. After that, we use the results of parametric and machine learning methods to evaluate biomarkers that may be used to identify patients who will benefit from a specific treatment from observational data. In chapter 4, the positivity assumption, which states that the propensity score must be constrained away from 0 and 1, is a crucial criterion for inverse probability weighting estimation. However, when the positivity assumption is violated in propensity score distributions between treatment groups, some weights can be approximately 0 and 1. These weights led to uncertainty, bias and large variance in estimators. We study various techniques to eliminate poor overlap. We propose different levels of nonoverlap scenarios to examine the performance of balance weighting family and generalized propensity score matching across true propensity model and misspecified propensity score models in multiple treatment cases. We present results of different methods of variance estimation when estimating the causal effect.

INDEX WORDS: Treatment Selection, Biomarker, Machine Learning, Balance Weighting, Lack of Overlap

TOWARDS DEVELOPMENT OF BEST PRACTICE METHODS OF CAUSAL INFERENCE TO ASSESS TREATMENT SELECTION BIOMARKERS FROM NON-RANDOMIZED DATA

by

HULYA KOCYIGIT

B.S., Selcuk University, TURKEY,2011

M.S., Georgia State University,2015

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

TOWARDS DEVELOPMENT OF BEST PRACTICE METHODS OF CAUSAL INFERENCE TO ASSESS TREATMENT SELECTION BIOMARKERS FROM NON-RANDOMIZED DATA

by

HULYA KOCYIGIT

Major Professor:	Kevin K. Dobbin
Committee:	Stephen L. Rathbun
	Shaying Zhao
	Ye Shen

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2021

DEDICATION

To mom, AYSE

ACKNOWLEDGEMENTS

I want to extend my most sincere gratitude to the people who have made my Ph.D. study at UGA an amazing and memorable experience. First, I would like to express my deepest gratitude to my advisors, Dr. Kevin K Dobbin, for their support, guidance, patience, and encouragement. Without his persistent support, this thesis would not have been possible. I owe a huge thanks to Dr. Stephen Rathbun for giving me continuing advice and support. I am appreciated to Dr. Ye Shen for his mentorship and guidance throughout my doctoral program. I offer special thanks to Dr. Shaying Zhao for taking the time to read this dissertation and for serving on my committee. I was fortunate to meet so many wonderful friends in graduate school that naming all of them would not be possible. Finally, my sincere appreciation goes to Biostatistics Department and UGA for financial support and making resources available during my study.

My lovely dad and my mom have always been a source of love, support, and inspiration, and my debt to you both is immeasurable but gratefully acknowledged. Most of all, my sister and brothers provided humor and love, which were pivotal to my success during these past years. Finally, my special and affectionate thanks to my mother who has encouraged me all my life for her ceaseless support and devotion during my studies. I want to dedicate this dissertation to my mom, AYSE. Finally, I say that “ Sen benim hayatimda gordugum en guclu rol model kadinsin. Iyi ki varsin,annecigim!”.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Literature Review

1.1 Introduction

The thesis's purpose has been to work within the framework of causal inference and biomarker research to eliminate gaps in the literature in terms of theory and application. Two main parts included in this chapter will provide literature reviews. The first section introduces the history of randomized and non-randomized studies --the framework of potential outcomes, identifiability assumptions under a different types of treatment (i.e., binary, multiple, and continuous treatment scenarios) are offered. The cornerstone tool of causal inference is propensity score adjustment, which removes imbalance between treatment groups, which is reviewed in terms of its definition and estimation. While different methods have explored propensity score evaluation, the discussions of methods' pros and cons in literature are examined in this chapter.

This chapter includes a literature review in which causal inference topics are presented. The first part of this chapter is motivated by causal inference in section 1.1, including a review of history in causal effect in section 1.2, an overview of the potential outcomes framework in section 1.3, describing identifiability assumptions in section 1.4 , revising propensity score and its theory in section 1.5, implementation of propensity score estimation process based on the different methods in section 1.5.1, presentation of the type of propensity score methods (i.e., matching on the propensity score, sub-classification, inverse probability of treatment and covariate adjustment) in section 1.6.

1.2 Causal Inference in Observational Studies

Biostatistics has been an important discipline that guides many researchers in many disciplines, such as epidemiology, health science, health economics, pharmacy, and other fields. There is a great deal of statistical literature that has addressed various methodologies over the last century. Classic statistical analysis includes regression analysis, estimation of parameters, hypothesis testing and examining the asymptotic distribution of parameter estimates. Even though we can compute the probability of past and future events using such standard statistical analysis, classic statistical analysis may not provide estimates with a causal interpretation. In other words, researchers may desire to understand the causal relationships that are beyond the information present in the observed likelihood. Causal analysis is one crucial tool of many disciplines. The objective of causal analysis is not to only make inferences based on the probability of events, but also to examine causal relationships among variables of interest. Thus, Pearl (2010) revealed the difference between causation and association and Pearl's framework shows how "correlation does not imply causation". There is a rich statistical literature in causal inference for both observational and randomized studies. Randomized experiments have been considered as the gold standard to make inferences about causal relationships. Randomized Controlled Trials (RCT) cannot be used in many instances because of being non-feasible, unethical, reasons of timeliness, and cost. Hence, observational data is an alternative to RCT's for use in medical research. The observational study is sometimes called a non-randomized experiment or quasi-experimental in literature.

However, there may be an imbalance between treatment groups due to lack of randomization in the observational study, and bias in the estimated treatment effect can be the result.

Moreover, confounder variables can induce a relationship with treatment or outcome or both treatment and outcome. These difficulties in causal inference have led us to formulate different frameworks of potential outcomes in observational studies. Thus, the remarkable question arises as to how a covariate's characteristic influences other covariate's characteristics. I will consider the context of potential outcomes in observational experiments in Neyman(1923) and Rubin(1978) through this dissertation.

1.3 A History of Causal Inference

Scientists in the biomedical field aimed to predict causal effects of binary, continuous, or multiple treatments on an outcome. They have utilized observational or randomized control trial data to investigate causal effects. Different types of data sets (i.e., randomized and observational studies) can lead to varying results in terms of estimation of treatment effects. In medical science, observational experiments are frequently used to estimate the treatment impacts on the outcomes. Owing to the lack of random treatment assignment of subjects in observational experiments, there can be an existing differences between the two groups. As a result, these differences may create bias in estimates of the treatment effect. In this way, statistical methods are a needed to eliminate or reduce the effects of confounding variables. In literature, researchers have conducted many causal effect studies using observational data sets (such as Cochran and Chambers,1965; Campbell and Stanley,1963,1966; Cochran,1965,1968; Cochran and Rubin,1973; Rubin,1970,1973a,1973b,1973c) and randomized experiments data sets (Fisher, 1935; Anscombe ,1974; Kempthorne,1952,1955). The foundation of potential outcomes in the context of randomized experiments, not in observational studies, was introduced by Neyman (1923). After Neyman's seminal works on the notation of potential outcomes, Fisher (1925) put forward the necessity of physical randomization in addition to Neyman's study to examine causal effects. However, there was no scientific development on potential outcomes for more than half a century, between 1923 and 1974. Rubin (1974) extended Neyman's(1923) idea that reinvented the framework's notation defining causal effects to examine potential outcomes in observational study settings.

At the end of the 70’s, Rubin's work using observational datasets brought to forefront widespread methods to assess causal effects. Holland(1986) called it the Rubin-causal model in a series of papers that provided a general framework of potential outcomes in observational studies. Another approach that is alternative to potential outcomes: Directed acyclic graphs(DAGs), introduced by Judea Pearl(2012). DAG method offered the formulation of causal models that extract confounding from the estimates.

1.4 Potential Outcome Framework

Holland (1986) coined the term the Rubin Causal Model (RCM), which defines the causal inference framework based on article series (Rubin, 1976,1979,1980 and 1983). RCM focuses on two main objectives -- first modeling the 'potential outcome' to estimate causal effect and, secondly, defining ‘assignment mechanism' to approximate a designed experiment from observed data. We work with binary or dichotomous treatment in Chapter 2 and Chapter 3. So, we have two possible treatments arms, which we call treatment and control groups. We assume that a pair of potential outcomes denoted $Y_i(0)$ and $Y_i(1)$, represent outcomes on treated and untreated for an individual i. Difference between outcomes on treated and outcomes on untreated express the treatment's causal effect on individual i:

$$\Delta_i = Y_i(1) - Y_i(0)$$

Then, we can propose the potential outcomes for observed one unit as :

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

where is $T = 0$ for control group vs $T = 1$ for treatment group. Imbens (2004) specifies average treatment effect (ATE):

$$E[Y_i(1) - Y_i(0)]$$

where treatment effect is known as $Y_i(1) - Y_i(0)$. Besides, average treatment effect among treated (ATT) is defined by

$$E[Y_i(1) - Y_i(0) | T = 1]$$

focusing on the subjects who received treatment as the target population. In RCT, there is no differences in covariates distribution between treatment and control arms because of randomization. Thus, ATE and ATT can be directly applied because their estimates are unbiased. We want to find an unbiased estimate of ATE , which will have mean $E[Y_i(1) - Y_i(0)]$. Besides, if we assume an RCT:

$$E[Y|T = 1] = E[TY(1) + (1 - T)Y(0)|T = 1] =$$

$$E[TY(1)|T = 1] + E[(1 - T)Y(0)|T = 1] = E[Y(1)|T = 1] = E[Y(1)] \qquad (1.1)$$

In the same way,

$$E[Y|T = 0] = E[TY(1) + (1 - T)Y(0)|T = 0] =$$

$$E[TY(1)|T = 0] + E[(1 - T)Y(0)|T = 0] = E[Y(0)|T = 0] = E[Y(0)] \qquad (1.2)$$

Equality (1.1) and (1.2) are valid if treatment assignments are independent of outcome: $(Y(0), Y(1)) \perp T$), where ⊥ signifies statistical independence. Thus, we can rewrite mathematical notation for estimate of ATE without bias as in following:

$$E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] \tag{1.3}$$

However, treatment effects are systematically different for covariates between treatment groups in observational studies. In other terms, treatment exposure T may not be independent of potential outcomes (i.e., $Y_i(1)\ and\ Y_i(0)$). Hence, covariate characteristics can be associated with exposure or outcome, or both of these. So, Equality (1.3) does not hold in observational studies which can produce bias in estimates of because $E[Y|T = 1] \neq\ E[Y(1)]$ or $E[Y|T = 0] \neq\ E[Y(0)]$.

1.5 Assumptions

Two assumptions that allow appropriate causal inference were recommended by Rubin and Rosenbaum (1983): strong ignorable treatment assignment (SITA) and stable unit treatment value assumptions (SUTVA). Hernan and Robins (2018) state that SUTVA is known as consistency. The first principal assumption in the estimation of a causal effect is stable unit treatment value condition (SUTVA) (Rubin,1978,1980,1990a,1990b) that includes: *i-)* no interaction between subjects *ii-)* interference among subjects is unavailable. Thus, SUTVA stipulates that we can observe only one version of the outcome under each treatment case. In other terms, the potential outcome has consistently occurred for each subject when the treatment assignment is fixed. The nonexistence of interference also means that this treatment did not affect another subject's outcome when we applied it to one subject. Moreover, "no interaction between subjects" means no hidden variations of treatment, so the outcome is properly described.

SUTVA led to denoting the observed outcome for unit i as $Y_i(T)$,where treatment T is defined with control group (T=0) and treatment group (T=1) under binary treatment. This assumption alludes that we express the observed outcome as $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$.However, If assumption of SUTVA is infringed, it produces inconsistent causal effect estimates. In other phrases, we do not receive a unique potential outcome of each subject under each treatment status. The major cause of this circumstance emerges when the "treatment variant" is present. Often, discrepancies of treatment assessments and uncertainties in the treatments received have led to the emergence of various treatment variants. The second assumption of estimates of causal effect is exchangeability that claimed that treatment and outcomes, given variables, are independent. The exchangeability condition is known as unconfoundedness. We can claim this assumption :

$$Y_i(1), Y_i(0) \ \perp T_i | X_i$$

The second is the principal assumption of positivity means that every unit in the sample of interest is capable of being assigned to all treatment levels. The positivity is sometimes called an overlap assumption in the literature. This assumption is written in mathematical notation as follows:

$$0 < P(T_i = 1 | X_i) < 1$$

which stipulates that each subject has a nonzero probability of obtaining treatment. Also, $P(T_i = 1 | X_i)$ is referred to propensity score (i.e., e_i) in next sections.

If strong ignorability assumption is met, it means that we can measure all confounders and then estimate the unbiased treatment effect. Moreover, this implies that overlap between treatment and control groups is encountered at least. Unfortunately, this assumption can be frequently violated, i.e. treatment or covariates can be effected by covariates' characteristics, because we cannot sometimes control the impact between exposure and covariates or outcome and covariates in non-randomized studies.

1.6 Overview of Propensity Score

Under the identifiability conditions, the exchangeability assumption might not be true given the absence of randomization in observational studies. Causal estimation based on a direct comparison of outcomes may be deceptive. In other words, there could be substantial variations between the observed covariates in the two groups, and these differences also may bias estimates of treatment effects. Rosenbaum and Rubin's seminal paper defines the propensity score in 1983 as the probability of treatment assignment conditional on observed baseline covariates. In the two groups, the propensity score can be utilized to balance the variables and thereby decrease this bias in observational research in many areas. In particular, the applications discussed in articles have come from a range of areas, including epidemiology, research in medical care, biostatistics, economics, and social sciences. There are different processes to estimate propensity scores based on binary, multiple, continuous, and ordinal treatments. Binary treatment denoted as T_i(e.g., T_i=1 if individual i is in treated condition versus T_i=0 if the individual i is in the untreated condition) for individual i, where specify index for number of units, i =1,2,…, n and also, X_i represent observed covariates:

$$e_i = e(X_i) = P(T_i = 1|X_i)$$

where we assume

$$pr(T_1 = t_1, \dots, T_N = t_N|X_1 = x_1, \dots, X_N = x_N) = \prod_{i=1}^{N} e(x_i)^{t_i}\{1 - e(x_i)\}^{1-t_i}$$

The remarkable article of Rosenbaum and Rubin summarized propensity score with five fundamental theorems as follows :

1. Propensity score is known as a balancing score.

2. Any score, which is better than propensity score, is a balancing score. Thus, observed covariates can be "best" balancing score and propensity score is referred as " coarsest".

3. The treatment assessment is strongly ignored if presented x, then it is strongly ignored provided any balancing score.

4.The discrepancy between treatment group at every value of the balancing score is an unbiased estimation of the mean treatment effect at that value of the balancing score if the treatment assignment is strongly ignorable.

5.Sample balance on observed covariates can be generated by using sample estimates of balancing scores.

1.6.1 Estimating Propensity Score

The propensity score can be employed in both randomized trials and observational research. Even though the true propensity score is typically established in randomized experiments and is determined by the study's design, the probability e(X) is unknown outside of randomized experiments and must be estimated using the study's data. Given the estimated propensity score, a significant part of the propensity score analysis is to verify whether pretreatment regressors have been balanced. Suppose we desire to use the method in the context of a binary treatment variable. In that case, the most common and traditional approach to obtain propensity scores is logistic regression, in which a parametric model is suggested. Let the binary treatment assignment be T_i (e.g. particular treatment T_i =1 versus nontreatment T_i =0), consider a collection of p independent covariates described by $X' = (X_1, X_2, \dots, X_p)$, and the vector of unknown parameters of interest be β. So, propensity score is defined as

$$Pr(T_i = 1 \mid X_i, \beta) = e(X_i) = e_i$$

The conditional likelihood of receiving treatment relies on using logistic regression can be expressed as:

$$Pr(T_i = 1 \mid X_i, \beta) = e(X_i) = \frac{exp(X_i^T \beta)}{1 + exp(X_i^T \beta)}$$

Generalized linear model is described by using logit function. We can make transformation covariates variable through the logit function to obtain a linear function of X:

$$log_e\left(\frac{P}{1 - P}\right) = X_i^T \beta$$

where P represents $P(T_i = 1)$.

Moreover, Covariate Balance Propensity Score(CBPS) has been another alternative parametric model to estimate the probability of treatment given observed variables. Unlike the well-known approach logistic regression, more advanced strategies have been adopted to eliminate the conflict of potential model misrepresentation in the parametric model, such as bagging, boosting, random forests, recursive partition regression trees, neural networks, and bayesian additive regression tree (BART). Generalized boosted model, established by McCaffrey et al. (2004), has been a popular method in machine learning and nonparametric techniques when population studies have contained a large number of covariates to estimate treatment effects. GBM is an iterative algorithm that relies on generating the number of trees using tuning parameters to estimate treatment effects from many covariates. Another common machine learning method is bagged (or bootstrap aggregated) CART that uses the original sample to match a CART to a bootstrap sample to replace it and repeat it multiple times(Breiman, 1996). Although random forest and bagged CART (Classification And Regression Tree) resemble to resemble the context of the application process, the random forest method considers subgroups of predictors in every CART structure (Breiman, 1996).BART represents a Bayesian approach utilizing the sum of regression trees to predict a nonparametric function (Chipman,2010).

The most widespread way to estimate PS values is logistic regression, even though there is increased interest in multiple treatments, especially in health science. Imbens (2000) suggested the extension of causal effects based on the binary case to multiple (more than two) treatment arms. The framework of identifiability assumptions (especially exchangeability and positivity assumptions) to estimate causal effect is expanded for multiple treatments. Also, propensity score can be estimated using different methods, such as boosted regression, CART, and random forest in multiple treatment cases.

PS models depend on a model of exposure or treatment, unlike traditional statistical approaches that focus on a model of the outcome under examination. A critical point confronting scientists utilizing PS techniques is how to choose the parameters to be employed in the PS model. Theoretically, the model specification would be directed by the subject matter experience, such as a thorough understanding of how patients are referred to a given procedure. Imbens and Rubin(2015) discussed how to choose the covariates and interactions. They said that in many empirical studies, the number of variates is large relative to the number of units. As a result, it is not always feasible to include all covariates in a propensity score model. Moreover, it may not be sufficient for some of the most critical covariates to include them only linearly. We may wish to have functions, such as logarithms, and higher-order terms, such as quadratic terms, or interactions between the primary covariates. Here we describe a stepwise procedure for selecting the covariates and higher-order terms for inclusion in the propensity score. Thus, Imbens and Rubin(2015) follow a stepwise process with three stages: primary covariates, additional linear/quadratic terms, and interaction terms.

The next phase of analysis often involves some control participants' methods based on the estimated propensity scores after we estimated the propensity score. Four propensity score methods have been expressed and applied in the next section: propensity score matching, sub-classification, covariate adjustment, and inverse probability of treatment weight. In other words, those propensity score methods would be critical analysis techniques to remove imbalances between the groups and for removing confounding between the treatment effect and other covariate effects.

1.7 Propensity Score Methods

Researchers have developed several propensity score-founded approaches for treatment effect measurement in many fields over the past three decades. The most preferred and well-known technique in propensity score analysis is matching. The purpose of the PS matching technique is to create a new sample of individuals with similar propensity score values or covariates values for treatment and control groups. Then the unmatched individuals are excluded from the sample(Rosenbaum & Rubin, 1980). Hence, the PS matching process basically may not use all of the data. The implementation of PS matching has included a variety of approaches, including the following.

Pair Matching: In literature, the most preferred propensity score matching methods are "pair matching " or "1:1 matching," consisting of couples of treated and untreated subjects. There are similar propensity scores for treatment and control participants in the matched pairs (Rubin and Thomas,1996; Austin,2011). Other less preferred alternative methods than 1:1 matching are many-to-one (M:1) based on the propensity score. M:1 matching on the propensity score method indicates that a treatment subject matched control subjects. Also, M represents a number greater than one (Austin and Elizabeth,2014).

Mahalanobis matching: Even though pair propensity matching has been frequently used, the Mahalanobis metric matching technique was discovered prior to that (Cochran and Rubin,1973; Rubin,1976(a), Rubin,1980). We randomly ordered individuals and then calculate the distance between the first treatment individual and all untreated. After selecting the minimum distance value as a match for the treated individual, a pair of subjects is discarded from the potential matching set. This process is repeated until all treated individuals are matched. Also, the distance is expressed as

$$d(i,j) = (x-y)^T \sum\nolimits^{-1} (x-y)$$

where covariate values x and y for treated individual i and untreated individual j. The sample covariance matrix Σ specifies the matching variables from sets of the treated group and untreated population individuals. There are some drawbacks to using Mahalanobis metric matching. This method is based on the high-dimensional score, making it difficult when the model contains many covariates (Guo and Fraser,2015).

Nearest Neighbor Matching: A treatment unit is chosen, and then the control unit with the propensity score that is nearest to the treatment unit is picked as a matched control unit(Austin and Schuster,2016). If many control participants have propensity scores comparable to the treatment participants, we randomly prefer one of the control participants. The most important point in this method is that there is no use of any maximum threshold value between matches participants' propensity score values(Rosenbaum et al.,1985).

Caliper Matching: Caliper matching and NN matching techniques are identical methods in terms of implementation (Cochran and Rubin, 1973). This technique stipulates that the absolute difference between matched participants' propensity scores must be less than a certain threshold. A researcher has preferred different pre-determined thresholds in literature; for example, used 0.20 threshold (Austin,2011,2012) and 0.25 threshold (Rosenbaum and Rubin,1985) for standard deviation on the propensity score.

Optimal Matching: Optimal matching aims to construct matched pairs with the smallest average between the difference in propensity scores. An advantage of optimal matching is that the implementation of the network flow principle to improve matching. Austin(2014) indicates that NNM and optimal matching have been more biased than caliper matching. Moreover, caliper matching is preferable to the other two methods because optimal matching was discovered late compared to greedy matching techniques.

The combination of the caliper and nearest-neighbor matching led to creating new techniques: Nearest neighbor matching within a caliper starts by ordering the treatment and control participants and then select first treatment participant *i* and provide the control participant *j* as matched for *ith* participant; after that, one selects the minimum absolute difference between *i* and *j* participants within prespecified caliper value. So, one discards *i* and *js* participants from the sample considered for matching. As indicated before, the caliper threshold is decided by researchers. The nearest available Mahalanobis metric matching within calipers specified by the propensity scores is produced by modifying nearest neighbor matching within a caliper(Guo and Fraser, 2015).

The central idea of employing sub-classification to balance data was developed by Cochran (1968) and formulated even before the development of propensity score analysis. However, exact sub-classification suffers from the same dimensionality problem as matching. Cochran states that the number of subclasses grows exponentially when variables numbers are increased. According to Cochran, as the count of variables grows, the number of subclasses or strata rises exponentially, i.e., we would have 2^s subclasses for s covariates when binary variables in the population are considered. For some subclasses, which include only units from the treatment group, estimating a treatment effect is difficult. Besides, Cochran employed stratification on the quintiles of a continuous variable, and then removing almost ninety percent of the bias due to imbalance between treatment and control groups. Even though Rosenbaum et al. and Cochran agreed on removing bias at 90 percent using the confounding variables between treated and untreated groups, the study of Rosenbaum et al. applied stratification based on propensity score values. If the propensity score is estimated correctly based on the model in observational studies, the distribution of variables within the same strata will be similar. Then, between treated and untreated subjects in the same strata there would not be bias in comparisons (Cochran,1968; Rosenbaum and Rubin,1983,1984; Imbens and Rubin,2015; Guo and Fraser 2015). A remarkable number of studies on sub-classification methods have been conducted in literature (such as Hullsiek and Louis,2002; D'agostino,1998;Rubin1983,2007; Rosenbaum,1991; Austin and Mamdani,2005; Austin,2007,2012; Austin and Schuster,2016).

IPTW aims to obtain a weighted population, which has a similar distribution of observed baseline covariates between treatment and control individuals, to remove imbalance between two groups (or more than two groups). In a paper published in the field of surveys, the concept of probability score weighting was originally suggested by Horvitz and Thompson (1952); the paper focused on sample averages and their method is commonly used in the weighted regression. In 1987(a), Rosenbaum recommended the inverse probability of treatment weighting, rely on the model direct optimization. Although researchers have published many papers about matching on propensity score because it is the oldest propensity score techniques, IPTW method has been more attractive in many fields in recent years compared to matching, sub-classification, and covariate adjustment (see Xie and Liu,2005; Lee et al.,2009; Shen et al.,2011; Austin,2011; Li et al.,2013; Austin and Stuart,2015). One of the most important purposes of the weighting is that no sacrifice in the data set matching is required, such as a trimming step.

The last propensity score approach is covariate adjustment using a propensity score that estimates linear treatment effects for continuous outcomes. The outcome on two covariates is regressed: estimated propensity score and indicator covariate for treatment case. The selection of model would is chosen according to a state of the outcome variable, i.e., we could select logistic regression for binary(or dichotomous) treatment, or linear model could be preferred for a continuous outcome.

CHAPTER 2

Evaluation a Biomarker for Treatment Selection in Observational Study

2.1 Introduction

Globally, growing income disparities have been followed by rising inequality in health outcomes. Dickman et al. (2017) states that the wealthiest Americans have a 10-to-15 year longer life expectancy because of receiving better health care than poorer Americans. The rising health needs have led people to buy more health insurance. However, growing premiums and burden sharing also stifled income growth for those with private health insurance driving more households into debt; and, for patients who do not hold insurance, bankruptcy can result from medical expenses. According to the WHO[1] (2020), cancer has been registered as the sixth of the top 10 causes of death. Cancer is seen as a significant world public health problem because of the high mortality and morbidity rates in the world (Favoriti et al., 2010), with about nearly 10 million deaths in 2020, and 70 % of these deaths occurring in low and middle-income countries. Cancer is a remarkable

[1] https://www.who.int/health-topics/cancer

disease that grows uncontrollably in any tissue or organ of the body and invades nearby in the body to other organs(WHO) .It has been observed that smoking, lifestyle disorders, unhealthy diet, alcohol use, air pollution, and late age at first births in women increase the risk of cancer in middle or less developed countries (Torre et al., 2012), which are least prepared to handle the cancer burden. According to WHO's report in 2020, lung (1.8 million deaths,18 % of total), colorectum and rectum (935 000 deaths,9.3%), and liver (830 000 deaths, 8.3%) are recognized as the most common cancer types worldwide, as seen in the below table. Fortunately, biotechnology, chemistry, and software accumulate resources to reduce the disease's side effects and decrease death rates (Pothur,2002).

Precision medicine has progressed due to improvements in our knowledge of disease molecular biology and treatment response pathways, as well as raised patient genetic profiling capabilities. The identification and clarification of treatment selection markers is one part of such personalized care (Janes,2011). It is thus critical to recognize and evaluate signs capable of guiding clinical decisions in order to prevent specific types of events(e.g., disease development, recurrence, or mortality) within a given post-treatment period(Blangero et al.,2019). Thus, biomarkers have taken an essential place in the medical field. Biomarkers can ensure the integration of a reliable indicator of effectiveness for a particular mechanism-depending on medication, or guide treatment selection for each case relying on the tumor's biological makeup and the patient's genotype. But, owing to the variety of biomarker evaluation techniques, the accessibility of collecting samples, the efficacy and reproducibility of the trial, and the increased expenses associated with evaluating the marker status on each patient, the confirmation of biomarkers by clinical testing, leading to effective utilization of the biomarker in clinical settings, remains a significant obstacle (Mandrekar and Sargent,2009; Dobbin et al., 2016; Mandrekar et al.,2015). Blangero et al. (2019) state that when comparing the success of two treatments (advanced vs. standard), such markers are supposed to maximize clinical outcomes by recognizing people who will improve the most from the advanced intervention and eliminating those who will not.

The terms "prognostic" and "predictive" have been utilized to describe markers of diverse components. Simon(2010) made clear the difference between markers. A prognostic marker is a measurement that is linked to a patient's health outcome in the absence of treatment or with the use of a conventional medication that they are expected to undergo. A predictive markers is a measurement that is related to the response or absence of response to treatment. However, we will not examine the prognostics-type biomarkers in this study. Besides, there is no agreement about how to name such a marker, which are called "treatment selection", "prescriptive," and "predictive" in literature (Holly et al.,2014). One of the best examples of treatment selection in the medical field is KRAS gene expression in colorectal cancer. Patients who do not have KRAS variants have illustrated better anti-epidermal growth factor receptor therapy performance than patients who have KRAS mutations. As a result, the expression of KRAS can be used to influence treatment selection. The US Food and Drug Administration has updated the labeling of two EGFR inhibitors, which are cetuximab and panitumumab, to state that they are not eligible for colorectal cancer treatment in patients with KRAS mutations in codon 12 or 13. It indicates that some markers related to treatment selection and require effective methods to measure how well they do. At the same time, KRAS mutation illustrates a strong association with treatment selection.

There are many different perspectives in the literature to evaluate treatment selection markers. Some papers studied descriptive analysis for treatment effect modeling(Cai et al.,2011; Claggett et al.,2011; Zhao et al.,2013); meanwhile, other articles focused on assessing individual measures for markers (Song and Pepe,2004; Vickers et al.,2007; Janes et al.,2011).The current popular method of biomarker assessment in the health literature has been utilized for statistical interaction between the marker and treatment variables in randomized control trials (Sargent et al., 2005; Simon, 2008; Huang et al.,2012). Janes et al.(2011) emphasize that interaction terms between treatment and biomarker based on the model might lead to inaccurate evaluations of markers performance as well as assessments that might not be scientifically helpful. The measurement metric known as (Θ) is globally utilized to determine the performance measure of the marker(Bonetti and Gelber,2004; Song and Pepe,2004; Cai et al., 2011.; Janes et al., 2014b; Janes et al., 2015). These studies were led by using data from randomized and controlled trials. However, the researchers can find themselves in a challenging situation because of a number of studies of biomarkers that use observational study data.

We aim to recommend a suitable method by designing and testing a treatment selection process based on data obtained from non-randomized trial settings where the subject's characteristics may influence treatment, outcomes, or both. So, we look at how treatment selection is evaluated in situations that raise some specific questions, such as:

1. This research's target question is whether causal inference adjustment is necessary to evaluate biomarker performance in lung cancer.
2. What type of causal inference techniques should we choose to eliminate bias on covariate characteristics and then assess the biomarker's performance?
3. Which features may affect treatment assignment and therefore need to be taken into account as confounders?.
4. Is there any remarkable difference between whether or not using causal inference for implementation of treatment selection?.

We emphasize developing and assessing treatment selection in cancer datasets, and questions that have not previously been studied in the literature. So we discuss the paper's unique contributions.

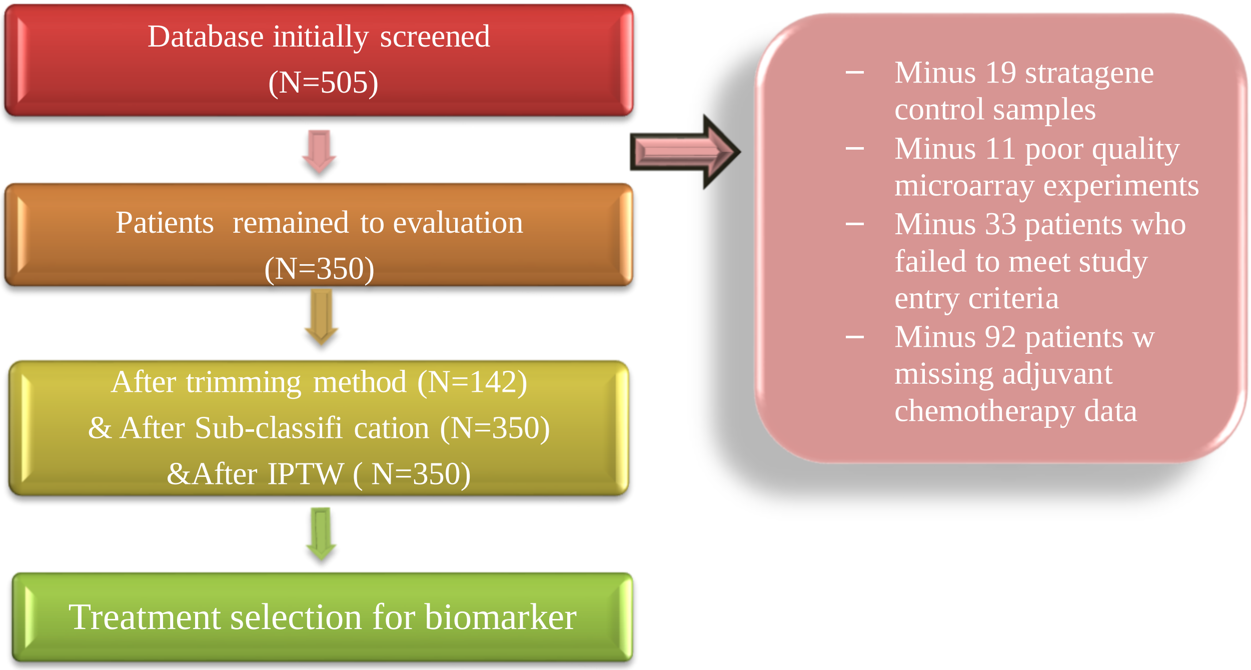
The remainder of the article is structured as follows. We present comprehensive data information on our motivation data in Section 2.2. We construct a statistical framework for causal inference and treatment selection in cancer biomarker from non-randomized settings in Section 2.3. We have developed our assumptions based on existing literature techniques while creating the framework of methodology part in Section 2.4. Then, we implement the methods on lung cancer experiments with adjuvant chemotherapy treatment in Section 2.5. Finally, we end with a discussion of our conclusions and potential future study subjects in Section 2.6.

