ABSTRACT

Title of dissertation:	PERFORMANCE OF PROPENSITY SCORE METHODS IN THE PRESENCE OF HETEROGENEOUS TREATMENT EFFECTS
	Kathleen M. Stepien, Doctor of Philosophy, 2016
Dissertation directed by:	Laura M. Stapleton, Associate Professor
	Department of Human Development and Quantitative Methodology; Measurement, Statistics and Evaluation Program

Estimating an average treatment effect assumes that individuals and groups are homogeneous in their responses to a treatment or intervention. However, treatment effects are often heterogeneous. Selecting the most effective treatment, generalizing causal effect estimates to a population, and identifying subgroups for which a treatment is effective or harmful are factors that motivate the study of heterogeneous treatment effects. In observational studies, treatment effects are often estimated using propensity score methods. This dissertation adds to the literature on the analysis of heterogeneous treatment effects using propensity score methods. Three propensity score methods were compared using Monte Carlo simulation: single propensity score with exact matching on subgroup, matching using group propensity scores, and multinomial propensity scores using generalized boosted modeling. Methods were evaluated under various group distributions, sample sizes, effect sizes, and selection models. An empirical analysis

using data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) is included to demonstrate the methods studied. Simulation results showed that estimating group propensity scores provided the smallest MSE, MNPS performance was comparable to GBM, and including the group indicator in the propensity score model improved treatment effect estimates regardless of whether group membership influenced selection. In addition, subclassification performed poorly when one group was more prevalent in the extremes of the propensity score distribution.

PERFORMANCE OF PROPENSITY SCORE METHODS IN THE PRESENCE OF HETEROGENEOUS TREATMENT EFFECTS

by

Kathleen M. Stepien

Dissertation proposal submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2016

Advisory Committee:

Professor Laura M. Stapleton, Chair
Professor Gregory R. Hancock
Professor Jeffrey Harring
Professor Frauke Kreuter
Professor Tracy Sweet

DEDICATION

To my spouse who encouraged me,
made me laugh when I needed it most,
and patiently endured 8 years of distraction.

To my advisor who unselfishly shared
knowledge, time, and advice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

Journals in various fields such as social and behavioral science, policy intervention, medical research, and criminology are filled with empirical studies that investigate the effect of a treatment, policy, or exposure. In addition to determining whether a treatment or intervention works on average, the goal is often to investigate effects at individual or group levels (Green & Stuart, 2014; Rothwell, 2005). This is because whether a treatment or intervention works on average is only sufficient if it is reasonable to assume that effects are truly homogeneous in the population. However, individuals and groups often have different responses to a treatment or intervention. It is these differences or heterogeneous treatment effects that are of most interest to researchers (Pocock, Assmann, Enos, & Kasten, 2002).

Selecting the most effective treatment, generalizing causal effect estimates to a population, and identifying subgroups for which a treatment is effective or harmful are aspects that motivate the interest in heterogeneous treatment effects (Imai & Ratkovic, 2013). Also, Abrahamowicz, Beauchamp, Fournier, and Dumont (2013) highlighted the importance of investigating heterogeneous treatment effects to avoid missing important treatment effects that may be unique to a particular subgroup. Their study showed that a non-significant average treatment effect does not imply non-significant subgroup effects.

Additional examples that motivate the study of heterogeneous treatment effects can be found in the literature on Patient-Centered Outcomes Research (PCOR) and Comparative Effectiveness Research (CER) as well as in non-medical disciplines such as econometrics (Hayward, Kent, Vijan, & Hofer, 2005; Heckman, Urzua, & Vytlacil, 2006; Luo, 2011; Varadhan, Segal, Boyd, Wu, & Weiss, 2013; Willke, Zheng, Subedi, Althin, & Mullins, 2012).

As Imai and Ratkovic (2013) highlight, inferring cause and not merely identifying association is one of the primary aims of the study of heterogeneous treatment effects. To infer cause, a researcher must be able to claim that the treatment is the only explanation for the effect. C*eteris paribus,* or holding all factors save one constant, is the ideal condition to isolate the causal effects of a specific treatment or intervention (Holland, 1986). Holding all factors save one constant improves the validity of causal claims because it removes the potential for covariates to confound the estimates of treatment effects. This ideal condition happens occasionally in the scientific community and almost never in the behavioral sciences primarily because of the increased potential for unobserved covariates to confound the estimates of treatment effects. As a result, researchers often resort to experimental studies or randomized controlled trials (RCTs) because randomization of units into treatment and control groups renders the groups equal in expectation on all observed and unobserved characteristics. In other words, randomization essentially creates groups with no systematic differences therefore getting as close as possible to the ideal condition needed to claim causation. In instances where RCTs may be unethical or cost prohibitive observational data may be analyzed using propensity score methods to create groups that are equal in expectation on all observed characteristics. Causation is inferred by including an additional assumption of no unobserved characteristics that could confound treatment effect estimates (Guo & Fraser, 2010).

The popularity of propensity score methods has increased exponentially in the last 15 years (Shadish, 2013; Westreich, Lessler, & Funk, 2010). The number of Google scholar hits for a search for “propensity score” was relatively flat and near zero prior to

1999. From 1999 to 2011, the number of hits increased steadily at a rate of 500 hits per year to slightly more than 6,000 in 2011. Figure 1 shows a similar trend for propensity score publications and citations. Propensity score methods are also used in numerous fields of study. Sekhon (2011) provides examples of the use of propensity score methods in statistics, medicine, economics, political science, sociology, and law.

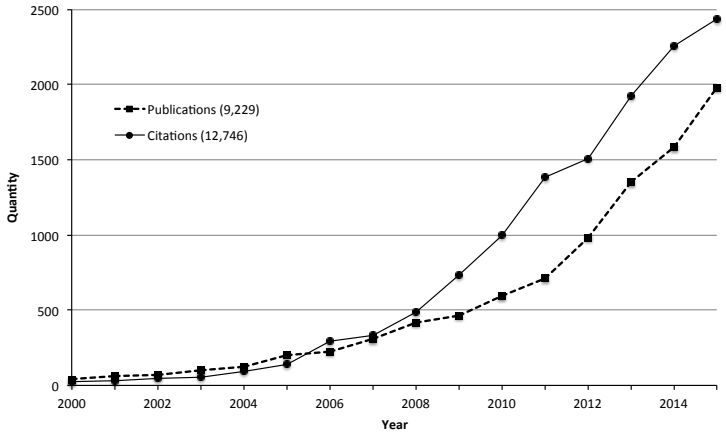


Figure 1. The number of publications and citations for the keyword "Propensity Score". Retrieved from http://scicurve.com/trends/Propensity%20Score.

Propensity score methods are most often used in observational studies to estimate treatment effects. Numerous studies highlight the mechanics of implementation and the advantages of propensity score methods (e.g., Austin, 2009c; Austin, 2009d; Austin, 2011a; Caliendo & Kopeninig, 2008; D'Agostino, 1998; Harder, Stuart, & Anthony, 2010; Ho, Imai, King, & Stuart, 2007; Rhodes, 2010; Shadish, 2013; Stuart & Rubin, 2008; Thoemmes & Kim, 2011). Although many of these studies address propensity score methods in general, few provide a systematic analysis of implementing propensity score methods in the presence of heterogeneous treatment effects.

The study described in this dissertation adds to the literature on the analysis of heterogeneous treatment effects in observational studies using propensity score methods.

Specifically, three propensity score methods were compared under various group distributions, sample sizes, effect sizes, and selection models. The dissertation starts with an overview of the potential outcomes framework including assumptions for unbiased treatment effects, followed by a description of treatment effect estimation in experimental studies versus observational studies. A more detailed review of propensity score estimation and methods that use propensity scores is then presented including descriptions of relevant studies. Chapter 3 provides the research design and the proposed Monte Carlo simulations. The dissertation concludes with a summary of the results in Chapter 4 and a discussion of the implications and value of the key findings and recommendations in Chapter 5.

Chapter 2: Literature Review

2.1 Framework

One of the main objectives in program evaluation or epidemiological research is to identify causal relations between an exposure or treatment, and outcomes. The ultimate goal is to be able to estimate or predict the effect of treatment for individuals or groups (Imbens & Wooldridge, 2009). The treatment effect for an individual is simply the difference in the outcome for the individual exposed to treatment, Y_{1i}, and the outcome if the individual had not been exposed to treatment, Y_{0i}. The individual treatment effect (ITE) is shown in Equation 1.

$$ITE = Y_{1i} - Y_{0i} \quad (1)$$

The challenge in program evaluation and epidemiological studies when attempting to define causal effects is the inability to observe an individual in both the control state and the treatment state. This issue is often referred to as the "fundamental problem of causality" (Holland, 1986). In statistics, the fundamental problem of causality is formalized using a model attributed to Neyman (1923) and Rubin (1974). Their model, the Neyman-Rubin Causal Model, provides a counterfactual or potential outcomes framework for data analysis and assumes that each individual, i, in the population has two potential outcomes, Y_{ti}, for the control state, $t = 0$, and for the treatment state, $t = 1$. However, each individual is observed in only one state at any moment in time. The unobserved, hypothetical, or potential outcomes are counterfactual. Table 1 (Morgan & Winship, 2007) demonstrates this concept. For example, individuals who are members of the treatment group, $T_i = 1$, have observed outcomes, Y_{1i}. Although members of the treatment group have the potential outcomes of no treatment, Y_{0i}, these outcomes are not

observed. They are considered hypothetical or counterfactual. The opposite is true for members of the control group, $T_i = 0$. For example, a researcher interested in analyzing the effects of a private school education on math outcomes might compare test scores for students that attend private school with test scores for students that attend public school. Test scores for students that attend private school are considered observed while test scores for the same students had they attended public school are considered counterfactual. The observed and counterfactual indicators would be reversed for students that attend public school.

Table 1
The Fundamental Problem of Causation

Group	Y_{1i}	Y_{0i}
Treatment (T_i=1)	Observed	Counterfactual
Control (T_i=0)	Counterfactual	Observed

Equation 2 represents the Neyman-Rubin causal model where T_i indicates whether individual, i, is in the treatment group (T_i =1) or the control group (T_i =0).

$$Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i} \quad (2)$$

Equation 2 further highlights the fundamental problem of causality and shows that the observed outcome variable, Y_i, for individual i, is the observed value in either the treatment state or the control state but not both. Because both outcomes are not observed for each person, the ITE as well as the average treatment effect (ATE), $E[Y_{1i} - Y_{0i}]$, cannot be directly calculated and must be estimated. The unobserved counterfactuals, $E[Y_{1i}|T_i = 0]$ and $E[\,Y_{0i}|T_i = 1]$, are estimated using the observed values for each group, $E[Y_{1i}|T_i = 1]$ and $E[Y_{0i}|T_i = 0]$, respectively. The ATE is then calculated as the difference between the average outcome of the treatment group and the average outcome of the control group:

$$ATE = E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 0] \quad (3)$$

Other parameters of interest in program evaluation and epidemiological studies include the average treatment effect on the treated (ATT), represented by Equation 4,

$$ATT = E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 1] \quad (4)$$

and the average treatment effect on the untreated or control group (ATC), represented by Equation 5.

$$ATC = E[Y_{1i}|T_i = 0] - E[Y_{0i}|T_i = 0] \quad (5)$$

The ATE describes the treatment benefit for an individual randomly selected from the population and averages across units that might never receive treatment (Wooldridge, 2002). ATT describes the treatment benefit for those individuals who were actually treated whereas ATC describes the potential treatment benefit for those individuals who did not receive treatment. Geneletti and Dawid (2011) provide examples where one estimate would be preferred over the other. For example, ATE would be a more appropriate estimate in an epidemiological trial to determine the effect of a drug. ATT would be a more appropriate estimate of the benefit of a government program such as a math refresher course for adults with no higher education. The treatment effect of interest is the benefit for those adults who chose to participate in the program. ATC would be of interest if the objective were to estimate the impact that a policy or intervention would have had on those who were not treated. Of these three parameters of interest, ATE and ATT are likely more applicable in medical and policy intervention studies (Austin, 2012).

Morgan and Winship (2007) refer to Equation 3 as the naïve estimate of the ATE. Equation 3 estimates the ATE for the population; however, because the unobserved potential outcomes are estimated there are assumptions that must be considered. The

naïve estimate of the ATE requires three assumptions in order to be an unbiased representation of the true treatment effects. First, the potential outcome must be independent of the treatment assignment, $Y_{ti} \perp T_i$. In other words, $E[Y_{0i}|T_i = 1] = E[Y_{0i}|T_i = 0]$ and $E[Y_{1i}|T_i = 1] = E[Y_{1i}|T_i = 0]$. This first assumption is often referred to as the ignorable treatment assignment assumption or conditional independence assumption (Morgan & Winship, 2007; Rubin, 2004). In an experimental setting, the conditional independence assumption is met because units in both groups are randomly assigned and therefore equally likely to receive treatment. In an observational study, the researcher must identify the observed covariates, **X**, that are associated with treatment assignment. Rosenbaum and Rubin (1983) showed that the potential outcomes are independent of treatment assignment conditional on these observed covariates. Once observed covariates that are associated with treatment assignment are identified, Equation 3 in observational settings becomes:

$$ATE = E[Y_{1i}|\boldsymbol{X}_i, T_i = 1] - E[Y_{0i}|\boldsymbol{X}_i, T_i = 0] \quad (6)$$

The second assumption, the common support or overlap assumption, is an extension of the ignorable treatment assignment assumption and requires that each individual with treatment condition, T_i, and observed covariates, $\boldsymbol{X}_i$, has a nonzero probability of being assigned to the treatment group and the control group (i.e., $0 < P(T_i|\boldsymbol{X}_i) < 1$). In addition, the range of the covariates must be similar in both the treatment and control groups (i.e., overlap). In the literature, the assumptions of ignorability and common support are often referred to collectively as strong ignorability (Morgan & Winship, 2007; Rosenbaum & Rubin, 1983; Rubin, 2004).

The third assumption necessary for the naïve estimate of the ATE to be unbiased in the Neyman-Rubin Causal Model is the stable unit treatment value assumption (SUTVA). SUTVA requires that the treatment or intervention is the same for each individual and that no interference exists between individuals. It adds to the ignorable treatment assignment assumption by assuming no interaction between individuals and no variation in treatment (Morgan & Winship, 2007; Rubin, 2004). Peer effects are an example of a violation of SUTVA. A violation of SUTVA due to peer effects occurs if an individual's math score is higher when a close friend joins the math intervention program because they form a study group. Treatment that affects the balance of supply and demand enough to alter outcomes is another example of a violation of SUTVA. For example, a job-training program that produces more qualified candidates than hiring companies can use is an example of a violation of SUTVA. Saturation of the job market interferes with the true effect of the program. Qualified candidates will be unable to find jobs because of market saturation not because the job-training program was ineffective.

A nonparametric method of conditioning on the observed covariates that influence treatment assignment is to match units with similar values of the observed covariates. The main objective is to create groups that are similar on all relevant characteristics except treatment condition. Relevant characteristics or covariates are those that are related to treatment assignment or selection. This method may work when relatively few covariates are associated with treatment assignment; however, there are typically more than a few variables that need to be matched. As the number of covariates increases, matching becomes exponentially more difficult. For example, 10 binary covariates will have 2^{10} or 1,024 possible values to match (Guo & Fraser, 2010).

To overcome the problem of dimensionality, Rosenbaum and Rubin (1983) introduced the propensity score. The propensity score reduces multiple covariates to a single dimension or balancing score. The propensity score, $e(x) = \Pr(T_i \mid \boldsymbol{X}_i)$, is essentially the probability of treatment assignment conditional on this specific vector of covariates and could be substituted in Equation 6 for $\boldsymbol{X}_i$.

Matching methods include those that match on covariates, propensity scores, or a combination of both. Regardless of the algorithm or variables used to match, in addition to conditioning on the observed covariates, both covariate balance and overlap must be verified to ensure that no systematic differences remain between the groups and to verify that the common support assumption has been met. No test exists to verify the conditional independence assumption. Covariates that are identified as having significant differences between treatment and control should be included in the propensity score model. The researcher must use theory and prior studies to defend her claim that the covariates measured are the complete set of variables that might confound treatment assignment (Stuart & Rubin, 2008).

2.1.1 Treatment effect estimate bias. If covariates that are associated with treatment assignment and outcome are omitted from the propensity score model, the strong ignorability assumption is violated and the naïve estimate of the treatment effect (as obtained with Equation 3) will be biased. There are two potential sources of bias in the estimation of ATE when covariates that are associated with treatment assignment are omitted from the model: baseline bias and sorting bias. Pre-treatment or baseline bias consists of attributes, like gender or intelligence, which could explain differences in the average outcome between the treatment and control group prior to treatment. For

example, courses completed prior to a student participating in a math intervention program could influence test scores. Therefore, conditioning on a measure of previous math education would isolate the true treatment effect from potential confounding from prior courses completed.

Sorting bias or differential treatment effect bias consists of situations where those who are more likely to "sort" into treatment benefit more (or less) from the treatment than the average person (Brand & Xie, 2010; Brooks & Fang, 2009; Winship & Morgan, 1999; Xie, Brand, & Jann, 2012). For example, college bound students may recognize the potential benefit of participating in a math intervention program which means they might be more likely to participate. Therefore, conditioning on a measure of interest in college would isolate the true treatment effect from potential confounding from any unmeasured differences such as motivation and focus on success that could be inherent in those interested in college versus those that are not interested in college.

Both types of bias, collectively referred to as selection bias, can be seen if the counterfactuals are not "missing" and can therefore be included in the equation to estimate the ATE as shown in Equation 7 (Morgan & Winship, 2007; Xie, et al., 2012). Equation 7 represents the ATE with all counterfactuals where p is the proportion of individuals in the treatment group and q is the proportion of individuals in the control group:

$$\{E[Y_{1i}|T_i = 1]p + E[Y_{1i}|T_i = 0]q\} - \{E[Y_{0i}|T_i = 1]p + E[Y_{0i}|T_i = 0]q\} \quad (7)$$

Substituting (1-q) for p in Equation 7:

$$\{E[Y_{1i}|T_i = 1] - E[Y_{1i}|T_i = 1]q + E[Y_{1i}|T_i = 0]q\}$$
$$-\{E[Y_{0i}|T_i = 1] - E[Y_{0i}|T_i = 1]q + E[Y_{0i}|T_i = 0]q\} \quad (8)$$

Rearranging terms results in Equation 9, which shows that the estimate of ATE has three components: the naïve estimate of the ATE (Equation 3) and two potential sources of selection bias (baseline bias and sorting bias).

$$\underbrace{\{E[Y_{1i}|T_i = 1] - E[Y_{0i}|T_i = 0]\}}_{\text{naïve estimate}} - \underbrace{\{E[Y_{0i}|T_i = 1] - E[Y_{0i}|T_i = 0]\}}_{\text{baseline bias}} - \underbrace{\{ATT - ATC\}q}_{\text{sorting bias}} \quad (9)$$

Brand and Xie (2010) refer to two manifestations of "sorting bias": positive selection and negative selection. Positive selection or sorting gain occurs when individuals with a higher probability of treatment benefit more from the treatment (ATT > ATC). Negative selection or sorting loss occurs when individuals with a lower probability of treatment benefit more from the treatment (Morgan & Todd, 2008; Winship & Morgan, 1999).

As previously mentioned, omitting covariates that are associated with treatment assignment and outcome means that the conditional independence assumption is violated (i.e., the potential outcomes are not independent of treatment assignment). This implies that the variance explained by the omitted covariates will be captured in the error term, u_i, of the model used to estimate the propensity score. These errors, u_i, will be correlated with the errors associated with outcomes, ε_i. Breen, Choi, and Holm (2015) presented results from Greene (2003) and Powell (1994) that show if key covariates are omitted from the propensity score or selection model, the size of the bias varies according to the probability of treatment, $p(T)$. Equation 10 represents this omitted variable bias when the selection and outcome errors are bivariate normal, with ϕ and Φ as the probability density and cumulative distribution functions respectively for $p(T)$.

$$\rho_{\varepsilon,u}\sigma_{\varepsilon}\frac{\phi\left(\Phi^{-1}(p(T))\right)}{p(T)\left(1-p(T)\right)} \quad (10)$$

Equation 10 shows that omitted variable bias is a function of the correlation of the selection and outcome errors ($\rho_{\varepsilon,u}$), and the variance of the probability of treatment. If the correlation is positive, the bias reaches a minimum at $p(T) = 0.5$ and a maximum at the extremes of the probability of treatment distribution. If the errors are negatively correlated, the shape of the graph of the bias across the distribution of probability of treatment is inverted. The implications of the relationship between the size of the bias and probability of treatment are discussed in Section 2.1.2.

2.1.2 Heterogeneous treatment effects. The causal inference framework and assumptions needed to estimate unbiased causal effects were reviewed in Section 2.1.1. The focus of this study is specifically the unbiased estimation of heterogeneous treatment effects. Heterogeneous treatment effects, also known as effect modification, occur when the treatment effect varies depending on the level of a third variable such as age, gender, or propensity for treatment. In other words, treatment effects vary for identifiable subgroups (Wang, Lagakos, Ware, Hunter, & Drazen, 2007). Effect modification is modeled by including an interaction term in the outcome model.

Heterogeneous treatment effects can be quantitative or qualitative. Quantitative effect modification occurs when the treatment works for all subgroups but by various degrees. For example, an afterschool study group helps improve test scores for both boys and girls but girls improve more than boys. Qualitative effect modification occurs when the treatment is better for some and worse for others. For example, a medical intervention may improve a condition in younger patients and make the condition worse in older patients. Figure 2 demonstrates these two types of effect modification (Wang et al., 2007).

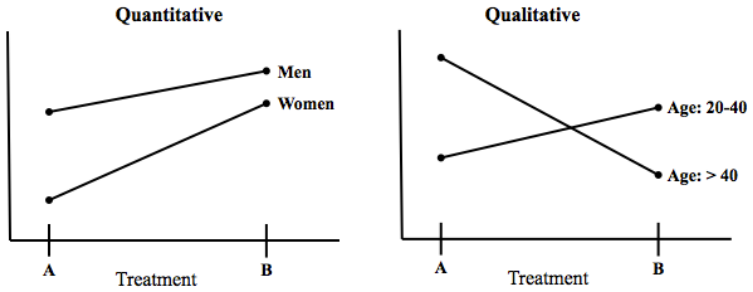


Figure 2. Types of effect modification.

Section 2.1.1 described how key covariates could be used to create "matched" data sets to eliminate bias in treatment effect estimates. The objective is to model selection so the treatment group is as similar as possible to the control group on all potential measured confounding covariates. Creating "matched" data sets enables the researcher to better defend a causal claim of a treatment effect in the population because by conditioning on confounders the only observed difference between the treatment group and control group is the effect of the treatment.

Covariates that confound selection may also identify groups that respond differently to treatment. Continuing with an earlier example, conditioning on a measure of previous math education would isolate the true treatment effect from potential confounding from prior courses completed. The same measure of previous math education could also identify groups of students that respond differently to treatment. Previous math education may moderate the effects of a math intervention program on test scores. Students with a stronger background in math may respond better to a math intervention program than students with a weaker background in math.

In addition to baseline covariates, the probability of treatment many also moderate treatment effects. For example, college bound students who recognize the potential

benefit of participating in a math intervention program may be more likely to participate and may also show more improvement than those who are less likely to participate. Because of the relation between the size of omitted variable bias and the probability of treatment as shown in Equation 10, heterogeneous treatment effects that vary according to the probability of treatment can be difficult to identify. Breen et al. (2015) replicated a study done by Brand and Xie (2010) that showed if both selection bias and treatment effects that vary according to the probability of treatment were present, further assumptions would be needed to discern one from the other.