2.2 Motivational Context

We demonstrate our approaches in the lung cancer treatment context. Patients with lung cancer are treated with or without adjuvant chemotherapy following the cancer surgery. A limited proportion of patients receive benefit from adjuvant chemotherapy. In contrast, the rest of the patients endure chemotherapy' toxic side effects, not to mention the stress and expenditure of unnecessary treatment. So, the top priority of public health is to define biomarkers that can be used to determine whether or not patients benefit from this specific chemotherapy.

As data were obtained from an observational study, we use it to illustrate the methods in this section for studying the effect of exposure variable (chemotherapy) on lung cancer outcomes in non-randomized setting. The data contains N=505 patients, and research records whether or not patients received adjuvant chemotherapy treatment. Then the researchers collected clinical variable and outcomes data on each patient. After some of the missing responses for the adjuvant chemotherapy and other variables were cleaned, 350 individuals remained, $N_{treatment} = 94$ men and women had been exposed to adjuvant chemotherapy; meanwhile, the comparison group consisted of $N_{control} = 256$ individuals from the same cohort who were not exposed to adjuvant chemotherapy. The data set involves twenty-seven covariates, although many covariates did not relate to the adjuvant chemotherapy variable in the sense of lung cancer. Then, ten covariates remained, including: “gender” based on the sex of a patient , “race” which identifies the racial origination, “adjuvant RT” which presents yes/no indicator for whether the patient had adjuvant radiation therapy as part of the primary treatment plan, “Smoking history” variable represents that the smoking history of a patient, “pathologic n stage” illustrates pathologic N(nodal) stage of lung cancer, using the AJCC TNM system of numbered categories for representation of data, “site” from which patient sample came and at which the microarray assay was performed, “age at diagnosis” is that age at which condition or disease was diagnosed, “surgical margins” represents the degree of cancer involvement of the surgical margins, “pathologic t stage” is that code for pathologic T(tumor) stage of lung cancer and using the AJCC TNM system of numbered categories for representation of the data and “histological grade” represents histologic grade. The lung cancer data set contains the total of 350 patients that were measured for the expression of 22500 biomarkers. The flow chart of lung cancer is taken place as follows in Figure-2.1.

Figure 2.1: Lung cancer data flow


Database initially screened
(N=505)
Patients remained to evaluation
(N=350)
After trimming method (N=142)
& After Sub-classification (N=350)
&After IPTW (N=350)
Treatment selection for biomarker
– Minus 19 stratagene control samples
– Minus 11 poor quality microarray experiments
– Minus 33 patients who failed to meet study entry criteria
– Minus 92 patients w missing adjuvant chemotherapy data

2.3 Methods

2.3.1 Conceptual Framework of Treatment Effect

In the Rubin Causal Model (RCM), causal effects are described using three fundamental principles: supposed n observations $(T, X, Y\ )$. We denote T as observed treatment under the binary case circumstances (T=1 if treatment and T=0 if control) and let X signify a vector of observed covariates. Also, let Y be a binary outcome variable. We propose two alternative treatments and outcomes in the future consequences setting. Let i specify the number of units, i=1,2,3…,n. Each individual has two possible outcomes: $Y_i(0)$ and $Y_i(1)$ for treated and untreated outcomes, respectively. So, an outcome is written as: $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$. There are two ways to estimate treatment effects in causal inference: the first is average treatment effect(ATE) that is defined to be $E[Y_i(1) - Y_i(0)]$. The second measure of treatment effect is the average treatment effect for the treated (ATT) that is defined to be $E[Y_i(1) - Y_i(0)|T = 1]$(Rubin,2008; Imai and Ratkovic, 2013).

2.3.2 Propensity Score

Rosenbaum and Rubin (1983) defined the propensity score as the estimated probability of each patient getting treatment based on the patient's variables:$e_i = e(X_i) = P(T_i = 1|X_i)$. Using propensity scores is different between randomized and observational studies, even though propensity scores can be applied in both settings. Because we know the true propensity score value in randomized studies (at least if randomization and blinding are perfect), however, it is not always needed; but in observational studies we don't know the true propensity score and therefore must estimate propensity score from the fitted logistic regression model using the data set. The second significant difference between those studies is whether all covariates related to treatment (T_i) or outcome (Y_i) are present in the collected data set.

We do not know the true propensity score in observational studies, and so we estimate propensity score using data. An adequate examination of the propensity score has been highly critical in observational studies. Use of the standardized difference can be beneficial to compare binary and continuous variables between control and treatment groups. Besides, Austin (2009e) proposed using a set of binary variables to demonstrate a standardized difference for multilevel categorical variables. Austin (2011) define the standardized difference for continuous variable:

$$d = \frac{(\bar{x}_{treatment-}\bar{x}_{control})}{\sqrt{\frac{s^2_{treatment+}\, s^2_{control}}{2}}}$$

where the sample means of covariates in treatment and control subjects are denoted as $\bar{x}_{treatment}$ and $\bar{x}_{control}$ and while let $s^2_{treatment}$ and $s^2_{control}$are sample variance of treatment and control subjects, respectively.

Standardized difference for dichotomous variable is defined as:

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treatment}(1 - \hat{p}_{treatment}) + \hat{p}_{control}(1 - \hat{p}_{control})}{2}}}$$

where prevalence or means of dichotomous variables are denoted in treatment and control subject as $\hat{p}_{treatment}$ and $\hat{p}_{control}$,respectively. Besides, the sample size does not influence the standardized difference, and it is also used to compare the balance of variables measured between groups. In the literature, the researchers do not agree upon specific criteria to determine the standardized mean difference threshold. However, Normand et al.(2001) suggest that the imbalance may not be important if the standard mean value is less than 0.1.

2.3.3 Related Work in Literature

In 1983, Rosenbaum and Rubin published a seminal paper on propensity score analysis. That paper articulated the theory and application principles for a variety of propensity score models. Ever since this work, the propensity score method has grown rapidly and moved in various directions. Before we start to look over these propensity techniques, we should emphasize the need to assess balance of the covariates. Rubin and Imbens in 2015 state that such a lack of covariate balance creates two problems. Firstly, it might lead to sensitivity in subsequent inferences where small changes in methods are made and produce large variation. Secondly, it is difficult to obtain an accurate estimate of treatment effects in the limited number of control or treatment groups in covariates. So, propensity score methods are highly critical analysis techniques to remove imbalances between the groups and help remove confounding between the treatment effect and other covariate effects. Many research fields such as statistics, economics, education, epidemiology, medical care have been focused on the propensity score to achieve balance between distributions of treated and untreated groups. The propensity score method is examined under the four main techniques: propensity score matching, the propensity score stratification, covariate adjustment using the propensity score, and inverse probability of treatment weighting (IPTW).

In the past decades, many articles have studied propensity score methods through matching (e.g. Cochran and Rubin,1973; Rosenbaum and Rubin ,1985; D’Agostino ,1998;Heckman et. al ,1998; Dehejia et al.,2002; Stuart,2010; Subroto et al.,2010), sub-classification on propensity score (Cochran ,1968; Rosenbaum and Rubin,1984; Lunceford and Davidian,2004), covariate adjustment on propensity score(Speroff,1996; Austin,2011), and inverse probability of treatment weighting (IPTW)(Rosenbaum and Rubin,1983; Rosenbaum,1987; Hirano and Imbens ,2001; McCaffrey et al.,2004 and Austin,2015). These four main propensity score methods have aimed to eliminate bias in estimates of the treatment effect between treatment and control groups and achieve overlap of covariates distributions. In other words, assumptions of unconfoundedness and overlap of distribution are key point to estimate causal effects. Besides, there are notable studies that look at limitations of overlap in covariates among groups (Dehajia and Wahba ,1999; King and Zeng,2005;Crump et al. ,2009). If propensity score values are close or equal to zero or one, these extreme values can cause bias in estimates of causal effects. Thus, it is essential to reduce the impact of extreme values through the trimming method. In literature, there are limited number of studies of trimming methods (see Crump et al. 2009; Sturmer et al. ,2010; Lee et all. ,2011; Rothman ,2018).

2.3.4 Propensity Score Trimming

A crucial purpose of applied statistical methods is to understand the causal relationship between treatment and outcome. In studies where treatment assignments are assigned randomly(and perfectly blinded with 100% compliance), the researcher can directly apply to estimate causal effects under the unconfoundedness assumptions. We desire to assess the covariate balance and provide that any propensity methods lead to comparable in terms of the assessed covariates. However, observational studies have violated this assumption due to an absence of randomization. In other words, covariates' characteristics can influence the treatment, outcome, or both. So, different distributions between treatment and control groups in covariates can produce the limited overlap issue. We stated that propensity score has some assumptions such as SUTVA, positivity, or exchangeability (see details in chapter 1 of this thesis) to infer appropriate causality. One assumption is the positivity assumption that looked at PS distributions' overlap between treated and untreated groups. Suppose there is limited overlap between treatment groups. In that case, it means that the absence of overlap may indicate a failure of the positivity principle, which could lead to propensity scores very near to zero or one. We employ matching, sub-classification, weighting, or covariate adjustment to eliminate bias between treatment groups. But few samples of extreme values are present, and estimators may be overly skewed, resulting in biased and unstable performance. Unfortunately, estimators of propensity score methods might be extremely affected by some covariates and cause biased and inconsistent results. Another technique has been recommended to address this problem intrinsic to estimators of propensity score techniques: the trimming method. Unluckily, researchers have rarely been concerned with utilizing propensity score trimming to estimate causal effects in the literature. A number of approaches have been advanced for determining the trimming method that has helped the overlap between treatment groups. Individuals whose propensity score values drop below the limit of propensity score values in the subpopulation with the opposing treatment are often excluded.

All trimming methods aim to identify a sample population that is as inclusive as possible but still has adequate overlap that extrapolation is redundant and the overall treatment effect can be accurately measured. Even though all researchers, who studied the trimming method, aim to thrive a systematic approach to target the absence of overlap in covariate distributions between treated and untreated groups, the implementation of all trimming techniques has employed different processes. Stürmer et al.(2010) proposed an asymmetric trimming method that relies on the distribution of propensity score values in two treatment groups. The bound of threshold on propensity scores to employ trimming method is determined based on the treatment and control groups, separately. Stürmer et al.(2010) defined the trimming method as follows:

$$I = \left\{ i \in I : e_i \in \left[F^{-1}_{e_i|T_i}(\alpha|1), F^{-1}_{e_i|T_i}(1-\alpha|0) \right] \right\}$$

Thus, 100*q *th* percentile of the propensity score in the treatment arm represents for lower bound of the trimming method (L). Besides, 100*(1-q) *th* percentile of the propensity score in control arm indicates upper bound (U). After the bounds [L,U]] are established, outside bound propensity score values are removed from the treatment and control groups' data sample. Another trimming method is suggested by Walker that recommended a technique for measuring covariate overlap that also acts as a trimming tool. Walker et al.(2013) suggested a technique for assessing covariate overlap that also acts as a trimming tool.

The last trimming method, which is utilized throughout this paper, is recommended by Crump et al. (2009). This study aims employ this method in some circumstances, such as extreme propensity score values, a large variance, poor finite sample properties, or bias. Thus, the propensity score trimming intervenes to ensure balance in distribution between two groups by excluding close to 0 and 1 values. Computing the asymptotic sampling variance for each subset's average treatment effect (ATE) is more appropriate because one cannot compute the exact sample variance. ATE is defined as

$$\tau_{\mathbb{C}}= \mathbb{E}_{sp}[\tau(X_i)|X_i \in \mathbb{C}] \tag{2.1}$$

where X is in some subset $\mathbb{C}$ of the covariate space, $\tau_{\mathbb{C}}$.

So, we focus on the asymptotic sampling variance for the efficient estimator for average treatment effect, which is

$$\mathbb{AV}_{fs}^{eff}(\mathbb{C}) = \frac{1}{q(\mathbb{C})} . \mathbb{E}_{sp}\left[\frac{\sigma_t^2(X_i)}{e(X_i)} + \frac{\sigma_c^2(X_i)}{1-e(X_i)} \mid X \in \mathbb{C}\right] \tag{2.2}$$

where $q(\mathbb{C}) = Pr_{sp}(X_i \in \mathbb{C})$ is covariate probability in subset of $\mathbb{C}$.The question in here is how we can make minimized the asymptotic sampling variance of an efficient estimator. If there is homoscedasticity, we define that the optimal sampling variance as

$$\mathbb{AV}_{fs}^{eff}(\mathbb{C}) = \frac{\sigma^2}{q(\mathbb{C})} . \mathbb{E}_{sp}\left[\frac{1}{e(X_i)} + \frac{1}{1-e(X_i)} \mid X \in \mathbb{C}\right]$$

where

$$\mathbb{V}(Y_i|X_i) = \sigma^2$$

So, $\mathbb{C}^{\star}$ is defined by optimal $\mathbb{C}$. So we have two possibilities to minimize asymptotic sampling variance under all subset $\mathbb{C}$ of X (Imbens and Rubin,2015).

Firstly, if we consider that,

$$sup_{x \in X} \frac{1}{e(x).(1-e(x))} \leq 2\ \mathbb{E}_{sp}\left[\frac{1}{e(X_i)} + \frac{1}{1-e(X_i)}\right] \tag{2.3}$$

then, entire covariate space $\mathbb{C}^{\star} = X$ and the optimal $\mathbb{C}$ is same. On the other hand, we can define the optimal $\mathbb{C}^{\star}$:

$$\mathbb{C}^{\star} = \{x \in \mathbb{X} | \alpha \leq e(x) \leq 1 - \alpha\},$$

where the threshold α is equal to

$$\alpha = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{\gamma}}$$

where γ is defined as

$$\gamma = 2\ \mathbb{E}_{sp}\left[\frac{1}{e(X_i)(1-e(X_i))} \mid \frac{1}{e(X_i)(1-e(X_i))} \leq \gamma\right] \qquad (2.4)$$

This procedure implementation can be done step by step as follows.

Step 1: We should estimate propensity score , $\hat{e}(X_i)$ as discussed it in section 2.3.2.

Step 2: After estimated propensity scores $\hat{e}(X_i)$, we need check in (2.5) inequality as taken in below.

$$max_{i=1,\ldots,N}\ \frac{1}{\hat{e}(X_i)(1-\hat{e}(X_i))} \leq 2\ \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{e}(X_i)(1-\hat{e}(X_i))} \qquad (2.5)$$

When (2.5) inequality holds, then $\hat{\mathbb{C}}$=X.

Step 3: if (2.5) inequality does not hold, so we consider in (2.6) inequality to solve for a value of γ satisfying,

$$\frac{\gamma}{N}\sum_{i=1}^{N} 1_{\widehat{(e}(X_i)(1-\hat{e}(X_i)))^{-1}\leq\gamma} = \frac{2}{N}\sum_{i=1}^{N}\frac{1}{\hat{e}(X_i)(1-\hat{e}(X_i))}.\ 1_{\widehat{(e}(X_i)(1-\hat{e}(X_i)))^{-1}\leq\gamma} \qquad (2.6)$$

Step 4: If inequality (2.6) does not hold and then $\gamma = min_i(\hat{e}(X_i)(1-\hat{e}(X_i)))^{-1}$, right hand side is larger than left-hand side in inequality (2.6). Thus, we will get largest value of γ and then it's called as $\hat{\gamma}$. Finally, we compute in following

$$\hat{\alpha} = \frac{1}{2} \text{-} \sqrt{\frac{1}{4} - \frac{1}{\hat{\gamma}}} \quad \text{and} \quad \hat{\mathbb{C}} = \{x \in \mathbb{X} | \hat{\alpha}\ \leq \hat{e}(x) \leq 1 - \hat{\alpha}\} \qquad (2.7)$$

$\hat{e}(X_i)$ value of outside $\hat{\mathbb{C}}$ will be discarded and then we will focus on balance and estimate average treatment effect for subset class.

2.3.5 Subclassification on Propensity Score

On the basis of propensity score ranking, stratification is sometimes called sub-classification, divides all sample into equal subclasses. Some researchers (Cochran 1968; Rubin and Imbens,2015) emphasized that employing percentiles of estimated propensity score values to split into five subclasses. Then those subclasses have illustrated to eliminate ninety percentile of bias due to calculated confounding variables. Each subclass will have a similar propensity score value for treatment and control groups. Thus, we will eliminate bias between treatment and control groups regarding the distribution of evaluated variables.

2.3.6 Propensity Score Weighting

We define inverse probability of treatment weighted as $w = \frac{T}{e} + \frac{1-T}{1-e}$. Lunceford and Davidian (2004) discuss theory of inverse probability of treatment effect to estimate ATE and ATT. We have that estimate of ATE as $\frac{1}{n}\sum_{i=1}^{n} \frac{T_i Y_i}{e_i} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-T_i)Y_i}{e_i}$ .There is another alternative way to define estimate of ATE as following: $\left(\sum_{i=1}^{n} \frac{T_i}{e_i}\right)^{-1} \sum_{i=1}^{n} \frac{T_i Y_i}{e_i} - \left(\sum_{i=1}^{n} \frac{1-Z_i}{1-e_i}\right)^{-1} \sum_{i=1}^{n} \frac{(1-T_i)Y_i}{e_i}$ .In addition, an estimate of ATT is $T_i + \frac{(1-T_i)e_i}{1-e_i}$ receive one weight and besides, estimate average effect of treatment in the controls is that $(1 - T_i) + \frac{T_i\ (1-e_i)}{e_i}$ (Lunceford and Davidian, 2004; Morgan & Todd,2008; Austin, 2011). Weights can be sometimes large and highly influential.

Robins (1998 and 1999) defines stabilization in weights that provide to decrease variability of estimation. Using stabilization for estimating ATE in IPTW is $Pr(T = 1)\frac{T}{e} + \Pr\,(T = 0)\frac{1-T}{1-e}$,where $Pr(T = 1)$ and $Pr(T = 0)$ represent marginal of probability in treated and control groups. To sum up, we can apply propensity weighting step by step as following : i) Estimate propensity score using covariates in fitted logistic regression. ii) There are two types of weights to compute estimates: weights for ATE and weights for ATT. If we have large weights, stabilization is applied. iii)After computing weights, we need to assess balance of baseline covariates in treated and untreated subjects in weighted sample.

2.4 Proposed Approach to Evaluate Biomarker Performance

Janes et al. (2015) performed a comprehensive review of an earlier study and suggested Θ parameter as a marker performance metric. This study suggest related to each biomarker versus linear regression model with interaction term between treatment as defined T and biomarker as defined B as follows:

$$logit\ P(Y = 1|T, B) = \delta_0 + \delta_1 T + \delta_2 B + \delta_3 TB \qquad (2.8)$$

where the δ_0, δ_1, δ_2, and δ_3 represent model parameters and let denote Y as the outcome of interest. Absolute treatment effect provided biomarker value is defined as $\Delta(B) = P(Y = 1|T = 0, B) - P(Y = 1|T = 1, B)$. So, The rule is described to reduce the incidence of population events as $\Delta(B) < 0$.Moreover, Janes et al. (2014) identify Θ parameter to enhanced outcomes by reduction in population incidence rate under biomarker-based treatment selection as following :

$$\Theta = [\Pr(\mathrm{Y} = 1|\mathrm{T} = 1, \Delta(B) < 0) - \Pr(\mathrm{Y} = 1|\mathrm{T} = 0, \Delta(B) < 0)] * \Pr\,(\Delta(B) < 0)$$

But alternative equations for estimating Θ parameter is established through this paper. We use the Cox regression model because this dataset has survival outcomes and not simple binary outcomes; the model has interaction between T and B,

$$h(w|T,B) = h_0(w)Exp[\beta_1 B + \beta_2 T + \beta_3 BT]$$

where $h_0(w)$ is the baseline hazard, B is the biomarker, and T is the treatment assignment ($T = 0$ for control arm, and $T = 1$ for active treatment arm) and also, $\beta_1, \beta_2\ and\ \beta_3$ are coefficient parameters of cox model. In addition, the overall survival time is the endpoint. When the biomarker ,B is not associated with patient outcomes in control group, we can inform that $\beta_1 = 0$. However, if the biomarker is not associated with patient outcomes on the treatment group, $\beta_1 + \beta_3 = 0$ is hold. Under the additional assumption of an exponential baseline hazard, we can write the baseline hazard as

$$h_0(w) = \lambda$$

for some $\lambda > 0$. Then,

$$h(w|T,B) = \lambda Exp[\beta_1 B + \beta_2 T + \beta_3 BT]$$

The cumulative hazard is then,

$$H(w|T,B) = \int_{s=0}^{s=w} h(w|T,B)ds$$

$$= \int_{s=0}^{s=w} \lambda\, Exp[\beta_1 B + \beta_2 T + \beta_3 BT]$$

$$= t\lambda Exp[\beta_1 B + \beta_2 T + \beta_3 BT]$$

Now , recalling that $S(w) = Exp[-H(w)]$ we have,

$$S(w) = Exp[-w\lambda Exp[\beta_1 B + \beta_2 T + \beta_3 BT]]$$

$$= \Pr\,(W > w)$$

Now if we set $w = w_0$ we have

$$\Pr(W > w_0|T, B) = Exp[-w_0\lambda Exp[\beta_1 B + \beta_2 T + \beta_3 BT]]$$

So, for an individual assigned to control the probability would be:

$$\Pr(W > w_0|T = 0, B) = Exp[-w_0\lambda Exp[\beta_1 B]]$$

and for an individual assigned to treatment

$$\Pr(W > w_0|T = 1, B) = Exp[-w_0\lambda Exp[\beta_1 B + \beta_2 + \beta_3 B]]$$

The treatment hazard ratio is described as

$$\frac{h(w|T = 1, B)}{h(w|T = 0, B)} = e^{\beta_2 + \beta_3 B}$$

Hence, I can re-write optimal strategy as:

$$T_{opt}(B = b) \Rightarrow \begin{cases} if\ :\ \beta_3 < 0 \begin{cases} T = 1: \quad b > \dfrac{-\beta_2}{\beta_3} \\ T = 0: \quad b \leq \dfrac{-\beta_2}{\beta_3} \end{cases} \\ if\ :\ \beta_3 > 0 \begin{cases} T = 1: \quad b < \dfrac{-\beta_2}{\beta_3} \\ T = 0: \quad b \geq \dfrac{-\beta_2}{\beta_3} \end{cases} \end{cases}$$

After the biomarker cut-off point is determined, we have

$$Pr\left(W > w_0 | T_{opt}\right) = \int_{c_1}^{d_1} Exp\left[-w_0 \lambda Exp[\beta_1 B + \beta_2 + \beta_3 B]\right] + \int_{c_0}^{d_0} Exp[-w_0 \lambda Exp[\beta_1 B]]$$

where (c_0, d_0) and (c_1, d_1) are the intervals for control and treatment assignments, respectively. The parameters of interest is developed as follows :

$$\Theta_0 = \int_{c_1}^{d_1} Exp\left[-w_0 \lambda\, e^{\beta_1 B + \beta_2 + \beta_3 B}\right] - Exp\left[-w_0 \lambda\, e^{\beta_1 B}\right] db$$

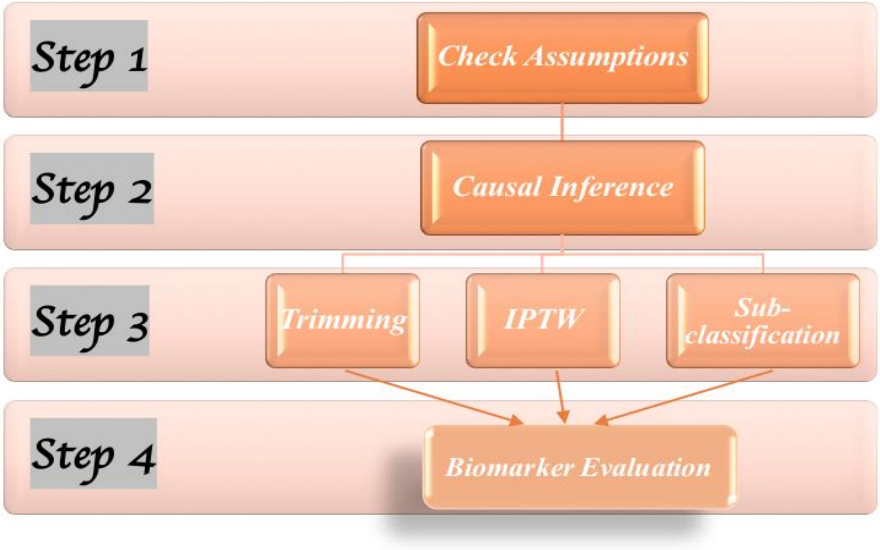
and

$$\Theta_1 = \int_{c_0}^{d_0} Exp\left[-w_0 \lambda\, e^{\beta_1 B}\right] - Exp\left[-w_0 \lambda\, e^{\beta_1 B + \beta_2 + \beta_3 B}\right] db$$

2.5 Application to Lung Cancer Dataset

To provide a proposed method, this paper has included an overview of the broad scientific areas. The first objective in the article is the assessment of causal inference using data on lung cancer, which is from an observational study. Since treatment assignment processes are neither known nor random, obtaining causal effects from retrospective trials in non-equivalent populations is difficult. Treated and untreated samples often vary systematically in both measured and unmeasured baseline characteristics. When estimating the impact of treatment on results using observational data, methodological approaches must be used to account for systemic discrepancies between treated and untreated subjects. So, those issues lead us to implement methodologies based on the propensity score. After we eliminated imbalance between groups using a variety of propensity score methods, we assess the performance of predictive biomarkers based on the results of causal inference techniques. Thus, this paper's second aim is to focus on the treatment selection process for lung cancer. Figure 2.2 clearly illustrates the flow diagram of the application processing.

Figure 2.2: Flow diagram of analysis



We used lung cancer data on 350 patients who were treated with chemotherapy. Overall, 256 patients were not treated by chemotherapy treatment, whereas 94 of whom were exposed to chemotherapy treatment for lung cancer. We also have ten covariates, age, gender, adjuvant RT, race, surgical margin, site, historical grade, smoking history, path N stage, and path T stage collected from patients' medical archives. The propensity score was estimated using a logistic regression model to regress individual assignments on the ten covariates that might affect the outcome. This approach has been demonstrated to perform better by including some variables that influence treatment selection. We can use a stepwise procedure for selecting the covariates for inclusion in the propensity scores. Race, adjuvant RT, path N stage, site, age, path T stage, and their interaction terms are particularly important covariates and inclusion of estimated propensity score. Table 2.1 presents the mean of dichotomous/ categorical baseline variables between treated and untreated groups. We can use the Wilcoxon rank-sum test/chi-squared test to check the prevalence of variables between groups for continuous and categorical variables, respectively.

Also, we report the standardized difference for each of the ten baseline variables in the original lung cancer data, untrimmed or unweighted, in table 2.1. There are standardized difference values in the original data that exceeded 0.10, with an adjuvant RT covariate having the greatest standardized differential (1.045). It is emphasized that there are most of the covariates are particularly unbalanced. One of the imbalanced issues could be that there is a remarkable difference in the percent between chemotherapy and non-chemotherapy groups. Such differences may bias a simple comparison of outcomes by treatment status and suggest that, at the very least, adjustments for pre-treatment differences are required to obtain credible inferences for the causal effect of chemotherapy exposure on outcomes. That's why we apply propensity score trimming that removes the imbalance of those differences between groups in the next step.

Table 2.1: Baseline characteristics of treatment and control subjects in lung cancer dataset

	CONTROL **N=256**	**TREATMENT** **N=94**	**SMD**	**P value**
AGE**=TRUE(%)**	127 (49.6)	36(38.3)	0.229	0.078
GENDER**=2(%)**	140(54.7)	49(52.1)	0.051	0.760
ADJUVANT RT**=Yes (%)**	24(9.4)	49(52.1)	1.045	< 0.0001
RACE **(%)**			0.496	< 0.0001
White	216(84.4)	66(70.2)		
Unknown	25(9.8)	26(27.7)		
Other	15(5.9)	2(2.1)		

SURGICAL MARGINAL **(%)**			0.112	0.678
All	242(94.5)	91(96.8)		
Microscopically	5(2.0)	1(1.1)		
Unknown	9 (3.5)	2(2.1)		
SITE **(%)**			0.531	< 0.0001
DFCI	23 (9.0)	25(26.6)		
HLM	79 (30.9)	19(20.2)		
MI	78 (30.5)	19(20.2)		
MSKCC	76 (29.7)	31(33.0)		
HISTORICAL GRADE **(%)**			0.437	0.437
Moderate	118(46.1)	41(43.6)		
Poorly	85(33.2)	39(41.5)		
Unknown	13(5.1)	4(4.3)		
Well	40(15.6)	10(10.6)		
SMOKING HISTORY **(%)**			0.118	0.802
Currently	21(8.2)	8(8.5)		
Never	31(12.1)	15(16.0)		
Smoked	197(77.0)	69(73.4)		
Unknown	7 (2.7)	2(2.1)		
PATH N STAGE **(%)**			0.413	< 0.0001
N0	183(71.5)	43(45.7)		
N1	37(14.5)	26(27.7)		
N2	27(10.5)	23(24.5)		
Unknown	9(3.5)	2(2.1)		
PATH T STAGE **(%)**			0.583	<0.0001
T1	97(37.9)	13(13.8)		
T2	134(52.3)	71(75.5)		
T3	15(5.9)	5(5.3)		
Unknown	10(3.9)	5(5.3)		

Table 2.2 presents summary statistics and standardized mean difference in trimming sample. One observes that trimming on the propensity score has diminished or eliminated many systematic differences in means or prevalence between treated and untreated subjects reported in Table 2.2, which compares it with Table 2.1. In this case, trimming the sample by removing units with extreme values of the estimated propensity score to improve overlap should lead to more robust inferences at the subsequent analysis stage.

Table 2.2: Baseline characteristics of treatment and control subjects in the trimming sample

	CONTROL **N=83**	**TREATMENT** **N=59**	**SMD**
AGE=**TRUE(%)**	44(53.0)	25(42.4)	0.214
GENDER=**Male(%)**	34(41.0)	33(55.9)	0.033
ADJUVANT RT=**Yes(%)**	20(24.1)	22(37.3)	0.289
RACE (%)			0.282
WHITE	64(77.1)	38(64.4)	
UNKNOWN	19(22.9)	21(35.6)	
OTHER	0	0	
SURGICAL MARGINAL (%)			0.132
ALL	78 (94.0)	56(94.9)	
MICROSCOPICALLY	3 (3.6)	1 (1.7)	
UNKNOWN	2 (2.4)	2 (3.4)	
SITE (%)			0.371
DFCI	22(26.5)	24(40.7)	
HLM	17(20.5)	14(23.7)	
MI	44(53.0)	21(35.6)	
MSKCC	0	0	
HISTRIGOCIAL GRADE (%)			0.190
Moderate	4(4.8)	3(5.1)	
Poorly	29(34.9)	24(40.7)	
Unknown	40(48.2)	23(39.0)	
Well	10(12.0)	9(15.3)	
SMOKING HISTORY (%)			0.089
Currently	10(12.0)	6(10.2)	
Never	12(14.5)	8(13.6)	
Smoked	59(71.1)	44(74.6)	
Unknown	2(2.4)	1(1.7)	
PATH N STAGE (%)			0.269
N0	2(2.4)	2(3.4)	
N1	49(59.0)	27(45.8)	
N2	21(25.3)	20(33.9)	
Unknown	11(13.3)	10(16.9)	
PATH T STAGE (%)			0.383
T1	2(2.4)	2(3.4)	
T2	10(12.0)	2(3.4)	
T3	63(75.9)	51(86.4)	
Unknown	8(9.6)	4(6.8)	

After we present the implementation of the covariates selection procedure on the lung cancer data set, Figure 2.3 in left side illustrates the propensity score value identification with adjuvant chemotherapy versus without adjuvant chemotherapy in original sample. They display a limited overlap in preintervention characteristics between treated and control groups. In other words, Figure 2.3 reveals a considerable imbalance between treatment and control groups. As Figure 2.3 in left sides indicates, two groups' density differs from each other on the distribution of estimated propensity scores, and so, the common support region is especially problematic. However, the key to applying the trimming method is that if the true propensity score values are equal to zero or one, it is supposed that there are no counterparts with alternative treatment for such units. Hence, We cannot credibly and accurately estimate the effect of treatment. We perform a propensity trimming score technique to reduce the impact of confounding variables on the model. Right side of Figure 2.3 illustrates a remarkable performance in terms of the removing bias between treatment groups. In other terms, overlap between chemotherapy and non-chemotherapy groups seems to be satisfied. It seems that propensity score trimming works well and is reasonably balanced.