Crump, Hotz, Imbens, and Mitnik (2008) provide parametric and nonparametric tests that can be used to determine if there is statistical evidence against the null hypothesis that the treatment effect is zero on average for all subpopulations. These tests provide a first step to identify whether further subgroup analyses are warranted. The literature and research on subgroup analyses is extensive (e.g., Fink, McConnell, & Vollmer, 2014; Pocock et al., 2002; Varadhan et al., 2013; Wang et al., 2007). The focus of this paper is not a thorough review of subgroup analyses. Therefore, additional considerations when effect modification is of interest, such as the advantages of planning subgroup analyses during the design, increase in the familywise error rate from multiple interaction tests, and sufficient sample size, are left for the reader to explore.

The previous sections provided an introduction to key ideas that provide the foundation for estimating unbiased treatment effects in observational studies. In addition, an overview of heterogeneous treatments effects was provided. Because experimental studies are the standard which propensity score methods attempt to mirror, an overview of examining effect modification in experimental studies is provided next along with a

brief transition to estimating treatment effects in observational studies using propensity score methods.

2.1.3 Treatment effect estimation. Randomized controlled trials are often referred to as the reference standard of causal evidence (Austin, 2011a; Austin, Grootendorst, & Anderson, 2007; Rubin, 1974). The random assignment component of an RCT improves internal validity. If random sampling is also used, external validity or generalizability of results is improved as well. For large sample sizes, random assignment increases the likelihood that groups are similar in both observed and unobserved characteristics on expectation (Austin, Manca, Zwarenstin, Juurlink, & Stanbrook, 2010). The similarity of groups on expectation ensures that the potential outcomes are independent of treatment assignment, therefore meeting the conditional independence assumption. The remaining assumptions of common support and SUTVA can also be addressed by the study design in RCTs. Although common support is an assumption that is more often verified in an observational study (Green & Stuart, 2014), common support can be examined in an RCT as well.

Because RCTs are designed with the three assumptions in mind, an unbiased estimate of the ATE can be estimated as a simple difference in the treatment and control group means (Murnane & Willett, 2011). In situations where the researcher is interested in improving the precision of the estimate, covariates can be added to the model. If more complex study designs are planned, such as randomization of intact groups, more complex methods such as blocking and multilevel modeling can be used (Tabachnick & Fidell, 2013).

In addition to an unbiased estimate of the ATE, unbiased main and interaction effects can be estimated as a simple difference in means (Dong, 2015). In an RCT, assuming dichotomous groups and random assignment to treatment and group, unbiased estimates of main and interaction effects can be obtained using a 2×2 factorial design as shown in Figure 3. For example, if a researcher was interested in whether test type (i.e. paper-and-pencil, computer-delivered) moderated the effects of a math intervention program, students could be randomly assigned to one of two test-types and to the intervention. Through randomization four groups are created with observed characteristics that are equivalent on expectation except for treatment and group assignment. The equivalent groups are represented in Figure 3 by the dotted circles.

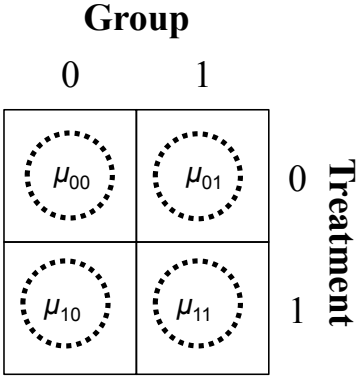


Figure 3. 2×2 factorial design for estimating treatment effects in RCTs.

The estimands of interest in a 2×2 factorial balanced design include the average treatment effect, ATE; the average treatment effect for each subgroup, ATE_g; and the interaction of treatment and group, $INT_{t\times g}$. In an RCT where both treatment and group are randomly assigned, unbiased estimates of main and interaction effects can be obtained using Equations 11 through 13, where μ_{tg} represents the mean outcome for individuals assigned into treatment *t* and group *g*.

$$ATE = \frac{\mu_{11}+\mu_{10}}{2} - \frac{\mu_{01}+\mu_{00}}{2} \quad (11)$$

$$ATE_g = \mu_{1g} - \mu_{0g} \tag{12}$$

$$INT_{t\times g} = \frac{\mu_{11}+\mu_{00}}{2} - \frac{\mu_{10}+\mu_{01}}{2} \tag{13}$$

The 2×2 factorial balanced design is the standard for analyzing effect moderation for a binary moderator and a binary treatment. The 2×2 factorial design can also be represented using Ordinary Least Squares (OLS) regression as shown in Equation (14) where *T* indicates treatment assignment and *G* indicates group assignment.

$$Y = \beta_0 + \beta_t T + \beta_g G + \beta_{tg} TG + \varepsilon, where\ \varepsilon \sim N(0, \sigma^2) \tag{14}$$

The previous paragraphs reviewed treatment effect estimation in experimental studies. The next several paragraphs follow Rhodes (2010) and describe the link between OLS regression and the probability of treatment for effect estimation in observational studies. Rhodes discusses the causal interpretation of treatment effect estimates for both homogeneous and heterogeneous treatment effects if the random assignment condition, inherent in experimental designs such as the 2×2 factorial design, is modified.

Rhodes (2010) provides the derivation of Equation 15, which shows that OLS regression estimates a weighted average of individual treatment effects, δ_i, with weights proportional to the variance of the propensity score or probability of treatment.

$$\hat{\delta} = \frac{\sum \delta_i var(propensity\ score_i)}{\sum var(propensity\ score_i)} = \frac{\sum \delta_i P(T_i=1|\boldsymbol{X}_i)[1-P(T_i=1|\boldsymbol{X}_i)]}{\sum P(T_i=1|\boldsymbol{X}_i)[1-P(T_i=1|\boldsymbol{X}_i)]} \tag{15}$$

Rhodes outlined the implications of Equation 15. First, OLS regression will provide consistent estimates of ATE (=ATT) when treatment effects are constant or random such as those found in an RCT. Second, OLS regression will provide consistent estimates of ATT when treatment effects are heterogeneous and the probability of treatment is fixed within the study as found in RCTs. Third, in the presence of heterogeneous treatment effects and heterogeneous treatment assignment probabilities, OLS regression provides

an unbiased estimate of a conditional-variance-weighted estimate of the treatment effect, which is interpretable as neither ATE nor ATT (Angrist, 1998; Morgan & Winship, 2007; Rhodes, 2010; Sloczynski, 2014). Also, the conditional-variance-weighted estimate of the treatment effect varies based on the distribution of the probability of treatment for different samples from the same population. This occurs because the variance of the probability of treatment is largest at 0.5, so the weights applied to the estimate of the treatment effect are largest at 0.5 and will become smaller as the probability of treatment moves away from 0.5 and approaches 0 or 1.

Heterogeneous treatment probabilities are typical in situations where RCTs may be unethical or cost prohibitive (Harder et al., 2010). For example, many conditions of interest such as drug use and smoking are unethical to randomly assign. In these observational settings, the researcher does not have control over treatment assignment; therefore, the probability of treatment will vary across individuals. Rhodes (2010) showed that estimates with causal interpretations could be obtained using OLS regression by reweighting the data. The "reweighted" least squares regression provides consistent estimates of ATE and ATT in the presence of heterogeneous treatment effects and heterogeneous treatment probabilities. Methods for reweighting using the probability of treatment are presented in Section 2.2.2.3. Also, if the model to estimate the probability of treatment is misspecified, the final treatment effect estimates may be biased (Harder et al., 2010). Methods for estimating the probability of treatment and considerations for specifying the treatment assignment model are presented in Section 2.2.1.

There are other methods available to estimate unbiased treatment effects in observational designs. Regression discontinuity and instrumental variables estimation are

a few examples. The focus of this paper is propensity score methods, which will be reviewed in Section 2.2.

2.2 Propensity Score Methods

There are five steps to estimate treatment effects using propensity score methods. The first step is to identify and measure covariates that are associated with both treatment assignment and outcome. These confounders are then used in the second step to estimate the propensity score or probability of treatment. Estimating the propensity score is an iterative search for the model that produces the best covariate balance and overlap. Using the propensity score to create groups with similar characteristics on expectation and balance checks are the next two steps. The final step is to undertake sensitivity analyses to determine how robust treatment effect estimates are to violations of the assumptions that must be invoked for observational studies (Zubizaretta, 2012). Sensitivity analyses are particularly important if the researcher suspects the presence of heterogeneous treatment effects that vary according to the probability of treatment (Breen et al., 2015).

Some algorithms handle several steps in a single procedure. For example, model specification and covariate balance could be combined in a single algorithm. Genetic matching (Sekhon, 2011), covariate balancing propensity score (Imai & Ratkovic, 2014) and generalized boosted modeling (McCaffrey, Ridgeway, & Morral, 2004) are examples of algorithms that combine several steps by automating the iterative search for a propensity score model that optimizes covariate balance.

Propensity score methods have several advantages over other methods in estimating unbiased treatment effects in observational studies. One often cited advantage is that the method or process of achieving covariate balance is separated from the

outcome analysis (Ho et al., 2007). This means that the researcher can iteratively create, test, and modify the matching process in order to obtain the best covariate balance. Other methods such as instrumental variables estimation and regression discontinuity start with the creation of a research design that considers the outcome of interest. In contrast, propensity score methods create a research design without considering the outcome of interest. This allows the researcher to iteratively refine the matching part of the study design finding the optimal covariate balance and overlap without "fishing." Optimal covariate balance is one of the critical components to achieving unbiased treatment effect estimates (Austin, 2009c).

Other advantages of propensity score methods include no parametric assumptions and no assumptions of homogeneity of treatment effects. Also, as previously mentioned, Rhodes (2010) showed that propensity score methods provide treatment effect estimates with causal interpretations in the presence of heterogeneity.

Although propensity score methods are simple to implement and have advantages over other methods for estimating treatment effects in observational studies, there are several considerations that must be understood and pitfalls that should be avoided when using propensity score methods (Shadish, 2013). As previously mentioned, no test exists to determine if the assumption of strong ignorability is met. This is a "strong" assumption and one of the main foundations for conclusions reached in propensity score studies. The researcher should recognize that achieving optimal covariate balance only creates groups balanced on the observed or measured attributes (Austin, 2011a; Guo & Fraser, 2010; Harder et al., 2010; Stuart, 2010; Shadish, 2013). Unlike RCTs, which balance both observed and unobserved characteristics through randomization, propensity score

methods can only address balance for observed characteristics. Other considerations include avoiding covariate selection based solely on "predictors of convenience," understanding the limitations of using archival data, and sample size (Shadish, 2013).

The next two sections detail the major steps for using propensity score methods. Specifically, Section 2.2.1 reviews the considerations and methods for estimating the propensity score followed by Section 2.2.2 which outlines methods that use the propensity score: matching, stratification, weighting and covariate adjustment.

2.2.1 Propensity score estimation. The propensity score is a single number summary of the covariates that are associated with treatment assignment. As previously mentioned, Rosenbaum and Rubin (1983) defined the propensity score as the probability of treatment conditional on the covariates that are associated with treatment assignment. Once observed covariates that are associated with treatment assignment are identified, the propensity score can be estimated and used to match, prune, or weight a non-experimental treatment group and a non-experimental control group in order to claim that the groups are as good as random such that the remaining differences are attributable solely to the effect of treatment (Morgan & Harding, 2006). If the strong ignorability assumption is met, the propensity score could be used to obtain unbiased treatment effects. Methods that use the propensity score will be reviewed in Section 2.2.2.

Shadish (2013) cites several studies that highlight the importance of covariate selection in maximizing bias reduction (Steiner, Cook, & Shadish, 2011; Zhao, 2004). Cuong (2013) used Monte Carlo simulation to show that covariates that are associated with both outcomes and treatment assignment, as well as covariates that are associated with outcomes but not treatment assignment, should be included in the propensity score

model. Cuong (2013) also showed that including covariates that are only associated with treatment assignment increases the mean squared error of the treatment effect. These results are in line with those in Leacy and Stuart (2013) and Brookhart et al. (2006), which emphasize the influence of covariates with high prognostic importance in reducing bias. These results indicate that if effect modification is suspected, the propensity score model should include covariates that are hypothesized to moderate the effects of treatment on outcome regardless of whether they are related to selection.

In addition to confounders, there are also covariates related to the outcome but only through treatment assignment (West et al., 2014). These are often referred to as instruments. Wooldridge (2009) showed that including instruments in a propensity score analysis would either increase the standard error of the treatment effect if all confounders are included or increase bias if confounders were omitted. The literature on covariate selection is vast and not the focus of this dissertation. The reader is referred to Brookhart et al., 2006; Hansen, 2008; Kelcey, 2011; Leacy & Stuart, 2013; Shadish, 2013; West et al., 2014; Wooldridge, 2009. The next three sections describe parametric and nonparametric methods of estimating the propensity score.

2.2.1.1 Logistic regression. The propensity score is usually estimated using a confirmatory method such as logistic regression (Luellen, Shadish, & Clark, 2005; McCaffrey et al., 2004; Westreich et al., 2010). Logistic regression estimates the propensity score or probability of treatment, P_{treat}, as a function of the covariates, $\boldsymbol{X}_i$, that confound selection or treatment assignment as shown in Equation 16. Equation 17 provides the logit form of the logistic regression function.

$$P_{Treat} = \Pr(T_i = 1|\boldsymbol{X}_i) = \frac{e^{\boldsymbol{X}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{X}_i\boldsymbol{\beta}}} \quad (16)$$

$$Logit\ (P_{Treat}) = \log(odds_{Treat}) = ln\left(\frac{P_{Treat}}{1-P_{Treat}}\right) = \boldsymbol{X}_i\boldsymbol{\beta} \quad (17)$$

Logistic regression is well understood and easy to implement; however, assumptions such as linearity, additivity, and proper model specification must be met. The number of predictors evaluated is also limited by sample size. Failure to meet these model assumptions can lead to poor model fit and biased effect estimates (Lee, Lessler, & Stuart, 2010; Westreich et al., 2010). Also, although logistic regression is a confirmatory method, it is used in an exploratory fashion when estimating the propensity score. The researcher essentially recursively refines the propensity score model until acceptable covariate balance is achieved. Metrics and criteria for determining acceptable covariate balance are reviewed in Section 2.2.3.

2.2.1.2 Multinomial logistic regression. The framework reviewed thus far is entirely in the context of a single treatment level, a context in which logistic regression can be used to create propensity scores. The generalized propensity score introduced by Imbens (2000) extends the propensity score framework to continuous treatments. Imai and Van Dyk (2004) further expanded the generalized propensity score to include multilevel ordinal and categorical treatments.

Imbens (2000) defined the generalized propensity score, $r(t,x)$, as the conditional probability of a particular level of treatment given the vector of covariates that are associated with treatment assignment:

$$r(t,x) = \ prob(T_i = t|\boldsymbol{X}_i) \quad (18)$$

Unbiased average treatment effects are estimated by conditioning on the generalized propensity score provided the assumption of weak ignorability holds. Weak ignorability is defined by Imbens (2000) as pairwise independence of the treatment assignment and

potential outcomes as opposed to the Rosenbaum and Rubin (1983) ignorability assumption where the treatment assignment is assumed to be independent of all potential outcomes. These concepts will be explained further by comparing the implementation of the generalized propensity score for multiple categorical treatments to the single treatment scenario.

For multiple categorical treatments, multinomial logistic regression is used to estimate the generalized propensity score for each of the potential outcomes. For *T* treatment levels, each individual will have *T* generalized propensity scores. The generalized propensity scores for *T-1* treatment levels are estimated using Equation 19 and the generalized propensity score for the reference category is estimated using Equation 20.

$$\Pr(T_i = t|\boldsymbol{X}_i) = \frac{e^{\boldsymbol{X}_i\boldsymbol{\beta}}}{1+\sum_{t=1}^{T-1} e^{\boldsymbol{X}_i\boldsymbol{\beta}}}, t = 1, \dots, T-1 \tag{19}$$

$$\Pr(T_i = T|\boldsymbol{X}_i) = \frac{1}{1+\sum_{t=1}^{T-1} e^{\boldsymbol{X}_i\boldsymbol{\beta}}} \tag{20}$$

In the single treatment scenario comparisons between treatment and control at specific propensity scores have causal interpretations. Propensity scores are conditional expectations that partition the individuals into subpopulations with similar characteristics. In the single treatment scenario, the subpopulations are the same in both the treatment and control group; therefore, individuals with the same propensity score in the treatment group have similar measured characteristics on expectation as individuals in the control group. Because both groups have the same conditioning set of characteristics, unbiased estimates of treatment effects for individuals with the same propensity score can be obtained by comparing the mean outcome of individuals in the treatment group with the mean outcome of individuals in the control group (Imbens, 2000).

In situations where there is more than one treatment level, comparisons cannot be made between individuals with the same propensity score that are in different treatment levels because the subpopulations are not the same. In other words, characteristics that predict an individual to have a generalized propensity score of 0.22 in one treatment level will not be the same as the characteristics for an individual with a predicted generalized propensity score of 0.22 in another treatment level. However, Imbens (2000) showed that if weak ignorability conditional on the vector of covariates that are associated with treatment assignment holds, an unbiased estimate of the treatment effect for each treatment level could be obtained. Weak ignorability assumes that within treatment levels, treatment assignment and potential outcomes are independent (i.e., pairwise independence). The weak ignorability assumption implies a binary condition within each treatment level that can be described as the probability of being in a specific treatment level versus not being in that particular treatment level. If weak ignorability holds, the inverse of the generalized propensity scores at each treatment level can be used as weights to estimate the mean potential outcome, μ_t, for each treatment level as shown in Equation 21. Comparisons of the

$$\hat{\mu}_t = \frac{\sum_{i=1}^{N} T_i[t] Y_i w_i(t)}{\sum_{i=1}^{N} T_i[t] w_i(t)}, where\ w_i(t) = \frac{1}{\Pr(T_i=t|\boldsymbol{X}_i)} \quad (21)$$

weighted means, $\hat{\mu}_t$, are then used to obtain average treatment effects of interest.

In addition to the weak ignorability assumption, the common support assumption discussed in Section 2.1 for the single treatment scenario also applies to situations with multiple treatment levels. Each individual must have a nonzero probability of receiving each treatment. The third assumption discussed in Section 2.1, SUTVA, must also hold.

Dong (2015) and Eeren, Spreeuwenberg, Bartak, de Rooij, and Busschbach (2015) adapted the generalized propensity score to examine main and interaction treatment effects assuming no randomization for two binary covariates as shown in Figure 4. The claim is that a 2×2 factorial design can be replicated as a 4×1 design. Both studies are described in more detail in Section 2.2.4.

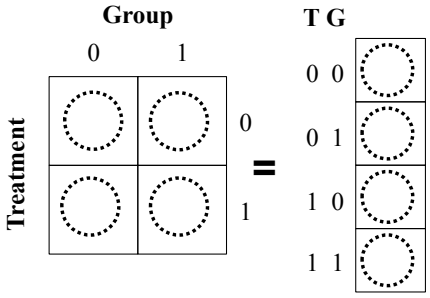


Figure 4. 2×2 factorial design using the generalized propensity score.

2.2.1.3 Machine learning algorithms. In addition to logistic regression for a single treatment and multinomial logistic regression for multiple treatments, more recent studies have drawn attention to the benefits of exploratory methods such as machine learning algorithms to estimate the propensity score (e.g., Austin, 2012; King & Resick, 2014). Machine learning algorithms, also known as statistical learning algorithms or data mining algorithms, use recursive partitioning to explore the data and identify patterns. Examples of machine learning algorithms include Classification and Regression Trees (CART), bagged CART, boosted CART, random forests, neural networks, and support vector machines. Generalized boosted modeling (GBM) is an ensemble method that creates multiple trees to improve prediction and reduce overfitting (Luellen et al., 2005). GBM is implemented in the twang package in R version 3.2.2 and can be used for research designs with single treatments or multilevel treatments (McCaffrey et al., 2004; McCaffrey et al., 2013; Ridgeway, McCaffrey, Morral, & Burgette, 2015).

Advantages of machine learning algorithms include the ability to analyze high dimensional data and the ability to handle categorical, ordinal, continuous, and missing data (Breiman, 2001; Ellis et al., 2013; King & Resick, 2014; Lee et al., 2010). Several studies have demonstrated that machine learning algorithms also produce estimates that are closer to the true propensity score particularly when modeling nonadditivity and nonlinearity (Austin, 2012; Lee et al., 2010; Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008).

For example, Setoguchi et al. (2008) compared logistic regression, regression trees, and neural networks to estimate the propensity score when using matching to estimate odds ratios. Regression trees and neural networks produced the least biased results with neural networks performing the best. Lee et al. (2010) used Monte Carlo simulation to compare logistic regression, regression trees, bagged regression trees, random forests, and boosted regression trees to estimate propensity scores. Their results showed that machine learning algorithms performed better than logistic regression when the treatment selection model had nonaddivitity and nonlinearity regardless of sample size. Boosted CART and random forests were superior to other methods. Their findings support those found in Setoguchi et al. (2008).

For the simulation conditions, Setoguchi et al. (2008) used pharmacoepidemiologic studies to model sample size, outcome probabilities, treatment probabilities, correlations among covariates, and effect sizes. Many of the same conditions were also simulated in Lee et al. (2010). One exception worth noting is that Lee et al. explored performance under smaller sample size conditions (i.e., 500 and 1,000) versus the larger sample sizes in Setoguchi et al. of 2,000 and 10,000. Neither

study explored the performance of machine learning algorithms in the presence of heterogeneous treatment effects.

Machine learning algorithms are not completely ideal. The results can be challenging to interpret, are not as effective in modeling main effects and are sensitive to overfitting (Berk, 2006; Lee et al., 2010). However, the benefits of machine learning algorithms, particularly when modeling complex selection mechanisms, indicate that machine learning algorithms may provide the best option for estimating propensity scores in the presence of heterogeneous treatment effects. If treatment effects vary for different levels of a subgroup, the covariates that influence selection may also differ for each level of a subgroup. For example, health concerns may have more influence on participation in an exercise program for individuals over 40 years of age as opposed to younger individuals.

In summary, logistic regression is the most common method of estimating the propensity score in the literature (Luellen, et al., 2005; McCaffrey et al., 2004; Westreich et al., 2010). For multiple categorical treatment levels, multinomial logistic regression can be used to estimate generalized propensity scores. However, both approaches assume a properly specified propensity score model. Machine learning algorithms are a nonparametric method of estimating the propensity score and outperform logistic regression if complex patterns of selection are present. The next section describes four methods that use the estimated propensity score.

2.2.2 Using the propensity score. After the propensity score is estimated, balanced groups are constructed for estimating the treatment effect using one of the following methods: matching, stratification, covariate adjustment, and weighting (Austin,

2011a; Austin, 2014; Lee et al., 2010). All methods assume that the researcher has chosen all covariates that are associated with treatment assignment (i.e., no unmeasured confounders) and that the researcher has correctly modeled the propensity score. The latter assumption has been found to be robust to misspecification (McCaffrey et al., 2004). In fact, Rosenbaum (1987) showed that using an estimate of the propensity score often produces better estimates of the treatment effect than using the true propensity score. This occurs because weighting by an estimate of the propensity score adjusts for both systematic and chance imbalances versus weighting by the true propensity score which accounts for only systematic imbalances between groups (Hirano, Imbens, & Ridder, 2003; Joffe & Rosenbaum, 1999; McCaffrey et al., 2004; Rosenbaum, 1987). In the subsequent sections, four methods of using the propensity score are described.

2.2.2.1 Matching. Matching is a popular method that is used frequently in the medical and social sciences (Austin, 2014; Thoemmes & Kim, 2011; Wu, Ding, Wu, & Hou, 2015). Matched sets of treated and untreated individuals with similar attributes (i.e., covariates, propensity scores) are created according to various guidelines designed to maximize covariate balance and overlap. Decisions made by the researcher are shown in Figure 5. These include the size of the matched sets (i.e., 1 treatment unit to *K* control units), matching algorithm, distance restrictions, reuse of the sample members, and the order for choosing matches from the sample. The size of the matched sets refers to the number of control units matched with each treatment unit. Austin (2010) recommends using a treatment to control ratio of 1:2 to provide the optimal improvement in both bias and variance of the estimated treatment effect.

The researcher can also determine whether to restrict matches to within a specific value or caliper. Treatment units that do not have a match within the specified caliper would be excluded from the final treatment effect estimate. Austin (2011b) showed that the optimal caliper width for estimating differences in means for continuous outcomes was 0.2 of the standard deviation of the propensity score distribution.

The main matching algorithms include greedy and optimal matching. Briefly, these algorithms differ by the function that is minimized. For example, greedy nearest neighbor matching minimizes treatment and control unit differences for matches according to the order and replacement strategy specified by the researcher. In contrast, optimal matching minimizes the average within pair differences across the entire sample. For a full description of these various manifestations of matching, the reader is referred to Guo and Fraser (2010) or Austin (2014).

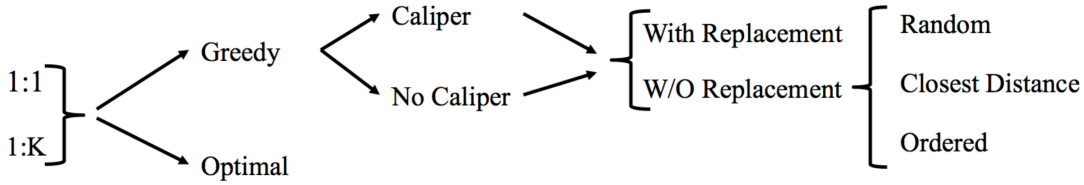


Figure 5. Matching decisions.

In addition to the decisions outlined in Figure 5, the researcher must also determine the distance metric that will be used to determine the closeness of the matches. Options include the propensity score (Equation 22), Mahalanobis distance (Equation 23), or a generalization of Mahalanobis distance (Equation 24). Equation 22 is the difference between the propensity score in the treatment group, *t*, and the control group, *c*. Equations 23 and 24 provide a measure of the closeness of the matches based on

differences in covariate values, $\boldsymbol{x}$, between the treatment and control groups where $\boldsymbol{S}$ is the covariance matrix.

$$|e_t(x) - e_c(x)| \tag{22}$$

$$MD(\boldsymbol{x}_t - \boldsymbol{x}_c) = \sqrt{(\boldsymbol{x}_t - \boldsymbol{x}_c)^T \boldsymbol{S}^{-1} (\boldsymbol{x}_t - \boldsymbol{x}_c)} \tag{23}$$

$$MD_{Gen}(\boldsymbol{x}_t, \boldsymbol{x}_c, \boldsymbol{W}) = \sqrt{(\boldsymbol{x}_t - \boldsymbol{x}_c)^T (\boldsymbol{S}^{-1/2})^T \boldsymbol{W} \boldsymbol{S}^{-1/2} (\boldsymbol{x}_t - \boldsymbol{x}_c)} \tag{24}$$

The generalization of Mahalanobis distance shown in Equation 24 uses a weight matrix, $\boldsymbol{W}$, that provides a definition of distance that varies based on the specific covariates. It adjusts to optimize covariate balance and overlap. This adjustment is shown in Figure 6 where panel (a) shows $\boldsymbol{W}$ if matching on the propensity score maximizes balance, panel (b) shows $\boldsymbol{W}$ if matching using different weights across the propensity score and the covariates maximizes balance, and panel (c) shows $\boldsymbol{W}$ if matching on Mahalanobis Distance maximizes balance.

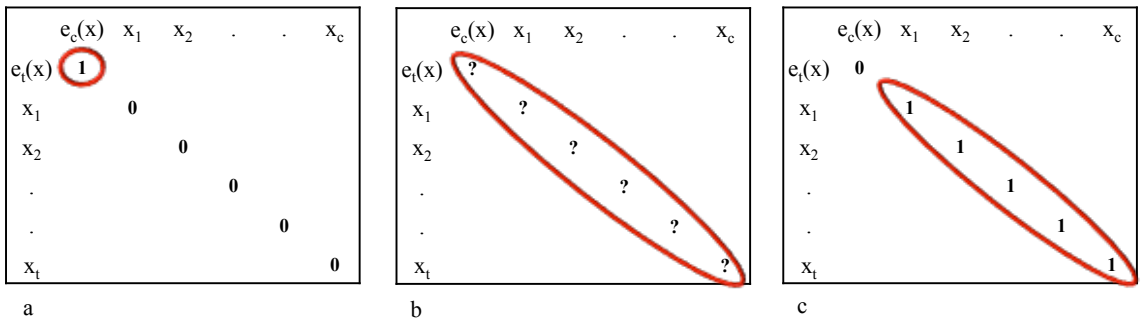
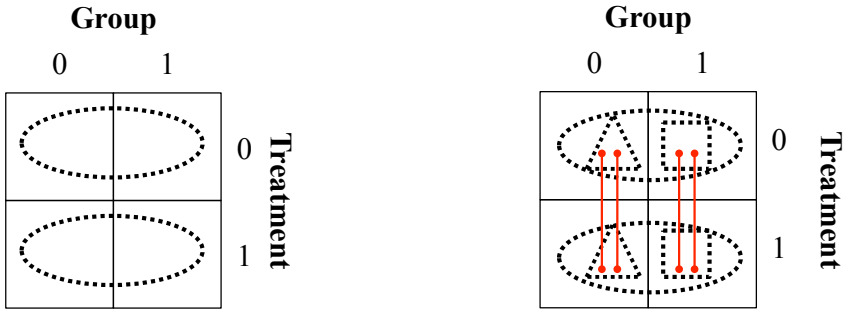


Figure 6. Variations of $\boldsymbol{W}$ that maximize covariate balance.

The genetic matching algorithm developed by Diamond and Sekhon (2013) uses the generalization of Mahalanobis distance. The algorithm iteratively searches for weights that optimize a loss function. Potential loss functions include paired *t* tests, nonparametric Kolmogorov-Smirnov tests, or any loss function specified by the

researcher. The genetic matching algorithm directly optimizes covariate balance and has been shown to outperform other methods in the presence of nonlinearity and nonadditivity (Diamond & Sekhon, 2013). In addition, Ramsahai, Grieve, and Sekhon (2011) demonstrated the flexibility of the algorithm by specifying a loss function that prioritizes covariates by prognostic importance.

Matching designs used to study effect modification, specifically when the interest is in a binary treatment and binary subgroup, need to create groups that mirror those shown in Figure 3 (Ho et al., 2011). Simply matching on several covariates or the propensity score will create balanced treatment and control groups as shown in Figure 7(a). Adding an additional matching condition, such as exact matching on group shown in Figure 7(b), is expected to create groups that are more similar to those in an experimental study (Figure 3) than the groups created by matching using only the propensity score (Figure 7(a)). Groups that are expected to be similar on all relevant characteristics are represented in Figure 7 by the dotted circles.



(a) Matching (b) Propensity score with exact matching on group

Figure 7. Matching to mirror 2 x 2 factorial design.

Austin (2014) used Monte Carlo simulation to compare 12 combinations of the matching parameters shown in Figure 5. The study highlighted the bias-variance tradeoff that occurs when using algorithms that maximize the number of subjects retained versus

algorithms that optimize the match. Algorithms such as optimal matching maximize the number of subjects used, which increases precision and generalizability of results while sacrificing bias reduction. Algorithms such as nearest neighbor matching within calipers discard subjects that are outside of a pre-specified caliper, which maximizes bias reduction while sacrificing generalizability and precision. Austin found that greedy nearest neighbor caliper matching without replacement with subjects chosen for matching in random order resulted in minimal bias and a negligible increase in variability across a wide range of scenarios. However, none of the conditions explored included the performance of matching algorithms in the presence of heterogeneous treatment effects. Austin (2014) did not provide an exhaustive comparison of the various matching conditions. For additional references and comparisons see Austin (2009d), Austin (2010), Austin (2011a), Austin (2011b), Augurzky and Kluve (2007), Caliendo and Kopeinig (2008), Gu and Rosenbaum (1993), Pan and Bai (2015), Stuart (2010), and Stuart and Rubin (2008).

2.2.2.2 Stratification. Stratification on the propensity score involves partitioning individuals into strata according to their estimated propensity score. This method is also known as interval matching or subclassification. Individuals within each propensity score stratum would have similar propensity scores and therefore similar distributions of covariates (Rosenbaum & Rubin, 1983). Treatment effects are estimated within strata as a difference in means, providing ATE_s for the s^{th} strata. Overall treatment effects can be estimated as a weighted average of the treatment effects within each strata, s, as shown by Equation 25, where N represents the total sample size and N_s represents the sample size within strata (Rosenbaum & Rubin, 1984; Schafer & Kang, 2008).

$$ATE = \sum_s \left(\frac{N_s}{N}\right) ATE_s \quad (25)$$

Equation 25 can be easily modified to estimate ATT by weighting the ATE_s by the proportion of treated individuals in each stratum. Cochran (1968) showed that using five strata based on a single covariate eliminated approximately 90% of the bias, a result that was later extended to include stratifying on the quintiles of the propensity score (Rosenbaum & Rubin, 1984).

Stratification has an advantage over the weighting methods reviewed next, in section 2.2.2.3, when individuals have probabilities of treatment near the extremes of the propensity score distribution. Stratification has been shown to lessen the impact of extreme weights that occur when individuals have propensity scores near 0 or 1 (Kang & Schafer, 2007). Matching and stratification also rely less on the precise value of the propensity score than the methods reviewed in Sections 2.2.2.3 and 2.2.2.4 (i.e., weighting and covariate adjustment), and are therefore less sensitive to misspecification of the propensity score model (Kang & Schafer 2007; Lee et al. 2010; Rubin 2007).

2.2.2.3 Weighting. Weighting is a nonparametric method for estimating treatment effects using propensity scores. The most common estimators are inverse probability of treatment weighting (IPTW) and weighting by the odds. IPTW is used to estimate ATE and, in this process, data for individuals in the treatment group are weighted by the inverse of the propensity score, e_i, and data for individuals in the control group by the inverse of one minus the propensity score. The weights, w_i, are estimated using Equation 26 and ATE is estimated using Equation 27.

$$w_i = \frac{T_i}{e_i} + \frac{1-T_i}{1-e_i} \quad (26)$$

$$ATE = \frac{1}{N}\sum_{i=1}^{N} \frac{T_i Y_i}{e_i} - \frac{1}{N}\sum_{i=1}^{N} \frac{(1-T_i)Y_i}{1-e_i} \quad (27)$$

Weighting by the odds is used to estimate the ATT where individuals in the treatment group are assigned a weight equal to one and individuals in the control group are assigned a weight equal to the propensity score divided by one minus the probability of treatment. The weights, w_i, are estimated using Equation 28 and ATT is estimated using Equation 29.

$$w_i = T_i + \frac{(1-T_i)e_i}{1-e_i} \tag{28}$$

$$ATT = \frac{1}{N}\sum_{i=1}^{N} T_i Y_i - \frac{1}{N}\sum_{i=1}^{N} \frac{(1-T_i)e_i Y_i}{1-e_i} \tag{29}$$

One of the advantages of weighting to estimate treatment effects is that the estimates are based on data for all individuals in the sample, which improves the generalizability of the study results. For this reason, in situations with few extreme propensity scores, weighting estimators like IPTW and weighting by the odds often result in less bias than stratification. In addition, stratification relies on comparing individuals with the same propensity score value. In practice, few exact matches are available; therefore, strata are often formed with individuals who have similar as opposed to exact propensity scores resulting in some residual confounding within strata. Disadvantages of weighting include sensitivity to misspecification of the propensity score model and the influence of individuals at the extremes of the propensity score distribution (Kang & Schafer, 2007; Lunceford & Davidian, 2004).

2.2.2.4 Covariate adjustment. Covariate adjustment using the propensity score is regression where the outcome variable is regressed on the treatment status and propensity score as shown in Equation 30. Covariate adjustment assumes a properly specified regression model and could also be implemented by including all confounding covariates in the regression model. There are two advantages to using the propensity score as

opposed to including all covariates in the regression equation. The first is that the separation of design and analysis is maintained. The second is that a more parsimonious model can be posited for the outcome model while the propensity score model can include complex interactions and nonlinearities as needed (D'Agostino, 1998).

$$Y = \beta_0 + \beta_t t + \beta_1 \hat{e} + \varepsilon, \;\; \varepsilon \sim N(0, \sigma^2) \tag{30}$$

The success of covariate adjustment using the propensity score depends on a properly specified propensity score model. Adjusting for a poor approximation of the propensity score may increase bias (Rubin, 1973; Rubin, 1979). In addition, regression cannot adjust for differences in observed covariates when the distributions of the observed covariates in the treatment group are substantially different from the distributions of the observed covariates in the control group (Rubin, 2001). Rubin (2001) describes the distributional conditions that must be met in order for covariate adjustment to reliably estimate treatment effects. These distributional considerations relate to balance and overlap of the covariate distributions between the treatment and control groups. If covariate balance is not adequate, a doubly robust approach should be considered where covariate adjustment is used in conjunction with one of the other propensity score methods that balance the covariate distributions between groups. Covariate balance and overlap will be reviewed in Section 2.2.3.

2.2.3 Balance statistics. As previously mentioned, propensity score methods match or prune a non-experimental treatment group and a non-experimental control group in order to claim that the groups are as good as random. Regardless of the algorithm or variables used, both covariate balance and overlap must be verified to ensure that no systematic differences remain between the treatment and control groups and to

verify that the common support assumption has been met (Guo & Fraser, 2010; Harder et al., 2010; Ho et al., 2007).

Model adequacy is measured by covariate balance to determine if the groups are as good as those formed under the conditions of an experimental study. There is no single measurement available that addresses all aspects of covariate balance; therefore, several measurements should be used to support claims of sufficient balance and overlap. There is also no consensus on what constitutes success in terms of achieving balance (Caliendo & Kopeinig, 2008; Pan & Bai, 2015). This decision is left to the researcher primarily because of the differences in the prognostic importance of the covariates. The researcher must also recognize when sufficient balance cannot be achieved and choose another method (Stuart, 2010).

A simple and commonly used balance diagnostic in the literature is the absolute standardized mean difference also known as Cohen's effect size index (Austin, 2009a). It has the advantage of not being influenced by sample size like *t* tests and other hypothesis tests. It also allows for the comparison of variables measured in different units because the difference is standardized. One limitation is that there is no consensus on what value determines that a covariate is considered balanced, a common theme for all balance diagnostics. Various recommendations include below the values of 0.10, 0.20, and 0.25 (Austin, 2011a; Caliendo & Kopeinig, 2008; Ho et al., 2007). Graphical plots like those in Figure 8 provide a quick overview of balance improvements in individual covariates.

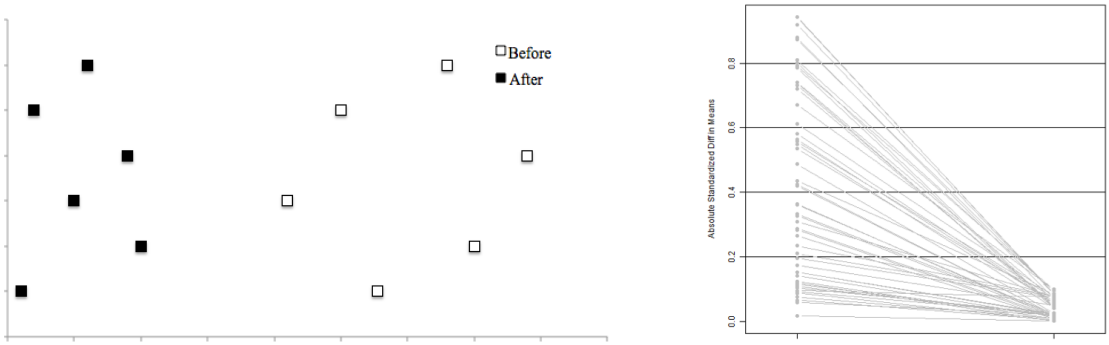


Figure 8. Graphical representation of absolute standardized mean difference.

Equations 31 and 32 represent the absolute standardized difference in means for continuous and dichotomous variables, *x* and *p,* respectively. In Equation 31, the variance of the treatment group is represented by s_t^2 and the variance of the control group is represented by s_c^2.

$$d = \frac{|\bar{x}_t - \bar{x}_c|}{\sqrt{\left(s_t^2 + s_c^2\right)/_2}} \tag{31}$$

$$d = \frac{|\bar{p}_t - \bar{p}_c|}{\sqrt{\frac{\bar{p}_t(1-\bar{p}_t)+\bar{p}_c(1-\bar{p}_c)}{2}}} \tag{32}$$

Other graphical displays such as side-by-side boxplots, empirical quantile–quantile plots, and nonparametric estimates of the density functions can inform both balance and overlap on the distribution of each group as opposed to univariate summaries like Cohen's effect size and variance ratios. Austin (2009a), Ho et al. (2007), and Imai, King, and Stuart (2008) recommend these broader descriptions of balance and overlap to compare the distribution of the covariates between the treatment group and the control group.

There are several measurements that are not recommended. Austin (2009a) showed that comparing the distribution of the propensity score is uninformative. The distributions of the propensity scores in the treatment and control groups are informative

of balance only when the researcher is sure that the model is properly specified. Similarity of the distribution of the propensity score does not imply balance or proper model specification; therefore, balance must be checked by examination of the individual covariates and not the propensity score.

Significance testing is also not recommended. Imai et al. (2008) provided the most compelling argument for not using the *t* test as a measure of balance. They conducted a simulation study where control units were randomly dropped from the study. At each iteration, balance was checked and the number of dropped controls was increased. As expected, power was reduced because of the smaller sample size resulting in a reduced *t* statistic; however, balance did not improve. Imai et al. (2008) also argue that balance is a property of a particular sample and not a "super-population."

Once matched groups are created and the researcher is satisfied that the treatment variable is as close as possible to being independent of the observed covariates, normal parametric analysis methods such as multivariate analysis of treatment effects may be used to estimate treatment effects. Sensitivity analyses should also be conducted to see how sensitive the study results are to a potential unobserved covariate not included in the model. It is also important to realize that a skeptical observer could always find some variable that differs systematically between the treatment and control groups (Rubin 1974); therefore, it is critical to support any claim of the ignorable treatment assignment assumption with related studies and theory.

Despite the simplicity of the measurements to demonstrate balance, reporting balance has been absent or poorly demonstrated in most of the literature regardless of the discipline. For example, Thoemmes and Kim (2011) did a systematic review of

propensity score methods in the social sciences. They searched three databases, Web of Science, ERIC and PsycINFO. They restricted their results to papers published in the areas of psychology, education, and social science. Their review included a review of balance checks among other metrics. A total of 86 studies were found using propensity score methods published between 2003 and 2009. Sixty-two of the 86 had at least one balance check, and only 30 checked for common support. Forty-one of the 62 used a *t* test and 8 of the 62 checked propensity score distributions only. Diamond and Sekhon (2013) wrote a similar review of the economics literature, and Austin (2008) provided a review of the medical literature. The outcomes were the same as those of the Thoemmes and Kim systematic review: balance results post matching were insufficient or entirely missing from the study results.

The absence of focus on a key component of matching methods is likely due to the lack of updated guidelines and no consensus on what value determines that a covariate is balanced. In addition, recent research emphasizes the value of prioritizing covariates by prognostic importance to maximize bias reduction (Leacy & Stuart, 2013). For example a researcher with evidence to support a hypothesis of teacher assignment moderating the effects of a math intervention program might put a higher priority on covariate balance for teacher assignment than other covariates with less prognostic importance. However, covariate balance is not the only consideration for maximizing bias reduction. The performance in minimizing the bias of treatment effect estimates varies for the four propensity score methods. The next Section, 2.2.4, provides an overview of research that compares the bias reduction capabilities of the four propensity score methods.

2.2.4 Method comparison. The four propensity score methods, matching, stratification, weighting and covariate adjustment, vary in their ability to reduce selection bias. Luellen et al. (2005) identify several studies that show that the bias in and efficiency of the treatment effect estimate depends on several factors with these methods. These include the characteristics of the sample, the propensity score estimation and matching method, the choice of covariates and their relation to outcome and treatment, and the potential for omitted variable bias (Harder et al., 2010). More recent studies continue to support the findings of differences in bias reduction based on the conditions described in Luellen et al. (2005). For example, Austin (2009c) compared all four propensity score methods across various degrees of covariate imbalance and overlap assuming normally distributed predictors. Propensity score matching and IPTW produced better covariate balance than stratification and covariate adjustment across all scenarios explored. Austin (2009b) also compared the performance of the four propensity score methods for estimating risk differences. The comparisons included a doubly robust approach using IPTW. The results showed that the IPTW doubly robust estimator outperformed all other methods even with a misspecified outcome model. These findings support those in Lunceford and Davidian (2004). Within the propensity score method framework, the majority of the simulation studies that compare various propensity score methods assume homogeneous treatment effects (e.g., Austin, 2014; Austin et al., 2007). There are few simulation studies in the literature that focus directly on the performance of propensity score methods in the presence of heterogeneous treatment effects (Green & Stuart, 2014; Leacy & Stuart, 2013).

Kreif et al. (2012) provide a rare example of a simulation study that directly compares the performance of propensity score matching, IPTW, and genetic matching in estimating subgroup effects. Kreif et al. modeled heterogeneous treatment effects, nonlinearities in the propensity score, and separate treatment assignment mechanisms for each subgroup. Two continuous correlated covariates were used and propensity scores were estimated for each group. Kreif et al. found that all three methods produced unbiased estimates when the propensity score model was properly specified and weights were stable or not near the extremes of the propensity score distribution. Inverse probability of treatment weighting did not perform well under conditions of unstable propensity score weights or with a misspecified propensity score model. Genetic Matching was found to be robust under all conditions. These results were obtained for one large sample size condition of 2,000 participants. Group effects were obtained using separate propensity scores for each group; however, overall treatment effect estimates and interaction effects were not evaluated.

Rassen, Glynn, Rothmas, Setoguchi, and Schneeweiss (2012) also used simulated data. Their study investigated whether propensity scores estimated using a full cohort could serve as a proxy for propensity scores estimated within a subgroup. The primary outcome measure used by Rassen et al. was the difference in the treatment effect estimated using a propensity score for the full sample and the treatment effect estimated using a propensity score for a subgroup. Rassen et al. found that both methods were similar for subgroup sizes > 1,000 and for correctly specified propensity score models. The conclusions in Rassen et al. supported the conclusions in Rosenbaum and Rubin

(1983), which are that misspecified propensity score models increase bias and a correctly specified propensity score for the full cohort remains valid within subgroups.

Two recent simulation studies investigated using the generalized propensity score to estimate main and interaction effects. The first study, Dong (2015), compared the generalized propensity score estimated using multinomial logistic regression with subclassification using group propensity scores. The generalized propensity score was used in matching, subclassification, and weighting. Results showed that the generalized propensity score using IPTW and subclassification using group propensity scores had the best bias reduction and the smallest mean squared error for the coefficients of the treatment, group, and interaction of treatment and group terms. The proportion of treated individuals was one of the modified conditions in the study and was shown to have little effect on the results. The majority of the conditions explored had sample sizes > 6,000. In addition, complex selection conditions were not simulated.