Figure 2.3: Density plot of treated and untreated groups in original sample (left) and trimming sample (right)

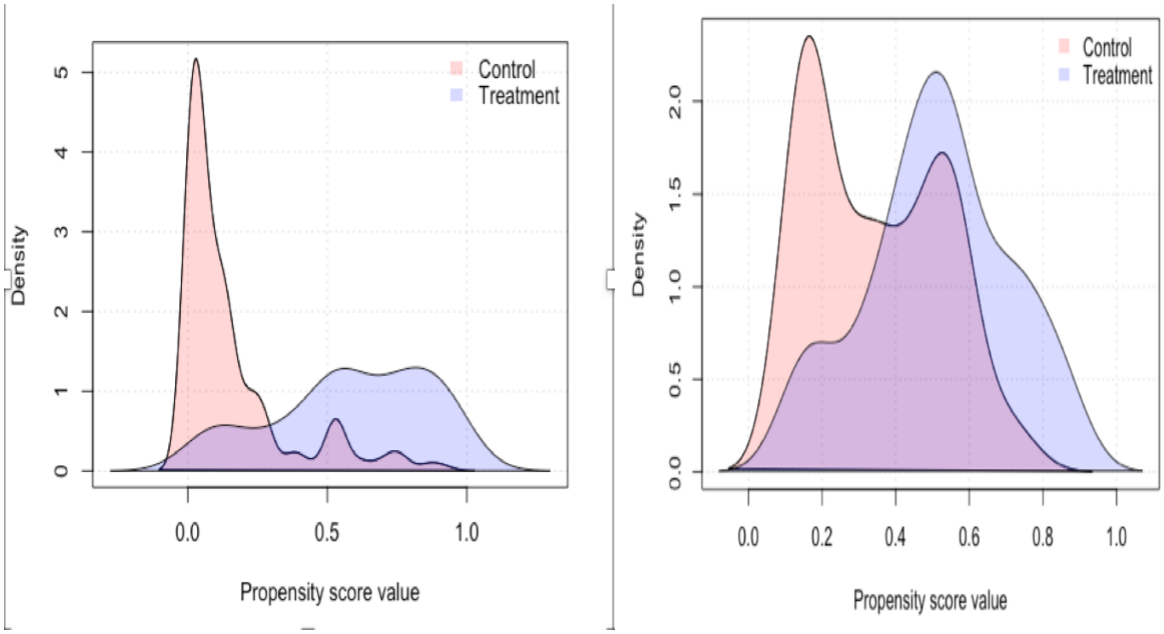


Table 2.3 shows our methods in the lung cancer treatment context. Lung cancer patients are typically treated with adjuvant chemotherapy (treatment arm) and no chemotherapy groups. In other terms, we want to identify a biomarker that can be used to predict which patients are and are not likely to benefit from adjuvant chemotherapy. The final biomarker would have been the gene with expression values corresponding to the “Max” column, and the cutoff value which is a function of the Cox regression parameters.

Table 2.3: Summarizing marker performance that depend on PS trimming and IPTW method

	Parameter	Min	1^{st} quantile	Median	Mean	3^{rd} quantile	Max
Trimming Method	Θ_0	0.0000	0.0000	0.0000	0.0036	0.0035	0.0901
	Θ_1	0.0909	0.1486	0.1529	0.1544	0.1592	0.2432
	No chemo prob	0.6618	0.7106	0.7140	0.7127	0.7157	0.7489
	Chemo prob	0.4968	0.5591	0.5628	0.5619	0.5655	0.5971
	Optimal prob	0.6762	0.7126	0.7150	0.5619	0.7183	0.7914
Sub-classification Method	Θ_0	0.0000	0.0000	0.0000	0.0007	0.0002	0.0352
	Θ_1	0.1252	0.1471	0.1490	0.1504	0.1527	0.1959
	No chemo prob	0.6937	0.7109	0.7118	0.7117	0.7126	0.7271
	Chemo prob	0.5225	0.5602	0.5621	0.5621	0.5649	0.5841
	Optimal prob	0.6937	0.7113	0.7120	0.7125	0.7133	0.7424
IPTW Method	Θ_0	0.0000	0.0000	0.000	0.0006	0.0001	0.0303
	Θ_1	0.1339	0.1573	0.1592	0.1603	0.1627	0.2082
	No chemo prob	0.6941	0.7138	0.7148	0.7147	0.7157	0.7309
	Chemo prob	0.5115	0.5534	0.5561	0.5550	0.5574	0.5774
	Optimal Prob	0.6974	0.7142	0.7150	0.7154	0.7163	0.7421

Abbreviations: PS: Propensity Score , IPTW: Inverse Probability of Treatment Weight

As seen in Table 2.3, Θ_0 , Θ_1, "No chemotherapy probability", "chemotherapy probability" and "optimal biomarker probability" are presented to examine biomarker that explains whether or not chemotherapy is necessary to treat the lung cancer. The optimal biomarker for target population is biomarker treatment that maximizes the probability of surviving past the prespecified time w_0 =41weeks: Θ_0, if we consider the standard of care treatment (no chemotherapy); Θ_1 if the default treatment is "adjuvant chemotherapy treatment"(i.e. T=1 arm). Besides, we presented a method based on the Cox hazard model that provides an unbiased estimate of Θ_0 and Θ_1 in the presence of right censoring. We studied the observational data set herein. So, it is hard to examine relevant results in biomarker because covariates' characteristics can impact treatment assignments. So, we needed to apply appropriate causal effect techniques (i.e., trimming, sub-classification and IPTW) to eliminate bias in estimates of the treatment effect between treatment and control groups. Table 2.3 shows that each method causes getting different results that correspond to Θ_0 and Θ_1 parameters. Θ_0 has a larger range in using the trimming technique than in IPTW. Similarly, the largest probability value of Θ_1(i.e., $\Theta_1 =$0.2432) is hold by trimming methods. We understood from table 3 that different propensity score technique can lead to different the result when assessing the performance of the biomarker. In other words, Table 2.3 illustrates that using the different propensity score methods has influenced the assessment of the biomarker's performance. We understand that removing bias between treated and untreated has been vital to making appropriate therapy decisions.

2.6 Summary & Discussion

This study proposed a descriptive analysis and a summary measure to evaluate cancer biomarkers using observational studies. This paper is a rare study in observational studies investigating causal inference and presenting summary measures for cancer markers. Few studies in the understanding of biomarker rely on observational studies instead of randomized control trial. Using randomized control trial data for evaluating treatment selection guarantees that there should be no systematic difference in treated or untreated covariates between units assigned to the different treatments. But, in a non-randomized observational study, researchers have no control over the treatment. Thus, we perform various propensity techniques, such as propensity score trimming, sub-classification on propensity score and inverse probability of treatment weighting, to eliminate confounding effects when used in observational lung cancer data. According to three methods, Propensity score trimming is an essential and appropriate technique to look at the marker's impact and understand the average benefit of treatment policies suggested.

Dobbin and Song (in revision) proposed that no genes had both Θ_0 and Θ_1, values greater than 0.001 when any causal inference methods were not considered in their research proposal. In addition, Shedden et al.(2008) studied various approaches, not rely on the propensity score adjustment, establish simply weak signal genes. However, we found a group of genes with a value roughly 0.1 or over 0.1 for Θ_0 and Θ_1, respectively. According to proof of results in Section 2.5, propensity score methods have been remarkably significant for describing treatment selection. To be clear, each propensity method, i.e., trimming, sub-classification and weighting, had different performance in removing bias between adjuvant chemotherapy and standard therapy. The results suggest that treatment selection based on the propensity score trimming method has performed slightly better than treatment selection on other propensity scores. To sum up, we strongly recommend using any propensity score techniques that give sensible results in observational studies. In other terms, we understand that propensity score adjustment has been vital to employ in the non-randomized trial.

2.7 Appendix

Figure : Histogram-based estimate of the distribution of propensity score for treated and untreated group for original sample

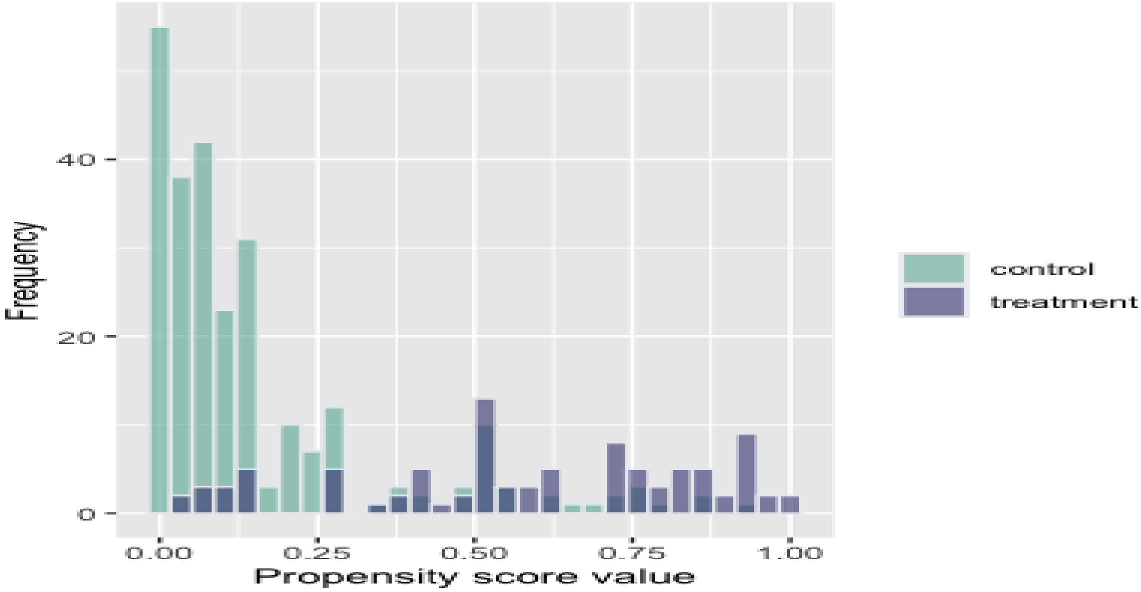
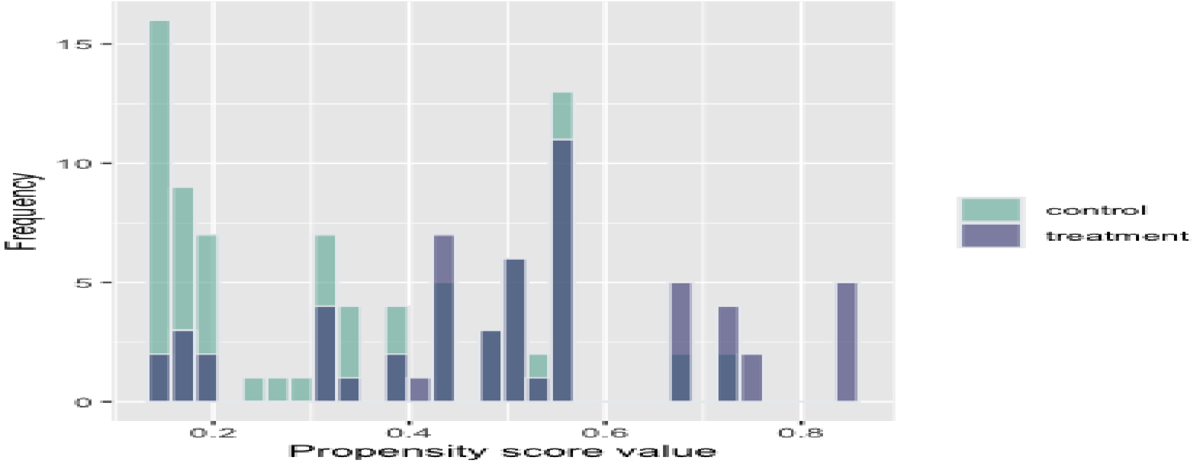


Figure: Histogram-based estimate of the distribution of propensity score for treated and untreated group for trimming sample



CHAPTER 3

Development of Biomarker in Propensity Score-Adjusted of Parametric and Machine Learning Methods

3.1 Introduction

Treatment selection biomarkers are indispensable tools for determining whether or not a participant improves from a specific treatment. Numerous research on tumor markers has been conducted in oncology throughout the years, but the number of indicators that have been found to be medicinally valuable is few. In addition, while markers in initial studies often illustrate great promise, large inconsistencies are observed in subsequent studies conducted with the same marker. Discrepancies have been attributed to various factors, including general methodologic variations, inadequate research design, non-standardized or non-reproducible assays, and improper or misleading statistical analyses. So, choosing the proper treatment selection biomarker is highly important to ensure benefits for the patient. In literature, there are various approaches to illustrate treatment selection biomarkers. Some researchers have focused on the descriptive analysis of modeling treatment selection (see Bonetti and Gelber ,2004; Cai et al.,2011; Janes et al.,2014) while others have studied optimizing markers for treatment selection (Lu et al.,2011; Gunter et al.,2011; McKeague and Qian,2013).

The main question in the assessment of treatment selection is what statistical method we should use to identify essential markers. A well-known methodology in the development of treatment selection markers in the literature is the statistical interaction between marker and treatment arm as a fundamental measure of biomarker performance(see Sargent et al. ,2005; Simon,2008; Janes,2011; Janes et al,2013). Even though the interaction between marker value and treatment assignment is essential, it is not sufficient to determine marker performance(Huang et al. ,2012; Janes,2012). Most of the studies on treatment selection have been based on data coming from a randomized clinical trial. The treatment effect can be relatively simple to estimate in a randomized study. However, when we consider observational studies, we never know what determined the treatment selection process because it is non-randomized.

In many scientific circumstances, researchers want to know how an intervention affects an outcome. In many cases, allocation of the intervention and evaluation in a randomized clinical trial can offer rigorous assessment, but, in other cases, such research is not possible owing to ethical restrictions. This has stimulated a lot of study in causal inference, especially using the potential outcomes approach. Causal inference is required due to imbalances in baseline factors between treated and untreated, which can act as confounders. In contrast to a randomized trial, the assignment system in an observational study is not controlled by the scientist. Therefore it is unknown how subjects' attributes impact their chances of being assigned to the treatment or control group in the observational study. As a result, the data must be used to assess the participants' chances of obtaining treatment. A coherent approach for examining causal effects of the treatment effect on the outcome recommended by Rubin (1974) and Rosenbaum and Rubin are based on the potential outcomes proposed by Rosenbaum and Rubin (1983). Propensity score aims to reduce confounding bias in treatment average effect estimates. So, it assists in achieving this aim by predicting the exposure's probability provided individual covariates and, especially, by employing a propensity score to establish balance on confounders. The fundamental assumption is that, given similar propensity scores for exposed and unexposed individuals, treatment assignment for these two individuals is independent of all confounding variables. So the two observations may be utilized as counterfactuals for causal inference. This method reduces the need to balance a multivariate set of observable features to the simpler task of adjusting based on a one-dimensional propensity score.

An important topic covered in the paper is the selection of the propensity score modeling approach. It's feasible that alternative methods for estimating the propensity scores will result in different treatment effect estimates. Propensity scores have traditionally been evaluated using logistic regression. Some papers (see Westreich et al.,2010; Lee,2010; Wyss et al., 2014) address some of the pros and drawbacks of logistic regression for propensity score prediction. This paper will discuss and examine the relative advantages and disadvantages of various representative methods. Imai and Ratkovic (2014) suggested covariate balance propensity score as an alternative parametric approach to logistic regression. CBPS technique substitutes MLE with an extended form of moments estimation to improve treatment assessment estimation and covariate balancing simultaneously. Despite the fact that the CBPS has been demonstrated to work well in some particular circumstances, it has not been used in medical settings or with a large dataset. Even though parametric methods, especially logistic regression, are more preferable techniques, parametric methods, on the other hand, need assumptions about the variable selection process, covariates' distributions, and interaction term definition. So, there can be an imbalance between treated and untreated groups in covariates if the models do not meet the assumptions or are defined wrong. Note that throughout this thesis, variable selection refers to the variables used in the model fitting and not to the process of selecting variables from the full list of all variables for inclusion in the model. So, there can be an imbalance between treated and untreated groups in covariates if the models do not meet the assumptions or are defined incorrectly.

Machine learning techniques might be used in place of parametric approaches, i.e., logistic regression (LR) or covariate balance propensity score (CBPS). In health care research, machine learning is a machine algorithm that is growing more widespread. Even though many studies have been focused on prediction issues, the latest advances in machine learning (ML) have expanded their implementations from predictive models into the statistical inference field, allowing for more widespread use in the area(Connell and Lindner ,2019; Fang et al. ,2011).

Machine learning is a broad concept that encompasses a wide range of categorization and prediction algorithms with uses ranging from economics to health care, engineering, accounting fields. While statistical techniques to modeling presume a data model with parameters determined from the data, machine learning uses an algorithm to identify an association between the result and a predictor without utilizing any data model. We apply some machine learning techniques, such as generalized boosted model (GBM), random forest (RF), bagging (BAG), and classification and regression trees (CART) models, to estimate propensity scores. Then those values are employed to perform one-to-one matching.

When these methodologies aim to reduce the influence of treatment selection bias, the researcher seeks to identify which factor/covariates to include in or exclude from the estimation model. Any factors may be characterized in terms of two features for a particular treatment and outcome. These relationships are defined as associated with treatment, outcome, both of them or neither of them. So, it is highly critical to determine whether included variables should be connected with treatment, outcome, both of them, or confounders variables. There are limited studies for variable selection (Austin et al., 2007; Brookhart et al.; 2004). However, there is no comprehensive review to explore variable importance in several parametric and machine learning methods.

This paper is motivated by two fundamental purposes. Firstly, we investigate the performance of one-to-one matching techniques depending on LR, CBPS, GBM, RF, CART, and BAG methods when considering different sets of variables in the models. Secondly, we present a detailed framework for evaluating markers in the context of treatment selection. The proposed tools (Janes et al.,2012) are utilized to identify descriptive analysis and summary measurements. In literature, all studies are based on a randomized trial. However, we present the methods that rely on using the results of causal inference under many scenarios. At the end of the study, we reveal a general framework for the performance of causal inference and then treatment selection biomarkers.

The remainder of this paper is laid out as follows. Section 3.2 highlights the major fundamental principles of propensity score methodologies and the disadvantages or advantages of employing machine learning techniques for propensity score estimates and balance diagnostics. We discuss about variable selection importance and previous studies in Section 3.3.Also, the treatment selection process is summarize in Sect. 3.4. We present the comprehensive results from the simulation study to examine the performance of parametric and non-parametric methods and then performed marker assessment for treatment selection in Sect 3.5 .We present a summary of our results as well as some recommendations for professionals interested in using propensity score approaches in Section 3.6.

3.2 Parametric and Machine Learning Methods

3.2.1. The Framework of Causal Effects

In the potential outcomes framework, let T be the treatment variable, Y signify outcome (or sometimes called response), and X be the vector of baseline covariates. Thus, we define data as $(Y_i, T_i, X_i), i = 1, \ldots, n$, and random sample (Y, T, X). A pair of potential outcomes: $Y(0)$ and $Y(1)$ represent the potentially unobserved response under the control and treatment groups, respectively. We denote T=1 if an individual is assigned to active treatment, T=0 if control. We can observe an outcome as $Y\ \big(Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)\big)$ in the case of a binary treatment.

3.2.2. Methods to Estimate the Propensity Score

Researchers in various fields have recently introduced several propensity score estimation approaches that emphasize reducing the covariates' imbalance. There are two main guidelines as parametric and nonparametric methods for predicting propensity scores under any of the propensity score approaches in the literature. Logistic regression is a parametric approach commonly preferred to estimate propensity score. Later, CBPS is recently proposed alternative methods to logistic regression. Nevertheless, parametric methods require meeting with assumptions concerning variable selection, covariates' distributions, or defining interaction terms. Also, Nonparametric techniques, i.e., random forest, generalized boosted model, bagging and classification, and regression trees methods, are reviewed in terms of the theoretical and application process in this section.

3.2.2.1 Logistic Regression

A great deal of effort has been taken into developing techniques for estimating propensity scores. In fact, parametric approaches along with logistic regression are most frequently used to predict propensity scores. Logistic regression is also a well-known technique to estimate the conditional probability of receiving treatment when there are two treatment conditions. Logistic regression starts including main effects for supplied covariates characteristics. Logistic regression starts including main effects for supplied covariates characteristics and then adding squared terms of variables and interaction terms of covariates to enhance propensity score values if adequate balance is not achieved. The logistic regression is written to estimate propensity score :

$$\text{logit}\ (T_i = 1|X) = \delta_0 + \delta_1 X_{1i} + \cdots + \delta_p X_{pi}$$

where let covariates be $X_i, \ldots, X_p$ with p independent variables. δ_0 is described as an intercept and δ's are unknown parameters. The log odds of the probability is as follows:

$$\text{logit}\ (T_i = 1|X) = log\left(\frac{\Pr\,(T_i=1)}{1-Pr(T_i=1)}\right) \tag{3.1}$$

Equation (3.1) is re-written for the estimated propensity scores

$$\mathrm{e_i(X)} = \frac{\exp(\mathrm{logit(T_i} = 1|\mathrm{X})}{1 + \exp\big(\mathrm{logit(T_i} = 1|\mathrm{X})\big)}$$

We can predict the propensity score using the maximum likelihood estimator:

$$\hat{\beta}_{\mathrm{MLE}} = \arg\max_{\beta\in\Theta} \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{T_i}\log\{\mathrm{e(X_i)}\} + (1 - \mathrm{T_i})\log\{1 - \mathrm{e(X_i)}\}$$

where $e(.)$ is twice differentiable with respect to β maximizes the likelihood function,

$$\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{s}_{\beta}(\mathrm{T_i}, \mathrm{X_i}) = 0, \mathrm{s}_{\beta}(\mathrm{T_i}, \mathrm{X_i}) = \frac{\mathrm{T_i}\ \mathrm{e}'(\mathrm{X_i})}{\mathrm{e}(\mathrm{X_i})} - \frac{(1-\mathrm{T_i})\mathrm{e}'(\mathrm{X_i})}{1-\mathrm{e}(\mathrm{X_i})}, \tag{3.2}$$

where $\mathrm{e}'(\mathrm{X_i}) = \partial\ \mathrm{e}(\mathrm{X_i}) / \partial\beta^{\mathrm{T}}$ is the gradient. So, we state that equation 2 represents first derivation of $e(X_i)$.

3.2.2.2 Covariate Balance Propensity Score

There may be deviations in estimated propensity score when considering parametric models such as logistic regression due to the incorrect model specification, for example, when the true propensity score model is not logistic. This has led researchers to use different parametric models to minimize imbalance in treated and untreated groups and reduce bias and variability. One of the popular alternative approaches to estimating the propensity score was discovered by Imai and Ratkovic in 2014.Various CBPS models to estimate causal effects have been presented in the literature (see Hainmueller,2012; Graham et al. ,2012). However, the difference between Imai and Ratkovic's paper and other papers is that it is based on a single model for determining treatment assignment and covariate balancing weightings. Estimating treatment assignment based on the CBPS model is usually implemented with the generalized method of moments or empirical likelihood framework.

CBPS technique is a parametric model and has remarkable advantages for estimation of causal effects. CBPS estimation helps reduce the causal effects misidentification in the parametric model by choosing parameter values that make important covariates balance. Even if the CBPS model is correctly determined, the CBPS method may further improve the balance of covariates in observational data compared to using logistic regression. In addition to maximizing the model likelihood, the covariate balancing technique includes a balance requirement for the weighted averages of the factors in the variable prediction process. As mentioned earlier, the crucial challenge of standard approaches e.g., logistic regression, is that misidentification of models can lead to biased estimates in treatment effects. It may be appropriate to use more complex non-parametric models, but covariate X's with high dimensionality can challenge the estimating propensity score. . In this case, CBPS estimation is a robust method chosen to mitigate a parametric model's misrepresentations.

Imai and Ratkovic (2014) proposed a logistic regression model, i.e.,

$$e(X) = e_\beta(X) = \frac{1}{1 + exp\{-\beta' X\}}$$

Then β is solved by satisfying the following condition:

$$E\left\{\frac{T\tilde{X}}{e_\beta(X)} - \frac{(1-T)\tilde{X}}{1-e_\beta(X)}\right\} = 0 \tag{3.3}$$

where $\tilde{X} = f(X)$ is measurable function of X. Choosing $f(.)$, i.e. $\tilde{X} = \frac{\partial\, e_\beta(X)}{\partial\beta}$ is solved the maximum likelihood estimator (MLE) of β because equation (3.3) is the score function of MLE. The above balancing condition is for the estimation of ATE. Besides, we use CBPS method to estimates the parameters of propensity score by solving estimating equation:

$$\overline{g_\beta}(T,X) = \frac{1}{n}\sum_{i=1}^{n} g(T_i\,, X_i) = 0,$$

where

$$g(T_i\,, X_i) = \left(\frac{T_i}{e(X_i)} - \frac{1-T_i}{1-e(X_i)}\right) f(X_i) \tag{3.4}$$

for some covariate balancing function $f(.)\colon \mathbb{R}^p \to \mathbb{R}^n$ when we hold that parameter number, p is equal to (3.4) equation numbers, n. However, if $n > p$, we estimate $\hat{\beta}$ by optimizing the covariate balance using the generalized method of moments (GMM) method :

$$\hat{\beta} = argmin_{\beta\in\Theta}\, \bar{g}_\beta(T,X)^T \widehat{W}\, \bar{g}_\beta(T,X)$$

where Θ is the parameter space for β in $\mathbb{R}^p$ and $\widehat{W}$ is an $(n \times n)$ positive define weighting matrix. For estimating ATT, the balancing condition gets

$$E\left\{T\tilde{X} - \frac{e_\beta(X)(1-T)\tilde{X}}{1-e_\beta(X)}\right\} = 0$$

3.2.2.3 Generalized Boosted Model

Researchers prefer using different approaches when they acknowledge logistic regression has disadvantages in estimating the propensity score. Nonparametric methods (i.e., boosting) have been shown to outperform parametric methods (i.e., logistic regression or CBPS methods) to estimate propensity score for dichotomous or multiple treatment factors. Boosting is an automated and data-adaptive algorithm. It can be used with many pretreatment covariates to fit several models through a regression tree and predict treatment assignment. It is an ensemble method that combines simple models into a nonparametric approach. There are many variants of boosting studies in machine learning, such as the AdaBoost algorithm by Freund and Schapire(1997), generalized boosted models by Ridgeway in 1999, LogitBoost by Friedman et al. in 2000, and gradient boosting machine by Friedman in 2001. McCaffrey et al. (2004) recommended one of the versions in machine learning to estimate propensity score using a generalized boosted model (GBM). So, GBM derives propensity scores by fitting numerous regression trees given the covariates repeatedly. After that, it is linearly merging all of the regression trees to obtain a smoothed function for the overall estimate of propensity scores.

Moreover, interactions between covariates and the treatment variables, or between covariates and nonlinear variables can be systematically incorporated because all machine learning algorithms are nonparametric structures. Like logistic regression, GBM models can be written as

$f(X) = log\frac{Pr(T=1|X)}{1-Pr(T=1|X)}$. Then, we start algorithm with log-odds of treatment as $f(X) = log\frac{\widehat{Pr}(T=1|X)}{1-\widehat{Pr}(T=1|X)}$,where let $\widehat{Pr}(T = 1)$ is the average probability of the treatment indicator variable in the sample. Let $f(X)$ denote iteration updating to $f(X) + \gamma . h(X)$, where γ represents a shrinkage factor and fitting regression trees estimate $h(X)$. Also, shrinkage parameter helps decreasing variance with small adjustments without growing bias and so, small shrinkage parameter might give a more accurate fit for the model.

3.2.2.4 Random Forest

There has been much interest in "ensemble learning" methods based on decision trees, representing classification or regression. The most known methods are boosting (Shapire et al., 1998) and bagging(Breiman,1996) of decision trees and random forest (Breiman,2001). The most famous tree-based algorithm is random forests that first recommend by Breiman (2001). It corresponds to the class of nonparametric methods, which build multiple classification trees rather than just one. So, it selects a random subgroup of the variables at every node of the tree, and then, a node is divided utilizing the optimal split among the chosen variables.1 It corresponds to a class of nonparametric methods, which build multiple classification trees rather than just one. It selects a random subgroup of the variables at every node of the tree, and the node is divided. Each tree is individually constructed, relying on a bootstrap of the data set's sample. Finally, a simple majority vote is utilized to make a prediction. Each tree was generated using all of the data. Consequently, each observation's propensity score was predicted and between any pair of observations, respectively, based on each tree. The paper results present using 500 trees and nodes to use a predetermined minimum size of 25, according to Zhao et al.(2016). Breiman and Cutler (2016) provide a "randomForest" package in R, and it is very user-friendly in the sense that easy implementation.

3.2.2.5 Classification and Regression Trees (CART)

One of best-known and oldest machine learning techniques is classification and regression trees developed by Breiman et al.,1984. CART is a kind of decision tree, which is known with regards to easy implementation and interpretation. While some non-parametric models, such as neural networks or support vector machines, do not offer the probabilities of class membership, fortunately, CART is eligible to supply probabilities (Westreich et al. 2010). Hastie et al.(2001) provide that CART, on the other hand, is classified as an unstable learner due to its bias towards overfitting. CART is a recursive automated system for identifying the most relevant explanatory factors (x) in deciding the dependent variables (y) to be interpreted from a vast number of explanatory factors (x). CART is constructed based on a classification tree and a regression tree. So, the split at each phase is determined by selecting the variable that minimizes the prediction error or classification error. So, Relative error is specified to minimizes the sum of a square as follows:

$$RE(d) = \sum_{k=0}^{K} (y_k - \overline{y_K})^2 + (y_s - \overline{y_S})^2$$

let y_k and y_s defined as left and right partition corresponds to K and S observations of y in each step with means $\overline{y_K}$ and $\overline{y_S}$.

3.2.2.6 Bagging

Bootstrap aggregating, also known as bagging, is an ensemble approach. Breiman(1996b) recommended the method of bagging. The method's aim seeks to decrease the variance of a predictor in order to increase prediction efficiency. Also, it can be utilized to increase the predictability and consistency of classification and regression trees. However, it has not had any restriction to advance tree-based predictions. Bagging is a technique for generating independent classification trees from a set of bootstrap samples selected randomly from a set of results. The data will differ significantly from the prior bootstrap study for each new sample. Besides that, every tree will differ significantly from the one before it. The algorithm then averages the expected category participation probability over the whole set of classification chains. When the baseline regression or classification technique being bagged is not particularly reliable, bagging performs well. In addition, bagging can yield a notable decrease in average prediction error when minor modifications in the learning sample can often result in considerable variations in the predictions made using a defined technique (Sutton ,2004; Breiman et al. , 1984;Breiman 1996b; Breiman 2001a).

3.2.3 Propensity Score Matching

Matching is an intriguing statistical tool for estimating the impact of treatments owing to its clarity and simplicity whereby the results may be presented. Considerable diversity of different matching algorithms have been explored in literature. This study intends to review matching on propensity score (pair matching), which is focused on a scalar function of the covariates, and is utilized to balance all variables and mimic randomization. There are two main objectives in applying propensity score matching. The primary purpose is to remove systematic biases associated with differences in observed covariates when adjusting differences in propensity score between treatment and control groups. Secondly, it is easier to determine near matches on scalar variables than to get close matches on all variables jointly.

Let N_t represents the total number of treated units, indexed by $i = 1, \dots, N_t$, and a set of potential controls , of size N_c' , larger than N_t. We select $N_c < N_c'$ units from this set to establish a sample of size $N = N_c + N_t$ that will be utilized to estimate treatment effects. Let $\mathbb{I}_c'$ symbolize the pool of indices for set of possible controls, $\mathbb{I}_c' = \{N_t + 1, \dots, N_t + N_c'\}$. We concentrate on the difficulty of selecting a subgroup $\mathbb{I}_c$ of the total control sets, $\mathbb{I}_c \subseteq \mathbb{I}_c'$, that has better balance with respect to the treated units than a random sample of the full set of possible controls. For the sake of clarity, this procedure will be applied to case M=1 throughout this paper. Fixing $N_c = N_t$ may be a reasonable choice if we consider the effect of N_c on the sampling variance of estimators for causal effects. We simply denote $\mathbb{I}_t = \{1, \dots, N_t\}$ ordered set of indices for treated units. Assume that the treated units are sorted depending on the propensity score value. The largest average propensity score is matched first, which corresponds to matching the units that are a priori the most difficult to match first in many real data problems. $d(x, x')$ describe a measure of "distance" between two vectors of covariates. Let $\mathcal{M}_i^c \subseteq \mathbb{I}_c'$ denote the set of matched controls for treated unit i . $\mathcal{M}_i^c = \{m_i\}$, where $\{m_i\}$ is the index of control units that is matched to treated unit i . For the i^{th} treated unit, the set containing the closest match is

$$\mathcal{M}_i^c = \left\{ j \in \mathbb{I}_c' - \cup_{i'=1}^{i-1} \mathcal{M}_{i'}^c \mid d\left(X_i , X_j\right) = \ min_{j' \in \mathbb{I}_c' - \cup_{i'=1}^{i=1} \mathcal{M}_{i'}^c} d\left(X_i, X_{j'}\right) \right\}$$

where $\mathbb{I}_c' - \cup_{i'=1}^{i=1} \mathcal{M}_{i'}^c$ is subset of $\mathbb{I}_c'$ excluding the set of all the control units used as matches, $\mathbb{I}_c = \cup_{i'=1}^{i-1} \mathcal{M}_i^c$ with N_t distinct elements.