The second study, Eeren et al. (2015), compared covariate adjustment using a single propensity score as a predictor in a regression model that included the group, treatment, and treatment/group interaction with covariate adjustment using the generalized propensity score. Conditions investigated included nonadditivity of treatment assignment, various covariate correlations, and sample size. Logistic regression was the only method used to estimate the propensity scores and three normally distributed continuous covariates were simulated. Across all conditions, the generalized propensity score adjusted estimate provided more efficient results than the single propensity score adjusted results as compared by average bias and mean squared error

across three coefficients of each model. ATE, main, and interaction effects were not directly compared.

Analyses of subgroup effects or treatment effect heterogeneity comparing propensity score methods using empirical data are more common in the literature than studies using simulated data. Although these studies provide some insight regarding the performance of propensity score methods, their conclusions have limited generalizability. For example, Green and Stuart (2014) compared five propensity score methods to evaluate whether major depressive disorder (treatment) increases the risk for later substance use disorders (outcome) among men and women. The five methods varied by full matching across or within gender, propensity score estimates for the entire sample or one per gender, and propensity score models with or without interactions. Green and Stuart found that the best balance was achieved when separate propensity scores were estimated for men and women, whereas the worst balance was achieved when a single propensity score was estimated with no interaction terms. Green and Stuart recommended that the results should be tested with a simulation study that assesses the bias in the treatment effect estimate instead of using balance as a proxy to determine the method that minimizes bias.

Several empirical studies in the literature analyze how well propensity score methods estimate the treatment effects of individuals who are treated contrary to prediction (Ellis et al., 2013; Kurth et al., 2006; Sturmer, Rothman, & Glynn, 2006). Individuals who are treated contrary to prediction are individuals at the extremes of the propensity score distribution.

Ellis et al. (2013) compared propensity score matching and weighting by the odds to estimate the treatment effect in the treated using data from the Sequenced Treatment Alternatives to Relieve Depression effectiveness trial. They found that both methods balanced covariates but produced different treatment effect estimates. Their study included sensitivity analyses that systematically removed observations from the extremes of the propensity score distribution. As extreme observations were removed, the weighted estimate of the treatment effects approached the matched estimate of the treatment effects. The treatment effect was also no longer significant, which indicated possible treatment effect heterogeneity where only individuals in the extremes of the propensity score distribution benefited from treatment. They concluded that if heterogeneous treatment effects are present, matching removes observations that may be key in the identification of the true treatment effect and weighting is sensitive to extreme observations. They recommended that studies should include sensitivity analyses particularly for individuals who are treated contrary to prediction.

Kurth et al. (2006) compared propensity score stratification, covariate adjustment, propensity score adjustment, IPTW, and weighting by the odds using data from the German stroke registry. A statistically significant interaction of the propensity score and treatment effect indicated qualitative effect modification. Specifically, the control group mortality increased across propensity score, while the treatment group mortality decreased across propensity score. Kurth et al. found that in the presence of heterogeneous treatment effects, population treatment effect estimates varied across propensity score methods. They concluded that this variation was a result of either explicitly or implicitly excluding individuals with low propensity scores. Explicit

exclusion occurs when the method provides different estimates such as IPTW estimating ATE versus weighting by the odds estimating ATT. Implicit exclusion occurs when the method excludes certain observations such as individuals with no match. Similar to Ellis et al. (2013), Kurth et al. also recommended that studies should include sensitivity analyses particularly for individuals who are treated contrary to prediction.

In addition to Kurth et al. (2006), several empirical studies highlight the importance of understanding the different effect estimates provided by propensity score methods (Lundt et al., 2009; Morgan & Todd, 2008; Sturmer et al., 2006). In the presence of homogeneous treatment effects, recognition of these differences is moot. However, as previously described, in the presence of heterogeneous treatment effects, ATE ≠ ATT ≠ ATC.

Sturmer, Rothman, Avorn, and Glynn (2010) also focused on individuals treated contrary to prediction; however, they investigated bias and root mean squared error (RMSE) when treatment effects were confounded by an unmeasured covariate operating only in the tails of the propensity score distribution. Treatment assignment that is confounded by patient frailty when treatment is provided as a "last resort" for low propensity score individuals and treatment assignment that is confounded by patient frailty when treatment is "withheld" from high propensity score individuals are two examples of an unmeasured covariate operating only in the tails of the propensity score distribution. Sturmer et al. used a simulation study to demonstrate that when treatment is confounded by a covariate related to "treatment withheld" or "last resort treatment," restricting the range of the propensity score distribution reduces the bias of treatment effect estimates. Their study provides an example of one of the many benefits of

propensity score methods in estimating treatment effects. Although no test exists to determine if unmeasured confounding or true heterogeneity of treatment effects is present, restricting the range of the propensity score distribution provides data that could lead to identification of a defined subgroup.

Sturmer, Wyss, Glynn, and Brookhart (2014) recommend the approach outlined in Lunt et al. (2009) to identify heterogeneity of treatment effects and situations where restricting the range of the propensity score distribution might be appropriate. Sturmer et al. (2014) also provide a more comprehensive review of the application of propensity score methods in medical interventions. They identify common themes found in medical studies such as heterogeneity of treatment effects across age and gender.

Lunt et al. (2009) used data from a United Kingdom-based registry of subjects treated with anti-tumor necrosis factor drugs for rheumatoid arthritis to compare treatment effects estimated by stratification, IPTW, weighting by the odds, matching with replacement, and matching without replacement. They found that average treatment effects varied across methods; however, treatment effects estimated within quintiles of the propensity score distribution were "broadly similar" across methods. They recommended examining the change in the treatment effect estimates within quintiles of the propensity score distribution after creating matched groups. The goal was to assess the impact of remaining covariate imbalances. Analyzing the change in the treatment effect estimates or outcome to address covariate imbalances does not maintain the separation of design and analysis inherent in most propensity score methods (Harder et al., 2010). However, recent additions to the propensity score literature suggest that ignoring imbalances in variables that are strongly related to outcome can lead to

increased bias and decreased precision in treatment effect estimates (Brookhart, et al., 2006; Leacy & Stuart, 2013).

The empirical studies described provide a few examples of the challenges and considerations associated with estimating treatment effects when effects vary by individual or subgroup. This dissertation aims to add to the few studies that use Monte Carlo simulation to analyze the performance of propensity score methods when treatment effects are heterogeneous.

Chapter 2 started with an overview of the causal modeling framework, namely the Neyman-Rubin causal model. Sources of treatment effect bias were described along with an overview of methods for estimating and using the propensity score to create matched groups specifically in the presence of heterogeneous treatment effects. Studies that compare these methods were summarized. Few simulation studies were found that compare propensity score methods in the presence of heterogeneous treatment effects. Chapter 3 describes the research design to investigate which propensity score method produces estimates of main and interaction effects with the smallest bias and mean squared error. Specifically the following research questions are addressed:

1. Which of the proposed propensity score methods produces estimates of main and interaction effects with the smallest bias and variance in the presence of complex selection and heterogeneous treatment effects?
2. Do the proposed propensity score methods produce similar estimates of main and interaction effects when one subgroup is more prevalent in the extremes of the propensity score distribution?

3. Is generalized boosted modeling more effective than logistic regression in estimating propensity scores in the presence of complex selection and heterogeneous treatment effects as measured by main and interaction effect estimates?

Methods were evaluated under various group distributions, sample sizes, effect sizes, and selection models. Qualitative and quantitative effects were simulated as well as simple and complex selection models.

Chapter 3: Methods

As previously mentioned, the primary focus of this study is to add to the literature on the analysis of heterogeneous treatment effects using propensity score methods. The specific scenario investigated is an observational study where a researcher suspects that group membership moderates the effect of a single level treatment or intervention. For example, the effects of a math intervention program on test scores are suspected of being moderated by family support of education where individuals are grouped into two levels of family support. Two Monte Carlo simulations are proposed and are intended to extend the results of several studies (Dong, 2015; Eeren et al., 2015; Green & Stuart, 2014; Kreif et al., 2012; Rassen et al., 2012).

The simulations were designed to answer three research questions. The objective for the first research question was to compare propensity score methods that directly address the performance of propensity score methods in the presence of heterogeneous treatment effects:

1. Which of the proposed propensity score methods produces estimates of main and interaction effects with the smallest bias and variance in the presence of complex selection and heterogeneous treatment effects?

Research question 1 was investigated in two parts. The separation of the outcome analyses from the research design is one of the benefits of propensity score methods. As previously mentioned, this separation allows the researcher to refine the propensity score model without "fishing". This separation also implies that the bias of treatment effect estimates is the same because the matched samples or weights created will be the same regardless of the size of the interaction effect. The performance of methods in the

absence of effect moderation for estimates of the coefficient of treatment is expected to be similar in the presence of effect moderation. Therefore, the first area under investigation for research question 1 was to confirm that the findings from previous research hold when effect moderation is present. Specifically, do methods that perform well when estimating the coefficient of treatment, also perform well when estimating the coefficient of the interaction term? Are previously established guidelines for including covariates in the propensity score model similar in the presence of effect moderation?

The second area under investigation for research question 1 was to determine if methods that provide additional focus on group information when estimating the propensity score have better performance than methods that do not focus on group information when estimating the propensity score. Specific methods and hypotheses are described in Section 3.2.

The objective of the second research question was to compare the performance of propensity score methods when one subgroup is more prevalent than another in the extremes of the propensity score distribution:

2. Do the proposed propensity score methods produce similar estimates of main and interaction effects when one subgroup is more prevalent in the extremes of the propensity score distribution?

The objective of the third research question was to confirm that generalized boosted modeling outperforms logistic regression in the presence of complex selection and heterogeneous treatment effects:

3. Is generalized boosted modeling more effective than logistic regression in estimating propensity scores in the presence of complex selection and heterogeneous treatment effects as measured by main and interaction effect estimates?

Generalized boosted modeling is expected to have smaller bias in the presence of complex selection.

This chapter starts with a description of the simulation design, followed by a high level overview of the methods under investigation, and concludes with the metrics that were used to evaluate the results.

3.1 Simulation Methods

The following sections describe data generation and the specific conditions imposed for each simulation. Coefficients for all models are in Appendix A.

3.1.1 Data generation. Data were simulated with a binary subgroup indicator, g; a binary treatment indicator, t; and ten continuous covariates. Six covariates (x_1, ... x_6) were associated with both outcome and treatment assignment, x_7 and x_8 were associated with the outcome only, and x_9 and x_{10} were associated with treatment only. Covariates were generated from a normal distribution with mean 0 and standard deviation 1. The subgroup indicator, g, was randomly generated from a Bernoulli distribution with probability of 0.50. No correlation structure was imposed among the covariates to create a more difficult matching condition. One thousand replications of sample size 250, 500, and 1000 were generated following sample size studies and recommendations in Feng, Zhou, Zou, Fan, and Li (2012) and Shadish (2013).

The number of replications used for each simulation was confirmed as sufficient for stable results by estimating bias and mean squared error (MSE) at replication

increments of 50 for a sample size of 250 for three methods: regression, MNPS, and weighting by the odds. Figure 9 shows that a steady state is reached at approximately 500 replications. This outcome was consistent regardless of the research design and shows that the original proposal of 1000 replications was sufficient to provide stable results.

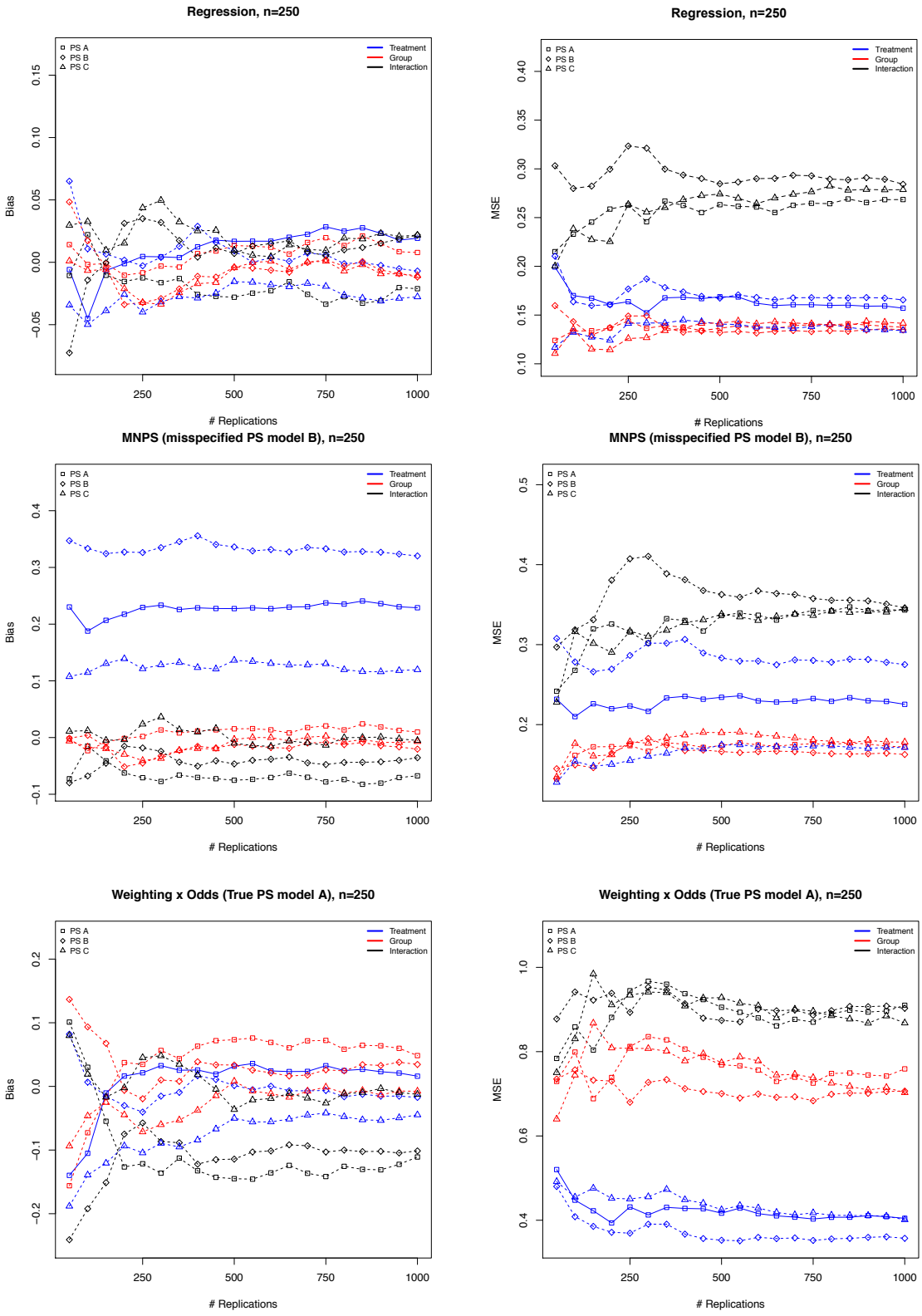


Figure 9. Bias and MSE as a function of the number of replications.

The treatment effect indicator, *t*, was generated as a function of one of five "true" propensity score models. A random number between 0 and 1 was generated from a Bernoulli distribution with a probability of 0.50. The treatment effect indicator was assigned a value of 1, indicating treatment, if the random number was less than the true propensity score and 0, indicating control, otherwise. The proportion of individuals assigned to treatment and control was balanced. This condition remained fixed in both simulations to focus the analysis on the variability of the group proportion conditions. Data were generated to reflect treatment assignment compliance.

Five true propensity score models were chosen to model situations where various covariates influence treatment assignment. Equation 33 (i.e., data generating model A) included covariates that were not related to outcome and where selection (i.e., treatment assignment) differed by group.

$$logit_A(e(x)) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_6 x_6 + \alpha_9 x_9 + \alpha_{10} x_{10} + \alpha_g g \quad (33)$$

Equation 34 (i.e., data generating model B) represented a model where selection differed by group and only covariates related to outcome were included.

$$logit_B(e(x)) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_8 x_8 + \alpha_g g \quad (34)$$

Equation 35 (i.e., data generating model C) was similar to data generating model B but included nonadditivity and nonlinearity.

$$logit_C(e(x)) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_6 x_6 + \alpha_7 x_7 x_8 + \alpha_8 x_8^2 + \alpha_9 x_2^2 + \alpha_{10} x_3 x_5 + \alpha_{9g} x_1^2 g \quad (35)$$

Equation 36 (i.e., data generating model D) represented a model where selection did not differ by group and only covariates related to outcome were included.

$$logit_D(e(x)) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_6 x_6 \quad (36)$$

The final model shown in Equation 37, (i.e., data generating model E) included nonlinearity and nonadditivity, included covariates not related to outcome, and did not model complex selection.

$$logit_E(e(x)) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_6 x_6 + \alpha_7 x_7 x_8 + \alpha_8 x_8^2 + \alpha_9 x_2^2 + \alpha_{10} x_3 x_5 \quad (37)$$

Five outcome values were generated as described in Equation 38 for each "true" propensity score model to simulate qualitative and quantitative effect moderation for both small and large effect sizes in addition to a condition of no effect moderation.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_8 x_8 + \beta_g g + \beta_t t + \beta_{tg} tg + \varepsilon \quad (38)$$

3.1.2 Propensity score conditioning methods. For simulation I, two misspecified propensity score models, Equations 39 and 40, were assessed under the assumption that the researcher expects that selection into treatment varies across subgroup; however, the true model for selection is unknown. Misspecified model mA excluded two covariates that influence outcome only, x_7 and x_8, and excluded two covariates that influence treatment assignment only, x_9 and x_{10}. Misspecified model mB excluded two covariates that influence treatment assignment, x_9 and x_{10}, and x_7 and x_8 were also modeled as an interaction. Both misspecified models are reasonable considering that the true propensity score model is rarely known (Ho et al., 2007).

$$logit_{mA}(e(x)) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_6 x_6 + \alpha_g g \quad (39)$$

$$logit_{mB}(e(x)) = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_8 x_8 + \alpha_g g + \alpha_7 x_7 x_8 \quad (40)$$

Propensity scores were estimated using logistic regression and generalized boosted models. as specified in Equations 33-37, 39, and 40. Generalized boosted modeling was chosen over other machine learning algorithms because it has been optimized to estimate treatment probabilities and has been adapted for multiple treatment

scenarios (McCaffrey et al., 2004; McCaffrey et al., 2013). Logistic regression was included for comparison because it is the most common method for estimating propensity scores (Luellen et al., 2005; McCaffrey et al., 2004; Westreich et al., 2010). However, as previously mentioned, it is not the best method for estimating propensity scores in the presence of nonlinearity and nonadditivity (Setoguchi et al., 2008).

3.1.3 Summary of simulation conditions. Tables 2 and 3 summarize the manipulated and fixed conditions for simulation I. Fixed conditions were chosen to focus the analyses on the specific research questions. For each dataset, 11 different propensity score methods were imposed. Four methods are designed to estimate an ATE and seven methods are designed to estimate ATT. Descriptions of each method are in Section 3.2.

Table 2
Manipulated Factors for Simulation I

Manipulated Factors	Levels
Between	
Sample Size	250, 500, 1000
Interaction Effect Size	Qualitative/Quantitative, Large/Small, None
True PS Models	Data generating models A-C
Within	
PS Estimation	Logistic Regression, Generalized Boosted Models
PS Methods ATE	GBM, MNPS, IPTW, Subclassification
PS Methods ATT	Group, Match (1:1), Match (1:2), ExMatch (1:1), ExMatch (1:2), GBM, Weighting x Odds
PS Models	true, mA,mB

Note. PS = propensity score

Table 3
Fixed factors for Simulation I

Fixed Factors	Condition
Measurement Error	none
Missing Data	none
Covariate Distributions	normal
Covariate Correlations	none
Treat/Control Proportions	50/50
Group Proportions	50/50

In simulation II, the majority of the manipulated conditions remained the same as simulation I. The main differences were that the group indicator was excluded from the true propensity score models and three additional conditions governing the simulation of g were added. The objective of the group proportion condition in the second simulation was to determine which propensity score method produces the least biased estimates of main and interaction effects when one group is more prevalent in the extremes of the propensity score distribution. The subgroup indicator, *g*, for the baseline condition was randomly generated from a Bernoulli distribution with probability of 0.50. Three additional subgroup conditions were determined by the cumulative probability distribution of the true propensity score. Cumulative probability cutoff values of 0.9, 0.8, and 0.7 were used to assign proportions where one subgroup is more prevalent ($p = 0.9$) in the upper extremes of the propensity score and less prevalent elsewhere ($p = 0.4$). As an example, for a cumulative probability cutoff value of 0.9, the group indicator for individuals in the upper 10% of the true probability distribution were generated from a Bernoulli distribution with probability of 0.9 and 0.4 otherwise. Table 4 summarizes the manipulated conditions and Table 5 summarizes the fixed conditions for simulation II.

Table 4
Manipulated Factors for Simulation II

Manipulated Factors	Levels
Between	
Sample Size	250, 500, 1000
Interaction Effect Size	Qualitative/Quantitative, Large/Small, None
True PS Models	Data generating models D & E
Group Proportions	50/50; True PS > x%ile (x = 90, 80, 70) p=0.9 else p=0.4
Within	
PS Estimation	Logistic Regression, Generalized Boosted Models
PS Methods ATE	GBM, MNPS, IPTW, Subclassification
PS Methods ATT	Group, Match (1:1), Match (1:2), ExMatch (1:1), ExMatch (1:2), GBM, Weighting x Odds
PS Models	true, mA

Note. PS = propensity score

Table 5
Fixed factors for Simulation II

Fixed Factors	Condition
Measurement Error	none
Missing Data	none
Covariate Distributions	normal
Covariate Correlations	none
Treat/Control Proportions	50/50

3.2 Treatment Effect Estimation

After data generation, several propensity score methods were used to identify the matched subsets or weights for the outcome analyses. The methods that include group information in the research design were hypothesized to have performance advantages over the other methods. The group-focused methods (i.e., exact matching, MNPS and group propensity scores), are variations of propensity score methods that are most similar to the experimental design shown in Figure 3. Propensity score methods were implemented to estimate ATT or ATE. Several methods were used to estimate both ATT

and ATE. For example, GBM was used to estimate both ATT and ATE, while matching was only used for ATT estimates.

The methods used to estimate ATT include several variations of matching. The first method, referred to as Match (1:1) in this study, is the optimal matching design recommended in Austin (2014) under homogeneous treatment effect conditions: 1:1 nearest neighbor caliper matching without replacement with subjects chosen for matching in random order. Nearest neighbor matching is also referred to as "greedy" matching. The algorithm, as specified in this study, randomly selects an individual from the treatment group and creates a "match" with the individual in the control group that is "nearest" as measured by the propensity score. Individuals in the control group that are not matched to an individual in the treatment group are excluded from the outcome analysis. In this implementation, the matched sets must also be within a caliper of 0.2 of the standard deviations of the logit of the propensity score. Treated individuals are also excluded from the outcome analysis if no match is found because their propensity score is not within the specified limit of the propensity score of any of the remaining individuals in the control group. Matching was also performed without replacement. This means that individuals in the control group were not matched to more than one individual in the treatment group.

Several variations to the matching algorithm described in the previous paragraph were included. Exact matching on group, referred to as ExMatch (1:1) in this study, was an added condition to the design recommended in Austin (2014). The condition to restrict matches to treatment and control units from the same subgroup was added to more closely mirror the experimental design in Figure 3.

In addition, as recommended in Austin (2010), a treatment to control ratio of 1:2 was implemented. Increasing the number of control units will typically increase bias and reduce precision. This is the bias / variance tradeoff referred to earlier. Austin (2010) found that a ratio of 1:2 improved precision with a slight but not comparable increase in bias. The treatment to control ratio of 1:2 was implemented both without exact matching on group (i.e., Match (1:2)) and with exact matching on group (i.e., ExMatch (1:2)). Replacement was allowed to provide more controls.