As indicated in the previous part, the estimated propensity score is denoted as $\hat{e}(x)$ and then we define

$$d_l(x,x') = \left(\hat{l}(x) - \hat{l}(x')\right)^2 = \left(ln\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right) - ln\left(\frac{\hat{e}(x')}{1-\hat{e}(x')}\right)\right)^2$$

where $\hat{l}(x)$ is linearized estimated propensity score (lps) or the logarithm of the odds ratio. This is the squared Euclidean distance between the linearized propensity scores (Hansen,2004; Austin et al.,2007; Imbens and Rubin ,2015).

3.3 Variable Selection for Propensity Score Models

The propensity score method seeks to decrease the effectiveness of treatment selection bias when assessing treatment effects in non-randomized studies. Estimated propensity scores are employed to assure that the distribution of measured risk indicators for the outcome between treatment and outcome groups is similar and adjust for confounding variables. Furthermore, the PS is not exclusively intended to predict treatment well. Balancing covariates so as to control confounding and create a model for the prediction of treatment are different goals that require different approaches to variable selection. This raises an important question: Which covariates in the propensity score model should be added or exempted. It might be that variables that influence the treatment selection should be included. One of the essential features of the propensity score is highlighted in these circumstances: It is a balancing score.

Consequently, there might not be equal importance for all covariates when we consider balancing scores. According to Austin et al. (2007), the true model (factors connected to treatment and result) and the confounder model (variables related to the outcome alone) can be preferable to models that solely incorporate variables that impact the treatment selection procedure in predicting propensity score. Myers et al. (2011), Wooldridge (2009), and Brookhart MA (2006) found that instrumental variables related to treatment but not the outcome can cause inflation in bias and variance of the treatment effect estimate.

Moreover, the treatment-outcome relationship based on the measured baseline covariates is crucial for identifying causal diagrams. (Austin and Stuart, 2015). This relationship is described for four categories of variables. First of all, if all covariates are linked with both treated and untreated assignments, it is called a true confounder model. The second definition is a true propensity confounder which is when covariates are related to only treatment but not the outcome. Third, some variables are related to the outcome but not treatment. This model is called a potential confounders model. Lastly, all measured variables are included is a full model.(Austin et al. , 2007).

3.4 Evaluate Biomarkers for Treatment Selection

3.4.1 Setting

In this study, two treatment alternatives are considered as "treatment"(T=1) and "Control" (T=0). Clinical binary outcome is denoted as Y, state whether or not outcome represents death after providing treatment/control. The marker, D is useful to explain the subgroup that can be avoided or defines the necessity of treatment.

3.4.2 Treatment Rule ,Estimation and Summary Measures

Janes et al.[2014] stated that the absolute treatment effect given marker value in randomized control trial in the following:

$$\Delta(D) = P(Y=1|T=0,D) - P(Y=1|T=1,D) \tag{3.5}$$

when there is no benefit from treatment, marker performances are considered based on the rule

$$\Delta(D) < 0$$

Δ(D)<0 and Δ (D)>0 are called as “marker negatives” and “marker-positive”, respectively. Generalized linear regression risk model is considered with an interaction between treatment and marker in following:

$$g\big(P(Y = 1|T, D)\big) = \beta_0 + \beta_1 T + \beta_2 D + \beta_3 TD \tag{3.6}$$

where g is denoted as logit function.

Following Janes et al. (2014), we propose a comparison of marker performance based on the characterization of the treatment rule at following:

- Proportion marker-negative,

$$P_{neg} = P(\Delta(D) = 0)$$

- The average utility of untreated among marker-negatives

$$B_{neg} = P(Y = 1|T = 1, \Delta(D) = 0) - P(Y = 1|T = 0, \Delta(D) = 0)$$

$$= E(-\Delta(D)|\Delta(D) = 0)$$

- The average utility of treated among marker-positives,

$$B_{pos} = P(Y = 1|T = 1, \Delta(D) = 1) - P(Y = 1|T = 0, \Delta(D) = 1)$$

$$= E(\Delta(D)|\Delta(D) = 1)$$

- Decreasing in the population event rate in marker-based treatment assignment

$$\Theta = P(Y = 1|T = 1) - [P(Y = 1|T = 1, \Delta(D) = 1)P(\Delta(D) = 1)$$

$$+ P(Y = 1|T = 0, \Delta(D) = 0)P(\Delta(D) = 0)$$

$$= [\mathrm{P}(\mathrm{Y} = 1|\mathrm{T} = 1, \Delta(\mathrm{D}) = 0) - \mathrm{P}(\mathrm{Y} = 1|\mathrm{T} = 0, \Delta(\mathrm{D}) = 0)]\mathrm{P}(\Delta(\mathrm{D}) = 0)$$

$$= B_{neg} P_{neg}$$

where we assume that $P(Y=1|T,\Delta(D)=0)=0\ if\ P(\Delta(D)=0)=0$.So, Θ is acceptable as measure of treatment selection performance. The risk and treatment effect estimates are written as

$$\hat{P}(Y=1|T=0,D)=\widehat{R\imath sk}_0\ (D)=g^{-1}\big(\hat{\beta}_0+\hat{\beta}_2D\big),$$

$$\hat{P}(Y=1|T=1,D)=\widehat{R\imath sk}_1\ (D)=g^{-1}\big(\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2D+\hat{\beta}_3D\big)$$

$$\text{So, }\hat{\Delta}_1(Y)=\widehat{R\imath sk}_0-\widehat{R\imath sk}_1$$

Equation (3.5) identifies the estimation of direct treatment effects given marker value. Because covariates characteristics have not affected the treatment estimation of treatment effects is straightforward. We can propose the absolute treatment effect given marker value:

$$\Delta(D)=P(Y=1|T=0,D,X)-P(Y=1|T=1,D,X)\ \ \ldots \tag{3.7}$$

Risk model involves one of covariates ,which is related to treatment subject:

$$g\big(P(Y=1|T,D,X)\big)=\beta_0+\beta_1T+\beta_2D+\beta_3TD\ldots \tag{3.8}$$

I consider risk and treatment effects predicts that result from fitting the model (3.8) is given

$$\hat{P}(Y=1|T=0,D,X)=\widehat{R\imath sk}_0\ (D)=g^{-1}\big(\hat{\beta}_0+\hat{\beta}_2D\big),$$

$$\hat{P}(Y=1|T=1,D,X)=\widehat{R\imath sk}_1\ (D)=g^{-1}\big(\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2D+\hat{\beta}_3D\big)$$

$$\text{So, }\hat{\Delta}_2(Y)=\widehat{R\imath sk}_0-\widehat{R\imath sk}_1$$

"Empirical" and "Model-based" estimators are written as

$$\widehat{B^{e}_{neg}}=\widehat{Pr}\big(Y=1\big|T=1,\hat{\Delta}(D)=0\big)-\widehat{Pr}\big(Y=1|T=\widehat{0,\Delta}(D)=0\big)$$

$$\widehat{B^{m}_{neg}}=\hat{E}\big(-\hat{\Delta}(D)|\hat{\Delta}(D)=0\big)$$

$$\widehat{B^{e}_{pos}} = \widehat{Pr}\left(Y = 1\middle|T = 0, \hat{\Delta}(D) = 1\right) - \widehat{Pr}\left(Y = 1|T = \widehat{0, \Delta}(D) = 1\right)$$

$$\widehat{B^{m}_{pos}} = \hat{E}\left(\hat{\Delta}(D)|\hat{\Delta}(D) = 1\right)$$

$$\widehat{Pr_{neg}} = \widehat{Pr}\left(\hat{\Delta}(D) = 0\right)$$

$$\widehat{\Theta^{e}} = \widehat{B^{e}_{neg}}\ \widehat{Pr_{neg}}$$

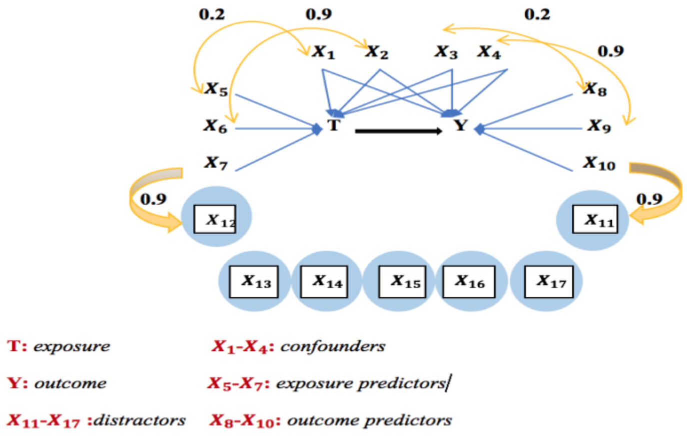
$$\widehat{\Theta^{m}} = \hat{B}^{m}_{neg}.\widehat{Pr_{neg}} = \int -\hat{\Delta}\,(D) I\left[\hat{\Delta}(D) = 0\ \right] d\hat{F}_{\Delta}$$

where e and m superscripts define empirical and model-based estimators, we denote $\widehat{Pr}$ to an empirical probability estimate and $\hat{E}$ to an empirical mean.

3.5 Simulation Study

In this section, I demonstrate results from a large simulation study that compares the statistical properties and performance of proposed methodologies with that of various alternative methods. I used modification of the simulation structure defined by Setouchi et al (2008). This study performs a set of Monte Carlo simulations. Twelve variables each varying in their association with the treatment and outcome were considered ; these are shown in the following diagram:

Figure 3.1: Simulation diagram



As seen in above Figure 3.1, covariates $X_1, X_2, X_3, X_4, X_5, X_6$ and X_7 are connected with exposure assignment, whereas seven variables $X_1, X_2, X_3, X_4, X_8, X_9$ and X_{10} are related to outcome variable(outcome predictors). Moreover, only the four covariates X_1, X_2, X_3 and X_4 are related with both treatment and outcome assignments, in that those four covariates are true confounders. But covariates X_{11} -X_{17} are not associated with treatment or outcome and so, those variables are called as distractors variables. But $X_{11}\ and\ X_{12}$ variables are correlated to X_{10} and X_7 variables, respectively.

Seventeen covariates are generated as a mixture of continuous and binary variables. I generate continuous predictors based on the standard normal distribution and the binary variables were dichotomized versions of standard normal distributions. So, $X_1, X_3, X_5, X_6, X_8, X_9$,X_{11} ,X_{13}, X_{14} ,$and\ X_{16}$ variables were dichotomize and X_2, X_4, X_7, X_{10} ,X_{12}, X_{15}, X_{17} variables were represented as continuous variables. Moreover, there are correlations between some of the variables with correlation coefficients varying 0.2 to 0.9. These correlation coefficients are defined before dichotomizing some of the covariates. Also, correlation coefficient set up between two covariates in following: $(X_1\ ,X_5) = 0.2$,$(X_2\ ,X_6) = 0.9$,$(X_3\ ,X_8) = 0.2$,$(X_4\ ,X_9) = 0.9$, $(X_{10}\ ,X_{11}) = 0.9$ and $(X_7\ ,X_{12}) = 0.9$. The data is simulated for a cohort study (n=1000) and also, datasets were generated 1000 times for all scenarios.

3.5.1 Simulation Design

Treatment Simulation and Scenarios

First of all, treatment assignments are generated using logistic regression, covariate balancing propensity score, generalized boosted model, random forest, classification and regression trees and bagging models with function of X_i.Six scenarios are considered for generating treatment assignments and the form is defined by $Pr[T = 1|X_i] = (1 + exp\{-(version + \tau\zeta)\})$. Version is shown below and τ is denoted for this variable's variability effect. ζ represents a random number , standard normal distributed and ζ variable is not associated with treatment or outcome. Various versions for generating treatment assignment are as follows:

A- Additive and linear (main effects terms only):

$Version_A$=α_1X_1+α_2X_2+α_3X_3+α_4X_4+α_5X_5+α_6X_6+α_7X_7

B- Moderate non-linearity (three quadratic terms):

$Version_B$=$Version_A$+$\alpha_1X_2^2$+$\alpha_4X_4^2$+$\alpha_7X_7^2$

C- Mild non-additivity (four two-way interaction terms)

$Version_C$=$Version_A$+$\alpha_1X_1X_3$+$\alpha_2X_2X_4$+$\alpha_3X_4X_5$+$\alpha_4X_5X_6$

D- Mild non-additivity and non-linearity (a quadratic and four two-way interaction terms)

$Version_D$=$Version_C$+$\alpha_2X_2^2$

E- Moderate non-additivity (ten two-way interaction terms):

$Version_E$=$Version_A$+α_1 X_1X_3+$\alpha_2X_2X_4$+α_3 X_3X_5+$\alpha_4X_4X_5$ + $\alpha_5X_5X_6$+ α_5 X_5X_7+ α_1 X_1X_6+$\alpha_2X_2X_3$+α_3 X_3X_4+$\alpha_4X_4X_6$

F- Moderate non-linearity and non-additivity (3 quadratic term and 10 two-way interaction terms):

$Version_F$=$Version_E$+$\alpha_2X_2^2$+$\alpha_4X_4^2$+$\alpha_7X_7^2$

Then, the number is generated based on the uniform distribution between 0 and 1. T is equal to 1 when $Pr[T = 1|X_i]$ (true propensity score value) is larger than random number that is generated based on uniform distribution. Otherwise, treatment assignment set to 0 value. $\alpha's$ variables are defines as $\alpha_1 = 0.8$, $\alpha_2 = -0.25$, $\alpha_3 = 0.6, \alpha_4 = -0.4$, $\alpha_5 = -0.8$, $\alpha_6 = -0.5$, $\alpha_7 = 0.7$, respectively. There is no intercept on the generating treatment model.

Outcome Simulation and Scenarios

The simulated data includes realizations of a dichotomous outcome. Two versions were of the form $Pr[Y = 1|T, X_i](1 + exp\{-(version\ outcome + \varphi\xi)\})$. $"version\ outcome"$ represents the complexity of association between outcome and treatment assignment. If $Pr[Y|T, X_i]$ value is larger than randomly generated number , I denote outcome as 1 value. Otherwise, outcome set up to 0. The" scenario outcome" was obtained in two different ways as follows. The first scenario in the outcome model is promoted to rely on the additive and linearity model with no intercept and treatment exposure -0.4.Also, non-linearity model is considered for second scenario, which includes exponential interaction among variables that associated with outcome and using the same values as the first scenario for intercept and treatment effect. φ terms is random error term in the outcome model. ξ is random error and term and it is generated with standard normal mean 0 and variance 1.5.The outcome versions are generated based on a range of covariates and treatment assignment versions.

Two versions for generating outcome assignments are as follows:

I. Additive and linear outcome model:

$Version\ outcome_i$=β_0+δ_1T+β_1X_1+β_2X_2+β_3X_3+β_4X_4+β_5X_8+β_6X_9+β_7X_{10}

II. Non-linearity outcome model:

$Version\ outcome_{ii}$=β_0+δ_1T+ exp (β_1X_1+β_2X_2+β_3X_3+β_4X_4+β_5X_8+β_6X_9+β_7X_{10})

β's parameters are defined as 0.3, -0.36,-0.73,-0.2,0.71,-0.19,0.26,-0.4 from β_1 to β_7, respectively.

Biomarker Simulations

The biomarker is simulated to compare the results under different scenarios, which rely on the results of propensity score analysis. A biomarker is generated based on the standard normal distribution and second stronger marker considered. This marker is similar in structure to the simulation described by Janes et al. (2015). Thus, each marker is associated to outcome, Y via a logistic regression model

$$logit\ P(Y=1|T,D)=\beta_0+\beta_1 T+\beta_2 D+\beta_3 TD$$

Propensity Score Estimation Scenarios

The Monte Carlo simulation examines how well various propensity score models can balance the seventeen variables between treated and untreated individuals. Estimation propensity scores are generated from an LR, CBPS, GBM, RF, CART, and BAG models using treatment scenarios and outcome versions. Then, estimated values are employed for assessing the pair matching technique. So, we can determine which variables to include in estimation processes. These approaches are modeled after some of those selections. The following alternatives were considered for the model strategies:

PSM1: This model encompasses X_1-X_{10} covariates, which are associated with both exposure or/and outcome.

PSM2: Seven covariates (i.e., X_1-X_7) are connected with treatment assignment and this model is called as a " true propensity score model".

PSM3: This model is referred to *"potential confounder model"* that includes X_1-X_4 and X_8-X_{10} covariates that are related with the outcome subject.

PSM4: The only four main X_1-X_4 covariates, are connected with outcome and exposure subjects at same time, included in model that is called as *"true confounder model"*.

PSM5: This model was created by including X_1-X_{12} covariates. So, it mean that this model covered only two distractor variables.

PSM6: All variables (X_1-X_{17}) are involved in the model is referred to "full model".

Performance metrics

The performance of the different propensity score models fitting approaches was measured using numerous metrics, including bias, SE,RMSE, relative bias and standardized difference.

Bias: We compute bias based on the difference between the mean estimated treatment effect and the true effect set at θ = -0.4.So, it is formulated as $\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\theta_\iota} - \theta\right)$

Empirical Standard Error (ESE): Standard deviation of treatment effects estimates for each simulated data in each scenario represent standard error.

$$ESE = \sqrt{\frac{1}{n}\ \sum_{i=1}^{n}\left(\widehat{\theta_\iota} - \bar{\bar{\theta}}\right)^{2}}$$

where n denoted for number of datasets.

Theoretical standard errors (SE): standard errors are produced based on the standard errors of average treatment effect and then taking average of n standard errors as

$$SE = \frac{1}{n}\sum_{i=1}^{n}\widehat{SE_\jmath}$$

RMSE: It is represented taking square root of means square error for each estimator. It’s formula is$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\theta_i - \theta)^2}$.

Relative Bias: It is computed as $100x\left(\frac{Bias}{\theta}\right)$

3.6 Results

3.6.1 The Results of Simulation in Causal Inference

We performed a comprehensive simulation study to assess the performance of our recommended techniques. We modified the simulation design of Setoguchi et al. (2008). Parametric and machine learning are employed to estimate propensity scores and investigate how the elements in the propensity score model impact the number of matched treatment and control individuals. We performed logistic regression(LR), and covariate balance propensity score (CBPS) as parametric approaches, while generalized boosted model (GBM), Random Forest(RF), classification and regression trees (CART), and bagging(BAG) are employed for the machine learning methods. When Propensity score model scenarios (i.e., PMS1-6) were fitted utilizing parametric and machine learning methods, we used six different treatment scenarios (called by Treatment A-F) and two outcome scenarios (called Outcome1-2) shown in below tables. All results examined rely on 1:1 matching on propensity score data.

Performance of bias and RMSE in Treatment A-D-F and Outcome-1 (Table-3.1):

The performance metric we use in Tables 3.1 and 3.2 below to compare models is the bias and RMSE of the propensity score estimates. The performances of LR and CBPS showed variability in the results in the linearity and additivity treatment (Treatment-A) and linear outcome assignment (i.e., outcome-1) scenarios accompanied by propensity score models (PSM1-6). The performance of CBPS was the slightly better with a mean of bias of 0.88,1.476 for PSM1 and PSM2 in the linearity and additivity treatment (Treatment A) and linearity and additivity outcome(Outcome 1) versions, whereas the remainder PSMs methods illustrate the smaller bias values in LR, which correspond to 1.022, 0.522, 0.943, and 1.064 for PSM3-6, respectively. There is a significant rise in RMSE values when more variables are added to the propensity score models or confounders are added in the LR and CBPS methods. For example, RMSEs of LR and CBPS were 0.230 and 0.237 in true confounder models; meanwhile, their values correspond to 0.270 and 0.273 in full models, respectively. The bias and RMSE of LR and CBPS techniques seem to be large across all PSMs in moderate nonlinearity and non-additivity treatment (Treatment F) compared to the simpler additivity and linear treatment setting (Treatment-A). Overall, the performance of LR is generally smaller biased in scenario of additivity and linearity (Treatment F) with a mean bias of 2.73,1.377,1.050, 1.576,1.509, and 1.873 across PSM1-6 compared to CBPS. In other words, we can state that it doesn't matter which variables are included in the PS model of LR method when using moderate nonlinearity and non-additivity treatment assignment in linearity outcome scenarios (Table 3.1). There is a significant rise in RMSE values when more variables or confounders are added to the propensity score models in the LR and CBPS methods. For example, RMSEs of LR and CBPS were 0.230 and 0.237 in true confounder models (PSM4); meanwhile, their values correspond to 0.270 and 0.273 in full models (PSM6), respectively. The bias and RMSE of LR and CBPS techniques seem to be large across all PSMs in moderate nonlinearity and non-additivity treatment (Treatment F) compared to the simpler additivity and linear treatment setting (Treatment-A).

The machine learning techniques reveal a large variety of results in evaluating bias and RMSE metrics. For example, the CART method has the lowest biases with a respective mean of 0.004 for PSM1 and PSM5 among the machine learning techniques even though BAG method corresponds to 4.666 and 4.206 for PSM1 and PSM5, respectively in linearity and additivity treatment. Also, CART presents the most downward bias in the all full propensity score model, which is included X1-X17 (i.e., full model) in all treatment scenarios. Moreover, there is a huge increasing trend in bias from Treatment-A to Treatment-F across in RF methods. When considering more complex treatment scenarios (i.e., Treatment-F), CART performs more reliable than the rest of the three machine learning methods across all PSMs models. However, exclusion or inclusion variables in models do not make a remarkable difference in assessing BAG model performance. So, BAG method performs poorest compared to other methods across all PSMs and all treatments and outcome scenarios. Moreover, the second poor performance among all methods is showed by GBM.

We can conclude from this result that the RF method estimates tended to make smaller bias when PSM models included all variables related to treatment, outcome, both them or confounders in simpler treatment (Treatment-A). . But CART managed to produce less bias if PSMs are formed by covariates X1-X10, which are associated with treatment or outcome assignments. Both RF and CART were pretty close RMSE values for all PSM versions according to the rest of the four methods (LR, CBPS, GBM, and CART), as seen in table 3.1. However, the bagging method performed poorly no matter which of PSMs were considered to assess bias and RMSE metrics in Treatment A-D-F versions.

Table 3.1: Bias and RMSE of LR, CBPS, GBM, RF, CART, BAG propensity score methods across all propensity score model scenarios in **Treatment A-D-F and Outcome-1 scenarios**

		PSM 1		**PSM 2**		**PSM 3**		**PSM 4**		**PSM 5**		**PSM 6**	
		Bias	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
TREATMENT A	**GLM**	0.943	0.250	1.665	0.246	1.022	0.231	0.522	0.230	0.943	0.250	1.064	0.270
	CBPS	0.880	0.257	1.476	0.252	1.505	0.229	1.026	0.237	1.505	0.229	1.168	0.273
	GBM	1.915	0.295	2.490	0.287	1.882	0.255	1.181	0.254	1.915	0.295	1.990	0.312
	RF	0.560	0.253	2.206	0.252	0.822	0.253	-0.030	0.245	0.793	0.238	0.451	0.254
	CART	0.004	0.262	0.243	0.260	1.745	0.248	1.080	0.239	0.004	0.262	1.323	0.269
	BAG	4.666	0.421	2.773	0.384	2.480	0.344	2.842	0.307	4.206	0.422	4.671	0.489
TREATMENT D	**GLM**	2.758	0.360	2.291	0.360	1.211	0.357	1.647	0.358	2.758	0.360	2.768	0.360
	CBPS	1.749	0.359	1.773	0.360	1.584	0.358	2.202	0.358	1.584	0.358	2.053	0.360
	GBM	1.976	0.364	2.716	0.362	1.055	0.356	1.220	0.333	1.976	0.364	2.060	0.366
	RF	1.373	0.360	1.204	0.360	1.362	0.359	1.460	0.358	0.942	0.360	1.878	0.361
	CART	2.064	0.171	1.636	0.169	0.885	0.164	1.805	0.158	2.064	0.171	1.570	0.177
	BAG	6.735	0.379	7.092	0.376	4.940	0.375	2.094	0.360	6.871	0.380	8.254	0.386
TREATMENT F	**GLM**	2.73	0.359	1.377	0.359	1.050	0.358	1.576	0.358	2.731	0.359	1.873	0.359
	CBPS	2.906	0.359	3.014	0.360	1.509	0.358	1.894	0.358	1.509	0.358	2.940	0.359
	GBM	3.842	0.365	3.230	0.364	1.769	0.356	3.534	0.334	3.842	0.365	2.969	0.368
	RF	2.338	0.359	2.902	0.360	1.298	0.358	1.780	0.358	2.981	0.359	2.692	0.359
	CART	1.926	0.176	1.931	0.172	1.134	0.160	1.415	0.154	1.926	0.176	1.626	0.180
	BAG	6.130	0.380	3.173	0.377	0.754	0.373	3.233	0.367	5.541	0.380	4.723	0.386

Performance of bias and RMSE in Treatment A-D-F and Outcome-2 (Table-3.2):

The difference between table 3.1 and table 3.2 is to use only different outcome versions: linearity and additivity outcome version (Outcome 1), and nonlinearity and non-additivity outcome version (Outcome-2). As illustrated in table 3.2, the LR consistently resulted in a smaller bias when PSM1-2-5-6 were considered, but the trend reversed for PSM3-4. There is no difference between PSM1 and PSM5 LR for biases that correspond to 2.537 percent in linearity and additive treatment version (Treatment-A) via nonlinearity and non-additivity outcome version (Table 3.2). It is observed that there is no effect by covariates X11 and x12, which are not associated with treatment nor outcome but related with X7 and X10, respectively. So, PSM1 and PSM5 displayed similar bias and RMSE measurement values in Treatment-A(Table 3.2). Machine learning techniques in linearity and additivity treatment, Treatment-A and non-linearity outcome,Outcome-2 are partly more complicated than the same techniques in Treatment-A and Outcome-1. Those results imply that CBPS produce better performance for bias in PSM1,2,5,6 while PSM3 and PSM4 were the slightest bias in Treatment-A and Outcome-2 for the parametric approach. Nevertheless, there is a remarkable growth in bias values from simple treatment assignment form (Treatment-A) to mild or moderate complex treatment assignment versions (Treatment -D and F). The CART yields less bias, especially in the presence of adding confounder and distractors variables on propensity score models. RF method in PSM3 and PSM4 appears to be lower bias in Treatment-A and Treatment-F, as seen in table-3.2.To conclude the findings of table 3.1 and table-3.2, random forest is the best for PSM3, and CART tends to produce less bias for PSM1-2-5 scenarios across all scenarios (i.e., treatment A via outcome-1 or outcome-2, treatment F via outcome-1 or outcome-2).In the sense of bias, permanently across all method and model versions, the BAG method demonstrated higher bias and RMSE for the simple, all treatment and outcome versions.

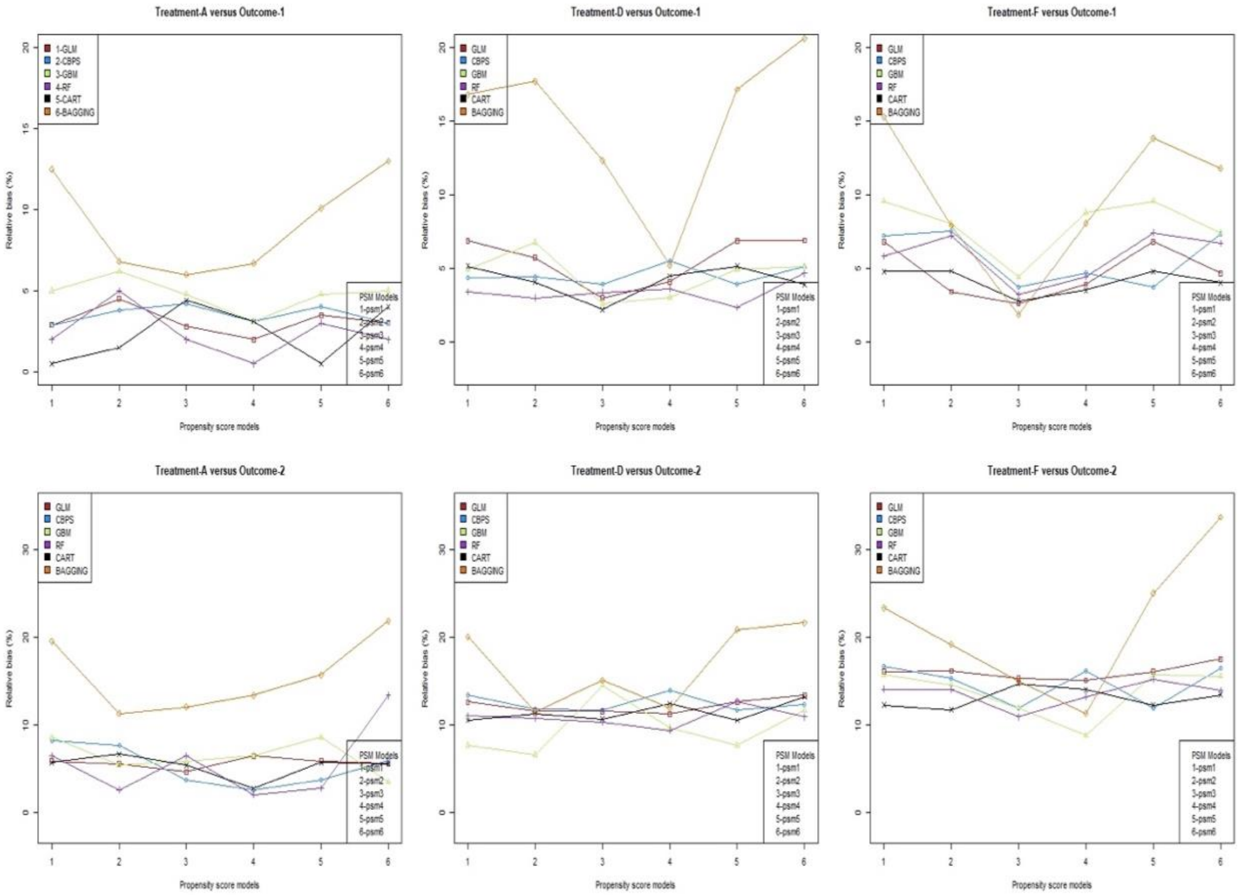
Table 3.2: Bias and RMSE of LR, CBPS, GBM, RF, CART, BAG propensity score methods across all propensity score model scenarios in **Treatment A-D-F and Outcome-2 scenarios**

			PSM1	**PSM 2**		**PSM 3**		**PSM 4**		**PSM 5**		**PSM 6**	
		Bias	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
TREATMENT A	**LR**	2.357	0.295	2.234	0.278	1.860	0.272	2.589	0.259	2.357	0.295	2.250	0.291
	CBP	2.274	0.289	3.087	0.282	1.493	0.271	1.043	0.254	1.493	0.271	2.384	0.299
	GBM	3.4370	0.341	2.181	0.321	2.337	0.283	2.599	0.291	3.437	0.341	1.395	0.343
	RF	2.589	0.278	2.241	0.278	1.397	0.268	0.829	0.279	2.453	0.275	1.703	0.275
	CAR	2.292	0.322	2.682	0.323	2.169	0.293	1.113	0.285	2.292	0.322	2.220	0.336
	BAG	7.824	0.481	4.530	0.433	4.832	0.374	5.353	0.345	6.297	0.474	8.764	0.532
TREATMENT D	**LR**	5.055	0.325	4.654	0.310	4.633	0.287	4.514	0.286	5.055	0.325	5.356	0.320
	CBP	5.371	0.317	4.737	0.325	4.697	0.289	5.562	0.287	4.967	0.289	4.925	0.333
	GBM	3.069	0.380	2.653	0.346	5.818	0.327	3.875	0.327	3.069	0.380	4.672	0.391
	RF	4.433	0.308	4.315	0.313	4.132	0.301	3.766	0.285	5.090	0.309	4.373	0.296
	CAR	4.227	0.339	4.520	0.325	4.286	0.305	4.989	0.299	4.227	0.339	5.288	0.346
	BAG	8.047	0.526	4.626	0.474	6.053	0.406	4.776	0.366	5.370	0.537	8.679	0.557
TREATMENT F	**LR**	6.450	0.289	6.479	0.282	6.158	0.274	6.052	0.284	6.450	0.289	7.014	0.268
	CBP	6.671	0.291	6.154	0.287	4.782	0.277	6.477	0.269	4.784	0.277	6.597	0.289
	GBM	6.295	0.390	5.857	0.374	4.792	0.312	3.523	0.301	6.295	0.390	6.244	0.404
	RF	5.635	0.303	5.633	0.291	4.380	0.288	5.304	0.292	6.111	0.306	5.570	0.285
	CAR	4.911	0.332	4.702	0.315	5.893	0.321	5.606	0.307	4.911	0.332	5.385	0.347
	BAG	9.358	0.520	7.692	0.481	6.004	0.423	4.531	0.376	10.023	0.521	13.510	0.560

Performance of absolute relative bias(%):

The relative bias is investigated based on the pair matching on propensity score across all scenarios. Figure 3.2 illustrates the relative bias percent for three treatment versions (A, D, F) versus two outcomes (i.e., linearity and additivity outcome (Outcome-1) and non-linearity and non-additivity outcome (Outcome-2)). Each propensity score method resulted in additivity and linearity treatment (Treatment-A), mild non-additivity and non-linearity (Treatment-D), and moderate non-linearity treatment versions (Treatment E) show various fluctuations for propensity score models (PSMs) output in Figure 3.2. According to panels from Treatment-A and Treatment-D, it is evident that the bagging method had tended to largest the relative bias. Indeed, the second-largest relative bias is represented by GBM across all propensity score models. In contrast, we can generally conclude that RF and CART partially presented better performance than others, and CBPS may not be an alternative for LR across all propensity score models when PSMs are constructed based on then additivity model.

Figure 3.2: Absolute relative bias (%) of propensity score matching on different parametric and machine learning techniques using various PSMs by Treatment A-D-F versus Outcome 1-2 scenarios



Abbreviations: GLM: Generalized Linear Model , CBPS: Covariate Balance Propensity Score, GBM: Generalized Boosted Model, RF: Random Forest, CART: Classification and Regression trees. PMS1: $X_1 - X_{10}$ covariates are included ,PMS2: $X_1 - X_7$ covariates are included ,PMS3: $X_1 - X_4$ and $X_8 - X_{10}$covariates are included , PMS4: $X_1 - X_4$ covariates are included, PMS5: $X_1 - X_{12}$ covariates are included, PMS6: $X_1 - X_{17}$ covariates are included. Treatment A: linearity and additivity, Treatment D: Mild non-additivity and non-linearity, Treatment F: Moderate non-additivity and non-linearity, Outcome-1: Additive and linear outcome model, Outcome-2: non-linearity outcome model

Performance of empirical standard error (ESE) and theoretical standard error (SE) :

We preferred to use standard errors, which investigate treatment effects on the outcome. Standard errors were computed based on the model-based for every simulated dataset (i.e., the 1000 standard error for 1000 simulated datasets). Then, we take averages of all estimated standard errors according to the simulated dataset's sample size. Model-based standard error represented as "SE" in tables.SE results for propensity estimation methods versus propensity score models are informed in table 3.7-3.10 (see Appendix) for all treatment and outcome scenarios. The second metric is the empirical standard error (ESE) that considered the sample standard deviation of average estimates of treatment effects. LR and CBPS methods obtained similar SEs for all estimation methods and PSMs under all treatment versions of nonlinearity outcome (Outcome-1).However, slightly increasing the SEs from outcome-1 to outcome-2 versions i.e., compared the SEs for each treatment version between table-3.7 and table-3.8 across all treatment assignments. Hence, outcome complexity may lead to increasing variation of estimated treatment effects.

In contrast, the CART performs the low mean SEs in all PMSs for linearity outcome, ranging from SEs: 0.027-0.030 for PSM1 across all treatment scenarios in linearity outcome (Outcome 1). Also, the ESE yields a remarkable difference between BAG methods and the remaining methods for all PMSs. For example, the ESE from the BAG method was almost two times larger than the other methods. The ESE performance for all PMSs is parallel among propensity score estimation methods, and so, there is a constant pattern throughout all PMSs. In other words, there are growing patterns in the ESEs from linearity outcome version (Outcome-1) to non-linearity outcome (Outcome-2) against the same treatment assignment.

3.6.2 The Results of Simulation in Treatment Selection

As seen in table 3.1-3.2 and 3.5-3.10 (in Appendix), to perform the different propensity score estimation methods utilizing several PSMs across different generated treatment and outcome assignments, we obtained the results from those tables. We then saved the datasets, which are based on each scenario. Finally, we aim to examine treatment selection markers based on those observational datasets in tables 3.3-3.4 and tables 3.11-3.12 (Appendix) in this section. Table 3.3 presents marker evaluation using B_{neg}, B_{pos}metrics for "linearity and additivity treatment" (i.e., treatment-A assignment) and "linearity and additivity outcome"(outcome-1 assignment). Θ, B_{neg}, B_{pos}parameters are beneficial for analyzing the marker's effects. To make comparisons, we first look at PSM1 using six alternatives propensity score methods. CART represents the smallest Θ, B_{neg}, B_{pos}metric in Table 3.3.CART has a 1.9 percent reduction in sample impact. In comparison, GBM has a 2.6 percent reduction in population impact as it has the bigger theta measure among of methods.

This is attributable to the fact that 50 percent (P_{neg} metric) of participants avoided specific treatment, resulting in a 5.1(B_{neg} metric) percent decrease in the occurrence rate. The treated group reduces the event rate by 13% for average among marker-positive patients according to consider the RF method in PSM1 (Table-3.3). This method corresponds to 0.002 as variance in treatment effect and 0.044 for width from the marginal exposure impact to exposure impact curve.

Table 3.3: Estimates of measures of marker performance based on the resulted LR, CBPS, GBM, RF, CART, BAG propensity score methods across **PSM 1-4-6** scenarios in **Treatment-A and Outcome 1 versions**

		Θ		$\mathbf{P_{neg}}$	$\mathbf{B_{neg}}$		$\mathbf{B_{pos}}$		$\mathbf{V_{\Delta}}$	**TG**
		Model	Emp.		Model	Emp.	Model	Emp.		
PSM1	*LR*	0.023	0.023	0.501	0.047	0.047	0.132	0.132	0.003	0.047
	CBPS	0.023	0.023	0.502	0.046	0.046	0.130	0.130	0.003	0.047
	GBM	0.026	0.026	0.498	0.054	0.053	0.157	0.157	0.003	0.056
	RF	0.025	0.025	0.498	0.051	0.051	0.128	0.128	0.002	0.044
	CART	0.019	*0.019*	0.500	0.038	0.038	0.123	0.123	0.003	0.042
	BAG	0.023	0.023	0.488	0.052	0.052	0.226	0.226	0.013	0.091
PSM4	*LR*	0.028	0.028	0.502	0.057	0.057	0.128	0.128	0.002	0.041
	CBPS	0.026	0.026	0.501	0.053	0.053	0.130	0.130	0.002	0.044
	GBM	0.031	0.031	0.494	0.064	0.064	0.149	0.149	0.003	0.046
	RF	0.20	0.021	0.498	0.041	0.042	0.119	0.119	0.003	0.044
	CART	0.028	0.028	0.499	0.057	0.057	0.130	0.130	0.001	0.036
	BAG	0.041	0.042	0.493	0.085	0.086	0.202	0.202	0.006	0.062
PSM6	*LR*	0.023	0.023	0.497	0.046	0.046	0.128	0.129	0.003	0.046
	CBPS	0.023	0.023	0.498	0.047	0.047	0.137	0.138	0.004	0.050
	GBM	0.027	0.027	0.492	0.056	0.056	0.168	0.168	0.005	0.060
	RF	0.021	0.021	0.500	0.043	0.042	0.120	0.120	0.003	0.044
	CART	0.021	0.021	0.503	0.042	0.042	0.120	0.128	0.002	0.042
	BAG	0.012	0.012	0.487	0.029	0.029	0.259	0.259	0.001	0.116

Abbreviations: LR: Logistic Regression , CBPS: Covariate Balance Propensity Score, GBM: Generalized Boosted Model, RF: Random Forest, CART: Classification and Regression trees. PMS1: $X_1 - X_{10}$ covariates are included,PMS4: $X_1 - X_4$ covariates are included, PMS6: $X_1 - X_{17}$ covariates are included. Treatment A: linearity and additivity, Outcome-1: Additive and linear outcome model

When looking performance of methods from PSM4 to PSM1 and PSM6, respectively,   we notice a growing trend for each method corresponding to each PSM for the mean benefit of treatment among people who have found evidence for a marker (B_{pos}) .In contrast, reduced trends for $B_{neg}$from PSM4 to PSM1, PSM6, respectively. For example, it was a consequence of 15 percent,16 percent,17 percent mean utility of treatment assignment in marker-positivities for GBM methods of PSM4, PSM1, and PSM4 models. BAG method in PSM6 produced bigger total gain(TG) than the others, with respective ranges of 0.062-0.116 across the three propensity score models. Indeed, we notice that the RF method performed the lowest value in examining the reduction in sample occurrence rate under marker-based treatment.

Table 3.4: Estimates of measures of marker performance based on the resulted LR, CBPS, GBM, RF, CART, BAG propensity score methods across **PSM 1-4-6** scenarios in **Treatment-F and Outcome 2 versions**

		Θ		**P_{neg}**	**B_{neg}**		**B_{pos}**		**V_{Δ}**	**TG**
		Mod.	**Emp.**		**Mod.**	**Emp.**	**Mod.**	**Emp.**		
PSM1	*LR*	0.021	0.021	0.495	0.044	0.044	0.123	0.123	0.003	0.044
	CBPS	0.022	0.022	0.495	0.044	0.044	0.113	0.114	0.002	0.040
	GBM	0.011	0.011	0.508	0.022	0.022	0.134	0.134	0.005	0.059
	RF	0.029	0.029	0.498	0.059	0.059	0.131	0.131	0.003	0.042
	CART	0.019	0.019	0.500	0.038	0.038	0.123	0.123	0.003	0.042
	BAG	0.018	0.018	0.489	0.039	0.039	0.214	0.215	0.012	0.090
PSM4	*LR*	0.020	0.020	0.498	0.041	0.041	0.119	0.119	0.003	0.044
	CBPS	0.023	0.023	0.500	0.045	0.045	0.122	0.122	0.003	0.044
	GBM	0.027	0.027	0.494	0.057	0.057	0.145	0.146	0.003	0.048
	RF	0.016	0.016	0.499	0.033	0.033	0.114	0.114	0.003	0.045
	CART	0.020	0.020	0.500	0.040	0.040	0.139	0.139	0.003	0.049
	BAG	0.027	0.027	0.496	0.056	0.056	0.181	0.181	0.006	0.066
PSM6	*LR*	0.022	0.022	0.493	0.046	0.046	0.127	0.127	0.003	0.046
	CBPS	0.017	0.017	0.494	0.036	0.036	0.122	0.122	0.003	0.047
	GBM	0.002	0.002	0.500	0.007	0.007	0.139	0.139	0.007	0.070
	RF	0.020	0.020	0.498	0.041	0.041	0.131	0.131	0.003	0.049
	CART	0.011	0.011	0.501	0.024	0.024	0.127	0.127	0.003	0.051
	BAG	0.006	0.006	0.493	0.007	0.007	0.222	0.222	0.021	0.115

Abbreviations: LR: Logistic Regression , CBPS: Covariate Balance Propensity Score, GBM: Generalized Boosted Model, RF: Random Forest, CART: Classification and Regression trees. PMS1: $X_1 - X_{10}$ covariates are included , PMS4: $X_1 - X_4$ covariates are included, PMS6: $X_1 - X_{17}$ covariates are included. Treatment F: Moderate non-additivity and non-linearity, Outcome-2: non-linearity outcome model.

3.7 Summary & Discussion

In literature, researchers in healthcare and public health studies frequently use propensity score estimation methods such as logistic regression, random forest, neural network, CART, pruned CART, etc. Unfortunately, there are significant challenges for the implementation procedure of these methods and comparing the performance of these methods. Therefore, we examine two main fields in this paper using Monte Carlo simulations: assessing the propensity score methods to eliminate bias and satisfy covariate balance between treated and untreated groups and examined the biomarker performance for treatment selection.

The first purpose of this paper is to evaluate six methods to estimate propensity score: logistic regression(LR) and covariate balance propensity (CBPS) as parametric approaches, and generalized boosted method (GBM), random forest (RF), classification and regression trees(CART) and bagging(BAG) as machine learning techniques. Secondly, the term " variable selection" refers to the variables included or excluded in the model, not the procedure by which variables were chosen and determine variables' effects on the model. Also, we performed these methods fitting different propensity score models across several treatment and outcome assignment versions throughout this study. We use six treatment scenarios (called Treatment A-F) and two outcome scenarios (called Outcome1-2) below tables. All results examined rely on 1:1 matching on propensity score data. We modified the simulation design of Setoguchi et al. (2008) along with this paper. Finally, we give the relevant recommendations for researchers: Firstly, Setodji et al. (2017) recommend that using logistic regression has been observed to have the drawback to employed estimation of the weighting on the propensity score. That's why CBPS always might not be an alternative technique for logistic regression. According to this paper, Setodji paper' recommendation might not be valid under different conditions and scenarios from simulations. We show that logistic regression might consistently be practical using treatment and outcome scenarios for employing different propensity score models (PSMs).Also, We cannot generalize about LR's good performance when considering complex propensity score models and complex treatment and outcome scenarios. However, even though the GBM method was not the best method among all the machine learning techniques, GBM might be an excellent alternative to estimate treatment effects instead of preferring bagging methods. Moreover, Throughout this paper, bagging performed the worst method, even utilizing different PSMs, considering complexity or simplest treatment and outcomes assignments. RF and CART can be comparable techniques in even parametric or non-parametric techniques. But we realize that choosing a variable selection on the models has been a critical thing for these two methods. In other words, results illustrated which including or excluded variables have been influenced the performance of RF and CART methods. Besides, there is a remarkable difference between different treatment assignments and outcome versions, i.e., the complexity of treatment or outcomes is highly crucial on propensity score estimation performance on RF and CART. Finally, we provided a statistical methodology for analyzing treatment selection markers based on using datasets of causal inference. This paper makes the first generalization about treatment selection implementation in observational studies under considering several versions. We used the same biomarker with different produced datasets from causal inference. We exactly make sure that there is a significant effect from which methods or models are preferred in causal inference application.

3.8 Appendix

Table 3.5: Bias and RMSE of LR, CBPS, GBM, RF, CART, BAG propensity score methods across all propensity score model scenarios in **Treatment B-C-E** and **Outcome-1** scenarios

		PSM 1		PSM 2		PSM 3		PSM 4		PSM 5		PSM 6	
		Bias	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
TREATMENT B	**LR**	2.844	0.359	0.174	0.359	2.969	0.357	2.793	0.357	2.844	0.359	1.690	0.359
	CBPS	1.260	0.359	1.968	0.359	1.667	0.357	2.271	0.358	1.667	0.357	1.335	0.359
	GBM	3.320	0.365	3.504	0.363	1.544	0.355	2.445	0.333	3.302	0.365	3.715	0.368
	RF	2.548	0.358	2.317	0.359	2.066	0.357	1.335	0.356	2.296	0.358	3.261	0.359
	CART	3.831	0.173	3.411	0.169	2.412	0.160	1.480	0.153	3.831	0.173	4.622	0.178
	BAG	7.137	0.379	7.281	0.376	3.936	0.375	5.256	0.367	7.891	0.379	7.877	0.385
TREATMENT C	**LR**	0.376	0.360	-0.639	0.361	-1.118	0.358	0.153	0.358	0.376	0.360	-0.443	0.360
	CBPS	-0.418	0.360	1.113	0.361	0.069	0.358	0.841	0.359	0.069	0.358	-0.532	0.361
	GBM	0.291	0.364	0.756	0.362	0.641	0.356	1.408	0.331	0.291	0.364	-0.368	0.365
	RF	-1.230	0.360	-0.658	0.361	0.182	0.359	1.120	0.359	-0.418	0.361	-0.538	0.361
	CART	-1.741	0.170	-1.151	0.167	0.758	0.162	0.350	0.156	-1.741	0.170	-1.234	0.175
	BAG	3.018	0.380	3.933	0.376	3.729	0.376	2.796	0.368	1.615	0.379	5.950	0.386
Treatment E	**LR**	2.409	0.359	2.477	0.360	2.098	0.358	1.826	0.359	2.409	0.359	0.785	0.359
	CBPS	1.741	0.359	1.698	0.360	3.812	0.358	2.414	0.358	3.812	0.358	2.201	0.360
	GBM	3.773	0.363	2.354	0.360	2.583	0.355	1.583	0.331	3.773	0.363	1.959	0.365
	RF	1.569	0.360	0.686	0.360	1.440	0.359	2.579	0.359	1.334	0.360	2.619	0.360
	CART	2.657	0.172	2.629	0.168	1.929	0.162	2.711	0.156	2.657	0.172	3.070	0.175
	BAG	4.953	0.380	2.637	0.376	4.561	0.374	4.462	0.367	6.946	0.380	5.547	0.387

Figure 3.6: Bias and RMSE of LR, CBPS, GBM, RF, CART, BAG propensity score methods across all propensity score model scenarios **in Treatment B-C-E and Outcome-2 scenarios**

	PSM1		**PMS2**		**PMS3**		**PSM4**		**PSM5**		**PSM6**	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
GLM	3.423	0.289	2.968	0.288	2.449	0.272	1.776	0.271	3.423	0.289	3.371	0.293
CBPS	2.862	0.285	3.698	0.281	1.529	0.272	1.373	0.265	1.529	0.272	3.279	0.298
GBM	2.806	0.381	1.901	0.374	0.316	0.299	-0.892	0.302	2.806	0.381	3.917	0.422
RF	1.589	0.300	3.089	0.303	1.379	0.278	2.763	0.274	1.687	0.294	3.578	0.296
CART	1.831	0.327	2.040	0.317	1.520	0.303	1.533	0.287	1.831	0.327	1.900	0.330
BAG	5.587	0.512	5.240	0.480	3.278	0.422	1.922	0.365	1.704	0.517	6.773	0.537
GLM	1.825	0.313	1.221	0.315	2.099	0.292	1.814	0.276	1.825	0.313	0.442	0.325
CBPS	0.829	0.306	0.471	0.319	1.893	0.285	2.233	0.275	1.893	0.285	0.607	0.306
GBM	0.498	0.358	1.653	0.327	0.750	0.296	0.281	0.304	0.498	0.358	2.631	0.365
RF	0.688	0.304	0.897	0.301	1.827	0.282	0.710	0.283	-0.021	0.291	0.253	0.300
CART	-0.052	0.309	-0.017	0.295	2.244	0.295	1.893	0.295	-0.052	0.309	-0.502	0.319
BAG	3.893	0.465	5.341	0.443	3.553	0.394	4.227	0.351	5.654	0.498	5.067	0.520
GLM	2.054	0.304	2.540	0.305	2.183	0.292	1.417	0.278	2.054	0.304	3.708	0.317
CBPS	2.833	0.317	2.990	0.309	2.578	0.289	1.353	0.279	2.578	0.287	2.331	0.303
GBM	3.119	0.350	3.171	0.341	2.386	0.304	0.961	0.305	3.119	0.350	2.649	0.357
RF	1.506	0.286	2.179	0.291	0.976	0.292	3.537	0.289	2.839	0.295	2.516	0.299
CART	3.092	0.329	2.952	0.324	3.537	0.293	1.274	0.287	3.092	0.329	3.505	0.341
BAG	4.553	0.506	5.532	0.455	1.274	0.394	2.011	0.359	7.107	0.482	6.525	0.539

Table-3.7: Performance of empirical standard error (ESE) and theoretical standard error (SE) of LR, CBPS, GBM, RF, CART, BAG propensity score methods across all propensity score model scenarios **in Treatment A-D-F and Outcome-1 scenarios**1

		PSM1		PMS2		PMS3		PSM4		PSM5		PSM6	
		E SE	SE	ESE	SE	E SE	SE	E SE	SE	E SE	SE	E SE	SE
TREATMENT A	GLM	0.250	0.129	0.246	0.129	0.231	0.128	0.230	0.128	0.250	0.129	0.270	**0.129**
	CBPS	0.257	0.129	0.252	0.129	0.229	0.127	0.237	0.128	0.299	0.127	0.273	**0.129**
	GBM	0.294	0.131	0.286	0.129	0.255	0.126	0.254	0.109	0.294	0.131	0.312	**0.132**
	RF	0.253	0.129	0.251	0.129	0.253	0.128	0.245	0.128	0.239	0.129	0.258	**0.129**
	CART	0.263	0.028	0.260	0.026	0.248	0.026	0.239	0.024	0.263	0.028	0.269	**0.030**
	BAG	0.418	0.143	0.383	0.140	0.344	0.141	0.306	0.136	0.420	0.143	0.487	**0.148**
TREATMENT D	GLM	0.274	0.129	0.272	0.130	0.248	0.128	0.255	0.128	0.274	0.129	0.264	**0.129**
	CBPS	0.272	0.129	0.270	0.130	0.256	0.128	0.247	0.128	0.256	0.128	0.273	**0.129**
	GBM	0.319	0.132	0.314	0.131	0.276	0.127	0.278	0.111	0.319	0.132	0.344	**0.134**
	RF	0.254	0.130	0.260	0.130	0.253	0.128	0.247	0.128	0.252	0.129	0.254	**0.130**
	CART	0.287	0.030	0.280	0.028	0.271	0.027	0.261	0.024	0.287	0.029	0.292	**0.031**
	BAG	0.440	0.144	0.406	0.141	0.366	0.141	0.345	0.135	0.446	0.144	0.487	**0.149**
TREATMENT F	GLM	0.258	0.129	0.252	0.129	0.244	0.128	0.253	0.128	0.258	0.129	0.271	**0.129**
	CBPS	0.273	0.129	0.259	0.129	0.241	0.128	0.253	0.129	0.241	0.128	0.269	**0.129**
	GBM	0.355	0.133	0.350	0.132	0.287	0.127	0.275	0.111	0.355	0.133	0.368	**0.135**
	RF	0.279	0.129	0.287	0.129	0.257	0.128	0.265	0.128	0.279	0.129	0.264	**0.129**
	CART	0.289	0.031	0.282	0.029	0.274	0.025	0.276	0.023	0.289	0.031	0.297	**0.033**

Table -3.8: Performance of empirical standard error (ESE) and theoretical standard error (SE) of LR, CBPS, GBM, RF, CART, BAG propensity score methods across all propensity score model scenarios in **Treatment A-D-F** and **Outcome-2** scenarios

		PSM1		**PMS2**		**PMS3**		**PSM4**		**PSM5**		**PSM6**	
		E SE	**SE**	**ESE**	**SE**	**E SE**	**SE**	**E SE**	**SE**	**E SE**	**SE**	**E SE**	**SE**
TREATMENT A	**GLM**	0.29	0.142	0.278	0.143	0.272	0.141	0.258	0.141	0.295	0.142	0.291	0.142
	CBPS	0.28	0.142	0.281	0.143	0.271	0.141	0.254	0.141	0.271	0.141	0.298	0.142
	GBM	0.34	0.145	0.320	0.143	0.282	0.139	0.290	0.120	0.340	0.145	0.343	0.146
	RF	0.27	0.142	0.278	0.142	0.268	0.141	0.279	0.141	0.274	0.142	0.274	0.142
	CART	0.32	0.031	0.322	0.029	0.292	0.029	0.285	0.026	0.322	0.031	0.335	0.033
	BAG	0.47	0.159	0.431	0.155	0.371	0.156	0.341	0.150	0.470	0.159	0.525	0.165
TREATMENT D	**GLM**	0.32	0.144	0.307	0.144	0.283	0.142	0.282	0.142	0.322	0.144	0.316	0.144
	CBPS	0.31	0.144	0.322	0.144	0.286	0.142	0.282	0.123	0.286	0.142	0.330	0.144
	GBM	0.37	0.147	0.345	0.145	0.323	0.141	0.325	0.142	0.379	0.147	0.388	0.148
	RF	0.30	0.143	0.311	0.144	0.299	0.142	0.283	0.027	0.305	0.144	0.293	0.144
	CART	0.33	0.032	0.322	0.031	0.303	0.029	0.295	0.149	0.336	0.033	0.342	0.035
	BAG	0.52	0.160	04.72	0.157	0.402	0.155	0.363	0.149	0.531	0.160	0.511	0.166
TREATMENT F	**GLM**	0.28	0.144	0.275	0.144	0.268	0.142	0.277	0.143	0.282	0.144	0.259	0.144
	CBPS	0.28	0.144	0.280	0.144	0.273	0.142	0.261	0.143	0.273	0.142	0.282	0.144
	GBM	0.38	0.149	0.370	0.148	0.308	0.141	0.299	0.124	0.385	0.149	0.400	0.151
	RF	0.29	0.144	0.286	0.144	0.285	0.142	0.287	0.142	0.301	0.144	0.280	0.144
	CART	0.32	0.034	0.312	0.033	0.316	0.029	0.302	0.027	0.329	0.034	0.343	0.036
	BAG	0.51	0.161	0.475	0.158	0.420	0.155	0.374	0.149	0.512	0.161	0.544	0.166

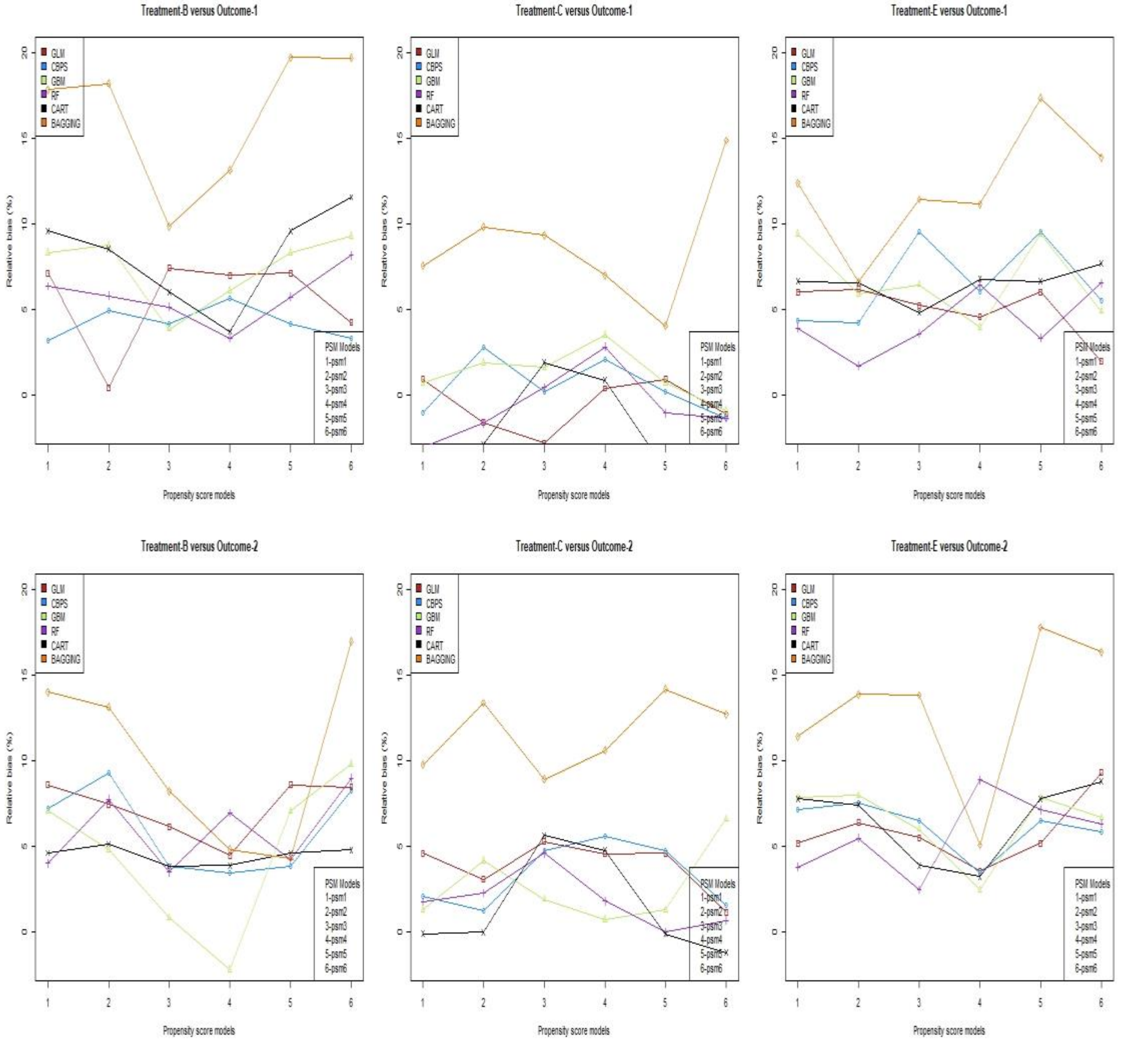
Table -3.9: Performance of empirical standard error (ESE) and theoretical standard error (SE) of LR, CBPS, GBM, RF, CART, BAG propensity score methods across all propensity score model scenarios in **Treatment B-C-E and Outcome-1 scenarios**

		PSM1		PMS2		PMS3		PSM4		PSM5		PSM6	
		E SE	**SE**	**ESE**	**SE**	**E SE**	**SE**	**E SE**	**SE**	**E SE**	**SE**	**E SE**	**SE**
TREATMENT B	**GLM**	0.270	0.128	0.246	0.129	0.239	0.127	0.234	0.128	0.270	0.128	0.256	0.128
	CBPS	0.258	0.128	0.245	0.129	0.235	0.127	0.240	0.128	0.235	0.127	0.265	0.128
	GBM	0.368	0.133	0.331	0.132	0.271	0.126	0.277	0.110	0.368	0.133	0.375	0.135
	RF	0.267	0.128	0.268	0.129	0.242	0.127	0.242	0.127	0.267	0.128	0.265	0.129
	CAR	0.279	0.029	0.272	0.028	0.282	0.025	0.262	0.023	0.279	0.029	0.291	0.031
	BAG	0.455	0.144	0.427	0.141	0.382	0.140	0.338	0.135	0.450	0.143	0.489	0.148
TREATMENT C	**GLM**	0.297	0.130	0.299	0.130	0.251	0.128	0.267	0.128	0.297	0.1301	0.296	0.130
	CBPS	0.299	0.130	0.302	0.130	0.256	0.128	0.261	0.128	0.256	0.128	0.301	0.130
	GBM	0.335	0.132	0.316	0.131	0.278	0.126	0.275	0.109	0.335	0.132	0.356	0.133
	RF	0.280	0.130	0.282	0.130	0.259	0.129	0.263	0.129	0.87	0.130	0.277	0.130
	CAR	0.305	0.029	0.301	0.027	0.285	0.026	0.273	0.024	0.305	0.029	0.309	0.030
	BAG	0.443	0.144	0.417	0.141	0.380	0.141	0.344	0.136	0.456	0.144	0.518	0.149
	GLM	0.267	0.129	0.269	0.130	0.236	0.128	0.240	0.128	0.267	0.129	0.261	0.129
TREATMENT E	**CBPS**	0.261	0.129	0.268	0.129	0.246	0.128	0.243	0.128	0.246	0.128	0.266	0.129
	GBM	0.304	0.131	0.288	0.130	0.257	0.126	0.250	0.109	0.304	0.131	0.324	0.133
	RF	0.248	0.129	0.250	0.129	0.245	0.129	0.255	0.129	0.249	0.129	0.257	0.130
	CAR	0.274	0.029	0.273	0.028	0.260	0.026	0.255	0.024	0.274	0.029	0.276	0.031
	BAG	0.422	0.144	0.417	0.141	0.359	0.140	0.308	0.135	0.423	0.145	0.493	0.150

Table -3.10: Performance of empirical standard error (ESE) and theoretical standard error (SE) of LR, CBPS, GBM, RF, CART, BAG propensity score methods across all propensity score model scenarios in **Treatment B-C-E and Outcome-2 scenarios**

		PSM1		PMS2		PMS3		PSM4		PSM5		PSM6	
		E SE	**SE**	**ESE**	**SE**	**E SE**	**SE**	**E SE**	**SE**	**E SE**	**SE**	**E SE**	**SE**
TREATMENT B	**GLM**	0.288	0.142	0.287	0.143	0.271	0.141	0.271	0.141	0.288	0.142	0.291	0.142
	CBPS	0.284	0.142	0.279	0.143	0.272	0.141	0.265	0.141	0.272	0.141	0.297	0.143
	GBM	0.380	0.148	0.374	0.146	0.299	0.139	0.302	0.122	0.380	0.148	0.420	0.150
	RF	0.277	0.142	0.278	0.143	0.268	0.141	0.279	0.141	0.274	0.142	0.274	0.142
	CART	0.322	0.032	0.322	0.031	0.292	0.028	0.285	0.026	0.322	0.032	0.335	0.034
	BAG	0.475	0.159	0.431	0.156	0.371	0.155	0.341	0.149	0.470	0.158	0.525	0.164
TREATMENT C	**GLM**	0.312	0.143	0.315	0.144	0.291	0.141	0.276	0.417	0.312	0.143	0.325	0.143
	CBPS	0.306	0.143	0.319	0.144	0.284	0.141	0.275	0.421	0.284	0.141	0.306	0.143
	GBM	0.358	0.146	0.327	0.144	0.297	0.139	0.304	0.121	0.358	0.146	0.364	0.148
	RF	0.304	0.143	0.301	0.143	0.281	0.142	0.283	0.142	0.291	0.143	0.300	0.144
	CART	0.310	0.032	0.295	0.030	0.294	0.029	0.294	0.026	0.327	0.032	0.319	0.034
	BAG	0.464	0.160	0.440	0.156	0.393	0.156	0.349	0.150	0.517	0.160	0.518	0.166
TREATMENT E	**GLM**	0.303	0.143	0.304	0.144	0.292	0.142	0.278	0.142	0.303	0.143	0.315	0.143
	CBPS	0.316	0.143	0.308	0.144	0.288	0.142	0.279	0.142	0.288	0.142	0.303	0.143
	GBM	0.349	0.146	0.340	0.144	0.304	0.140	0.305	0.121	0.349	0.146	0.356	0.148
	RF	0.286	0.143	0.291	0.144	0.292	0.143	0.287	0.142	0.294	0.143	0.298	0.144
	CART	0.327	0.032	0.323	0.031	0.293	0.029	0.287	0.026	0.327	0.032	0.340	0.035