The final variation on matching implemented in this study was matching using group propensity scores as described in Green (2014). The group matching method uses nearest neighbor matching with the same parameters as Match 1:1 except for the group indicator. The group matching method uses the group indictor to partition the groups and estimates separate propensity scores for each group.

Subclassification was one of four matching methods implemented to estimate ATE. Five subclasses were used per Cochran (1968). Treatment effects were estimated within quintiles and the overall treatment effect was estimated as a weighted average of the treatment effects within each quintile.

Weighting was described in Section 2.2.2.3 and was implemented for estimates of ATE (i.e. IPTW) and ATT (i.e. Wt x Odds). GBM was also implemented for estimates of ATE and ATT. GBM uses regression trees to find weights that provide the best covariate balance. The stopping parameter specified in the implementation of GBM for this study was “es.mean”. The stopping parameter defines the rules for summarizing across covariates (i.e., mean) and the balance metric used (i.e., effect size or absolute standardized mean difference). Additional model tuning parameters were set as follows.

Interaction depth, which describes the tree complexity or the level of the interactions fitted, was set to two because no condition in this study specified three-way or higher interactions. The maximum number of iterations, n.trees, was set to 3000 and shrinkage was set to 0.01 as described in McCaffrey et al. (2013).

The final method investigated was the optimal 4×1 design recommended in Dong (2015), weighting using the generalized propensity score. MNPS is the multiple treatment version of GBM and was implemented with the same parameters as GBM. All methods evaluated are summarized in Table 6.

Table 6

Summary of Treatment Effect Estimates and Methods Evaluated

Treatment	Method	Specification
	regression	outcome model used for data generating model
ATT	exact matching	1:1 NN caliper 0.2 without replacement
		1:2 NN caliper 0.2 with replacement
	matching	1:1 NN caliper 0.2 without replacement
		1:2 NN caliper 0.2 with replacement
	group PS	1:1 NN caliper 0.2 without replacement
	GBM	interaction.depth=2, shrinkage=0.01, stop=es.mean, n.trees=3000
	Weighting	
ATE	GBM	interaction.depth=2, shrinkage=0.01, stop=es.mean, n.trees=3000
	IPTW	
	MNPS	interaction.depth=2, shrinkage=0.01, stop=es.mean, n.trees=3000
	subclassification	subclass=5

Note. NN = nearest neighbor

Treatment effects were estimated using a weighted regression of a continuous outcome on the treatment indicator, subgroup indicator, and the treatment x subgroup interaction. Doubly robust estimates were not included to better isolate and compare

performance of the methods (Lee et al., 2010). All simulations and analyses were completed in *R*, version 3.2.2 (Ho, Imai, King, & Stuart, 2011; Ridgeway et al., 2015).

3.3 Criteria for Evaluating Results

Performance under the various conditions and estimation methods were evaluated using statistics associated with the treatment effect estimates. Metrics associated with treatment effect estimates included a measure of the difference between the true value of a parameter, θ, and an estimate of the parameter, $\hat{\theta}$, over R replications as shown in Equation 41.

$$\widehat{Bias}\left(\hat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_r - \theta, where\ r = 1, 2, \ldots, R \tag{41}$$

In addition, mean squared error (Equation 42), and the variance of the parameter estimate (Equation 43) were reported. In this study, the majority of the method comparisons use MSE because MSE provides an applied researcher with a method recommendation that represents both bias and variance of the parameter estimates.

$$MSE\left(\hat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \theta\right)^2 \tag{42}$$

$$var\left(\hat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{\theta}_r - \bar{\hat{\theta}}\right]^2 \tag{43}$$

Bias, mean squared error and variance were evaluated for the regression coefficients for the treatment, group, and interaction terms; β_t, β_g, and β_{tg} respectively. The metrics for the coefficients of the group propensity scores were represented in all summary tables as follows. The treatment effect for group 0 was reported as β_t. The difference in the treatment effects between group 0 and group 1 is reported as β_{tg}. Because this study evaluated the robustness in estimating heterogeneous treatment effects, the discussion of

the findings will focus mainly on the metrics (i.e., bias, MSE and variance of estimates) of β_t, β_g, and β_{tg}.

Based on the findings of previous studies that focused on the performance of propensity score methods in the presence of homogeneous treatment effects, several conclusions were expected. First, generalized boosted modeling was expected to produce estimates that are closer to "truth" than methods that use logistic regression to estimate the propensity score. Second, subclassification was expected to outperform the other methods when one group is more prevalent in the extremes of the propensity score because individuals within each strata will be more homogeneous and the impact of extreme weights should be reduced. Third, weighting was expected to perform best when the groups are randomly distributed throughout the propensity score distribution.

Chapter 4: Results

This chapter details relevant outcomes for both simulations. One section is dedicated to each simulation. The chapter ends with an empirical example of the methods under investigation.

4.1 Simulation I

The conditions investigated in simulation I were sample size, propensity score method, and the inclusion, in the propensity score model, of covariates related to either treatment or outcome but not both. These conditions were all investigated in the context of effect moderation by a baseline group characteristic. Also, group membership was simulated as related to selection into treatment. Therefore, the focus of simulation I was to determine the method with the smallest bias and highest precision under these conditions (i.e., research question 1) and whether GBM is more effective than logistic regression under these conditions (i.e., research question 3).

Three sample sizes were explored and were chosen based on previous studies. Shadish (2013) describes propensity score methods as large sample methods and includes references to several studies that indicate that a moderate sample size (≈250) is necessary; however, Shadish goes on to say that design conditions such as covariates included may moderate the sample size needed to reduce bias and that no extensive simulation studies have been published that investigate this issue. This dissertation does not attempt to explore sample size conditions in depth; however, it does illuminate directions for future research that will be reviewed in Chapter 5.

As expected, throughout all simulated models larger sample sizes result in estimates with less variability. Figure 10 illustrates this result for estimates of the

treatment coefficient for model A. For example, across all propensity score methods and models, MSE for a sample size of 1000 (i.e. red triangles) is consistently closer to zero compared with the smaller sample size of 500 (i.e. blue circles).

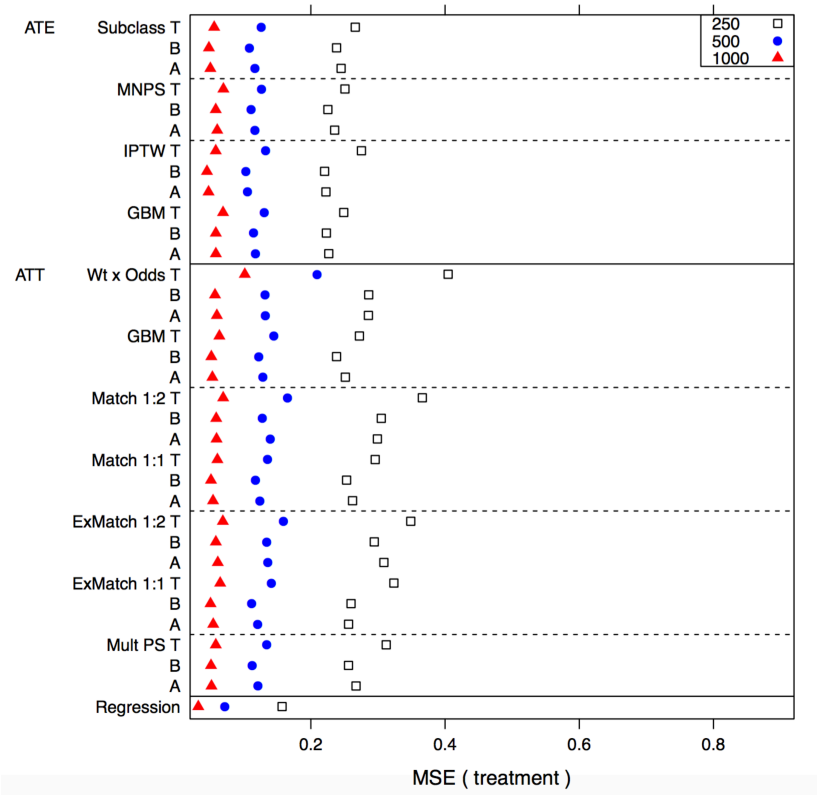


Figure 10. MSE for the coefficient of treatment for data generating model A.

Specific outcomes for each sample size will be reviewed in subsequent sections within the context of each model. The next three sections will present key observations for each of the three true propensity score models in simulation I. Recall that the true propensity score models are those that were used to generate the data. These observations will be organized by first presenting findings related to the explicit inclusion of covariates in the implemented propensity score model and the impact to each coefficient estimate in the model to estimate treatment effects. This will be followed by observations regarding the performance of method categories (e.g., weighting, matching, machine

learning algorithms). Comparisons will be made within the type of treatment effect estimated (ATT or ATE) because researchers would typically be interested in either ATT or ATE but not both. Finally, the optimal design to estimate coefficients of the treatment, group, and interaction terms in the presence of effect moderation by a baseline group characteristic under the conditions simulated will be described. Select results will be highlighted with tables and figures. Detailed results for all models and all outcome measurements are presented in Appendices B through D. Metrics for regression using a correctly specified model are provided as a reference.

For simulation I, recall that three propensity score models were implemented with each propensity score method. The first model, referred to as the "true" model, is the model that was used to generate the data. In addition to the data generating model, two misspecified models were also implemented: misspecified model mA and misspecified model mB. The main differences between models mA and mB are whether all covariates influencing outcome were included and whether all covariates influencing selection for the specified data generating model were included. Specifically, misspecified model mA excluded two covariates, x_7 and x_8, related to outcome and misspecified model mB included all covariates related to outcome. Because, the degree of misspecification for models mA and mB depends on the specific propensity score data generating model, I discuss results for each of the propensity score data generating models (A, B, and C) separately.

4.1.1 Data generating model A. Recall that propensity score data generating model A included two covariates, x_9 and x_{10}, related to selection but not related to outcome and that both misspecified propensity score models mA and mB excluded these

covariates. In addition, misspecified propensity score model mA also excluded two covariates related to the outcome (x_7 and x_8), while misspecified propensity score model mB did not exclude x_7 and x_8.

For this condition, I examined whether previous research regarding the performance of propensity score methods is similar in the presence of heterogeneous treatment effects and when selection differs by group (i.e., complex selection). Table 7 provides the mean MSE across all methods within each implemented propensity score model and shows that estimates using misspecified propensity score model mB (that included all covariates related to outcome) have the smallest MSE compared with estimates using misspecified propensity score model mA. Table 7 also shows that both misspecified models produce estimates with smaller MSE than estimates that use the true propensity score model.

Table 7
MSE Averaged Across All Methods for Data Generating Model A.

	n=250			n=500			n=1000		
Model	β_t	β_{tg}	β_g	β_t	β_{tg}	β_g	β_t	β_{tg}	β_g
true	0.2935	0.5512	0.3521	0.1397	0.2850	0.1807	0.0644	0.1239	0.0851
mA	0.2514	0.4705	0.2788	0.1192	0.2326	0.1393	0.0536	0.1051	0.0699
mB	0.2464	0.4684	0.2786	0.1140	0.2317	0.1397	0.0518	0.1029	0.0687

These results are consistent regardless of sample size and coefficient (i.e., treatment, group, interaction) estimated. They also support previous research that all coefficients related to outcome should be included in the propensity score model regardless of whether they influence selection (Cuong, 2013). In addition, previous research has shown that using an estimate of the propensity score often produces better

estimates of the treatment effect than using the true propensity score (Rosenbaum, 1987). Essentially, regardless of the method and under the conditions simulated, coefficient estimates had the smallest MSE when the propensity score model followed previously established guidelines governing the inclusion of covariates in the propensity score model. Because misspecified propensity score model mB had the lowest MSE for data generating model A, the remaining tables in this section will present results for misspecified model mB only. Results for the true propensity score model and misspecified model mA are in Appendix B.

Figure 11 represents the bias and variance of the estimates for the treatment coefficient. There are three things that stand out in Figure 11. First, matching methods (e.g., Match 1:1, Match 1:2, ExMatch 1:2) produce estimates that are closest to the true parameter on average in contrast to ATE methods, which consistently overestimate the coefficient of treatment. Because matching methods trim or discard units that are not matched, it is likely that extreme values that may shift bias in one direction are eliminated from the estimates.

Second, relative bias estimates of the treatment coefficient for misspecified model mB in Figure 11 range from 1.5% for ExMatch 1:2 to 77% for GBM. At n=250, only two methods, weighting by the odds and ExMatch 1:2, have relative bias values that are within an acceptable range (i.e., ± 10%). At n=1000, the majority of methods have relative bias values that are within an acceptable range (Muthen & Muthen, 2002).

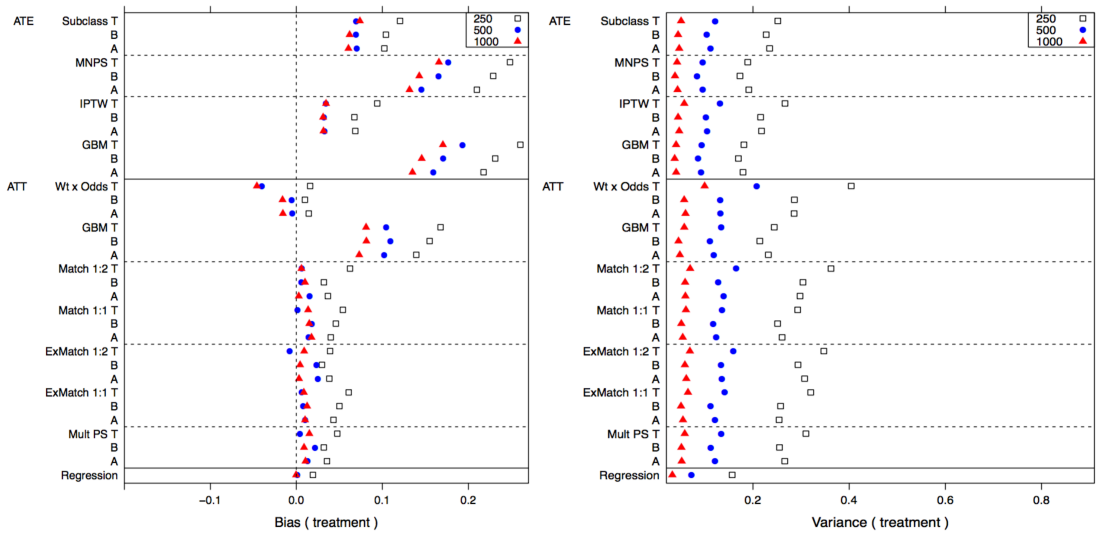


Figure 11. Bias, variance for the coefficient of treatment for data generating model A.

Third, on average, the variability of the ATE estimates for the treatment coefficient estimates is smaller than the ATT estimates. For example, for the n=250 condition, the mean sampling variance for the ATE estimates was 0.208 (SD=0.032), while the mean for the ATT estimates was 0.288 (SD=0.045). This difference was found at the other sample sizes as well. The smaller variability of the ATE estimates is also found for the group and interaction coefficient estimates. This result was expected as ATT estimates depend on who is treated in addition to the distribution of the propensity scores (Angrist, 2004).

The lower variability is also likely explained by the implicit inclusion or exclusion of individuals based on the method and the small number of outliers at the lower extreme of the propensity score distribution. For example, matching methods trim participants (i.e., reducing the sample size) compared with the ATE methods used, which retain all individuals. Also, outliers at both extremes of the propensity score distribution influence the variance of ATE estimates as opposed to outliers at the upper end of the propensity score distribution for ATT estimates (i.e., control units with a propensity score

close to 1) (Steiner & Cook, 2013). Figure 12 shows that under the conditions simulated, there are few outliers at the lower extreme of the propensity score distribution that would contribute to increased variance of the ATE estimates.

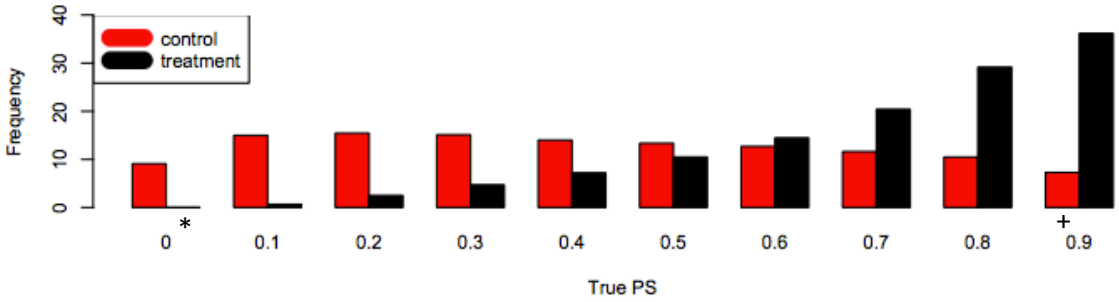


Figure 12. Distribution of propensity scores, data generating model A, n=250.
Note. * = influence ATE estimates; + = influence ATT and ATE estimates.

There are two additional observations that can be made related to the general performance of propensity score methods. First, propensity scores estimated using machine learning algorithms such as GBM and MNPS are expected to produce less biased treatment effect estimates than propensity scores estimated using logistic regression in the presence of nonlinearity and nonadditivity (Setoguchi et al., 2008). Recall that nonlinearity and nonadditivity were not included in the true propensity score design for data generating model A. Table 8 shows that for the ATE estimates of the coefficient of treatment, methods that used logistic regression to estimate propensity scores (i.e., IPTW and subclassification) were less biased on average than methods that used machine learning algorithms (i.e., GBM and MNPS).

Table 8 also shows that the estimates for the coefficient of the group and interaction terms for the machine learning algorithms had the smallest bias and variance compared with the other methods. These conflicting results were unexpected and are most likely explained by the implementation of logistic regression compared with

machine learning algorithms. Machine learning algorithms iteratively search for the weights that optimize balance across all covariates. For logistic regression, the researcher specifies the model. Although covariate balance is checked at each iteration in both machine learning algorithms and logistic regression, a researcher may stop at acceptable covariate balance; whereas, machine learning algorithms search for optimal balance. Section 4.3 provides an example of the improvement in covariate balance using methods that use machine learning algorithms to estimate the propensity score compared with methods that use logistic regression to estimate the propensity score.

Table 8

Metrics for Data Generating Model A, ATE Estimates, Misspecified Model mB.

		MSE		Bias		Variance	
Coefficient	Method	n=250	n=500	n=250	n=500	n=250	n=500
β_t	Subclass	0.238	0.109	0.104	0.069	0.227	0.104
	MNPS	0.225	0.111	0.229	0.165	0.173	0.084
	IPTW	0.220	0.103	0.067	0.032	0.216	0.102
	GBM	0.223	0.115	0.231	0.171	0.170	0.086
β_{tg}	Subclass	0.491	0.213	-0.073	-0.041	0.485	0.211
	MNPS	0.343	0.171	-0.068	-0.039	0.339	0.169
	IPTW	0.487	0.270	-0.124	-0.089	0.472	0.262
	GBM	0.364	0.191	-0.074	-0.055	0.358	0.188
B_g	Subclass	0.262	0.145	-0.137	-0.165	0.243	0.118
	MNPS	0.172	0.085	0.010	0.005	0.172	0.085
	IPTW	0.262	0.144	0.063	0.057	0.258	0.141
	GBM	0.182	0.094	0.008	0.008	0.182	0.093

Note. Metrics for n=1000 are in Appendix B.

Table 9 provides the MSE, bias, and variance of the ATT estimates. GBM is the only method in Table 9 that does not use logistic regression to estimate the propensity scores. Across all coefficients, Table 9 shows that methods that use logistic regression have less biased estimates on average than GBM. The results in Table 8 and Table 9

suggest that in the presence of effect moderation, the performance of methods that use logistic regression to estimate propensity scores compared with methods that use machine learning algorithms is consistent with previous research regarding bias except for ATE estimates of the coefficients of the group and interaction terms.

Table 9

Metrics for Data Generating Model A, ATT Estimates, Misspecified Model mB.

		MSE		Bias		Variance	
Coefficient	Method	n=250	n=500	n=250	n=500	n=250	n=500
β_t	Wt x Odds	0.286	0.132	0.010	-0.006	0.286	0.132
	GBM	0.238	0.122	0.155	0.109	0.214	0.110
	Match 1:2	0.305	0.128	0.032	0.006	0.304	0.128
	Match 1:1	0.253	0.118	0.046	0.018	0.251	0.117
	ExMatch	0.294	0.134	0.030	0.023	0.294	0.134
	ExMatch	0.260	0.112	0.050	0.008	0.257	0.112
	Group	0.256	0.113	0.032	0.022	0.255	0.112
	Regression	0.157	0.072	0.019	0.001	0.157	0.072
β_{tg}	Wt x Odds	0.634	0.358	-0.094	-0.088	0.625	0.351
	GBM	0.471	0.258	-0.077	-0.068	0.465	0.253
	Match 1:2	0.612	0.295	-0.063	-0.029	0.608	0.295
	Match 1:1	0.543	0.251	-0.069	-0.019	0.538	0.251
	ExMatch	0.610	0.291	-0.071	-0.041	0.605	0.290
	ExMatch	0.548	0.244	-0.070	-0.023	0.543	0.243
	Group	0.248	0.113	-0.003	-0.008	0.248	0.112
	Regression	0.269	0.126	-0.021	-0.017	0.268	0.126
β_g	Wt x Odds	0.462	0.252	0.032	0.029	0.461	0.251
	GBM	0.303	0.165	0.014	0.008	0.303	0.165
	Match 1:2	0.433	0.203	-0.010	-0.035	0.433	0.201
	Match 1:1	0.282	0.124	-0.048	-0.071	0.279	0.119
	ExMatch	0.423	0.205	0.015	-0.011	0.422	0.205
	ExMatch	0.286	0.133	-0.046	-0.079	0.284	0.127
	Group	0.138	0.062	-0.013	-0.023	0.138	0.061
	Regression	0.137	0.065	0.008	0.008	0.137	0.065

Note. Metrics for n=1000 are in Appendix B.

The second observation is the impact of increasing the number of controls in the matching methods. As previously mentioned, Austin (2010) found that increasing the number of control units typically increases bias and reduces variance. This bias/variance tradeoff was not evident in the conditions simulated. For example, Table 9 shows that a matching ratio of 1:1 had both smaller variance and larger bias than methods with a matching ratio of 1:2. The Austin simulation study used smaller proportions of treated individuals (e.g., 0.15) compared with the proportion of treated individuals in this study (0.50). Matching was also done without replacement in the Austin study.

The next few paragraphs shift the discussion from findings related to expectations based on previous research to findings related to the specific performance of group centric methods compared with methods that do not focus on group information. It was hypothesized that methods that emphasize group information in their design, such as MNPS, exact matching on group, and group propensity scores would perform better than methods that do not provide the same focus on group information (Dong, 2015; Eeren et al., 2015; Green & Stuart, 2014). This expectation was realized to a limited degree.

For example, Table 8 shows that although MNPS had the smallest MSE for ATE estimates of the group and interaction coefficients, the performance of GBM was comparable (i.e., within 0.02 or better). Essentially, under the conditions simulated MSE, bias and variance estimates across all coefficients for both MNPS and GBM were comparable. This suggests that GBM might be used, with similar effect, in situations where group size limits the use of MNPS. Although the use of IPTW produced estimates of the treatment coefficient with the smallest bias, the MSE for IPTW was only at most 0.012 better than MNPS or GBM.

For ATT estimates, Table 9 shows that under the conditions simulated there is no appreciable improvement in the bias or precision of the estimates that include exact matching on group in the propensity model design compared with matching. Again although under some conditions (e.g., estimates of the coefficient of treatment, n=500) exact matching (1:1) had the smallest MSE, matching (1:1) produced estimates with MSE within 0.006 of the MSE estimates for exact matching (1:1). Finally, MSE for estimates of the coefficient of group and interaction using group propensity scores was nearly 46% smaller on average than the next best method.

In summary, the conditions simulated in data generating model A show that the performance of propensity score methods in the presence of heterogeneous treatment effects and where selection differs by group is consistent with previous research regarding bias with the exception of matching methods with a ratio of 1:2 and ATE estimates of the group and interaction coefficients. For the methods used to estimate ATE, MNPS was optimal for estimates of the interaction and group coefficients and comparable to the IPTW for estimating the coefficient of treatment. For methods used to estimate ATT, estimating a propensity score for each group provided the best overall performance across all sample sizes and coefficients.

4.1.2 Data generating model B. Data generating model B included only covariates related to both selection and outcome whereas the results discussed in the prior section addressed propensity score data generating model A where two covariates informed the treatment selection but were not related to the outcome. The findings related to the effect on the treatment coefficient of excluding covariates that influence outcome from the propensity score model were similar to the observations found in data

generating model A. Essentially, Table 10 shows that the propensity score model that excludes covariates related to outcome (i.e., misspecified propensity score model mA) performs the worst in terms of MSE for the estimates of the treatment coefficient. This result is consistent regardless of sample size and supports previous research that shows that all coefficients related to outcome should be included in the propensity score model (Cuong, 2013). Because misspecified propensity score model mB had the lowest MSE for the treatment coefficient for data generating model B, the remaining tables in this section will present results for misspecified model mB only. Results for the true propensity score model and misspecified model mA can be found in Appendix C.

Table 10
MSE Averaged Across All Methods for Data Generating Model B.

	n=250			n=500			n=1000		
Model	β_t	β_{tg}	β_g	β_t	β_{tg}	β_g	β_t	β_{tg}	β_g
True	0.2797	0.5308	0.3299	0.1509	0.2794	0.1840	0.0758	0.1294	0.0891
mA	0.2906	0.4623	0.2739	0.1798	0.2470	0.1518	0.1076	0.1136	0.0712
mB	0.2796	0.5258	0.3297	0.1495	0.2790	0.1838	0.0747	0.1301	0.0895

The three patterns found in the results regarding data generating model A were also present in the results for data generating model B. First, the lower variability of the ATE estimates (M=0.202, SD=0.03, n=250) as compared with the ATT estimates (M=0.284, SD=0.042, n=250) for the treatment coefficient as well as the group and interaction coefficients were similar to those found in data generating model A. Second, Tables 11 and 12 show that treatment effect estimates obtained from methods that use logistic regression to estimate the propensity score are less biased than machine learning algorithms except for ATE estimates of the group and interaction coefficients. Finally,

Table 12 shows that a matching ratio of 1:2 data from data generating model B did not produce estimates with less variability than methods using a matching ratio of 1:1.

Table 11

Metrics for Data Generating Model B, ATE Estimates, Misspecified Model mB.

		MSE		Bias		Variance	
Coefficient	Method	n=250	n=500	n=250	n=500	n=250	n=500
β_t	Subclass	0.243	0.132	0.127	0.103	0.227	0.121
	MNPS	0.275	0.164	0.320	0.267	0.173	0.093
	IPTW	0.252	0.138	0.084	0.059	0.245	0.135
	GBM	0.286	0.173	0.336	0.289	0.173	0.090
β_{tg}	Subclass	0.485	0.245	-0.045	-0.028	0.483	0.244
	MNPS	0.346	0.180	-0.036	-0.049	0.345	0.178
	IPTW	0.638	0.370	-0.152	-0.146	0.615	0.349
	GBM	0.370	0.202	-0.058	-0.083	0.367	0.195
B_g	Subclass	0.269	0.177	-0.177	-0.199	0.238	0.138
	MNPS	0.163	0.089	-0.020	-0.003	0.162	0.089
	IPTW	0.372	0.224	0.087	0.080	0.364	0.217
	GBM	0.183	0.103	-0.016	0.001	0.183	0.103

Note. Metrics for n=1000 are in Appendix C.

Additional similarities were found when evaluating the performance of group-focused methods. As with data generating model A, Table 11 shows that MNPS was not optimal for estimating the treatment coefficient. Methods that use logistic regression to estimate the propensity score were clearly optimal across all sample sizes for ATE estimates of the treatment coefficient and MNPS was optimal for estimates of the interaction and group coefficients. For ATT, GBM had the lowest MSE for the treatment coefficient at n=250. However, at larger sample sizes, 1:1 matching methods and group propensity scores performed best. Also, no significant improvement was found in the bias or precision of the estimates that include exact matching on group in the propensity

model design compared with matching. Overall, estimating a propensity score for each group provided the best overall performance for methods used to estimate ATT.

Table 12

Metrics for Data Generating Model B, ATT Estimates, Misspecified Model mB.

		MSE		Bias		Variance	
Coefficient	Method	n=250	n=500	n=250	n=500	n=250	n=500
β_t	Wt x Odds	0.361	0.222	-0.015	-0.055	0.361	0.219
	GBM	0.260	0.149	0.196	0.134	0.221	0.131
	Match 1:2	0.329	0.163	0.004	0.001	0.329	0.163
	Match 1:1	0.283	0.137	0.018	0.010	0.282	0.137
	ExMatch $_{1:2}$	0.331	0.168	0.012	-0.006	0.330	0.168
	ExMatch $_{1:1}$	0.277	0.132	0.023	0.009	0.276	0.132
	Group	0.295	0.134	0.012	0.006	0.294	0.134
	Regression	0.166	0.080	-0.007	-0.008	0.166	0.080
β_{tg}	Wt x Odds	0.913	0.554	-0.105	-0.100	0.902	0.544
	GBM	0.513	0.298	-0.055	-0.074	0.510	0.292
	Match 1:2	0.674	0.357	0.009	-0.016	0.674	0.356
	Match 1:1	0.567	0.278	0.000	-0.014	0.567	0.278
	ExMatch $_{1:2}$	0.677	0.339	0.009	-0.013	0.677	0.339
	ExMatch $_{1:1}$	0.576	0.281	0.000	-0.019	0.576	0.281
	Group	0.265	0.116	0.034	0.001	0.264	0.116
	Regression	0.284	0.128	0.022	-0.005	0.284	0.128
β_g	Wt x Odds	0.720	0.445	0.038	0.018	0.718	0.445
	GBM	0.342	0.204	-0.012	-0.008	0.342	0.204
	Match 1:2	0.515	0.272	-0.087	-0.069	0.507	0.267
	Match 1:1	0.295	0.151	-0.118	-0.108	0.281	0.140
	ExMatch $_{1:2}$	0.520	0.257	-0.058	-0.060	0.517	0.253
	ExMatch $_{1:1}$	0.300	0.153	-0.108	-0.109	0.288	0.141
	Group	0.144	0.062	-0.027	-0.027	0.143	0.061
	Regression	0.135	0.068	-0.012	0.003	0.135	0.068

Note. Metrics for n=1000 are in Appendix C.

4.1.3 Data generating model C. Data generating model C also included only those covariates related to outcome. The main condition of interest for data generating

model C was the inclusion of nonlinearity and nonadditivity. Table 13 shows that the true propensity score model performed the worst in terms of bias and MSE of the resulting treatment effect estimates. This result is consistent regardless of sample size and supports previous research that using an estimate of the propensity score often produces better estimates of the treatment effect than using the true propensity score (Rosenbaum, 1987). Because misspecified propensity score model mA has the lowest MSE for data generating model C, the remaining tables in this section will present results for misspecified model mA only. Results for the true propensity score model and misspecified model mB can be found in Appendix D.

Table 13
MSE Averaged Across All Methods for Data Generating Model C.

	n=250			n=500			n=1000		
Model	β_t	β_{tg}	β_g	β_t	β_{tg}	β_g	β_t	β_{tg}	β_g
true	0.2451	0.4984	0.3089	0.1349	0.2445	0.1609	0.0688	0.1300	0.0866
mA	0.2006	0.4319	0.2442	0.1105	0.2019	0.1179	0.0488	0.0987	0.0588
mB	0.2170	0.4562	0.2722	0.1134	0.2141	0.1304	0.0516	0.1060	0.0654

Many of the results for data generating model C are similar to those found in data generating models A and B. Specifically, the observations regarding the variability of the ATE estimates compared with the ATT estimates, the central location of bias, comparison of matching ratios, and the performance of group-focused methods were similar to those found in data generating models A and B. Tables 14 and 15 show that the performance of methods that use logistic regression to estimate propensity scores compared with methods that use machine learning algorithms was also similar. This was unexpected considering the nonadditivity and nonlinearity simulated in data generating

model C and suggests that the level of nonlinearity simulated in data generating model C was not large enough to require the use of machine learning algorithms when estimating the coefficient of treatment.

Table 14
Metrics for Data Generating Model C, ATE Estimates, Misspecified Model mA.

		MSE		Bias		Variance	
Coefficient	Method	n=250	n=500	n=250	n=500	n=250	n=500
β_t	Subclass	0.196	0.104	0.012	0.045	0.196	0.102
	MNPS	0.184	0.104	0.103	0.111	0.173	0.092
	IPTW	0.178	0.096	-0.005	0.022	0.178	0.096
	GBM	0.170	0.103	0.111	0.117	0.158	0.090
β_{tg}	Subclass	0.441	0.199	0.000	-0.038	0.441	0.197
	MNPS	0.380	0.178	-0.003	-0.025	0.380	0.178
	IPTW	0.431	0.200	-0.047	-0.071	0.429	0.195
	GBM	0.369	0.182	-0.027	-0.050	0.368	0.180
B_g	Subclass	0.238	0.120	-0.112	-0.094	0.226	0.111
	MNPS	0.192	0.093	-0.004	0.005	0.192	0.093
	IPTW	0.214	0.101	0.020	0.029	0.214	0.100
	GBM	0.188	0.091	0.004	0.014	0.188	0.091

Note. Metrics for n=1000 are in Appendix D.

Table 15
Metrics for Data Generating Model C, ATT Estimates, Misspecified Model mA.

		MSE		Bias		Variance	
Coefficient	Method	n=250	n=500	n=250	n=500	n=250	n=500
β_t	Wt x Odds	0.213	0.114	0.002	0.032	0.213	0.113
	GBM	0.202	0.122	0.073	0.082	0.197	0.116
	Match 1:2	0.246	0.134	0.004	0.040	0.246	0.133
	Match 1:1	0.210	0.111	-0.005	0.038	0.210	0.110
	ExMatch 1:2	0.241	0.133	0.002	0.032	0.241	0.132
	ExMatch 1:1	0.215	0.116	-0.001	0.018	0.215	0.115
	Group	0.217	0.114	-0.011	0.019	0.217	0.113
	Regression	0.134	0.074	-0.028	0.004	0.133	0.074
β_{tg}	Wt x Odds	0.503	0.232	-0.022	-0.054	0.503	0.229
	GBM	0.476	0.230	-0.018	-0.047	0.475	0.228
	Match 1:2	0.542	0.272	0.006	-0.041	0.542	0.270
	Match 1:1	0.505	0.221	-0.010	-0.047	0.505	0.219
	ExMatch 1:2	0.536	0.252	0.013	-0.016	0.536	0.251
	ExMatch 1:1	0.489	0.221	0.002	-0.013	0.489	0.221
	Group	0.232	0.101	-0.008	0.001	0.232	0.101
	Regression	0.279	0.133	0.022	-0.003	0.278	0.133
β_g	Wt x Odds	0.326	0.150	0.002	0.015	0.326	0.149
	GBM	0.303	0.147	-0.002	0.008	0.303	0.147
	Match 1:2	0.358	0.182	-0.036	-0.003	0.357	0.182
	Match 1:1	0.239	0.118	-0.035	-0.023	0.238	0.118
	ExMatch 1:2	0.360	0.166	-0.032	-0.019	0.359	0.166
	ExMatch 1:1	0.252	0.118	-0.042	-0.042	0.250	0.116
	Group	0.117	0.058	-0.023	-0.035	0.116	0.057
	Regression	0.141	0.070	-0.011	0.001	0.141	0.070

Note. Metrics for n=1000 are in Appendix D.

4.2 Simulation II

Simulation II was designed to explore the performance of propensity score methods when one group is more prevalent in the extremes of the propensity score distribution. In simulation I, group membership was simulated uniformly throughout the

propensity score distribution. This condition was repeated in simulation II as the baseline comparison. Recall that three additional group conditions were also simulated based on the cumulative probability distribution of the true propensity score. Instead of a uniform group proportion of 0.5 throughout, group proportions were simulated as 0.9 in the upper extreme of the propensity score distribution and 0.4 elsewhere. The upper extreme of the propensity score distribution was determined by cumulative probability cutoff values of 0.9, 0.8, and 0.7. This changed the overall group membership as outlined in Table 16.

Table 16
Group Proportions for Simulation II.

	Percentile Condition			
Group	Baseline	70	80	90
0/1	50/50	45/55	50/50	55/45

Frequencies related to treatment and group assignment by model in the upper portion of the propensity score distribution are presented in Tables 17 (for data generating model D) and 18 (for data generating model E). Because the proportion of treated and control units remained fixed at 0.5, the nonuniform group distribution increased the number of treated units in the upper portions of the propensity score distribution for group 1 and decreased the number of treated units in the upper portions of the propensity score distribution for group 0. For example, the number treated in group 1 with propensity score 90 or greater increased from 9 to 17 for the 70^{th} percentile condition. The increase in the number treated was exacerbated in model E because of nonlinearity.

Table 17
Frequencies in the Upper Percentiles of the Propensity Score Distribution, Model D, n=250.

				Percentile Condition		
PS Range	Treatment	Group	Baseline	70	80	90
70-79	treatment	0	12	7	14	14
		1	11	16	9	9
	control	0	5	3	6	6
		1	5	7	4	4
80-89	treatment	0	13	3	5	14
		1	13	24	22	13
	control	0	4	1	1	4
		1	4	6	6	3
90-99	treatment	0	9	2	2	2
		1	9	17	17	16
	control	0	1	0	0	0
		1	1	2	2	2

Note. PS = propensity score.

Table 18
Frequencies in the Upper Percentiles of the Propensity Score Distribution, Model E, n=250.

				Percentile Condition		
PS Range	Treatment	Group	Baseline	70	80	90
70-79	treatment	0	10	11	11	11
		1	10	8	8	8
	control	0	6	7	7	7
		1	6	5	5	5
80-89	treatment	0	14	9	16	16
		1	14	19	11	11
	control	0	6	4	7	7
		1	6	7	5	4
90-99	treatment	0	22	4	6	16
		1	22	39	38	28
	control	0	5	1	2	4
		1	5	9	8	6

Note. PS = propensity score.

In addition to exploring the performance of propensity score methods when one group is more prevalent in the extremes of the propensity score distribution, the models for simulation II were chosen to determine whether the covariate indicating group membership should be in the propensity score model when group membership does not influence selection. Recall that misspecified propensity score model mA included the group indicator and was used in simulation II to provide a comparison to the true data generating models D and E which did not include the group indicator in the propensity score model.

The next two sections describe the results within the context of each data generating propensity score model. Select outcomes are highlighted with figures, and detailed results for all models and outcome measurements are presented in Appendices E and F. Also, metrics for regression using a correctly specified model are provided as a reference.

4.2.1 Data generating model D. The first question under investigation is whether the covariate indicating group membership should be in the propensity score model when group membership does not influence selection. Recall that propensity score data generating model D did not include the group indicator and misspecified propensity score model mA did include the group indicator. Figure 13 compares the absolute value of the differences between coefficient estimates using propensity score model mA and data generating model D. Figure 13 shows that at a sample size of 250, 60 of the 144 total coefficient and method combinations simulated had smaller average MSE when the grouping indicator was included in the propensity score model. The improvement in average MSE over a model that excluded the grouping indicator ranged from 0 to 0.05

(M= 0.006; SD=0.008). Eighty-four of the total combinations simulated had larger average MSE when the grouping indicator was included in the propensity score model. These differences ranged from 0 to 0.038 (M=0.005; SD=0.007). There was no pattern found in the method and coefficient combinations that had lower average MSE when the grouping indicator was included in the propensity score model. These results were consistent across sample sizes and show that under the conditions simulated in data generating model D including the group indicator in the propensity score model did not provide any appreciable advantage or disadvantage.

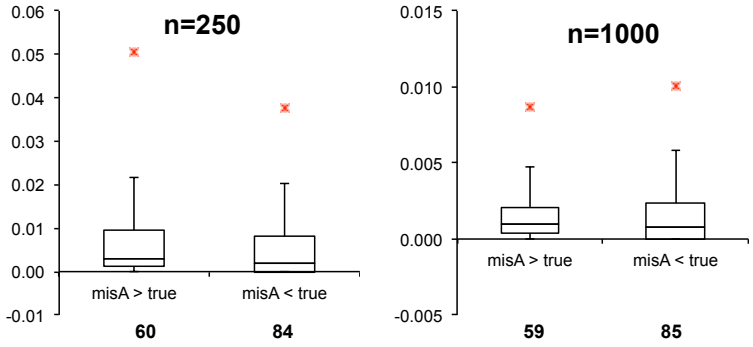


Figure 13. MSE improvement for data generating model D. × is an outlier defined as outside 1.5 times the interquartile range.

The second question under investigation is which methods perform best when one group is more prevalent in the extremes of the propensity score distribution. Tables 19 and 20 provide the metrics for estimates of ATE and ATT for data generating model D and a sample size of 250. Metrics for larger sample sizes are provided in Appendix E and are not discussed here because of their relative similarity.

Table 19 shows that methods that use logistic regression to estimate the propensity score (i.e., subclassification and IPTW) provided estimates with smaller bias on average compared with GBM and MNPS for all percentile conditions and all

coefficients. However, methods that use logistic regression to estimate the propensity score were less precise than GBM and MNPS. For example on average for the baseline condition, GBM overestimated the coefficient of treatment by 0.1857 compared with an overestimate of 0.0579 using subclassification. The "tradeoff" is that the variance of the estimate of the coefficient of treatment using subclassification compared with GBM was higher on average (≈ +0.03).

Table 19
Metrics for Data Generating Model D, ATE Estimates, Misspecified Model mA, n=250.

			Percentile Condition			
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.2132	0.2471	0.2025	0.1946
		MNPS	0.2213	0.2123	0.1909	0.1847
		IPTW	0.2439	0.2238	0.1995	0.1903
		GBM	0.2139	0.1983	0.1817	0.1790
	β_{tg}	Subclass	0.4628	0.5450	0.4860	0.4669
		MNPS	0.3690	0.3679	0.3463	0.3810
		IPTW	0.5247	0.5210	0.5433	0.5379
		GBM	0.3711	0.3724	0.3508	0.3945
	β_g	Subclass	0.2209	0.2544	0.2547	0.2471
		MNPS	0.1808	0.2060	0.1971	0.2042
		IPTW	0.2873	0.3817	0.4165	0.3273
		GBM	0.1855	0.2287	0.2304	0.2145
Bias	β_t	Subclass	0.0579	0.0724	0.0698	0.0832
		MNPS	0.1884	0.1562	0.1723	0.1935
		IPTW	0.0047	0.0089	0.0231	0.0324
		GBM	0.1857	0.1278	0.1486	0.1685
	β_{tg}	Subclass	0.0020	-0.0012	-0.0450	-0.0187
		MNPS	-0.0080	0.0485	0.0082	-0.0001
		IPTW	-0.0032	0.0010	-0.0715	-0.0228
		GBM	0.0040	0.0855	0.0487	0.0680
	β_g	Subclass	0.0019	-0.0010	0.0546	0.0258
		MNPS	0.0119	0.1073	0.1147	0.0516

		IPTW	0.0107	0.2840	0.2911	0.1481
		GBM	0.0031	0.1979	0.1895	0.0776
Variance	β_t	Subclass	0.2098	0.2418	0.1976	0.1877
		MNPS	0.1858	0.1880	0.1612	0.1472
		IPTW	0.2439	0.2237	0.1989	0.1892
		GBM	0.1794	0.1820	0.1596	0.1506
	β_{tg}	Subclass	0.4628	0.5450	0.4839	0.4665
		MNPS	0.3690	0.3656	0.3463	0.3810
		IPTW	0.5246	0.5210	0.5382	0.5374
		GBM	0.3710	0.3651	0.3484	0.3899
	β_g	Subclass	0.2209	0.2544	0.2517	0.2465
		MNPS	0.1806	0.1945	0.1839	0.2016
		IPTW	0.2872	0.3010	0.3318	0.3054
		GBM	0.1855	0.1896	0.1945	0.2085

Note. Metrics for n=500 and n=1000 are in Appendix E.

The performance of subclassification for the coefficient of treatment decreased more than the other methods as the group imbalance increased. For example, subclassification and GBM had the lowest MSE, 0.2132 and 0.2139, respectively, when the subgroups were evenly distributed (i.e., 50/50). This result shifts as the difference between the proportion of treated units in group 1 and group 0 in the upper end of the propensity score distribution increases. Essentially, subclassification shifts from having the lowest MSE to the highest MSE of the four methods when the difference between the number of treated units in group 1 and group 0 is largest at 70%. However, GBM remained optimal at the 70% condition for the coefficient of treatment.

Although MNPS was not the optimal method for estimating the coefficient of treatment, MNPS had the lowest MSE across all sample sizes and balance conditions for the estimates of the coefficients of the group and interaction terms. Table 19 shows that the bias of the estimates using MNPS was comparable (i.e., within 0.047 or better) to the

bias of the estimates using logistic regression to estimate the propensity score. However, the variance of the MNPS estimates was 30% smaller on average than the variance of the weighting methods. Although GBM did not have the lowest MSE for the estimates of the interaction and group coefficients, bias and variance estimates were similar to MNPS (within 0.02). Subclassification had the same pattern of performance found in the estimates of the treatment coefficient.

These results suggest and support previous research that subclassification is not ideal when group proportions are unbalanced (Pan & Bai, 2015). Under the conditions simulated for data generating model D, GBM was the best method for the coefficient of treatment and MNPS was optimal for the interaction and group coefficients.

Next, methods that estimate the ATT are compared. Table 20 provides the MSE for estimates of ATT for data generating model D and a sample size of 250. Bias and variance for a sample size of 250 are provided in Appendix E and are not discussed because the relative magnitude of bias and variance for each method is similar to that found in previous models already reviewed. Metrics for larger sample sizes are provided in Appendix E as well and are also not reviewed because of their similarity.

Table 20 shows that GBM had the lowest MSE for the estimate of the treatment coefficient when the groups were the most unbalanced (i.e., 70%). Group propensity scores performed best when the groups were balanced. As with simulation I, group propensity scores had significantly lower MSE compared with the other methods when estimating the interaction and group coefficients. Also, no advantage was found for using exact matching on group.

Table 20
MSE for Data Generating Model D, ATT Estimates, Misspecified Model mA, n=250.