Figure 3.3: Absolute relative bias (%) of propensity score matching on different parametric and machine learning techniques using various PSMs by Treatment B-C-E versus Outcome 1-2 scenarios



Abbreviations: GLM: Generalized Linear Model , CBPS: Covariate Balance Propensity Score, GBM: Generalized Boosted Model, RF: Random Forest, CART: Classification and Regression trees. PMS1: $X_1 - X_{10}$ covariates are included ,PMS2: $X_1 - X_7$ covariates are included ,PMS3: $X_1 - X_4$ and $X_8 - X_{10}$covariates are included , PMS4: $X_1 - X_4$ covariates are included, PMS5: $X_1 - X_{12}$ covariates are included, PMS6: $X_1 - X_{17}$ covariates are included. Treatment B: linearity and additivity, Treatment C: Mild non-additivity and non-linearity, Treatment D: Moderate non-additivity and non-linearity, Outcome-1: Additive and linear outcome model, Outcome-2: non-linearity outcome model

Table 3.11: Estimates of measures of marker performance based on the resulted LR,CBPS,GBM,RF,CART,BAG propensity score methods across PSM 2-3-5 scenario in Treatment-A against Outcome-1 assignments

		Θ		P_{neg}	B_{neg}		B_{pos}		V_{Δ}	TG
		Mod.	Emp.		Mod.	Emp.	Mod.	Emp.		
PSM2	*LR*	0.024	0.024	0.499	0.048	-0.048	0.130	0.130	0.003	0.046
	CBPS	0.023	0.023	0.501	0.045	-0.045	0.128	0.130	0.003	0.047
	GBM	0.033	0.033	0.496	0.066	-0.067	0.153	0.157	0.003	0.048
	RF	0.027	0.027	0.506	0.053	-0.053	0.126	0.126	0.003	0.042
	CART	0.018	0.018	0.498	0.036	-0.036	0.121	0.121	0.002	0.042
	BAG	0.036	0.036	0.487	0.075	-0.074	0.204	0.204	0.007	0.068
PSM3	*LR*	0.023	0.023	0.492	0.048	-0.048	0.129	0.129	0.003	0.045
	CBPS	0.020	0.024	0.494	0.048	-0.048	0.128	0.128	0.003	0.045
	GBM	0.029	0.029	0.502	0.057	-0.057	0.148	0.148	0.004	0.050
	RF	0.021	0.021	0.498	0.043	-0.043	0.118	0.118	0.003	0.042
	CART	0.027	0.027	0.502	0.055	-0.055	0.131	0.131	0.002	0.038
	BAG	0.031	0.031	0.486	0.070	-0.070	0.203	0.203	0.008	0.069
PSM5	*LR*	0.022	0.022	0.498	0.044	-0.046	0.128	0.129	0.003	0.047
	CBPS	0.027	0.027	0.495	0.054	-0.047	0.132	0.132	0.003	0.044
	GBM	0.032	0.032	0.499	0.065	-0.056	0.164	0.164	0.004	0.054
	RF	0.023	0.023	0.498	0.046	-0.042	0.122	0.122	0.003	0.043
	CART	0.018	0.018	0.498	0.035	-0.042	0.134	0.134	0.003	0.049
	BAG	0.016	0.017	0.500	0.037	-0.029	0.224	0.224	0.013	0.095

Abbreviations: LR: Logistic Regression , CBPS: Covariate Balance Propensity Score, GBM: Generalized Boosted Model, RF: Random Forest, CART: Classification and Regression trees. PMS1: $X_1 - X_{10}$ covariates are included , PMS4: $X_1 - X_4$ covariates are included, PMS6: $X_1 - X_{17}$ covariates are included. Treatment F: Moderate non-additivity and non-linearity, Outcome-2: non-linearity outcome model

Table 3.12: Estimates of measures of marker performance based on the resulted LR,CBPS,GBM,RF,CART,BAG propensity score methods across PSM 2-3-5 scenario in Treatment-F against Outcome-2 assignments

		Θ		$\mathbf{P_{neg}}$	$\mathbf{B_{neg}}$		$\mathbf{B_{pos}}$		$\mathbf{V_{\Delta}}$	**TG**
		Mod.	Emp.		Mod.	Emp.	Mod.	Emp.		
PSM2	*LR*	0.024	0.024	0.501	0.048	0.048	0.117	0.117	0.002	0.040
	CBPS	0.026	0.026	0.503	0.052	0.052	0.112	0.111	0.002	0.036
	GBM	0.015	0.015	0.448	0.031	0.031	0.147	0.147	0.006	0.062
	RF	0.021	0.021	0.501	0.042	0.043	0.131	0.131	0.003	0.049
	CART	0.020	0.019	0.499	0.041	0.041	0.120	0.120	0.002	0.040
	BAG	0.027	0.026	0.472	0.057	0.057	0.220	0.220	0.012	0.085
PSM3	*LR*	0.024	0.024	0.499	0.048	0.048	0.127	0.127	0.003	0.044
	CBPS	0.023	0.024	0.496	0.048	0.048	0.125	0.125	0.003	0.044
	GBM	0.028	0.028	0.502	0.056	0.056	0.148	0.148	0.004	0.050
	RF	0.019	0.019	0.500	0.038	0.039	0.128	0.128	0.003	0.049
	CART	0.020	0.020	0.503	0.040	0.040	0.144	0.144	0.003	0.052
	BAG	0.026	0.020	0.497	0.043	0.043	0.200	0.200	0.010	0.081
PSM5	*LR*	0.028	0.028	0.495	0.057	0.057	0.119	0.119	0.003	0.037
	CBPS	0.023	0.022	0.499	0.046	0.046	0.123	0.123	0.003	0.044
	GBM	0.001	0.001	0.497	0.004	0.005	0.150	0.151	0.007	0.076
	RF	0.020	0.020	0.496	0.041	0.042	0.127	0.127	0.003	0.047
	CART	0.015	0.015	0.499	0.031	0.031	0.127	0.127	0.003	0.048
	BAG	0.007	0.008	0.500	0.018	0.019	0.217	0.217	0.021	0.101

Abbreviations: LR: Logistic Regression , CBPS: Covariate Balance Propensity Score, GBM: Generalized Boosted Model, RF: Random Forest, CART: Classification and Regression trees. PMS1: $X_1 - X_{10}$ covariates are included , PMS4: $X_1 - X_4$ covariates are included, PMS6: $X_1 - X_{17}$ covariates are included. Treatment F: Moderate non-additivity and non-linearity, Outcome-2: non-linearity outcome model.

CHAPTER 4

The Performance of Propensity Score Weighting Methods under Limited Overlap and Model Misspecification

4.1 Introduction

Propensity score analysis has been frequently used to control for different kinds of bias in observational studies. The seminal study was by Rosenbaum and Rubin (1983). The propensity score theory and its application to various research areas' data sets have been fundamental in causal inference research. Most of the studies using propensity scores have focused on binary treatments/exposures, i.e., treatment and control groups. In the binary case, the definition of propensity score is the probability of treatment conditional on X, e(X)=P(T=1|X), where X is a covariate or vector of covariates and T is exposure or treatment assignment. Logistic regression is used to estimate these probabilities. Also, in the case of a binary treatment, subjects with similar estimated propensity score values have similar covariate vectors which removes imbalances between treated and untreated groups. For randomized studies, there is probabilistic balance between observed and unobserved covariates across treatment or exposure groups that eliminates bias and accurately estimate treatment effects so valid comparisons between groups can be made. While causal inference studies with two treatment groups common in the literature, assessing more than two treatment groups is vital in public health and medical research. But multiple treatments are more complicated than binary treatments for causal inference. Nonetheless, some papers have shown that propensity score methods can be extended to multiple treatment cases with three or more conditions. There are some advantages to using matching based on the propensity score. The main advantage of the propensity score matching is its reduction of dimensions. X includes covariates, which can have many dimensions, and the propensity score reduces all this dimensionality to a one-dimensional score. Secondly, the matching method considers not only strictly linear relationships between the outcome and propensity score, but also more complex relationships.

Researchers discuss the various matching methods in literature such as Mahalanobis metric matching, Mahalanobis metric matching including the propensity score, nearest-neighbor matching, caliper matching, nearest-neighbor matching within a caliper. One of the well-known papers examines all matching techniques (Austin, 2014), but only with two treatments. Later, propensity score weighting was generalized to the more than two treatment arms (Imbens, 2000). Even though there is an increasing demand for using various matching and weighting methods for any treatment assignment circumstances, researchers tend to prefer utilizing inverse probability weighting in literature. While having many advantages, inverse probability weights may suffer when too small or large propensity score values are present. Thus, it leads to estimate bias treatment effects and assessing inappropriate causal relationships between treatment and outcome, treatment and covariate or outcome and covariates.

There are two ways to assess whether a consistent estimator of ATE is obtained: checking covariate balance and overlap assumption. Different factors may have played a critical role in producing unsatisfied assumptions and unsatisfactory performance. In other words, reporting the inaccurate propensity score model, using missing covariates, or employing a small sample dataset may cause the assumptions to be violated. Trimming can sometimes be effective in reducing bias in nonrandomized studies. However, it may not be appropriate to perform trimming in small datasets. To address excessive PS weighting problems, Crump et al. (2009) recommend excluding large and small weights value from the sample of the estimated propensity score because they increase the variance of the estimators. The threshold for removing extreme weights is fixed as less than 0.1 and more than 0.9 propensity values. So, extreme weights might be less influential.

Overlap weights (OW), matching weights (MW), entropy weights (EW), treated weight(TW) , and inverse probability weighting(IPW) with trimming have been presented as alternatives to inverse probability weighting (IPW) for addressing overlap limitations (Crump et al.,2009; Li et al.,2018; Li et al.,2019; Mao et al.,2019; Yoshida et al.,2018; Zhou et al.,2020). Li et al.(2019) proposed extending those methods for multiple treatments and illustrated the improvement of overlap and covariate balance between each pair of treatment groups. However, their paper only gives a vague idea about relative performance of methods under the good overlap and lack of overlap conditions. Their paper did not discuss trimming with IPW, provided no information about the impact of the true propensity score models versus misspecified models, and included no study of what happens to covariate balance under violations of positivity assumptions. At the same time, Yoshida et al. (2018) conducted the study of trimming methods to deal with extreme propensity score values, but did not attempt to examine overlap of weights, matching weights, or other approaches. To address these gaps in the literature, this paper will examine eliminating positivity assumption violations under various scenarios. Furthermore, this study presents the in addition to the simulation study a study with real data from the "subset of alcohol and other drug treatment" dataset (AOD) where patients received one of three treatments. Lastly, we examined the performance of variance estimation based on the robust sandwich-type estimator and bootstrap variance estimator when using balance weights family (i.e., MW,OW,IPW,TW,EW and IPW) and GPSM methods.

The paper is organized as follows: we start Section 4.2 with our symbols and brief of background in multiple treatment. In Section 4.3,We discuss about limitation of the positivity assumption. Section 4.4 extensively addressed several approaches for eliminating the difference between treatment groups when utilizing Generalized Propensity Score Matching (GPSM), MW, OW, IPW, TW, EW, and IPW with trimming methods. We present datasets and its analysis in Section 4.5, and simulation strategies briefly Section 4.6. We illustrate a comprehensive set of monte Carlo simulations used to analyze the results of various algorithms. The results of simulations and data analysis are shown.

4.2 Background

4.2.1 The Framework of Potential Outcomes in Multiple Treatments

We offer a few notations to help explain the concepts of potential outcomes. N units are chosen from a large population ,indexed $i = 1, \dots, N$. Provide T_i is quantitative factors denoting which of the 3 or more treatments the ith subject received, and X_i is a vector of baseline covariates. Let Y_i indicate the outcome for individual i. We receive treatment $T_i = t$ if individual i is observed $t \in \mathfrak{I} = \{1, \dots, M\}$ in which M represents total of treatments. The set of potential outcomes are denoted as $\{Y(1), \dots, Y(M)\}$ for subject i considering all possible treatments, and to be clear, exactly one of those outcomes is observed for each subject. Thereby, the triple (Y_i, T_i, X_i) denoted for subject i throughout the paper. We also denote the indicator of observed treatment t for subject i as :

$$I_i(t) = \begin{cases} 1 & if \;\; T_i = t \\ 0 & otherwise \end{cases} \tag{4.1}$$

,where $I_i(t)$ represents the indicator function for receiving intervention t for individual i. In the Rubin Causal Model framework, there are M potential outcomes for each subject. The observed outcome Y_i is defined as

$$Y_i = \sum_{t=1}^{k} Y_i\,(t)\; I(T_i = t) \tag{4.2}$$

The individual treatment effect of treatment j vs treatment t ($j \neq t$) for subject i is illustrated in follows:

$$\Delta = Y_i(j) \; - Y_i(t) \tag{4.3}$$

There are two different frequently used estimands of treatment effect when employing propensity scoring for multinomial treatments: average treatment effect (ATE) and average treatment effect on the treated (ATT) . Considering ATE of treatment j versus treatment t in population is described as following :

$$\begin{aligned} \mu_{jt}^{ATE} &= E[Y(j) - Y(t)] = E[Y(j)] - E[Y(t)] \\ &= \mu_j - \mu_t \end{aligned} \tag{4.4}$$

Also, we can express the second parameter, which exemplifies the average treatment effect for the treated and as

$$\begin{aligned} \mu_{jk}^{ATT} &= E[Y(j) - Y(t)|T = t] = E[Y(j)|T = t] - E[Y(t)|T = t] \\ &= \mu_{j,t} - \mu_{t,t} \end{aligned} \tag{4.5}$$

Moreover, if we are interested in the binary outcome and the odds ratio to evaluate the treatment effect, we can employ a conditional causal treatment effect for estimands. We should point out a significant thing here that treatment cases possibly depend on the outcomes in non-randomized studies. In this way, there might be a remarkable difference for baseline covariates between treatment groups, leading us to get biased estimators for ATE. Thus, we should be aware of having overt and hidden biases in nonrandomized studies. Use of baseline covariates might provide an insight to acquire less biased estimators of ATE. In literature, some GPS approaches (such as IPTW, doubly robust estimators, etc.) have been used to decrease the bias of resulting estimator.

4.2.2. Generalized Propensity Score

The generalized propensity score (GPS) is introduced by Imbens (2000), and its theory is extended to the propensity score framework from the dichotomous exposure to multiple treatment setting (Imai and Van Dyk,2004; Rosenbaum ,1999).The generalized propensity score (GPS) is described as the likelihood of getting one of the treatments conditional on a given set of observed variables, e.g., $e_t(X) = r(t, X) = Pr(T = t|X) = E\{I(t)|X = X\}$ for T dimensional vector of probabilities $e(X) = \left(e_1(X), e_2(X), \dots e_t(X)\right)$.These propensity scores are subject to the constraint $\sum_{t\epsilon T} e_t(X) = 1$ for any value of covariates X. Because I can express each probability $e_t(X)$ as one minus the sum of the other probabilities under the $(M - 1)$ dimensional space. Also, Imbens (2000) extended the exchangeability, consistency, and positivity for multi-treatment. Estimating propensity scores in the presence of multinomial treatments have been based on the GPS vector that is generated from the multinomial regression model, e.g.,

$$log\left[\frac{Pr(T_i=t)}{Pr(T_i=M)}\right] = \theta_t + X_i'\beta_t \quad t = 1, \dots, M - 1 \tag{4.6}$$

where θ_t is an intercept, β_t is a vector of regression coefficients, T_i represents treatment and t=$\{1,2,\dots,M-1\}$ is total number of treatments. Thus, We rewrite the model to estimate generalize propensity score for $(M-1)$ treatment levels using equation (4.7) and the generalized propensity score for the reference category is estimated using equation (4.8).

$$Pr(T_i = t|X_i) = \frac{e^{\theta_t + X_i'\beta_t}}{1+\sum_{t=1}^{T-1} e^{\theta_t + X_i'\beta_t}}, \quad t = 1,\dots,M-1 \tag{4.7}$$

$$Pr(T_i = M|X_i) = \frac{1}{1+\sum_{t=1}^{T-1} e^{\theta_t + X_i'\beta_t}} \tag{4.8}$$

The existing applications have generally relied on the parametric estimation of the propensity score via the multinomial, nested, or ordinal logistic regression model for multiple treatments. We can use these models depending on the treatment values' characteristics to predict the generalized propensity score. For example, multinomial logit or probit regression are suitable for qualitative treatment values. Moreover, ordinal logistic regression can be used when there is ordering of treatment levels.

Some key assumptions have been generalized from binary treatment cases to more than two treatments cases for generalized propensity score (GPS). Also, these assumptions are indispensable for valid causal inference.

Assumption 4.1: Treatment assignment T, given the pre-treatment covariates X, is weakly unconfounded providing that,

$$I(t) \perp Y(t)|X$$

for all $t \in \Im$ and ,and $\perp$ refer to independence(Imbens 2000).

In other words, treatment indicator $I(t)$ is independent of a set of the outcome given identified covariates. Also, “strong unfoundedness” or “ignorability treatment” in dichotomous case is stronger of version than this assumption. This assumption is known as “weak unconfoundedness” in literature.

Assumption 4.2: One of the key assumptions for valid causal inference is an expanded version of the SUTVA assumption of Rubin(1978, 1980) and Rubin and Rosenbaum (1983) to the non-binary case as recommended by Imai and Dyk (2004).

$$\left(Y_i(1), \ldots, Y_i(M)\right) \perp T_s \quad for\ i \neq s$$

So, we can make something of assumption-2 that odds of interference between subjects are excluded.

Assumption 4.3: Positivity states that a non-zero likelihood of being appointed to every treatment(Rosenbaum and Rubin, 1983; Imai and Dyk,2004). Mathematical notation can be written as:

$$0 < \Pr(T_i = t | X_i = x) < 1 \quad \text{for all T,X.}$$

Positivity assumption implies that it is possible to have at least one similar individual in each treatment group. Thus, estimation of ATE can be made without needing to use extrapolation. Nevertheless, when we consider large numbers of treatments, or high dimensional of baseline covariates for estimating causal inference, the positivity assumption can be more difficult to satisfy than in binary settings. Besides, the positivity assumption is also known as overlap assumption in causal inference. So that, checking overlap between treatment groups of multiple cases can be difficult. By all means, there is a possibility that this assumption will be violated, but we can modify the population of interest to supply sufficient overlap between treatment groups.

Assumption 4.4: Imbens (2000) states that the treatment assignment indicator is independent of potential outcomes given generalized propensity score $e_t(X)$.

$$I(t) \perp Y(t)|\, e_t(X)$$

Lemma 1 : Assume that the assignment scheme is weakly unconfounded. Then,

$$\mathbb{E}[Y_i(t') - Y_i(t)] = \mathbb{E}\left[\mathbb{E}\left[Y_i^{obs}|\, T_i = t', e_{t'}(X)\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i^{obs}|\, T_i = t\ , e_t(X)\right]\right]$$

We construct subsets in which we may examine individuals at every level of treatment , resulting in

$$\mathbb{E}[Y_i(t') - Y_i(t)] = \mathbb{E}\left[\mathbb{E}\left[Y_i^{obs}|\, T_i = t', e_1(X), \dots, e_{t-1}(X)\ \right]\right] -$$

$$\mathbb{E}\left[\mathbb{E}\left[Y_i^{obs}|\, T_i = t, e_1(X), \dots, e_{t-1}(X)\ \right]\right] = \mathbb{E}\left[\mathbb{E}[Y_i(t') - Y_i(t)|\, T_i = t, e_1(X), \dots, e_{t-1}(X)\ ]\right]$$

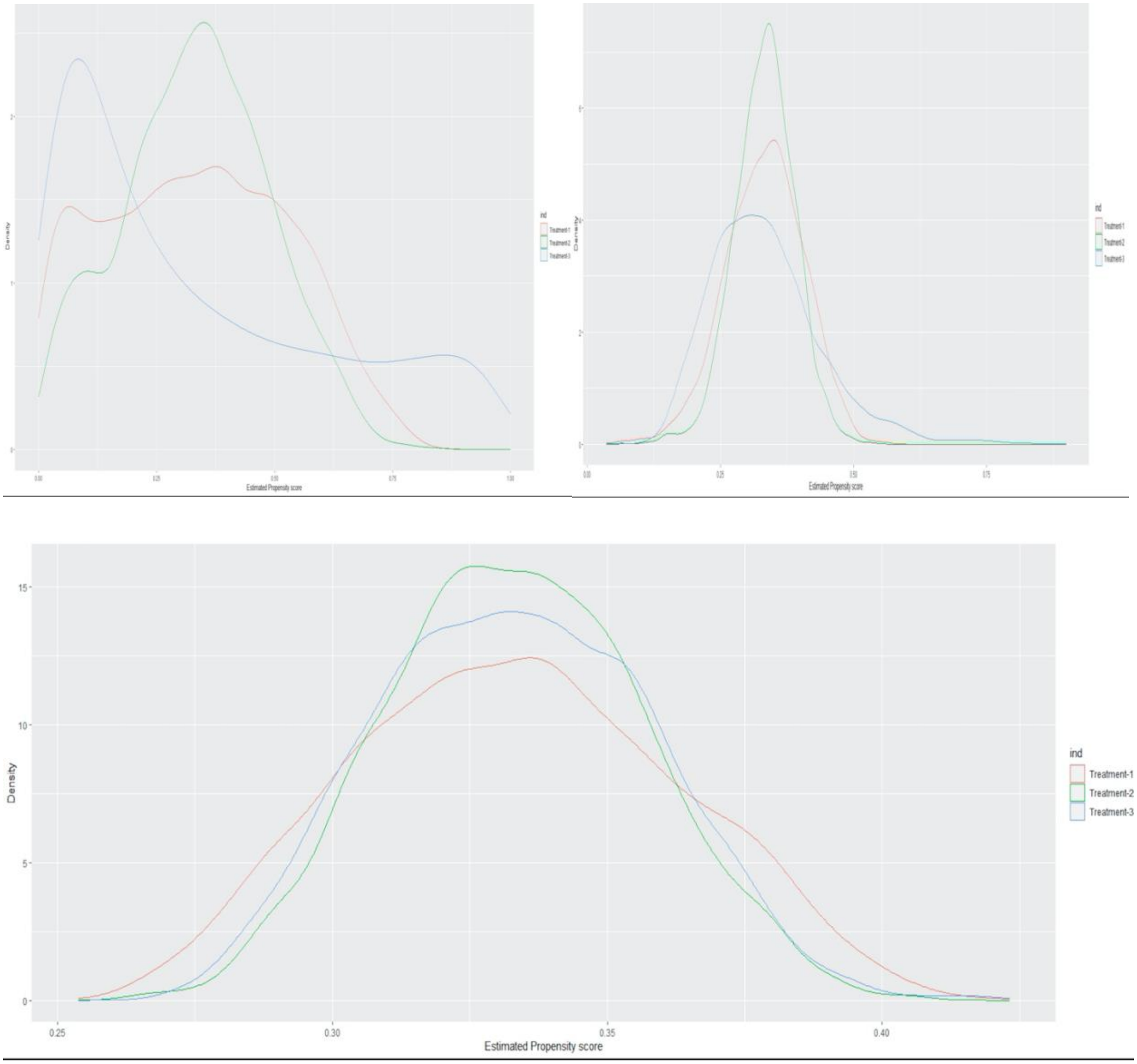
4.3 Limitation of the Overlap Assumption

When researchers desire to make inferences about causal effects in observational studies, ATE, ATT, or other estimands are used to estimate quantities such as the mean causal effects. When any fundamental assumptions in causal inference are violated, it may raise significant problems to assess causal effects. Each assumption violation may occur differently on causal inference, such as SUTVA may be violated when two different treatment versions are independently given while implementing it. Furthermore, the lack of overlap assumption has been crucial because it relies on the IPW method and this method versions in this paper. Because limited overlap between treatment and outcome may induce to occur extreme inverse probability weights. Violation of overlap assumption implies that between treated and untreated groups do not overlap, estimated propensity score values might close to 0 and 1 herein. So, those extreme propensity score estimation values are employed to examine weighting, and then, excessive propensity score values lead to having large weights (Crump et al. ,2009;Austin and Stuart,2017; Hu et al. ,2020).

We may not want excessive weights to occur because of becoming very imprecise for the causal effects. This violation can occur for various reasons, involving data constraints, a limited sample size, PS model misspecifications, and specified wrong relation among treatment/outcome and covariates. This confliction encourages researchers to think about other intended samples for whom exposure impact may be more meaningful and accurately investigated in terms of bias, RMSE, or variance(Li et al., 2018; Li et al., 2018) . So, we argued positivity assumption by conducting extensive simulation studies for multiple treatment cases across proposed IPW methods to assessing causal effects.

Numeric summaries (such as bias, variance ratios, or standard error) and some visualization (such as Q-Q plot, box plot, or density plots) are easy and good tools to display whether there is unbalanced between groups when considering binary cases. However, assessing tools have become more critical and complicated for three and more treatment cases .We have presented three different scenarios, i.e., strong lack of overlap, moderate lack of overlap, and good overlap as simulation structure and detailed results occur in Section 4.7. Figure 4.1 proposes the density plot as a graphical tool to illustrate those simulation scenarios. The horizontal axis is the plot of estimated propensity score values; meanwhile, the vertical lines of the plot represent the density values. As seen in the Figure 4.1, it is clear that difference the estimated PS values for each treatment group.

Figure 4.1: Distribution of estimated propensity score for strong lack of overlap (top-left panel) ,moderate lack of overlap(top-right panel) and good overlap (bottom)



4.4 Causal Estimand and Methods

Traditional propensity score methods, i.e., GPSM or IPW, have frequently been used in literature. However, there are various alternatives to limit the estimate of the treatment impact to an area of tolerable positivity , such as overlap weighting , matching weighting, entropy weighting, treated weighting or trimming methods (Crump et al., 2009; Li and Greene, 2013; Li et al., 2018; Mao et al., 2020).

We consider multiple treatments (i.e. $t \geq 3$) and denote the treatment for unit i as T_i. As seen in (4.1) , $\mathrm{I}_i(t)$ refers to a multinomial indicator array. From (4.2), Y_i represents potential outcome for indexes i under the exposure t as $Y_i(t)$. Also, Imbens et al. (2000) proposed generalized propensity score for potential outcome t as $e_t(X) = r(t, X) = Pr(T = t|X)$.To specify *t* th the expectation potential outcomes among target population (Li and Li, 2019):

$$\mu_t^h = \mathbb{E}[Y(t)] = \frac{\mathbb{E}\{h(x)\, m_t(x)\}}{\mathbb{E}\{h(x)\}}. \qquad (4.9)$$

Also, tilting function $h(x)$ is satisfied by ratio $h(x) = g(x)/f(x)$. Tilting function $h(x)$ means that pre-defined function of variables. To describe target population, $m_t(x) = \mathbb{E}[Y(t)|X = x]$ define the conditional expected potential outcomes in treatment *t.* Consistent estimates specified $\hat{\tau}_j^h = \frac{\sum_i^n T_i[j] Y_i w_i[j]}{\sum_{i=1}^n T_i[j] w_i[j]}$,where define $w_i[t] = \frac{1}{e_j(x)}$.Finally, we recommend causal target of inference:

$$\Delta_{j,j'}^h = \frac{\sum_i^n T_i[j] Y_i w_i[j]}{\sum_{i=1}^n T_i[j] w_i[j]} - \frac{\sum_i^n T_i[j'] Y_i w_i[j']}{\sum_{i=1}^n T_i[j'] w_i[j']} \text{ ,where } j \neq j \qquad (4.10)$$

As we introduced average treatment effect or average treatment on treated effect for multiple treatments, pairwise causal effect estimands describe $\tau_{t,t'}$ =μ_t-$\mu_{t'}$ between t and t'.

After define causal treatment effects for balancing weights, density function for each treatment group, t is provided by $f_t(X) = f(X|T = t)$.To target population, density of specific treatment group $f_t(X)$ is weighted as following :

$$\mathrm{w_j(X)} = \frac{\mathrm{f(X)h(X)}}{\mathrm{f(X)\ e_t(X)}} = \frac{\mathrm{h(X)}}{\mathrm{e_t(X)}}$$

GPSM: Yang et al. (2016) recommend the generalized propensity score matching (GPSM) for multiple treatment cases. The matching process of GPSM as following: i) Estimate propensity score based on the multinomial logit model. ii) Assume that three treatment levels are available (called t_1, t_2, t_3). We matched the treatment t_1 closest to treatment t_2 based on the estimated propensity score values without replacement. iii) matched treatment is removed from the sample and continues the same process for the rest of the unmatched observation.

IPW: The IPW is a frequently utilized technique which employs the propensity score. IPW is a common method that includes the weighting unit of each treatment level with the inverse of their assigned exposure probability. IPW purpose examining the mean weighted outcome covariate between treatment groups. The weighting literature is influenced by inverse probability weights, which originated with the Horvitz-Thompson weight in survey sampling (IPW). To target population, standard inverse probability weighting is defines as $\{1/e_t\}$.

Overlap weighting : Li et al. (2018) proposed overlap weighting (OW) which is a balancing weighting design to fix the problems of inverse probability weighting and trimming. There are available studies to examine the performance of overlap weighting when considering binary treatment cases (see Li et al.,2018; Maou et al.,2019; Thomas et al.,2020). The OW is very straightforward in terms of implementing techniques for the binary case. So, overlap weights e and (1-e) are remarked for treatment and control groups. Also, bootstrap variance estimator and robust sandwich variance estimator are available in Li et al. (2018); and Stefanski and Boos (2002). Furthmore, Li and Li (2019) extended the study of overlap weighting from binary to multiple treatment cases. The combination of IPW and harmonic mean of GPS lead to be established generalized OW method, which weighted each individual proportionate according to its probability of being assigned to the other group. The GOW aims to consider sub-population, which has probabilities of being assigned to all exposure groups. The most crucial feature of the GOW method is that illustrate good performance when tailing exists. It means that we can limit values between 0 and 1, and thus, we can eliminate excessive propensity score values when employing inverse probability weighting.

Li et al. (2018) introduced the OW , which weighted each individual proportionate to its probability of being assigned to the other group. Tilting function, $h(X)$ is defined as $\left(\sum_{k=1}^{M} 1/e_k(X)\right)^{-1}$. Balance weights for OW as

$$w_i^{\text{overlap}} = \frac{\left(\sum_{k=1}^{M} 1/e_k(X)\right)^{-1}}{e_t(X)}$$

Matching weighting: Another alternative to the IPW technique is matching weight (MW) methods. Even though the matching weighting technique is known as a balance weights technique, pair matching and matching weights techniques have nearly similar estimands. Contrary to the difficulties of applying pair matching techniques, matching weights are easy in terms of the implementation process and have eliminated the challenges of pair-matching in practice (Li and Greene, 2013).The matching weighting (MW) is recommended by Yoshida et al. (2017) :

$$w_i^{\text{matching}} = \frac{\{\min_{1\le k\le t}\{e_k(X)\}\}}{e_t(X)}$$

Entropy weighting: Hainmueller(2012) recommended entropy balancing(or called entropy weighting), which aims pre-processing approach to remove covariate imbalance. The most important difference of the entropy balancing technique is that entropy balancing is a reweighting technique that includes variables balance straightly in the weight function employed to the sample units. Entropy balancing enables to achieve a high degree of variable balance by applying the extensive set of balance requirements that include the first, second, and perhaps higher moments of variable distributions and interactions. Another critical feature in entropy weighting is that the reweighting method is flexible when we have weights around 0 and 1. Entropy weighting is defined by :

$$\mathrm{w}_{\mathrm{i}}^{\mathrm{entropy}} = \frac{-\sum_{\mathrm{k}=1}^{\mathrm{M}} \mathrm{e}_{\mathrm{k}}(\mathrm{x})\log\{\mathrm{e}_{\mathrm{k}}(\mathrm{x})\}}{\mathrm{e}_{\mathrm{t}}(\mathrm{x})}$$

Treated weighting **:** Horvitz–Thompson (HT) weight is different version of inverse probability weighting. HT weights can focus on the treatment effects on the treated group (Hirano and Imbens, 2001). Hirano and Imbens (2001) provide the treated weighting(TR):

$$\mathrm{w}_{\mathrm{i}}^{\mathrm{treated}} = \frac{\mathrm{e}_{\mathrm{k}}(\mathrm{X})}{\mathrm{e}_{\mathrm{t}}(\mathrm{X})}$$

<u>*Variance Estimation:*</u> Lunceford and Davidian (2004) and Li et al (2019) state that the empirical sandwich variance for PSW estimator rely on the M estimation theory. For multiple treatment cases. The parameters array is shown as $\theta = \left(\mathrm{v}_1, \ldots, \mathrm{v}_{\mathrm{J}}, \eta_1, \ldots, \eta_{\mathrm{J}}, \beta^{T}, \alpha^{\mathrm{T}}\right)^{\mathrm{T}}$.