			Percentile Condition		
Coefficient	Method	Baseline	70	80	90
β_t	Wt x Odds	0.3312	0.2620	0.2490	0.2335
	GBM	0.2766	0.2287	0.2136	0.2126
	Match 1:2	0.3143	0.2852	0.2724	0.2445
	Match 1:1	0.2661	0.2403	0.2391	0.2230
	ExMatch 1:2	0.3114	0.2671	0.2549	0.2329
	ExMatch 1:1	0.2648	0.2583	0.2174	0.2144
	Group	0.2446	0.2704	0.2247	0.2293
	Regression	0.1498	0.1582	0.1307	0.1287
β_{tg}	Wt x Odds	0.6992	0.6955	0.7259	0.7065
	GBM	0.5293	0.4861	0.4929	0.5480
	Match 1:2	0.6386	0.6029	0.6188	0.6675
	Match 1:1	0.5555	0.4993	0.5358	0.5532
	ExMatch 1:2	0.6346	0.5725	0.6358	0.6586
	ExMatch 1:1	0.5476	0.5515	0.5237	0.5970
	Group	0.2602	0.2455	0.2946	0.3431
	Regression	0.2658	0.2820	0.2618	0.2913
β_g	Wt x Odds	0.5435	0.6247	0.6935	0.5967
	GBM	0.3666	0.3811	0.4136	0.4050
	Match 1:2	0.4774	0.5094	0.5396	0.5493
	Match 1:1	0.2593	0.2901	0.2884	0.2931
	ExMatch 1:2	0.4505	0.5043	0.5819	0.5551
	ExMatch 1:1	0.2723	0.3053	0.2947	0.3157
	Group	0.1320	0.1479	0.1650	0.1824
	Regression	0.1317	0.1490	0.1404	0.1463

Note. Bias, variance for n=250 and all metrics for n=500 and n=1000 are in Appendix E.

4.2.2 Data generating model E. Recall that data generating model E included nonlinearity, which caused an increased difference between the proportion of treated units in group 1 compared with group 0 at the upper percentiles of the propensity score distribution. Figure 14 compares the absolute value of the differences between estimates

using a propensity score model that included the grouping indicator and estimates without the grouping indicator. The number of combinations with smaller average MSE with the grouping indicator included in the propensity score model increased to 91 (M=0.11, SD=0.12) as compared to 53 (M=0.001, SD=0.02) when no nonlinearity was included. This comparison essentially shows more improvement in average MSE over a model without the grouping indicator than was evident in data generating model D. This trend suggests that under the conditions simulated the grouping indicator should be included regardless of whether it influences selection and particularly if there are differences or imbalances in the groups. These results were consistent across sample sizes and show that under the conditions simulated, in data generating model E including the group indicator in the propensity score model provided an improvement in average MSE for the majority of the estimates. As with data generating model D, no coefficient, method or combinations thereof were found that would indicate including the grouping indicator was important in a specific method or in estimating a specific coefficient.

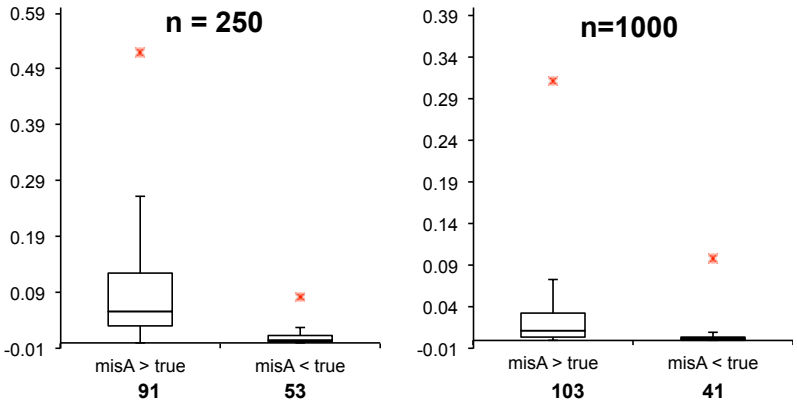


Figure 14. MSE improvement for data generating model E. × is an outlier defined as outside 1.5 times the interquartile range.

Table 21 provides MSE for estimates of ATE for data generating model E. There are a few interesting comparisons in the results for data generating model E compared with data generating model D. First, although subclassification produced estimates of the treatment coefficient that were less biased than GBM and MNPS for data generating model D, subclassification produced estimates of the coefficient of the treatment effect with the largest bias in data generating model E. Also, similar to the results found in data generating model D, GBM had the lowest MSE for the coefficient of treatment and MNPS was optimal for estimates of the group and interaction coefficients. These findings again support previous research that subclassification is not optimal when groups are imbalanced (Pan & Bai, 2015).

Table 21
Metrics for Data Generating Model E, ATE Estimates, Misspecified Model mA, n=250.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.2541	0.2875	0.2854	0.2252
		MNPS	0.1950	0.2129	0.1834	0.1743
		IPTW	0.2092	0.2127	0.1961	0.1644
		GBM	0.1943	0.1906	0.1794	0.1599
	β_{tg}	Subclass	0.4113	0.5691	0.4935	0.3959
		MNPS	0.3284	0.3451	0.3739	0.3327
		IPTW	0.4329	0.4212	0.4332	0.3568
		GBM	0.3424	0.3430	0.3838	0.2985
	β_g	Subclass	0.2543	0.3459	0.3023	0.2308
		MNPS	0.1713	0.1826	0.1903	0.1688
		IPTW	0.2218	0.2375	0.2525	0.2081
		GBM	0.1792	0.2029	0.2066	0.1655
Bias	β_t	Subclass	-0.2369	-0.2358	-0.2588	-0.2383
		MNPS	0.1439	0.1335	0.1005	0.1515
		IPTW	-0.0017	-0.0219	-0.0472	-0.0017
		GBM	0.1324	0.0823	0.0645	0.1179

	β_{tg}	Subclass	-0.0069	-0.0060	0.0118	-0.0317
		MNPS	-0.0199	-0.0054	0.0179	-0.0493
		IPTW	-0.0169	0.0456	0.0635	-0.0233
		GBM	-0.0115	0.0630	0.0963	0.0185
	β_g	Subclass	-0.1972	-0.3000	-0.2400	-0.1823
		MNPS	0.0208	0.0719	0.0561	0.0676
		IPTW	0.0145	0.1832	0.1562	0.1437
		GBM	0.0119	0.1505	0.1161	0.1056
Variance	β_t	Subclass	0.1980	0.2319	0.2184	0.1684
		MNPS	0.1743	0.1951	0.1733	0.1513
		IPTW	0.2092	0.2123	0.1939	0.1644
		GBM	0.1768	0.1838	0.1752	0.1460
	β_{tg}	Subclass	0.4112	0.5691	0.4934	0.3949
		MNPS	0.3280	0.3450	0.3736	0.3302
		IPTW	0.4326	0.4191	0.4291	0.3562
		GBM	0.3422	0.3390	0.3745	0.2982
	β_g	Subclass	0.2154	0.2559	0.2447	0.1975
		MNPS	0.1709	0.1774	0.1872	0.1642
		IPTW	0.2216	0.2039	0.2281	0.1874
		GBM	0.1790	0.1803	0.1931	0.1543

Note. Metrics for n=500 and n=1000 are in Appendix F.

Next, methods that estimate ATT are compared. Table 22 provides the MSE for estimates of ATT for data generating model E and a sample size of 250. Bias and variance for a sample size of 250 are provided in Appendix E and are not discussed because the relative magnitude of bias and variance for each method is similar to that found in previous models already reviewed. Metrics for larger sample sizes are provided in Appendix F as well and are also not reviewed because of their similarity.

Although using group propensity scores was the optimal method in data generating model F for the estimate of the treatment coefficient when groups were balanced, matching with a ratio of 1:1 was optimal for data generating model E. In data

generating model D with unequal group proportions, GBM was optimal; however, in model E weighting by the odds had the lowest MSE. The performance of weighting by the odds under the conditions simulated was unexpected because of the increased nonlinearity in model E which increased the number of extreme observations by a factor of 5 and increased the associated weights by a factor of 4 compared with data generating model D. Weighting typically does not perform well under conditions of extreme weights compared with GBM; therefore, this result warrants further investigation and will be discussed in Chapter 5. Finally, group propensity scores had significantly lower MSE compared with the other methods when estimating the interaction and group coefficients.

Table 22
MSE for Data Generating Model E, ATT Estimates, Misspecified Model mA, n=250.

		Percentile Condition			
Coefficient	Method	Baseline	70	80	90
β_t	Wt x Odds	0.2459	0.2217	0.2162	0.1881
	GBM	0.2379	0.2258	0.2180	0.1880
	Match 1:2	0.2526	0.2645	0.2520	0.2166
	Match 1:1	0.2259	0.2329	0.2180	0.1998
	ExMatch 1:2	0.2567	0.2508	0.2388	0.2136
	ExMatch 1:1	0.2440	0.2527	0.2238	0.1861
	Group	0.2316	0.2401	0.2345	0.1840
	Regression	0.1388	0.1470	0.1453	0.1294
β_{tg}	Wt x Odds	0.5225	0.4672	0.5158	0.4428
	GBM	0.4524	0.4463	0.4960	0.4019
	Match 1:2	0.5392	0.5242	0.5374	0.4958
	Match 1:1	0.4834	0.4564	0.4939	0.4348
	ExMatch 1:2	0.5351	0.5153	0.5324	0.5057
	ExMatch 1:1	0.4861	0.4754	0.5016	0.4244
	Group	0.2298	0.2120	0.2570	0.2743
	Regression	0.2575	0.2518	0.2812	0.2586
β_g	Wt x Odds	0.3682	0.3428	0.3875	0.3499
	GBM	0.3142	0.3327	0.3349	0.3025
	Match 1:2	0.3899	0.3674	0.4062	0.3880
	Match 1:1	0.2378	0.2364	0.2571	0.2178
	ExMatch 1:2	0.3687	0.3888	0.4051	0.3831
	ExMatch 1:1	0.2471	0.2650	0.2675	0.2217
	Group	0.1157	0.1161	0.1345	0.1437
	Regression	0.1327	0.1346	0.1462	0.1326

Note. Bias, variance for n=250 and all metrics for n=500 and n=1000 are in Appendix F.

4.3 Applied Example

Although the simulation results above provide information about best approaches in the presence of heterogeneous treatment effect estimation under fixed conditions, demonstration of the application of the various methods under typical data conditions can

be helpful. This section uses data from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K) to illustrate the methods studied and the differing results one obtains with them. The ECLS-K study collected data on individual, family, school, and community factors of children as they progressed from kindergarten through middle school. The applied example is based on a topic that permeates the educational research literature: the private school effect on learning.

For simplicity, the data preparation and analysis steps were minimized because the applied example is for illustration only and not intended to provide substantive conclusions. For example, missing data analysis and a comprehensive literature review and analysis of the covariates that pertain to the questions under study were not completed but would be required in order to defend any inference from this analysis to the population studied. In addition, weights were not used to adjust for problems (e.g., disproportionate selection, nonresponse) typically encountered with complex survey data. Excluding sampling weights when analyzing complex survey data may result in biased standard errors, incorrect inferences, and results that are not generalizable to the population (Hahs-Vaughn, 2005).

The covariates included in the model were based on Morgan (2001) who showed that educational success is often mediated by family background characteristics. The research question and group indicator were based on a study that investigated the SES-math achievement gap. Socioeconomic status (SES) is a combined measure of a person's or group's position in relation to others based on income, education and employment. Galindo and Sonnenschein (2015) studied the relation between SES and math

achievement, citing several studies that show that children with low SES score significantly lower than children with higher SES on standardized tests.

The model for the example in this study was proposed to evaluate whether SES moderates the effect of private school learning on standardized math test performance for 8th grade students. The treatment under evaluation was the effect of a private school education on learning. The outcome measure, math T score (M=52.3, SD=9.34), was based on a standardized assessment conducted in selected schools. Covariates hypothesized to influence the outcome included geographical region, type of location (i.e., rural, urban, suburban), gender, disability status, indicators of a student's interest in math, and an indicator of the student's household SES. The SES variable used in this analysis is a composite five measures: father/male guardian's education, mother/female guardian's education, father/male guardian's occupation, mother/female guardian's occupation, and household income (Tourangeau, Nord, Le, Sorongon, & Najarian, 2009). All covariates that were assumed to influence the outcome were included in the propensity score model. In addition, SES was hypothesized to moderate the effects of the private school learning on standardized math test scores. Descriptive statistics for all variables are in Appendix G.

All methods investigated in the simulation study were implemented in this applied example. Methods were evaluated after matching or weighting by first calculating the absolute mean differences between the treatment and control groups for each covariate. Absolute mean differences were calculated separately based on SES group and standardized for each SES group using the standard deviation of the original treatment group. Covariate balance was evaluated within each SES group because effect

moderation by SES group was hypothesized to moderate the effects of private school learning on standardized math tests. Finally, the mean absolute standardized mean distance (ASMD) was calculated and used to compare covariate balance across all methods.

The mean ASMD was used to evaluate method performance because the true effects in the population are unknown. As previously mentioned in Section 2.2.3, covariate balance is one of the methods used to assess whether propensity score methods are appropriate. Covariate balance has also been used to determine method performance when the true effects are unknown because smaller differences in means, particularly in covariates hypothesized to more strongly influence outcome, are expected to produce less biased estimates (Green & Stuart, 2014; Ho et al., 2007; Leacy & Stuart, 2013; Stuart, 2010).

The results for the methods under investigation that estimated the ATE will be reviewed first. Tables 23 and 24 show the ASMD for the low SES group and high SES group for all covariates before and after matching. Boxplots are provided in Figures 14 and 15. Recall that simulation I found that across all data generating models and coefficients MNPS and GBM had the smallest MSE in the majority of conditions. The evaluation of the methods using empirical data provides similar conclusions to those found in simulation I. The mean ASMD for MNPS and GBM were smaller than the other methods for both the low SES and the high SES group. In addition, both MNPS and GBM improved the balance between treatment and control of more covariates than IPTW and subclassification.

Table 23

Comparison of Absolute Standardized Differences Before and After Matching, Low SES, ATE Estimates (n=3,337)

Covariate	Before PS Methods	GBM	IPTW	MNPS	Sub-classification
Race	0.157	0.068	0.086	0.018	0.121
Gender	0.042	0.020	0.011	0.020	0.076
Siblings	0.079	0.145	0.141	0.045	0.346
Tutor	0.010	0.002	0.075	0.021	0.133
Enjoy Math	0.010	0.033	0.070	0.018	0.027
Like Math	0.008	0.038	0.101	0.007	0.024
Disability	0.186	0.124	0.079	0.050	0.067
Region	0.302	0.106	0.220	0.009	0.170
Urban	0.423	0.129	0.226	0.024	0.035
Mean ASMD	0.135	0.074	0.112	0.024	0.111
# ASMDs that decreased		6	5	7	4

Note. PS = propensity score.

Table 24

Comparison of Absolute Standardized Differences Before and After Matching, High SES, ATE Estimates (n=3,116)

Covariate	Before PS Methods	GBM	IPTW	MNPS	Sub-classification
Race	0.010	0.029	0.125	0.018	0.145
Gender	0.017	0.001	0.001	0.026	0.067
Siblings	0.134	0.060	0.097	0.048	0.159
Tutor	0.110	0.018	0.031	0.027	0.064
Enjoy Math	0.040	0.038	0.050	0.055	0.151
Like Math	0.045	0.035	0.059	0.033	0.161
Disability	0.042	0.010	0.034	0.043	0.048
Region	0.138	0.065	0.160	0.003	0.158
Urban	0.517	0.018	0.010	0.049	0.033
Mean ASMD	0.117	0.030	0.063	0.034	0.109
# ASMDs that decreased		8	5	5	2

Note. PS = propensity score.

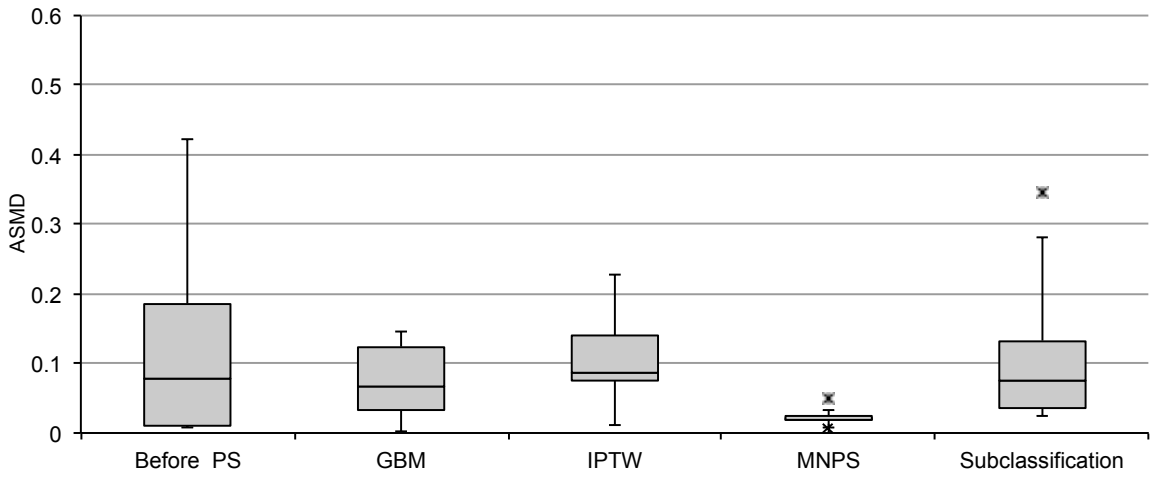


Figure 15. Covariate balance for the low SES group, ATE estimates. × is an outlier defined as outside 1.5 times the interquartile range.

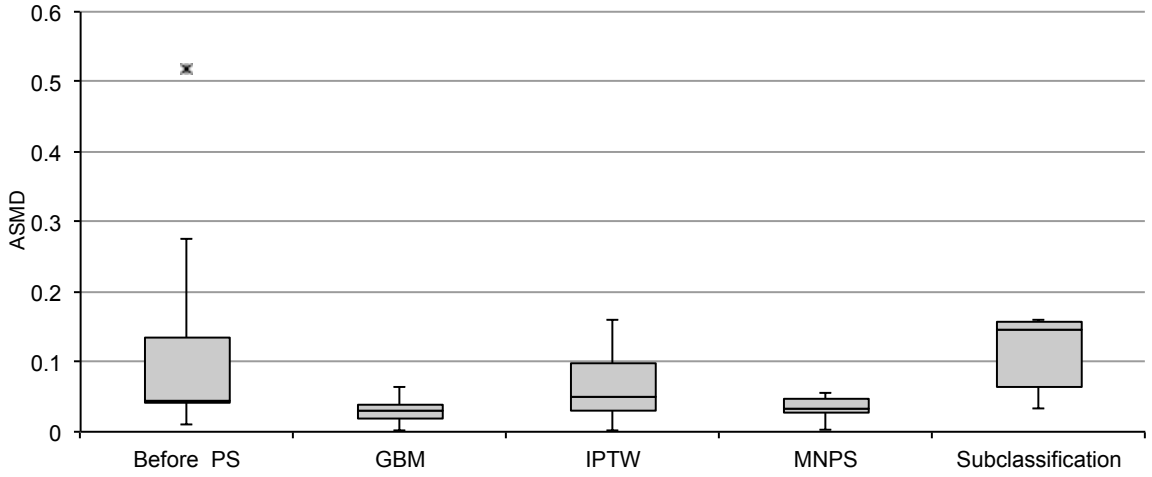


Figure 16. Covariate balance for the high SES group, ATE estimates. × is an outlier defined as outside 1.5 times the interquartile range.

Next, method comparisons for ATT will be reviewed. Results are summarized for the low SES group in Table 25 and Figure 17. The results for the high SES group are summarized in Table 26 and Figure 18. The evaluation of the methods using empirical data provided similar conclusions to those found in simulation I. Essentially, for ATT

estimates group propensity scores provided that largest improvement in covariate balance for both SES groups compared with the other methods.

Table 25
Comparison of Absolute Standardized Differences Before and After Matching, Low SES, ATT Estimates (n=3,337)

Covariate	Before PS Methods	Group	Match 1:1	Exact Match 1:1	Match 1:2	Exact Match 1:2	Wt x Odds	GBM
Race	0.157	0.015	0.113	0.078	0.166	0.155	0.167	0.046
Gender	0.042	0.000	0.025	0.171	0.048	0.088	0.029	0.012
Siblings	0.079	0.019	0.175	0.220	0.076	0.132	0.194	0.061
Tutor	0.010	0.051	0.152	0.144	0.068	0.101	0.094	0.027
Enjoy Math	0.010	0.003	0.083	0.012	0.018	0.005	0.021	0.038
Like Math	0.008	0.003	0.123	0.042	0.010	0.011	0.035	0.037
Disability	0.186	0.029	0.204	0.135	0.108	0.053	0.058	0.072
Region	0.302	0.022	0.298	0.206	0.298	0.311	0.302	0.099
Urban	0.423	0.008	0.046	0.031	0.008	0.018	0.136	0.127
Mean ASMD	0.135	0.017	0.135	0.116	0.089	0.097	0.115	0.058
# ASMDs that decreased		8	4	4	4	4	4	6

Note. PS = propensity score.

Table 26
Comparison of Absolute Standardized Differences Before and After Matching, High SES, ATT Estimates (n=3,116)

Covariate	Before PS Methods	Group	Match 1:1	Exact Match 1:1	Match 1:2	Exact Match 1:2	Wt x Odds	GBM
Race	0.010	0.014	0.079	0.119	0.106	0.078	0.067	0.028
Gender	0.017	0.015	0.049	0.020	0.029	0.023	0.019	0.029
Siblings	0.134	0.038	0.066	0.077	0.066	0.079	0.108	0.092
Tutor	0.110	0.009	0.034	0.006	0.013	0.006	0.055	0.041
Enjoy Math	0.040	0.032	0.005	0.018	0.002	0.003	0.007	0.009
Like Math	0.045	0.006	0.003	0.014	0.004	0.003	0.017	0.017
Disability	0.042	0.000	0.088	0.057	0.055	0.070	0.032	0.022
Region	0.138	0.068	0.033	0.065	0.061	0.064	0.116	0.018
Urban	0.517	0.021	0.016	0.047	0.014	0.046	0.111	0.035
Mean ASMD	0.117	0.022	0.041	0.047	0.039	0.042	0.059	0.032
# ASMDs that decreased		8	6	6	6	6	7	7

Note. PS = propensity score.

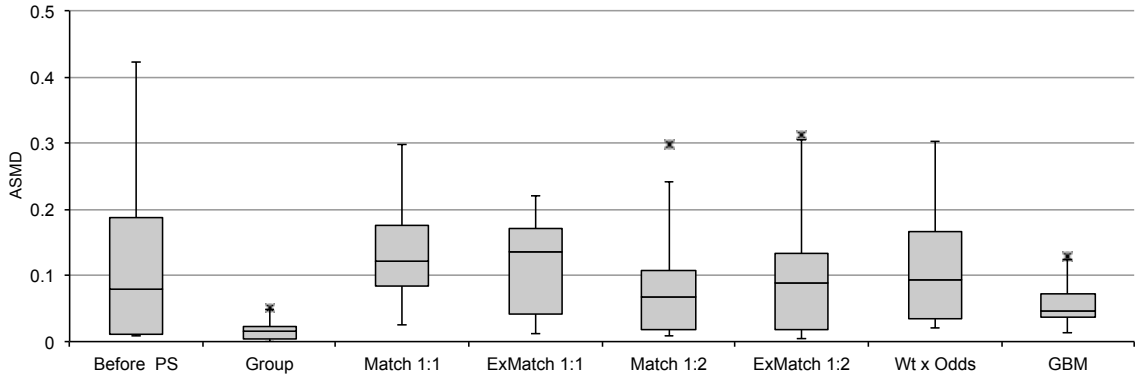


Figure 17. Covariate balance for the low SES group, ATT estimates. × is an outlier defined as outside 1.5 times the interquartile range.