Then$\left\{\mu_J = \widehat{v_J} + \hat{n}_J : j = 1, \ldots, J\right\}$ is solved as

$$\sum_{i=1}^{N} \psi_i(\theta) = \sum_{i=1}^{N} \begin{pmatrix} w_1\,(x_i)\;\; I_{i1}\{Y_i - m_1(x_i;\alpha) - v_1\} \\ . \\ . \\ w_J\,(x_i)\;\; I_{ij}\{Y_i - m_J(x_i;\alpha) - v_J\} \\ h(x_i)\{m_1(x_i;\alpha) - \eta_1\} \\ . \\ . \\ h(x_j)\{m_j(x_i;\alpha) - \eta_j\} \\ S_\beta(T_i, x_i, \beta) \\ S_\alpha(Y_i, T_i, x_i, \alpha) \end{pmatrix} = 0$$

where clearly that $S_\beta(T_i, x_i, \beta)$ and $S_\alpha(Y_i, T_i, x_i, \alpha)$ are score function for propensity score model and outcome model. Also, $m_j(x) = E[Y(j)|X = x]$. Empirical sandwich variance estimator is

$$\widehat{\mathbb{V}}(\hat{\theta}) = \left\{\sum_{i=1}^{N} \frac{\partial}{\partial \theta^T} \psi_i(\hat{\theta})\right\}^{-1} \left\{\sum_{i=1}^{N} \psi_i\,(\hat{\theta})\;\psi_i{}^T(\hat{\theta})\right\}\left\{\sum_{i=1}^{N} \frac{\partial}{\partial \theta^T} \psi_i{}^T(\hat{\theta})\right\}^{-1}$$

4.5 Application to Alcohol and Other Drug Treatment Dataset

4.5.1 Dataset

We used alcohol and other drug treatment dataset (AOD), which McCaffrey (2013) introduced. McCaffrey utilized AOD dataset to provide step-by-step guidance to estimate the average treatment effect based on the generalized boosted method for multiple treatment settings. Three treatment levels have been determined, including "traditional programs (community)," "motivational enhancement therapy plus cognitive behavior therapy (MET/CBT-5)", and "Strengthening Communities for Youth (SCY)." There are 600 individuals with five covariates in the twang package in R, even though McCaffrey (2013) used a larger sample. Also, the five pretreatment variables are illicit activities scale ("illact"), criminal justice involvement ("crimjust"),substance use problem scale (“subprob”),substance use dependence scale (“subpdep”) and race (“white”). The outcome variable “suf12” represents treating drug abuse following twelve months post-intake**.**

4.5.2 Results

Table 4.1 examines the features of patients who received one of three treatments(i.e., CBT-5,community, and SCY groups) for five covariates unweighted dataset and datasets using balance-weighted techniques. There are valuable metrics such as bias, standardized bias, or relative bias to measure balance in the binary or multiple treatment cases. These metrics allow analysis to examine the magnitude of the different exposure between treatment groups in the distributions of covariates. One of the preferable performance metrics is a standardized mean difference, which is computed as a difference in the average of a variable between treatment groups, divided by a pooled estimate of the variable's standard deviation. Even though researchers do not reach a consensus on what threshold for standard mean difference should be used, some articles (see Austin and Stuart,2015; Normand et al.,2001) suggest that SMD is smaller than 10% of SMD values is considered as evidence of balance between groups. However, McCaffrey et al. (2012) recommended that SMD keep within bounds of 20% might indicate a meaningful balance between treatment groups in covariates. When we prefer using the 10% threshold value to examine the balance between three treatment groups in AOD datasets in the distribution of five covariates, we see that four of 5 measured variables (i.e., illact, crimjust , subdep, and white variables) exceeded 10%, which means that significantly imbalance between treatment groups in the unweighted dataset (Table 4.1). Fortunately, all SMDs are less than 10%, which indicates a good balance for all covariates across all balance weights. When SMDs are close to 0, we should realize that it is evidence of perfect balance in distributions of covariates. The resulting of SMDs metrics reveal that using any of balance weight techniques provide improvements on the covariates.

In addition, there are significant means between community, CTB-5, and SCY treatment groups in the unweighted dataset, such as illact covariates. However, means in treatment groups are decreased difference. For example, means of illact variable for the community, CTB-5, and SCY groups have corresponded to 0.083,0.007 and 0.120 in the unweighted dataset, respectively, while illact covariate for those groups' means is 0.067, 0.082, and 0.078 in matching weight, respectively. Similar results were observed for other covariates where for mean, all methods tended to be similar. Thus, in general, all weighted methods perform close mean values for each to other groups. Therefore, in general, treatment group means are closer to each other in all weighted methods compared to treatment group means in the unweighted dataset.

Table 4.1: Averages for treatment groups in unweighted and weighted and standardized mean difference in AOD dataset

	Weighted means			Stand. mean diff. (SMD)
	Community	CTB-5	SCY	
Unweighted				
illact	0.083	0.007	0.120	0.112
crimjust	-0.033	0.037	-0.174	0.206
Subprob	-0.058	0.026	-0.013	0.085
subdep	0.052	0.058	-0.058	0.112
white	0.162	0.200	0.175	0.100
Matching weight				
illact	0.067	0.082	0.078	0.015
crimjust	-0.069	-0.067	-0.066	0.003
subprob	-0.018	-0.009	-0.009	0.010
subdep	0.007	0.022	0.008	0.015
white	0.178	0.179	0.182	0.010

Overlap weight				
illact	0.079	0.080	0.081	0.002
crimjust	-0.058	-0.066	-0.062	0.007
subprob	-0.021	-0.021	-0.016	0.006
subdep	0.018	0.022	0.010	0.012
white	0.179	0.181	0.184	0.015
Entropy weight				
Illact	0.080	0.080	0.080	0.000
Crimjust	-0.057	-0.066	-0.062	0.009
Subprob	-0.020	-0.022	-0.016	0.006
Subdep	0.019	0.021	-0.010	0.010
white	0.179	0.182	0.185	0.015
Treated weight				
illact	0.139	0.132	0.120	0.019
Crimjust	-0.167	-0.195	-0.174	0.028
Subprob	-0.013	-0.028	-0.013	0.016
subdep	-0.054	-0.066	-0.058	0.012
white	0.179	0.182	0.175	0.018
IPW				
illact	0.081	0.080	0.080	0.001
crimjust	-0.056	-0.067	-0.062	0.011
subprob	-0.019	-0.023	-0.015	0.007
subdep	0.020	0.019	0.010	0.009
white	0.179	0.182	0.185	0.015

Table 4.2 summarized the results of the causal estimands, standard error, and confidence intervals for group comparisons (i.e., CBT-5 vs. community, group SCY vs. community, and group SCY vs. CBT-5). Table 4.2 shows that average treatment effects in the GPSM method for three comparison groups are equal to 0.176,0.217, and 0.047, which provide larger values than all

weighted methods. It seems to receive a similar standard error for three treatment groups across all weighting methods and GPSM.

Table 4.2 :*Estimation of the treatment effect in the AOD data application employing various balance weighting methods*

	Estimand	***Std. error***	***95% CI***
Matching Weighting			
CBT-5 vs Community	0.1440	0.097	(-0.046,0.336)
SCY vs Community	0.098	0.096	(-0.090, 0.288)
SCY vs CBT-5	-0.045	0.103	(-0.248,0.157)
Overlap Weighting			
CBT-5 vs Community	0.136	0.097	(-0.053,0.326)
SCY vs Community	0.087	0.095	(-0.098,0.274)
SCY vs CBT-5	-0.049	0.102	(-0.250,0.151)
Entropy Weighting			
CBT-5 vs Community	0.134	0.097	(-0.055,0.324)
SCY vs Community	0.085	0.095	(-0.100,0.272)
SCY vs CBT-5	-0.049	0.102	(-0.250,0.151)
Treated Weighting			
CBT-5 vs Community	0.115	0.102	(-0.085,0.316)
SCY vs Community	0.086	0.096	(-0.102,0.275)
SCY vs CBT-5	-0.028	0.107	(-0.239,0.181)
IPW			
CBT-5 vs Community	0.133	0.097	(-0.057,0.323)
SCY vs Community	0.084	0.094	(-0.101,0.270)
SCY vs CBT-5	-0.048	0.102	(-0.250,0.152)
GPSM			
CBT-5 vs Community	0.176	0.100	(-0.342,-0.011)
SCY vs Community	0.217	0.103	(-0.386,-0.048)
SCY vs CBT-5	0.047	0.101	(-0.207,0.126)

4.6 Simulation Study

We perform a comprehensive simulation study to assess performance using a variety of methodologies. We modified the simulation structure define by Yang et al. (2016).We generate treatment with three levels, continuous outcome and ten covariates in the simulation. X_{1i}, X_{2i} and X_{3i} covariates are generated based on the multivariate normal distribution with mean of (0,0,0) , variances of (2,1,1) and covariances of (1,-1,-0.5) between X_1 and X_2, X_1 and X_3, and X_2 and X_3, respectively. Also, X_{4i} ~U[-3,3], X_{5i} ~χ_1^2, X_{6i}~Bernoulli(0.5),X_{7i}~Bernoulli(0.7), X_{8i}~U[-2,2], X_{9i}~Bernoulli(0.7), X_{10i}~U[-2,2], all independent of each other and X_1, X_2, and X_3 . The distribution of the treatments is

$$(T_i(1), T_i(2), T_i(3)) \sim Multinom\big(p(1|X_i), p(2|X_i), p(3|X_i)\big)$$

where exposure indicator $T_i(t)$ is defines as

$$p(1|X_i) = \left(\frac{exp(X_i^T \theta_t)}{\sum_{t=1}^{3} exp(X_i^T \theta_t)}\right)$$

where $\theta_1^T = \gamma_1 * (0,0,0,0,0,0,0,0)$, $\theta_2^T = \gamma_2 * (1,1,1,-1,-1,1,1,-1)$ and $\theta_3^T = \gamma_3 * (1,1,1,1,1,1,1,1)$. We should make reminder that covariates X_1- X_8 are related to treatment assignments. In other words, X_{9i} and X_{10i} covariates do not have association with treatment levels. We construct $(\gamma_1, \gamma_2, \gamma_3) = (0,0.2,0.8)$ for strong lack of overlap , $(\gamma_1, \gamma_2, \gamma_3) = (0,0.05,0.2)$ for middle lack of overlap and $(\gamma_1, \gamma_2, \gamma_3) = (0,0.1,0.1)$ for good overlap. We generate the outcome as following:

$$Y_i(k) = (1, X_i^T)\alpha_k + \epsilon_i$$

where $\epsilon_i \sim N(0,1)$, $\alpha_1^T = (-1.5,1,1,1,1,1,1,1,1)$, $\alpha_2^T = (-4,2,3,1,2,2,2,2)$ and $\alpha_3^T = (3,3,1\ ,2,-1,-1,-1,2)$. $N_t = 10000$ sample sizes and 500 iterations are considered with $t = 1,2,3$. Finally, to misspecify the true PS model, we delete the covariates X_3 and X_4 from full model

4.7 Results

We conducted a comprehensive simulation study to examine various weighting methods and generalized propensity score matching method in this article. Table 4.3 presents bias, RMSE, and empirical standard deviation. Bias is calculated by $\frac{1}{1000}\sum_{i=1}^{1000}(\theta_i - \theta)$ where θ is the population difference in response between two treatment groups and θ_i is the estimated difference for Monte Carlo run *ith* between the same two treatment groups; meanwhile, RMSE is specified as $\sqrt{\text{RMSE} = \frac{1}{1000}\sum_{i=1}^{1000}(\theta_i - \theta)^2}$. The empirical standard error is computed as the sample SD of the point estimates. We summarized the performance of matching weighting(MW), overlap weighting(OW), entropy weighting(EW), treated weighting (TW), IPW, IPW with trimming, GPSM and GPSM with trimming when considering good overlap, moderate lack of overlap, and substantial lack of overlap scenarios in Table 4.3, 4.4, 4.5. In addition, we present the results of misspecified PS models under the good overlap, mild lack of overlap, and strong lack of overlap scenarios. When all the methods are considered for good overlap and true PS model, MW, OW, EW IPW, and IPW with trimming offered adequate estimation in terms of absolute bias and RMSE. The TW method illustrates the most extensive bias, RMSE, and empirical SD in the true PS model and misspecified model(Table 4.3). The four remaining procedures, involving MW, OW, EW, and IPW, do a pretty good job of achieving a small bias for all treatment effects. GPSM results in greater bias than balance weighting methods in the presence of good overlap in PS distributions with 10000 sample size and generating 500 datasets. Table 4.3 provides the misspecified model, which represents removing covariates X_3 and X_4 from the full model. Table-3 provides a misspecified model created by subtracting the X_3 and X_4 covariates from the full model. MW, OW, and EW methods present no changes in the presence of bias from using the true

PS model to a misspecified model under the good overlap. Unfortunately, IPW and GPSM methods produce relatively larger bias in the misspecified PS model than the true PS model. This leads us to conclude that IPW and GPSM may be sensitive toward defining the model in terms of bias. Generally, MW, OW, EW, and IPW methods are related with the smallest and similar RMSE and empirical SD. So, those methods are more effective than TW and GPSM. However, when a moderate lack of overlap exists (Table 4.4), all methods show more bias, larger RMSE, and empirical SD compared to existing good overlap. TW and GPSM method (Table 4.4) illustrate huge bias, which similarly resulted. Finally, Table 4.5 reveals that all methods tend to be more extensive measured metrics. In other terms, there is a dramatic increase in terms of values of measured metrics from a good overlap scenario to a substantial lack of overlap (check Table 4.3 and Table 4.5).

We report the ratio of the average estimated standard error to the empirical standard deviation of estimated in Table 4.6-4.8. There are two different standard error estimators: *i)* bootstrap standard error estimator *ii)* robust sandwich-type standard error estimator in literature. So, we perform those estimators across six methods when considering good overlap, moderate overlap, and a strong lack of overlap. Efron and Tibshirani (1993) recommended that 200 bootstrap samples illustrate adequate to estimate the standard error. Austin (2015) provided that the performance of variance estimator is measured by $\frac{\bar{\bar{\gamma}}}{sd(\hat{\theta}_i)}$, where the mean standard error across the 500 iterations: $\bar{\bar{\gamma}} = \frac{1}{500}\sum_{i=1}^{500}\widehat{\gamma_i}$ and the standard deviation of the estimated across the 500 simulated datasets: $sd\left(\hat{\theta}_i\right)$. If this ratio is close to 1 in bootstrap variance estimation, it shows that the bootstrap estimation accurately approached the SD of the estimated empirical sample distribution. Robust standard error in IPW with trimming method overestimated the variability of estimates for either true PS model and misspecified PS model when considering a strong lack of overlap. However, GPSM, GPSM with trim, and TW methods perform underestimated the variability of estimates for both PS models (Table-4.8). To sum up, MW, OW, EW, and IPW methods in a ratio are approximately equal to one for three overlap scenarios. However, GPSM and TW methods in the ratio show the worst performance when there is an increasing lack of overlap.

We reported the standardized mean difference (SMD) to examine the balance of baseline variables before and after using MW, OW, EW, TW, IPW, and GPSM for each pairwise treatment. In literature, if SMD is below or around 0.1(10%) threshold, there is a weak imbalance between treatment groups. Austin (2008) provides that SMD does not depend on the measurement units and size of the dataset. So, we present the average SMD that calculate across the 500 simulated datasets for each of the 10 variables in Table 4.9-4.11. In this simulation study, SMD aims to compare the balance in baseline covariates between whether they get specific treatment.Table-4.9 presents that there is no crude imbalance in the original, unweighted sample. Because Figure 4.2 illustrates how to distribute the estimated propensity score values of each treatment group, which was called a good overlap scenario. We realize that there is a perfect overlap between treatment groups in Figure 4.2. After employing any weighting techniques or GPSM, It is expected to see perfect balance in covariates between treatment groups like Table-4.9. Because Rubin and Rosenbaum (1983) indicated that propensity score is a balancing score. Therefore, we can conclude that all covariates seem to be good balance when PS overlap is good. Even though six techniques decrease the SMD of covariates X_1-X_{10}compared with unweighted analysis (Table-4.10), all SMD for ten covariates across all methods(Table-4.10) illustrates increased when compared with Table-4.9. In particular, all weighting methods exact good covariate balance for moderate-overlap scenario (Table 4.10). However, the SMD of four covariates, i.e., X_1, X_3, X_4, and X_8, are more than 10% when employing the GPSM method. Due to a strong lack of overlap, there is a massive explosion in the SMD of baseline covariates in Table-4.11. While nine of ten measured baseline covariates in the GPSM method had SMD that overrun 10%, one covariate (i.e., X_1) in MW techniques and two covariates (i.e., X_1 and X_2) are more extensive than 10%, which means that there is an imbalance for those variables. The biggest observed SMDs are for X_1 (37.5%) and X_2 (27.5%) in the GPSM method. To conclude, if there is a substantial lack of overlap, matching weighting and overlap weighting lead to better balance in measured baseline covariates than EW, TW, IPW, and GPSM. However, all covariates perform well across all six methods when there is good PS overlaps.

4.8 Summary & Discussion

Researchers in medical, social science and public health studies frequently utilize propensity score-based techniques that aim to eliminate bias between treatment groups. Since Rosenbaum and Rubin introduced propensity score methods in 1983, binary treatment cases have generally been studied. While multiple treatment groups have become popular recently, multiple treatments might be more challenging to design, perform, and interpret. Imbens (2001) proposed the causal models and validated them utilizing propensity scores to eliminate bias in cases with more than two treatment. The use of the IPW methods in multiple treatment cases has posed some critical issues that are being discussed in the literature. The main problem of using IPW is that weights become large when the estimated PS value is approximately 0 and 1. So, large weights produce biased treatment effects and large variability of estimates. Alternative methods to IPW have been suggested recently to eliminate or alleviate this problem. These alternative approaches are treated weights (Hirano and Imbens,2001), matching weights(Li and Greene,2013), overlap weights (Li et al.,2018), entropy weight(Zhou et al.,2020). Those weighting techniques are called 'balance weighing. All those approaches are mainly employed based on binary treatment cases(Li et al. 2018; Mao et al.,2019; Zeng et al.,2020; Li and Greene) and multiple treatments (Yoshida et al.,2017; Li and Li,2019; Hu et al.,2020).

This paper used balancing weighting family and generalized propensity score matching approaches to derive causal inferences from observational studies with multiple treatment cases. We conducted comprehensive Monte Carlo simulations to explore these issues. When levels of violation of overlap assumption are increased, bias and RMSE metrics values across all methods also increased. The results of table 4.3-4.6 give a clue that there is a large variability of estimated treatment effects when the overlap assumption is violated. MW, OW, and EW methods perform nearly identically using both the true PS model and misspecified PS model when there is no violation of overlap assumption in Table 4.3. GPSM with or without trimming performed poorly in Table 4.3. So, we can conclude that applying GPSM may not be a good choice to eliminate bias in treatment effects estimates. However, MW and OW methods are consistently more effective than the remaining methods, including both PS models when moderate lack of overlap and substantial overlap existed. In other words, considering all measurements listed in Table 4.3-4.5, both OW and MW made the best out of the six techniques. As can be seen, OW, MW, and EW methods are not sensitive to model misspecification under good overlap assumption. There are no changes in the bias metric for a misspecified PS model.

We conduct standard error estimation using balance weighting family methods and GPSM with propensity score to estimate the treatment effect. We realize that the use of bootstrap and sandwich robust estimator tended to result in accurate estimated standard errors in MW, OW, EW. However, GPSM, IPW, and TW performed the worst for both bootstrap and robust sandwich-type standard error estimators under all overlap scenarios. However, Joffe et al. (2004) suggest using a robust standard estimator is more appropriate when employing IPW for binary cases. Abadie and Imbens (2008) recommend utilizing bootstrapping to estimate SE was improper. However, we conclude that either bootstrapping and a robust sandwich-type estimator can be used for OW, MW, and EW methods when assessing all balance weighting methods and GPSM in multiple treatment cases. This study sheds light on the evaluating of all balance weighting methods utilizing true propensity score model and misspecified propensity score model under different levels of overlap.

4.9 Appendix

Table 4.3: Performance of the various weighting and GPSM methods in simulation when both true propensity score model and the misspecified propensity score model with good overlap. Simulation results with t=3 groups in 1000 datasets

		GOOD OVERLAP					
		TRUE PS MODEL			**MISSPECIFIED PS MODEL**		
		τ_{12}	τ_{13}	τ_{23}	τ_{12}	τ_{13}	τ_{23}
|BIAS|	*Matching W*	0.084	0.107	0.135	0.084	0.109	0.138
	Overlap W.	0.080	0.100	0.123	0.080	0.100	0.124
	Entropy W.	0.080	0.099	0.122	0.080	0.099	0.123
	Treated W.	0.116	0.142	0.193	0.115	0.142	0.192
	IPW(No trim)	0.080	0.098	0.122	0.090	0.099	0.123
	IPW (at 0.1 trim)	0.080	0.099	0.121	0.091	0.100	0.125
	GPSM(No trim)	0.153	0.206	0.243	0.173	0.188	0.220
	GPSM (at 0.1 trim)	0.152	0.206	0.240	0.173	0.189	0.221
RMSE	*Matching W*	0.105	0.134	0.166	0.102	0.136	0.170
	Overlap W.	0.102	0.123	0.153	0.102	0.125	0.154
	Entropy W.	0.102	0.125	0.153	0.102	0.125	0.153
	Treated W.	0.146	0.183	0.241	0.144	0.183	0.240
	IPW(No trim)	0.102	0.124	0.152	0.102	0.124	0.153
	IPW (at 0.1 trim)	0.102	0.125	0.154	0.101	0.126	0.154
	GPSM(No trim)	0.105	0.134	0.166	0.218	0.241	0.286
	GPSM (0.1)	0.191	0.253	0.302	0.219	0.240	0.287
Emp. SD	*Matching W*	0.111	0.130	0.157	0.174	0.150	0.191
	Overlap W.	0.109	0.126	0.152	0.173	0.146	0.188
	Entropy W.	0.109	0.125	0.152	0.174	0.145	0.188
	Treated W.	0.167	0.184	0.250	0.205	0.172	0.225
	IPW(No trim)	0.109	0.125	0.152	0.173	0.145	0.188
	IPW (0.1)	0.108	0.126	0.153	0.172	0.146	0.189
	GPSM(No trim)	0.204	0.248	0.316	0.264	0.248	0.327
	GPSM (0.1)	0.207	0.266	0.309	0.261	0.271	0.322

Abbreviations: IPW: Inverse Probability Weighting, GPSM: Generalized Propensity Score Matching, Emp. SD: Empirical Standard Deviation

Table 4.4: Performance of the various weighting and GPSM methods in simulation when both true propensity score model and the misspecified propensity score model with good overlap. Simulation results with t=3 groups in 1000 datasets

		MODERATE LACK OF OVERLAP					
		TRUE PS MODEL			**MISSPECIFIED PS MODEL**		
		τ_{12}	τ_{13}	τ_{23}	τ_{12}	τ_{13}	τ_{23}
|BIAS|	*Matching W*	0.151	0.231	0.363	0.150	0.225	0.353
	Overlap W.	0.263	0.224	0.481	0.267	0.222	0.484
	Entropy W.	0.281	0.203	0.476	0.279	0.200	0.471
	Treated W.	0.738	0.480	1.219	0.738	0.486	1.224
	IPW(No trim)	0.251	0.167	0.404	0.256	0.164	0.402
	IPW (0.1)	0.269	0.194	0.455	0.269	0.189	0.450
	GPSM(No trim)	0.781	0.573	1.353	0.791	0.580	1.370
	GPSM (0.1)	0.545	0.464	0.863	0.538	0.450	0.843
RMSE	*Matching W*	0.178	0.266	0.400	0.174	0.259	0.391
	Overlap W.	0.283	0.255	0.504	0.286	0.253	0.507
	Entropy W.	0.301	0.235	0.500	0.299	0.230	0.494
	Treated W.	0.767	0.525	1.249	0.765	0.530	1.252
	IPW(No trim)	0.279	0.200	0.432	0.279	0.195	0.428
	IPW (0.1)	0.292	0.229	0.482	0.291	0.223	0.476
	GPSM(No trim)	0.795	0.612	1.375	0.803	0.625	1.394
	GPSM (0.1)	0.630	0.546	1.044	0.630	0.538	1.047
Emp. SE	*Matching W*	0.127	0.140	0.190	0.180	0.146	0.205
	Overlap W.	0.114	0.136	0.168	0.173	0.140	1.190
	Entropy W.	0.113	0.135	0.165	0.174	0.140	1.190
	Treated W.	0.252	0.211	0.298	0.272	0.186	0.296
	IPW(No trim)	0.125	0.137	0.170	0.181	0.140	0.195
	IPW (0.1)	0.154	0.171	0.249	0.181	0.183	0.242
	GPSM(No trim)	0.343	0.447	0.445	0.297	0.318	0.301
	GPSM (0.1)	0.127	0.140	0.190	0.180	0.146	0.205

Abbreviations: IPW: Inverse Probability Weighting, GPSM: Generalized Propensity Score Matching, Emp. SD: Empirical Standard Deviation

Table 4.5: Performance of the various weighting and GPSM methods in simulation when both true propensity score model and the misspecified propensity score model with strong lack of overlap. Simulation results with t=3 groups in 1000 datasets

		STRONG LACK OF OVERLAP					
		TRUE PS MODEL			**MISSPECIFIED PS MODEL**		
		τ_{12}	τ_{13}	τ_{23}	τ_{12}	τ_{13}	τ_{23}
|BIAS|	*Matching W*	1.054	1.392	2.447	1.049	1.304	2.354
	Overlap W.	1.234	1.311	2.545	1.247	1.290	2.538
	Entropy W.	1.552	1.116	2.661	1.501	1.170	2.671
	Treated W.	0.935	0.323	1.089	0.920	0.327	1.136
	IPW(No trim)	1.278	1.891	2.126	1.283	1.936	2.204
	IPW (0.1)	1.151	1.208	2.360	1.129	1.186	2.316
	GPSM(No trim)	2.460	2.027	4.488	2.475	2.114	4.590
	GPSM (0.1)	1.431	2.307	1.183	1.412	2.236	1.117
RMSE	*Matching W*	1.064	1.401	2.454	1.057	1.313	2.361
	Overlap W.	1.241	1.319	2.551	1.253	1.298	2.543
	Entropy W.	1.557	1.150	2.679	1.506	1.186	2.679
	Treated W.	1.246	0.438	1.352	1.184	0.456	1.352
	IPW(No trim)	1.323	1.945	2.183	1.317	1.969	2.237
	IPW (0.1)	1.241	1.319	2.551	1.257	1.164	2.411
	GPSM(No trim)	2.464	2.052	4.498	2.479	2.132	4.598
	GPSM (0.1)	1.459	2.416	1.213	1.433	2.328	1.147
Empr. SD	*Matching W*	0.154	0.172	0.210	0.171	0.152	0.179
	Overlap W.	0.137	0.164	0.195	0.162	0.149	0.176
	Entropy W.	0.128	0.299	0.309	0.162	0.200	0.220
	Treated W.	0.985	0.421	0.912	0.464	0.227	0.461
	IPW(No trim)	0.462	0.372	0.526	0.219	0.232	0.279
	IPW (0.1)	0.193	0.265	0.311	0.213	0.267	0.314
	GPSM(No trim)	0.154	0.346	0.363	0.205	0.274	0.285
	GPSM (0.1)	1.562	1.045	1.302	0.151	1.026	1.227

Table 4.6 : The proportion of average estimated standard error to empirical standard deviation of sampling variability of estimated across 1000 iterations of good overlap scenarios

Methods	**Pairwise of groups**	**Good Overlap**			
		True propensity score Model		**Misspecified PS Model**	
		Bootstrap	*Robust*	*Bootstrap*	*Robust*
Matching W.	τ_{12}	1.039	1.049	0.972	1.132
	τ_{13}	1.017	1.040	0.990	1.041
	τ_{23}	1.096	1.130	1.005	1.156
Overlap W.	τ_{12}	0.989	0.989	0.958	0.965
	τ_{13}	0.974	0.986	0.974	0.971
	τ_{23}	1.025	1.029	0.977	0.974
Entropy W.	τ_{12}	1.002	0.988	0.962	0.964
	τ_{13}	0.992	0.987	0.966	0.973
	τ_{23}	1.031	1.029	0.978	0.974
Treated W.	τ_{12}	0.970	0.955	0.974	0.815
	τ_{13}	1.004	1.010	0.998	0.821
	τ_{23}	1.007	1.006	1.014	0.816
IPW	τ_{12}	1.002	0.989	0.970	0.965
	τ_{13}	1.005	0.990	0.974	0.975
	τ_{23}	1.042	1.030	0.975	0.976
IPW (at 0.1 trim)	τ_{12}	0.998	0.989	0.976	0.965
	τ_{13}	1.004	0.990	0.977	0.975
	τ_{23}	1.042	1.030	0.981	0.975
GPSM	τ_{12}	0.956	0.963	0.747	1.002
	τ_{13}	0.644	0.657	0.630	0.702
	τ_{23}	0.721	0.749	0.676	0.760
GPSM (at 0.1 trim)	τ_{12}	0.942	0.950	0.749	0.957
	τ_{13}	0.601	0.659	0.606	0.660
	τ_{23}	0.736	0.773	0.669	0.780

Table 4.7 : The proportion of average estimated standard error to empirical standard deviation of sampling variability of estimated across 1000 iterations of moderate lack of overlap scenarios

Methods	**Pairwise of groups**	**Moderate Lack of Overlap**			
		True propensity score Model		**Misspecified PS Model**	
		Bootstrap	*Robust*	*Bootstrap*	*Robust*
Matching W.	τ_{12}	1.044	1.056	0.985	0.964
	τ_{13}	1.020	1.053	1.030	1.031
	τ_{23}	1.021	1.054	0.982	0.973
Overlap W.	τ_{12}	1.020	1.006	0.971	0.970
	τ_{13}	0.969	0.974	1.023	1.133
	τ_{23}	0.989	0.978	0.971	1.009
Entropy W.	τ_{12}	1.002	1.011	0.970	0.971
	τ_{13}	0.984	0.979	1.042	1.042
	τ_{23}	0.962	0.979	0.977	0.978
Treated W.	τ_{12}	1.019	0.989	0.947	0.622
	τ_{13}	1.034	1.024	1.061	0.784
	τ_{23}	1.007	0.988	0.898	0.627
IPW	τ_{12}	1.041	1.013	0.965	0.962
	τ_{13}	1.005	0.993	1.057	1.057
	τ_{23}	0.993	0.986	0.970	0.969
IPW (at 0.1 trim)	τ_{12}	0.990	1.067	0.974	0.998
	τ_{13}	0.971	0.987	1.024	1.037
	τ_{23}	0.935	0.971	0.958	0.981
GPSM	τ_{12}	0.920	0.985	0.910	0.794
	τ_{13}	0.659	0.683	0.668	0.576
	τ_{23}	0.859	0.863	0.841	0.656
GPSM (at 0.1 trim)	τ_{12}	0.412	0.424	0.410	0.573
	τ_{13}	0.364	0.375	0.371	0.519
	τ_{23}	0.417	0.422	0.401	0.472

Table 4.8: The proportion of average estimated standard error to empirical standard deviation of sampling variability of estimated across 1000 iterations of strong lack of overlap scenarios

Methods	**Pairwise of groups**	**Strong Lack of Overlap**			
		True propensity score Model		**Misspecified PS Model**	
		Bootstrap	*Robust*	*Bootstrap*	*Robust*
Matching W.	τ_{12}	1.057	1.060	1.052	0.887
	τ_{13}	1.056	1.066	1.067	0.951
	τ_{23}	1.043	1.046	1.097	0.972
Overlap W.	τ_{12}	1.085	1.088	1.031	1.038
	τ_{13}	1.061	1.063	1.064	1.068
	τ_{23}	1.026	1.027	1.075	1.089
Entropy W.	τ_{12}	1.071	1.068	1.022	1.018
	τ_{13}	0.910	0.905	0.993	0.988
	τ_{23}	0.897	0.896	1.015	1.016
Treated W.	τ_{12}	0.705	0.700	0.902	0.355
	τ_{13}	0.849	0.847	1.004	0.873
	τ_{23}	0.662	0.658	0.847	0.485
IPW	τ_{12}	0.728	0.714	0.972	0.967
	τ_{13}	0.909	0.897	0.968	0.962
	τ_{23}	0.785	0.781	0.984	0.968
IPW (at 0.1 trim)	τ_{12}	0.949	2.233	0.981	1.250
	τ_{13}	1.008	1.728	1.012	1.233
	τ_{23}	0.892	1.788	0.973	1.275
GPSM	τ_{12}	0.873	0.880	0.741	0.790
	τ_{13}	0.447	0.470	0.528	0.394
	τ_{23}	0.492	0.501	0.612	0.467
GPSM (at 0.1 trim)	τ_{12}	0.106	0.152	0.124	0.094
	τ_{13}	0.161	0.170	0.224	0.132
	τ_{23}	0.184	0.189	0.193	0.147

Table 4.9 : Average of SMD(%) across 1000 simulate dataset of good overlap scenarios

	Unweighted	Matching W.	Overlap W.	*Entropy W.*	Treated W	IPW	GPSM
X_1	5.716	0.451	0.388	0.389	0.640	0.392	0.944
X_2	5.552	0.461	0.356	0.356	0.580	0.358	0.912
X_3	5.632	0.479	0.369	0.377	0.640	0.377	0.997
X_4	5.968	0.442	0.352	0.352	0.560	0.353	0.995
X_5	5.955	0.531	0.405	0.412	0.732	0.423	1.002
X_6	5.883	0.411	0.326	0.328	0.546	0.331	0.957
X_7	5.423	0.421	0.318	0.317	0.522	0.318	0.901
X_8	5.600	0.439	0.356	0.355	0.596	0.357	0.920
X_9	5.664	0.398	0.325	0.328	0.560	0.332	0.939
X_{10}	5.514	0.409	0.320	0.323	0.545	0.327	0.924

Table 4.10 : Average of SMD (%) across 1000 simulate dataset of moderate lack of overlap scenarios

	Unweighted	Matching W.	Overlap W.	*Entropy W.*	Treated W	IPW	GPSM
X_1	27.475	2.364	2.75	2.989	4.668	3.063	10.091
X_2	17.744	2.114	2.87	3.169	4.100	3.246	8.042
X_3	21.136	2.177	2.984	3.297	4.342	3.360	11.672
X_4	32.003	1.121	0.967	1.032	2.314	1.137	15.834
X_5	25.975	1.415	1.157	1.276	4.741	1.929	9.971
X_6	9.785	1.202	1.017	1.067	2.491	1.169	5.098
X_7	9.709	1.269	0.994	1.026	2.222	1.119	7.086
X_8	21.641	1.143	0.944	0.990	2.369	1.087	12.023
X_9	5.486	1.234	1.053	1.106	2.491	1.237	8.567
X_{10}	5.742	1.230	1.027	1.063	2.409	1.152	5.395

Table 4.11: Average of SMD% across 1000 simulate dataset of strong lack of overlap scenarios

	Unweighted	Matching W.	Overlap W.	*Entropy W.*	Treated W	IPW	GPSM
$\boldsymbol{X_1}$	71.702	10.459	12.465	21.108	25.308	22.073	37.540
$\boldsymbol{X_2}$	33.992	6.321	10.105	20.479	18.690	21.201	27.497
$\boldsymbol{X_3}$	47.279	4.740	9.708	22.070	19.560	22.337	31.289
$\boldsymbol{X_4}$	95.279	3.156	2.798	5.318	13.688	7.876	21.509
$\boldsymbol{X_5}$	56.329	3.328	2.846	4.377	23.829	10.539	20.831
$\boldsymbol{X_6}$	25.058	3.245	2.893	4.697	12.109	6.691	17.845
$\boldsymbol{X_7}$	23.252	3.329	2.916	4.908	11.102	6.620	16.456
$\boldsymbol{X_8}$	59.970	3.161	2.703	4.959	12.618	7.078	19.673
$\boldsymbol{X_9}$	5.744	3.277	2.980	4.644	11.153	6.190	9.587
$\boldsymbol{X_{10}}$	5.782	3.428	3.010	4.650	12.676	6.804	12.456

Figure 4.2: Distribution of bootstrap standard error for all balance weightings across 1000 simulated for good overlap scenario

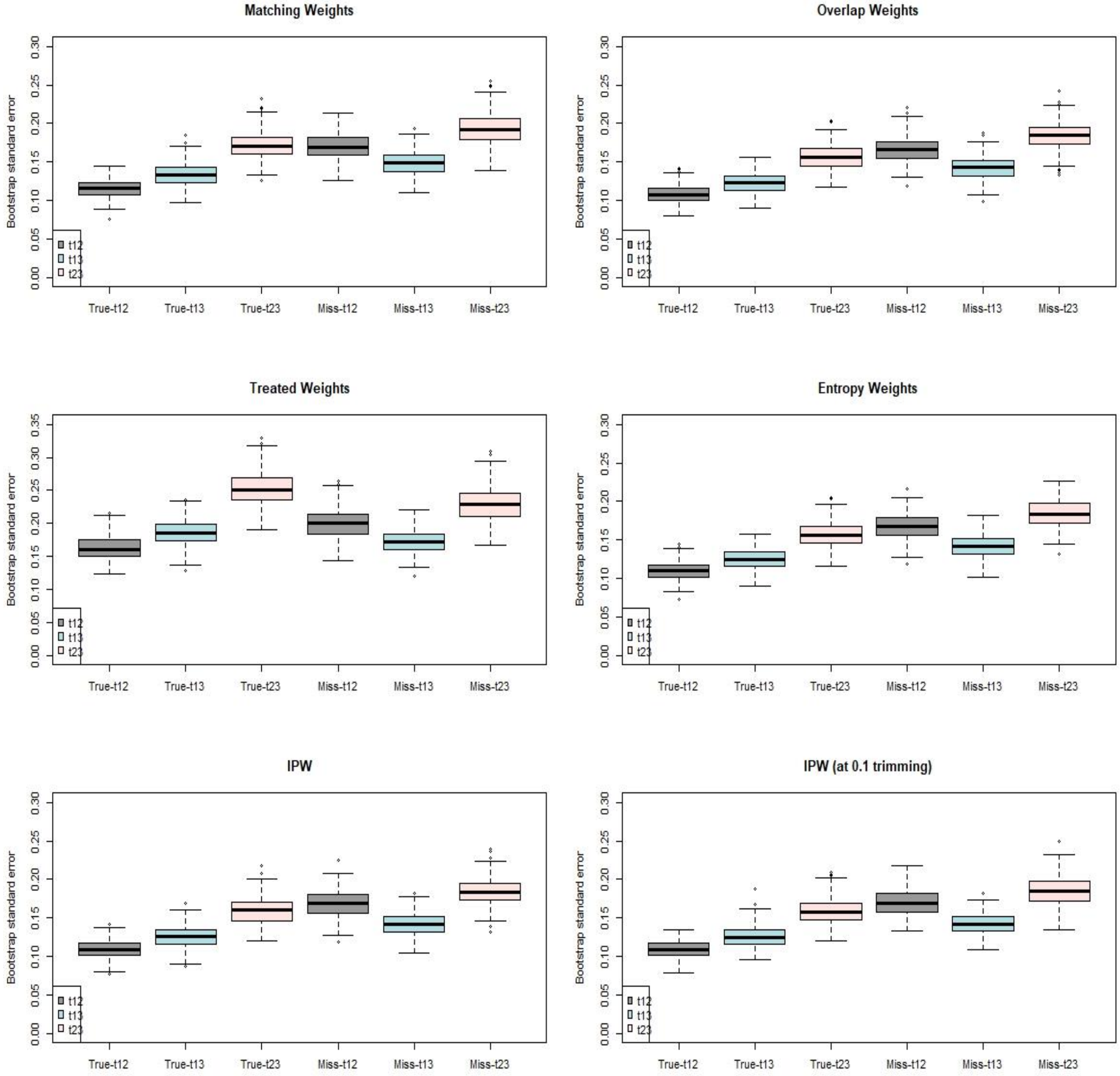


Figure 4.3: Distribution of bootstrap standard error for all methods across 1000 simulated for strong lack of overlap scenario

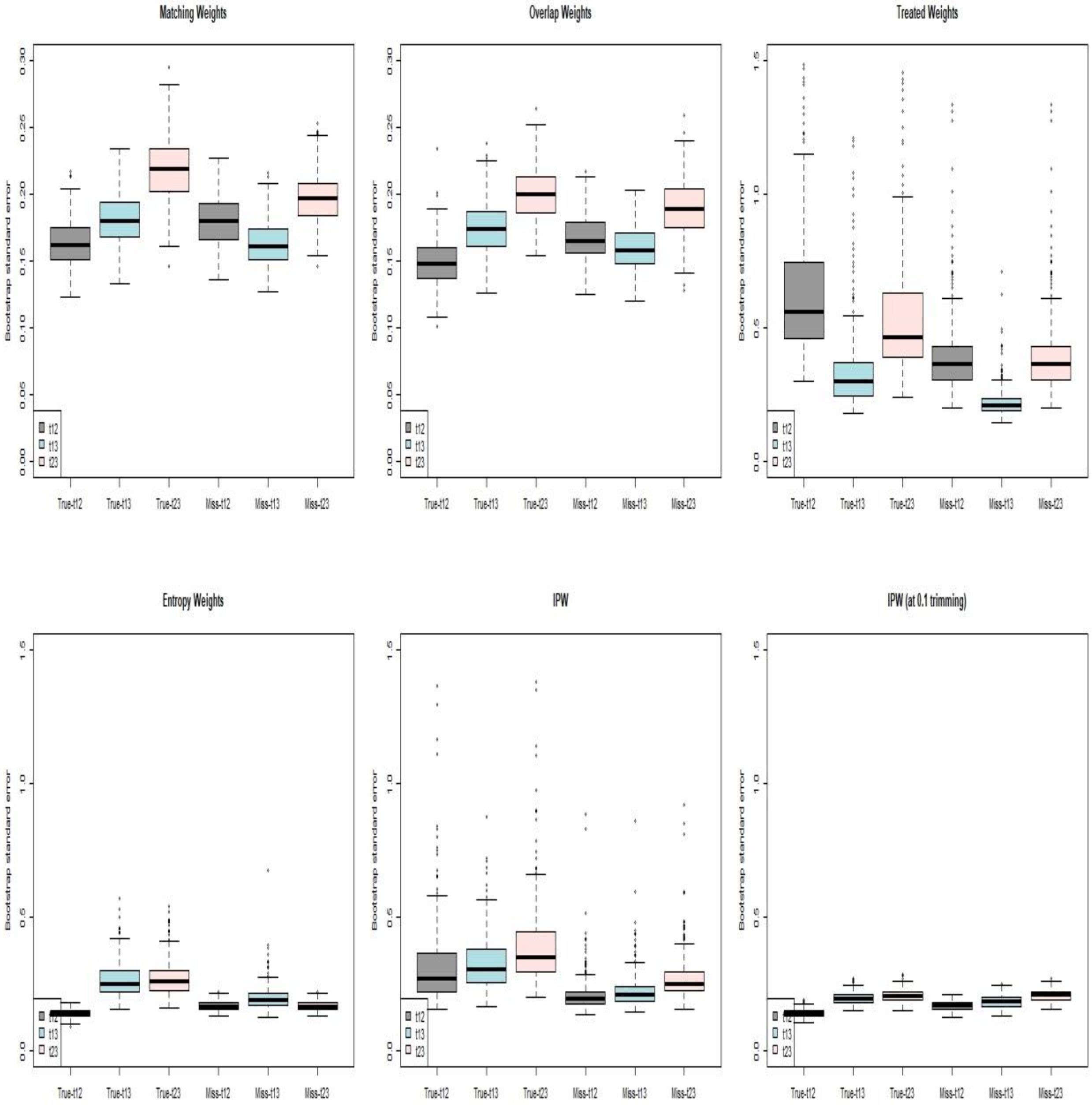


Figure 4.4: Distribution of robust sandwich standard error for all balance weighting methods across 1000 simulated for good overlap scenario

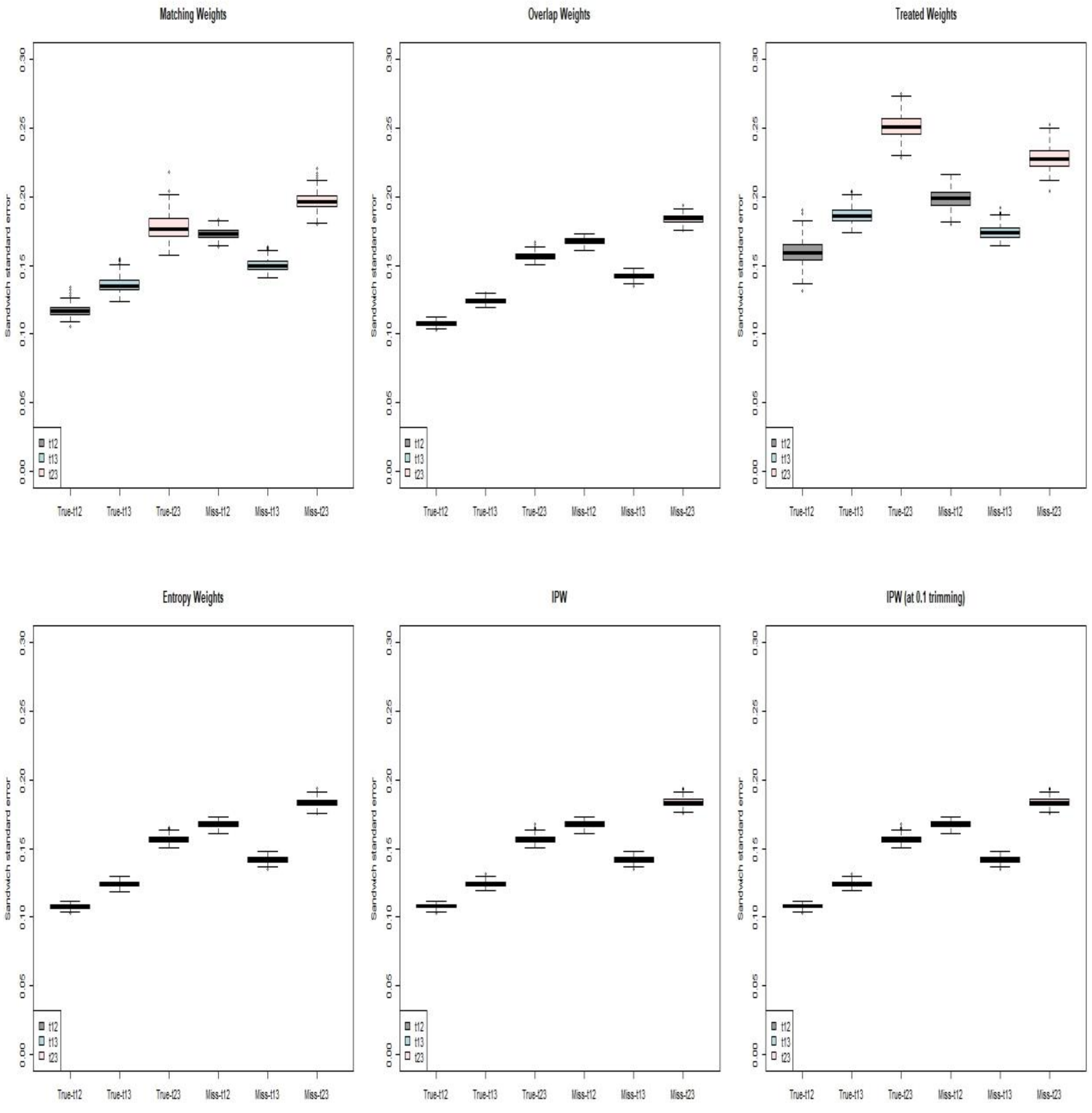
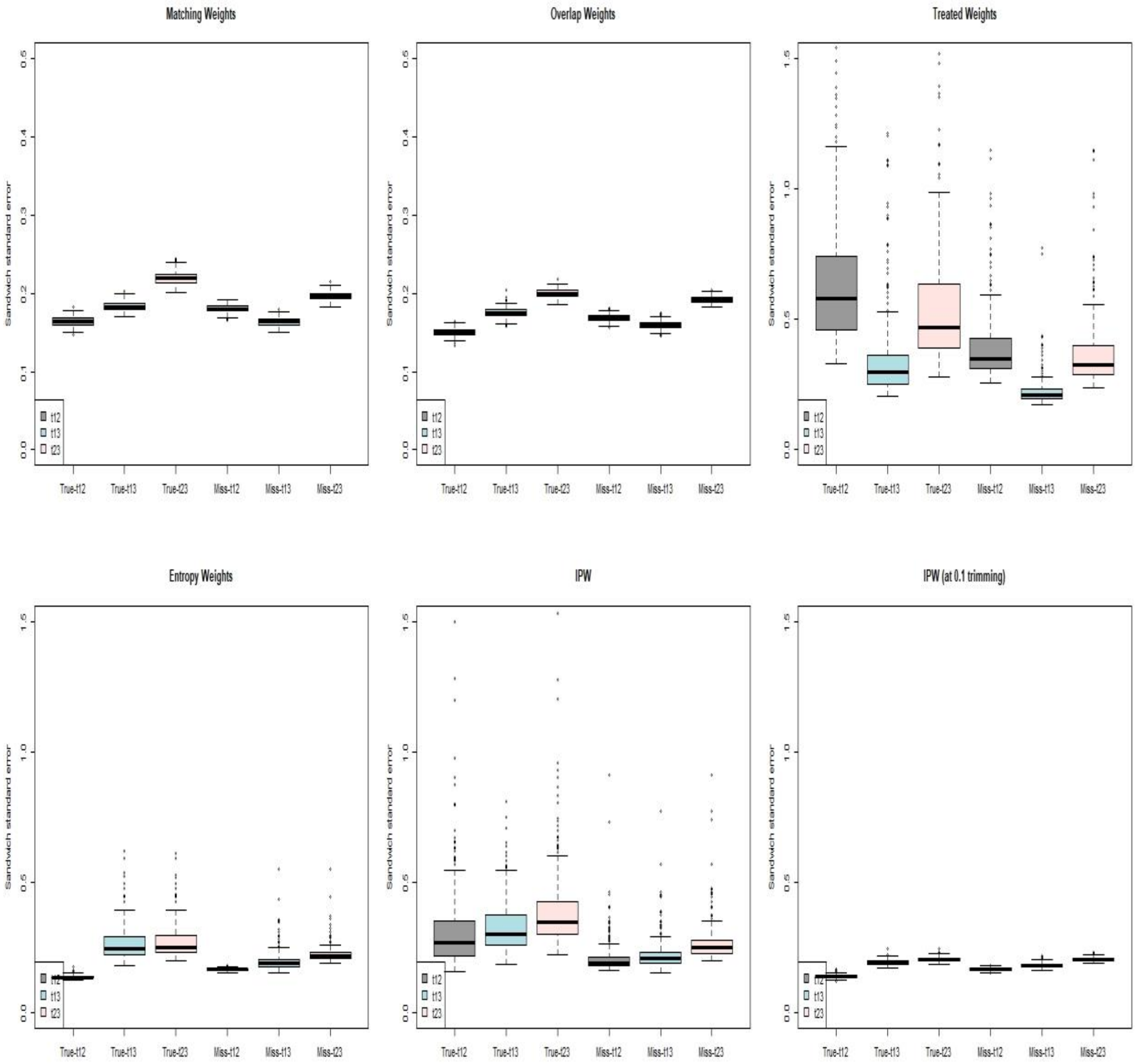


Figure 4.5: Distribution of robust sandwich standard error for all balance weighting across 1000 simulated in strong lack of overlap scenario



Chapter 5

Summary& Conclusion

This dissertation has offered three contributions to the literature. In these three papers, we presented various topics through three application areas of biostatistics: biomarkers, causal inference, and machine learning. Thus, this research expanded the understanding of these areas. In this chapter, we briefly summarize our findings and contributions in this dissertation.

The first paper's main goal has to examine techniques for adjusting for an observational dataset's treatment selection process. Identifying biomarkers that may be used to predict the potential benefits of a specific treatment for patients is a significant challenge in developing personalized or precision medicine techniques. Despite the extensive literature on improving biomarkers in randomized control trials, there is relatively limited research on developing a statistical methodology to assess the markers using observational datasets. Firstly, we derive $\Theta_1 \; and \; \Theta_0$ parameters metrics, which measure the performance of treatment selection for biomarkers' based on survival outcomes. Then, we performed causal inference techniques based on the proposed theta metric to examine how well the results of causal inference techniques impact the performance of biomarkers. We have concluded that observational studies without using causal inference techniques to evaluate biomarkers' effect may yield inaccurate results.

The true propensity score is not known in observational datasets. Rubin (1983) called that propensity score a balancing score. Even though logistic regression is the standard technique to estimate PS, there is recently growing interest in utilizing machine learning techniques. We implement machine learning techniques and parametric methods to estimate propensity scores in Chapter 3. However, implementing these approaches in practice raises several critical concerns that are currently being debated in the literature: i) how to estimate propensity score, ii) which variables are included/excluded from PS models iii) which methods (i.e., machine learning or parametric models) should be preferred, iv) how to estimate outcomes. The first purpose of Chapter 3 was to assess the bias of estimates derived from PS matching relying on the model utilized to estimate PS. Secondly, we used Monte Carlo simulations to illustrate how different combinations of covariates influenced the potential of matching on propensity score to construct subjects in which all measured baseline covariates were balanced between treatment groups, to address the lack of consensus on which variables to include the propensity score model. The best performing approaches for estimating the propensity score in our simulations were logistic regression, random forests, and CART. In comparison to the other techniques discussed, they frequently guaranteed significantly reduced bias and RMSE. The variations between the best-performing methods are generally rather minor. After that, we offered the results of parametric and machine learning methods to evaluate treatment selection biomarkers used to select a specific treatment in observational studies.

In chapter 4, the motivation was the increasing research on improving multiple (more than two) treatment effect estimators in observational studies. Inverse probability weighting (IPW), among the most preferred methods in the propensity scores literature, were used to minimize confounding effects and examine causal effects. One of the main assumptions is the positivity assumption, i.e., propensity score must be bounded away from 0 and 1. If the positivity assumption is violated, we can get inaccurate results, i.e., large bias, variance, RMSE, or imbalance in covariates between treatment groups. So, we study matching weighting, overlap weighting, entropy weights, treated weighting, inverse probability weighting with symmetric trimming, generalized propensity score matching, generalized propensity score matching with symmetric trimming that are alternatives to inverse probability weighting. We propose three different levels of overlap to investigate how positivity assumption is violated using true PS model and misspecified PS model. When good overlap exists, MW, OW, EW, and IPW performed similarly in terms of bias reduction. When the PS model is misspecified, MW, OW, and EW were not sensitive against the misspecification of the model. However, GPSM was more likely to be affected by a PS model misspecification when good overlap exists. In addition, a strong lack of overlap led to bias, large RMSE, and average SE across all methods. In addition, we discuss accurate standard error estimation using weighting methods and GPSM. We saw that utilizing bootstrap and sandwich robust estimator performed well in terms of an estimated standard error in MW, OW, and EW.

Bibliography

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in medicine, 28(25), 3083-3107.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical Statistics, 10, 150–161

Austin, P. C., & Schuster, T. (2016). The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: A simulation study. Statistical Methods in Medical Research, 25, 2214–2237.

Austin, P. C., & Stuart, E. (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, 3661–3679.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Statistics in medicine, 26(4), 734-753.

Austin, P. C., Grootendorst, P., Normand, S. L. T., & Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in medicine*, *26*(4), 754-768.

Bonetti, M., & Gelber, R. D. (2004). Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics, 5(3), 465-481.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24, 123–140.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L.,&Cutler, A. (2016). Random forests for scientific discovery. línea]. Available: https://www. stat. berkeley. edu/breiman/RandomForests/berkeley_files/frame. htm.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. American journal of epidemiology, 163(12), 1149-1156.

Cai, T., Tian, L., Wong, P. H., & Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics, 12(2), 270-282.

Cai, T., Tian, L., Wong, P. H., & Wei, L. J. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics, 12(2), 270-282.

CAMPBELL,D. and STANLEY, J. (1963). Experimental and quasi-experimental designs for research and teaching. In Handbook of Research on Teaching (N. L. Gage, ed.). Rand McNally, Chicago.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010), "BART: Bayesian Additive Regression Trees," The Annals of Applied Statistics, 4, 266–298. [1229]

Claggett, B., Zhao, L., Tian, L., Castagno, D., & Wei, L. J. (2011). Estimating subject-specific treatment differences for risk-benefit assessment with competing risk event-time data.

COCHRAN, W. (1965). The planning of observational studies of human populations. J. Roy. Statist. Soc. A 128 234–265

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. Sankhyā: The Indian Journal of Statistics, Series A, 417-446.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. Sankhya-A, 35, 417–446.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. Biometrika, 96(1), 187-199.

D'Agostino Jr, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Statistics in medicine, 17(19), 2265-2281.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American statistical Association, 94(448), 1053-1062.

Dobbin, K. K., Cesano, A., Alvarez, J., Hawtin, R., Janetzki, S., Kirsch, I., ... & Thurin, M. (2016). Validation of biomarkers to predict response to immunotherapy in cancer: volume II—clinical validation and regulatory considerations. Journal for immunotherapy of cancer, 4(1), 1-14.

Fang, K., & Huang, H. (2011). Variable Selection for Credit Risk Model Using Data Mining technique. J. Comput., 6(9), 1868-1874.

FISHER, R. (1935). Design of Experiments. Oliver and Boyd, Edinburgh.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political analysis, 20(1), 25-46.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association, 99(467), 609-618.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services and Outcomes research methodology, 2(3), 259-278.

HOLLAND, P. (1986). Statistics and causal inference. J. Amer. Statist. Assoc. 81 945–960. MR0867618

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. J. Am. Statist. Ass., 47, 663–685

Huang, Y., & Pepe, M. S. (2010). Assessing risk prediction models in case–control studies using semiparametric and nonparametric methods. Statistics in medicine, 29(13), 1391-1410.

Huang, Y., & Pepe, M. S. (2010). Semiparametric methods for evaluating the covariate-specific predictiveness of continuous markers in matched case–control studies. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(3), 437-456.

Huang, Y., Gilbert, P. B., & Janes, H. (2012). Assessing treatment-selection markers using a potential outcomes framework. Biometrics, 68(3), 687-696.

Huang, Y., Sullivan Pepe, M., & Feng, Z. (2007). Evaluating the predictiveness of a continuous marker. Biometrics, 63(4), 1181-1188.

Huppler-Hullsiek, K., & Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. Biostatistics, 2, 1–15

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

Imbens, G. and Rubin, D. (2015). An Introduction to Causal Inference in the Statistical, Biomedical and Social Sciences. Cambridge: Cambridge University Press.

Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

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

Janes, H., & Pepe, M. S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. Biometrika, 96(2), 371-382.

Janes, H., Brown, M. D., & Pepe, M. S. (2015). Designing a study to evaluate the benefit of a biomarker for selecting patient treatment. Statistics in medicine, 34(27), 3503-3515.

Janes, H., Brown, M. D., Huang, Y., & Pepe, M. S. (2014). An approach to evaluating and comparing biomarkers for patient treatment selection. The international journal of biostatistics, 10(1), 99-121.

Janes, H., Brown, M. D., Huang, Y., & Pepe, M. S. (2014). An approach to evaluating and comparing biomarkers for patient treatment selection. The international journal of biostatistics, 10(1), 99-121.

Janes, H., Brown, M. D., Huang, Y., & Pepe, M. S. (2014). An approach to evaluating and comparing biomarkers for patient treatment selection. The international journal of biostatistics, 10(1), 99-121.

Janes, H., Brown, M. D., Huang, Y., & Pepe, M. S. (2014). An approach to evaluating and comparing biomarkers for patient treatment selection. The international journal of biostatistics, 10(1), 99-121.

Janes, H., Pepe, M. S., Bossuyt, P. M., & Barlow, W. E. (2011). Measuring the performance of markers for guiding treatment decisions. Annals of internal medicine, 154(4), 253-259.

Janes, H., Pepe, M. S., McShane, L. M., Sargent, D. J., & Heagerty, P. J. (2015). The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. Journal of the National Cancer Institute, 107(8), djv157.

Kang, C., Janes, H., & Huang, Y. (2014). Combining biomarkers to optimize patient treatment recommendations. Biometrics, 70(3), 695-707.

KEMPTHORNE, O. (1952). The Design and Analysis of Experiments. Wiley, New York.MR0045368

Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am JEpidemiol. 2006; 163(3):262–270

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in medicine, 29(3), 337-346.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in medicine, 29(3), 337-346.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in medicine, 23(19), 2937-2960.

Mandrekar, S. J., Grothey, A., Goetz, M. P., & Sargent, D. J. (2005). Clinical trial designs for prospective validation of biomarkers. American Journal of Pharmacogenomics, 5(5), 317-325.

Mandrekar, S.J., Sargent, D.J. (2009). Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. Journal of Clinical Oncology 27:4027–4034

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods, 9, 403–425

McConnell, K. J., & Lindner, S. (2019). Estimating treatment effects with machine learning. Health services research, 54(6), 1273-1282.

Myers JA, Rassen JA, Gagne JJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol. 2011;174(11):1213–1222.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9. Translated in Statist. Sci. 5 465–480. MR1092986

Normand, S. L. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. Journal of clinical epidemiology, 54(4), 387-398.

Pearl, J. (2015). Causal thinking in the twilight zone. Introduction to Observational Studies and the Reprint of Cochran’s paper “Observational Studies” and Comments, 200

Ridgeway, G. (1999). The state of boosting. Computing science and statistics, 172-181.

Rosenbaum, P. R., & Rubin, D. B. (1986). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 40, 249–251.

ROSENBAUM,P. and RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70 41–55. MR0742974

RUBIN, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66 688–701.

RUBIN, D. (1975). Bayesian inference for causality: The importance of randomization. In The Proceedings of the Social Statistics Section of the American Statistical Association 233–239. American Statistical Association, Alexandria, VA.

RUBIN, D. (1976a). Inference and missing data. Biometrika 581–592. With discussion and reply. MR0455196

RUBIN, D. (1976b). Multivariate matching methods that are equal percent bias reducing, II: Maximums on bias reduction for fixed sample sizes. Biometrics 32 121–132. MR0400556

RUBIN, D. (1977). Assignment to treatment group on the basis of a covariate. J. Educ. Statist.

RUBIN, D. (1978). Bayesian inference for causal effects: The role of randomization. Ann. Statist. 6 34–58. MR047215

RUBIN, D. (1979a). Discussion of "Conditional independence in statistical theory" by A.P. Dawid. J. Roy. Statist. Soc. Ser. B 41 27–28.

RUBIN, D. (1979b). Using multivariate matched sampling and regression adjustment to control bias in observational studies. J. Amer. Statist. Assoc. 74 318–328.

RUBIN, D. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test" by Basu. J. Amer. Statist. Assoc. 75 591–593

RUBIN, D. (1984). William G. Cochran's contributions to the design, analysis, and evaluation of observational studies. In W. G. Cochran's Impact on Statistics 37–69. Wiley, New York. MR0758447

RUBIN, D. (1990a). Neyman (1923) and causal inference in experiments and observational studies. Statist. Sci. 5 472–480. MR1092987

RUBIN, D. (1990b). Formal modes of statistical inference for causal effects. J. Statist. Plann. Inference 25 279–292.

RUBIN, D. (1997). Estimating causal effects from large data sets using propensity scores. Ann. Internal Med. 127 757–763.

RUBIN, D. (2002). Using propensity scores to help design observational studies: Application to the tobacco litigation. Health Serv. and Outcomes Res. Methodol. 2 169–188

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. The Annals of Applied Statistics, 2(3), 808-840.

Sargent, D. J., Conley, B. A., Allegra, C., & Collette, L. (2005). Clinical trial designs for predictive marker validation in cancer treatment trials. Journal of Clinical Oncology, 23(9), 2020-2027.

Schapire, R. E. (1999, July). A brief introduction to boosting. In Ijcai (Vol. 99, pp. 1401-1406).

Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M. S., Yeatman, T. J., Gerald, W. L.,& Sharma, A. (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study: Director’s Challenge Consortium for the molecular classification of lung adenocarcinoma. Nature medicine, 14(8), 822.

Shen, C., Li, X., Li, L., & Were, M. C. (2011). Sensitivity analysis for causal inference using inverse probability weighting. Biometrical journal, 53(5), 822-837.

Simon, R. (2008). Development and validation of biomarker classifiers for treatment selection. Journal of statistical planning and inference, 138(2), 308-320.

Simon, R. (2010). Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. Personalized medicine, 7(1), 33-47.

Simon, R. M., Paik, S., & Hayes, D. F. (2009). Use of archived specimens in evaluation of prognostic and predictive biomarkers. Journal of the National Cancer Institute, 101(21), 1446-1452.

Song,X.,&Pepe,M.S.(2004). Evaluating markers for selecting a patient's treatment. Biometrics, 60(4), 874-883.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1), 1.

Stürmer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. American journal of epidemiology, 172(7), 843-854.

Vickers, A. J., Kattan, M. W., & Sargent, D. J. (2007). Method for evaluating prediction models that apply the results of randomized trials to individual patients. Trials, 8(1), 1-11.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. Journal of econometrics, 141(2), 1281-1301.

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Jonsson Funk, M., LoCasale, R., & Stürmer, T. (2014). The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. American journal of epidemiology, 180(6), 645-655.

Xie, J., & Liu, C. (2005). Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Statistics in medicine, 24(20), 3089-3110.

Zhao, L., Tian, L., Cai, T., Claggett, B., & Wei, L. J. (2013). Effectively selecting a target population for a future comparative study. Journal of the American Statistical Association, 108(502), 527-539.