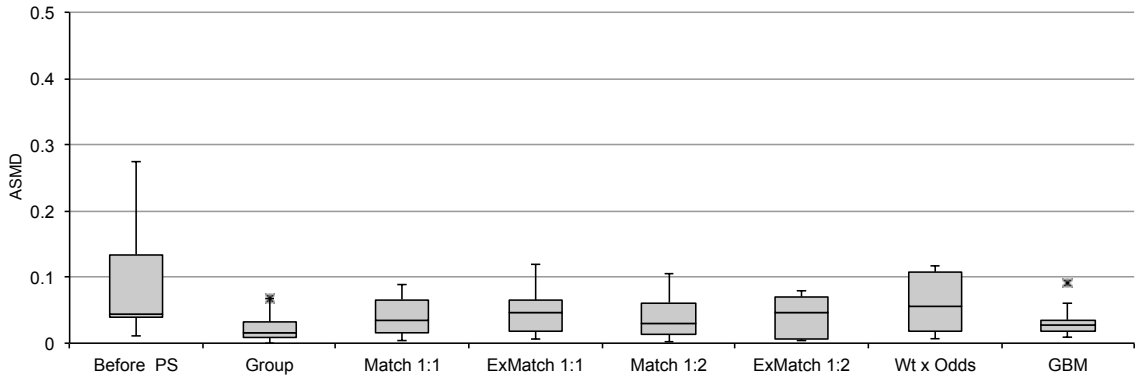


Figure 18. Covariate balance for the high SES group, ATT estimates. × is an outlier defined as outside 1.5 times the interquartile range.

Finally, the treatment effect estimates for the methods that were used to estimate the ATE are in Table 27. Table 28 provides the same information for the methods that were used to estimate the ATT. All coefficients were significant at $p<0.01$ providing evidence that SES moderates the effect of private school learning on standardized math test performance for 8th grade students.

Table 27
Summary of ATE Estimates

Method	β_t	*SE* β_t	β_g	*SE* β_g	β_{tg}	*SE* β_{tg}
Subclass	3.581	1.201	8.489	0.848	-4.459	1.351
MNPS	1.554	0.555	5.137	0.270	-1.729	0.351
IPTW	2.478	0.584	7.134	0.245	-3.244	0.677
GBM	2.147	0.296	7.067	0.291	-3.037	0.420

Note. All coefficients were significant at $p<0.01$

Table 28
Summary of ATT Estimates

Method	β_t	*SE* β_t	β_g	*SE* β_g	β_{tg}	*SE* β_{tg}
Wt x Odds	2.714	0.5006	7.199	0.2793	-3.392	0.6019
GBM	1.990	0.3880	6.484	0.3255	-2.677	0.4572
Match 1:2	2.912	0.5720	7.161	0.4064	-3.288	0.6798
Match 1:1	2.184	0.6512	6.456	0.5449	-2.538	0.7740
ExMatch 1:2	2.867	0.5749	7.104	0.4106	-3.303	0.6813
ExMatch 1:1	2.054	0.6426	6.518	0.5385	-2.665	0.7615
Group	1.383	0.6721	5.628	0.2841	-1.776	0.3690
Regression	2.595	0.5062	7.138	0.2416	-3.330	0.6168

Note. All coefficients were significant at $p<0.01$

In summary, the evaluation of the ASMD of the covariates in this dataset for the methods used in this study suggest that when covariate balance is used as a proxy to compare the performance of propensity score methods, the findings for the applied dataset align with the findings for the simulation. The next chapter will summarize the main conclusions and limitations of the current study, and recommend areas for future research.

Chapter 5: Discussion

This study extended the propensity score literature by investigating the performance of propensity score methods in the presence of effect moderation by a baseline characteristic. Specifically, three group-focused methods were evaluated: MNPS, exact matching on group, and group propensity scores. This chapter begins with a summary of key findings and concludes with suggestions for future research.

5.1 Summary of Key Findings.

Several recent studies, which investigated methods of combining propensity score methods with moderation analyses, provided the basis for the current study. The earliest study, Green and Stuart (2014), was an empirical study that showed that estimating separate propensity score models for each group of interest (i.e., group propensity scores) was superior over full matching and exact matching on group. Green and Stuart used empirical data and therefore used balance metrics as the criteria for model performance because the true effects were unknown.

The current study used Monte Carlo simulation and mostly supported the findings in Green and Stuart. Specifically, no significant advantage was found to including exact matching on group in the propensity score model design. Several other studies support this conclusion as well because forcing more focus on matches of a specific covariate may remove the focus from more prognostically valid covariates (Leacy & Stuart, 2013; King & Nielsen, 2016). This conclusion was consistent across all conditions in both simulations.

Green and Stuart (2014) also compared models with and without complex selection (i.e., models where treatment selection varies by group). Green and Stuart found

that including the theoretically relevant group interactions in the propensity score model provided better performance over models that excluded these interactions. Theoretically relevant group interactions were not assessed in the current study; however, the current study found that the group indicator should be included in the propensity score model regardless of whether group membership influenced selection.

Finally, the key finding in Green and Stuart (2014), that estimating separate propensity scores for each group of interest is the optimal method of combining propensity score methods with effect moderation, was supported in the current study. The current study showed that for the estimates of the coefficient of interaction and group, group propensity scores provided a significant reduction (≈50%) of MSE relative to other methods. This result was evident across all sample sizes and conditions in both simulations. It is important to note that using group propensity scores for estimates of the main treatment coefficient was not consistently the optimal method; however, no condition reflected more than a moderate increase (≈ +0.05) in MSE relative to the other methods. Therefore, considering that overall effects are a combination of all three coefficients and considering the significant improvement in MSE for the interaction and group coefficient, using group propensity scores was found to be the optimal method of combining propensity scores with effect moderation for ATT in this study.

Dong (2015) and Eeren et al. (2015), presented results advocating a multinomial model for estimating propensity scores in the presence of effect moderation. Both studies restructured a 2 x 2 factorial design into a 4 x 1 multiple treatment model and used multinomial logistic regression to estimate propensity scores. Dong used IPTW to estimate the coefficients of the treatment, interaction, and group terms. Covariate

adjustment was used in Eeren et al. Both studies found that bias and MSE for estimates of the treatment, interaction, and group coefficients were smaller compared with the methods evaluated in each study.

The current study used boosted regression and also found smaller bias and MSE for a multinomial implementation (i.e., MNPS) compared with a binomial implementation (i.e., GBM). Multinomial propensity score (MNPS) estimates for the group and interaction terms had the lowest MSE of all methods evaluated across all conditions. However, unlike group propensity scores, which exhibited a significant reduction in MSE compared with other methods, MNPS estimates were comparable to GBM estimates. This result suggests that any improvement realized using MNPS under the conditions simulated might be partly attributed to the simultaneous estimation of the coefficients rather than an increased focus on group. In addition, performance of MNPS for treatment coefficient estimates varied and aligned with previous research describing optimal conditions for propensity score methods. Therefore, in situations where machine learning algorithms like boosted modeling are advantageous for estimating the coefficient of treatment, this study found that the simpler binomial implementation (GBM) provided similar performance to MNPS in terms of bias and variance reduction.

This study also found that when one group is more prevalent in the extremes of the propensity score distribution, GBM had the lowest MSE for the estimate of the treatment coefficient across all conditions except one. Weighting by the odds had slightly lower (i.e., within 0.004) MSE than GBM in the presence of nonlinearity. Estimates for the other coefficients were consistent with previously mentioned results that MNPS and group propensity scores had the lowest MSE for ATE and ATE, respectively.

Three findings supported previous research. First, MSE for estimates of the treatment coefficient using propensity score models that included all covariates related to outcome regardless of their relationship to selection were consistently smaller than MSE for estimates using propensity score models that excluded covariates related to outcome. Simulation II further supported this finding and showed that including the group indicator in the propensity score model had little impact, positively or negatively, on MSE in propensity models without nonlinearity and improved MSE in propensity score models with nonlinearity.

Second, machine learning algorithms were optimal in the presence of nonadditivity and nonlinearity for the estimates of the treatment coefficient. Finally, as expected, subclassification did not perform well when groups were unbalanced. Results for estimates of the interaction coefficient were expected to align with those found in previous research for the treatment coefficient. This expectation was not met across all conditions and warrants a more comprehensive simulation study.

These results and the previous discussion are provided only in the context of the conditions simulated. Limitations associated with these conditions and future suggestions for research will be discussed in Section 5.2.

5.2 Limitations and Extensions.

There are several limitations of this study and possible extensions. First, the research design focuses on the moderating effects of a binary baseline covariate. In addition to baseline covariates, the probability of selection into treatment may also moderate treatment effects. Heterogeneous treatment effects identified by an interaction between the probability of treatment and treatment effect were described but not included

in the research design. Examples of studies that focus on the treatment effect and propensity score interaction include Brand and Xie (2010), Tsai and Xie (2008), and Tsai and Xie (2011).

The research design also included only one manifestation of matching. A more extensive simulation study similar to Austin (2014) would be needed to determine whether optimal matching conditions in the presence of heterogeneous treatment effects differ from those for homogeneous treatment effects. Also matching was only implemented for estimates of ATT. Comparisons for ATE would be helpful to determine under what conditions matching is ideal compared with other propensity score methods and whether the conditions align with previous research.

Similar to the single implementation of matching criteria, only one machine learning algorithm was implemented. Generalized boosted modeling was chosen because the tuning parameters (i.e., shrinkage, number of trees, tree complexity) have been studied and optimized for probability predictions (McCaffrey et al., 2013). However, algorithms such as genetic matching and Bayesian additive regression trees also have support in the literature for their performance in analyzing complex selection and heterogeneous treatment effects (Chipman, George, & McCulloch, 2010; Diamond & Sekhon, 2013; Hill, 2011; Hill, Weiss, & Zhai, 2011).

Also, the methods were not implemented or tested under accepted optimal conditions unique to each method according to previous research. For example, matching was not implemented in the current study with a large ratio of control units to treated units. Previous research showed that matching is best when there is a large pool of available control units compared with the pool of treated units (Caliendo & Kopeinig,

2008; Rubin, 1979; Stuart, 2010). Future studies could provide comparisons within the context of optimal conditions relative to each method in the presence of effect moderation.

The distribution and quantity of covariates used also presents opportunities for future research. Ten continuous normally distributed covariates were used. The few simulation studies related to propensity score methods in the presence of heterogeneous treatment effects assumed either 2 to 3 covariates or 10 covariates. Covariates were predominantly normally distributed. Empirical studies typically included more than 20 covariates. In a few simulation studies, the authors cited covariate quantities, distributions, and effect sizes related to a specific discipline or field of study such as pharmacoepidemiology (e.g., Austin, 2009b; Setoguchi et al., 2008; Sturmer et al., 2006). Future simulation studies could be designed to address conditions specific to a discipline or field of study.

Sample size is another area that could be investigated further. In the current study, larger sample sizes did not provide many additional insights. This was true particularly among the matching methods where estimates were very similar at n=500 and n=1000 (Caliendo & Kopeinig, 2008). As Shadish (2013) suggests, future research testing the performance of propensity score methods under sample sizes smaller than 250 could illuminate how covariates might moderate estimates or highlight methods that perform well at smaller sample sizes.

Finally, Tables 3 and 5 listed the fixed conditions for simulations I and II. These fixed conditions limited the generalizability of the results of this study. Therefore, there are opportunities for future research that could investigate the impact on modifying some

or all of the fixed conditions. In addition, although not mentioned in either Table 3 or Table 5, the number of groups was fixed to two with the same covariates influencing selection in both groups. Increasing the number of groups influencing selection or investigating the performance of methods when the covariates that influence selection are different for each group are opportunities for future research. Both conditions are prevalent in medical and psychological studies (Green & Stuart, 2014).

5.3 Conclusion

The goal of this study was to provide methodological recommendations, specific to the use of propensity score methods, for researchers who suspect effect moderation in observational studies. There are several considerations that are apparent from the results of this study. First, although the iterative process of determining the propensity score model that produces the best covariate balance is done without knowledge of the outcome model parameters, propensity score model design should not be done without hypotheses regarding the covariates that influence outcome. When implementing propensity score methods, researchers must provide sufficient evidence to support the research design for the outcome model and the propensity score model. This study demonstrated the importance of including covariates related to outcome regardless of their influence on selection. Therefore, a clear understanding and sound hypotheses regarding covariates that influence both outcome and selection are required to produce treatment effect estimates with the least amount of bias.

Another consideration is the performance of the group-focused methods in estimating the coefficient of the interaction term. I expected propensity score methods to have similar relative performance for each coefficient. For example, I expected that a

method that produces estimates of the treatment coefficient with the smallest bias would also produce estimates of the interaction coefficient with the smallest bias. This expectation was not realized. This study found that methods that perform best when estimating the interaction coefficient were not always the best method for estimating the treatment coefficient. However, across all conditions the optimal method for estimating the interaction coefficient provided estimates with the lowest overall MSE when considering both coefficients in the presence of effect moderation. For the methods compared in this study, group propensity score methods were best for ATT and MNPS were best for ATE and are recommended when researchers suspect effect moderation.

Exact matching on group was expected to perform better than methods that did not provide additional focus on the group information. This study found that exact matching on group should be avoided unless further research demonstrates that there are conditions under which exact matching on group would outperform group propensity scores or MNPS.

After covariates have been chosen, the propensity score must be estimated either by the researcher iteratively searching for the propensity score model that minimizes a loss function or by using a machine learning algorithm to find the propensity score model that minimizes a loss function. This study found that although logistic regression provided estimates with the smallest bias across some of the conditions, machine learning algorithms are recommended to estimate the propensity score when complex models are required (e.g., nonadditivity, nonlinearity, unequal group proportions).

Sample size is also a consideration regardless of the method chosen. Propensity score methods are often referred to as large sample methods. This study found that under

the balanced conditions simulated and with normally distributed covariates, method performance blurred or became more similar at larger sample sizes (i.e., 500, 1000). This suggests that under conditions similar to those in this study, the choice of the propensity score method may not be as critical when larger sample sizes (> 500) are available.

Finally, although propensity score methods are a popular and viable option for estimating treatment effects using observational data, propensity score methods may not be the best option. Researchers should seek alternative methods when covariate balance is not achievable or evidence supporting the strong ignorability assumption is insufficient.

Appendix A: Coefficients for All Models

Propensity Model Coefficients:

$\alpha_0 = \log(1.4)$
$\alpha_1 = \log(2)$
$\alpha_2 = \log(0.5)$
$\alpha_3 = \log(1.6)$
$\alpha_4 = \log(1.8)$
$\alpha_5 = \log(1)$
$\alpha_6 = \log(0.8)$
$\alpha_7 = \log(2)$
$\alpha_8 = \log(1.2)$
$\alpha_9 = \log(1.5)$
$\alpha_{10} = \log(2)$
$\alpha_g = \log(2)$
$\alpha_{9g} = \log(1.6)$

Outcome Model Coefficients

$\beta_0 = 0$
$\beta_1 = 0.2$
$\beta_2 = -0.4$
$\beta_3 = 0.5$
$\beta_4 = -0.1$
$\beta_5 = 0.2$
$\beta_6 = 0.3$
$\beta_7 = 0.2$
$\beta_8 = 0.6$

Table A1

Values for Treatment, Group and Interaction Coefficients.

		Coefficients		
Effect Size	Interaction	β_t	β_g	β_{tg}
	none	0.3	0.2	0
large	qualitative	0.3	0.3	-0.6
	quantitative	0.3	0.2	0.3
small	qualitative	0.1	0.1	-0.2
	quantitative	0.3	0.1	0.1

Appendix B: Simulation I Data Generating Model A

Table B1

MSE for Data Generating Model A, ATE Estimates, True Model A.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.266	0.126	0.056
	MNPS	0.251	0.127	0.070
	IPTW	0.275	0.132	0.058
	GBM	0.249	0.131	0.069
β_{tg}	Subclass	0.522	0.242	0.104
	MNPS	0.377	0.192	0.088
	IPTW	0.628	0.373	0.150
	GBM	0.380	0.202	0.091
β_g	Subclass	0.260	0.139	0.077
	MNPS	0.195	0.091	0.046
	IPTW	0.394	0.235	0.098
	GBM	0.197	0.100	0.048

Table B2

Bias for Data Generating Model A, ATE Estimates, True Model A.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.121	0.070	0.074
	MNPS	0.248	0.177	0.166
	IPTW	0.094	0.034	0.035
	GBM	0.260	0.193	0.170
β_{tg}	Subclass	-0.062	-0.026	-0.017
	MNPS	-0.061	-0.039	-0.035
	IPTW	-0.151	-0.113	-0.090
	GBM	-0.081	-0.063	-0.057
β_g	Subclass	-0.100	-0.119	-0.133
	MNPS	0.007	0.009	-0.003
	IPTW	0.082	0.071	0.049
	GBM	0.011	0.009	0.008

Table B3

Variance for Data Generating Model A, ATE Estimates, True Model A.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.252	0.121	0.050
	MNPS	0.189	0.095	0.042
	IPTW	0.267	0.131	0.057
	GBM	0.181	0.093	0.040
β_{tg}	Subclass	0.518	0.242	0.104
	MNPS	0.373	0.190	0.086
	IPTW	0.606	0.361	0.142
	GBM	0.373	0.198	0.088
β_g	Subclass	0.250	0.124	0.059
	MNPS	0.195	0.091	0.046
	IPTW	0.388	0.230	0.096
	GBM	0.196	0.099	0.048

Table B4

MSE for Data Generating Model A, ATE Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.245	0.117	0.050
	MNPS	0.235	0.117	0.060
	IPTW	0.222	0.106	0.048
	GBM	0.227	0.118	0.058
β_{tg}	Subclass	0.493	0.216	0.098
	MNPS	0.371	0.190	0.086
	IPTW	0.476	0.260	0.112
	GBM	0.374	0.195	0.089
β_g	Subclass	0.268	0.150	0.096
	MNPS	0.184	0.089	0.045
	IPTW	0.254	0.137	0.061
	GBM	0.186	0.095	0.046

Table B5

Bias for Data Generating Model A, ATE Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.102	0.070	0.061
	MNPS	0.210	0.145	0.132
	IPTW	0.069	0.033	0.032
	GBM	0.218	0.159	0.135
β_{tg}	Subclass	-0.066	-0.040	-0.014
	MNPS	-0.056	-0.032	-0.024
	IPTW	-0.118	-0.086	-0.068
	GBM	-0.073	-0.057	-0.049
β_g	Subclass	-0.149	-0.172	-0.194
	MNPS	0.009	0.008	-0.004
	IPTW	0.058	0.055	0.034
	GBM	0.009	0.012	0.006

Table B6

Variance for Data Generating Model A, ATE Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.235	0.112	0.047
	MNPS	0.191	0.095	0.043
	IPTW	0.218	0.105	0.047
	GBM	0.179	0.092	0.040
β_{tg}	Subclass	0.489	0.214	0.098
	MNPS	0.368	0.189	0.086
	IPTW	0.462	0.253	0.107
	GBM	0.369	0.192	0.086
β_g	Subclass	0.246	0.120	0.058
	MNPS	0.184	0.089	0.045
	IPTW	0.250	0.134	0.059
	GBM	0.186	0.095	0.046

Table B7

MSE for Data Generating Model A, ATE Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.238	0.109	0.048
	MNPS	0.225	0.111	0.058
	IPTW	0.220	0.103	0.045
	GBM	0.223	0.115	0.058
β_{tg}	Subclass	0.491	0.213	0.099
	MNPS	0.343	0.171	0.079
	IPTW	0.487	0.270	0.112
	GBM	0.364	0.191	0.086
β_g	Subclass	0.262	0.145	0.093
	MNPS	0.172	0.085	0.041
	IPTW	0.262	0.144	0.061
	GBM	0.182	0.094	0.044

Table B8

Bias for Data Generating Model A, ATE Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.104	0.069	0.062
	MNPS	0.229	0.165	0.143
	IPTW	0.067	0.032	0.031
	GBM	0.231	0.171	0.146
β_{tg}	Subclass	-0.073	-0.041	-0.014
	MNPS	-0.068	-0.039	-0.024
	IPTW	-0.124	-0.089	-0.067
	GBM	-0.074	-0.055	-0.048
β_g	Subclass	-0.137	-0.165	-0.190
	MNPS	0.010	0.005	-0.007
	IPTW	0.063	0.057	0.033
	GBM	0.008	0.008	0.003

Table B9

Variance for Data Generating Model A, ATE Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.227	0.104	0.044
	MNPS	0.173	0.084	0.038
	IPTW	0.216	0.102	0.044
	GBM	0.170	0.086	0.037
β_{tg}	Subclass	0.485	0.211	0.098
	MNPS	0.339	0.169	0.078
	IPTW	0.472	0.262	0.107
	GBM	0.358	0.188	0.084
β_g	Subclass	0.243	0.118	0.057
	MNPS	0.172	0.085	0.041
	IPTW	0.258	0.141	0.059
	GBM	0.182	0.093	0.044

Table B10

MSE for Data Generating Model A, ATT Estimates, True Model A.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.405	0.209	0.102
	GBM	0.272	0.145	0.064
	Match 1:2	0.366	0.165	0.069
	Match 1:1	0.296	0.135	0.061
	ExMatch 1:2	0.349	0.159	0.069
	ExMatch 1:1	0.324	0.141	0.065
	Group	0.312	0.134	0.058
	Regression	0.157	0.072	0.032
β_{tg}	Wt x Odds	0.910	0.574	0.242
	GBM	0.550	0.309	0.133
	Match 1:2	0.702	0.358	0.145
	Match 1:1	0.630	0.285	0.131
	ExMatch 1:2	0.714	0.334	0.147
	ExMatch 1:1	0.635	0.290	0.131
	Group	0.298	0.133	0.061
	Regression	0.269	0.126	0.063
β_g	Wt x Odds	0.759	0.463	0.207
	GBM	0.393	0.210	0.099
	Match 1:2	0.537	0.261	0.116
	Match 1:1	0.313	0.146	0.075
	ExMatch 1:2	0.549	0.248	0.114
	ExMatch 1:1	0.337	0.142	0.074
	Group	0.154	0.068	0.034
	Regression	0.137	0.065	0.032

Table B11

Bias for Data Generating Model A, ATT Estimates, True Model A.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.016	-0.040	-0.046
	GBM	0.168	0.104	0.081
	Match 1:2	0.062	0.006	0.006
	Match 1:1	0.054	0.001	0.014
	ExMatch 1:2	0.039	-0.008	0.009
	ExMatch 1:1	0.061	0.006	0.009
	Group	0.048	0.004	0.015
	Regression	0.019	0.001	0.000
β_{tg}	Wt x Odds	-0.111	-0.092	-0.057
	GBM	-0.089	-0.069	-0.051
	Match 1:2	-0.072	-0.026	0.004
	Match 1:1	-0.045	-0.007	-0.013
	ExMatch 1:2	-0.051	-0.008	-0.004
	ExMatch 1:1	-0.067	-0.011	-0.013
	Group	0.004	-0.013	0.003
	Regression	-0.021	-0.017	0.000
β_g	Wt x Odds	0.048	0.032	0.001
	GBM	0.026	0.010	-0.004
	Match 1:2	0.005	-0.035	-0.060
	Match 1:1	-0.051	-0.072	-0.071
	ExMatch 1:2	0.002	-0.043	-0.049
	ExMatch 1:1	-0.034	-0.060	-0.068
	Group	-0.004	-0.013	-0.014
	Regression	0.008	0.008	-0.004

Table B12

Variance for Data Generating Model A, ATT Estimates, True Model A.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.404	0.208	0.099
	GBM	0.244	0.134	0.057
	Match 1:2	0.362	0.165	0.069
	Match 1:1	0.293	0.135	0.061
	ExMatch 1:2	0.347	0.159	0.069
	ExMatch 1:1	0.320	0.141	0.065
	Group	0.310	0.134	0.058
	Regression	0.157	0.072	0.032
β_{tg}	Wt x Odds	0.898	0.566	0.239
	GBM	0.542	0.304	0.131
	Match 1:2	0.696	0.358	0.145
	Match 1:1	0.628	0.285	0.130
	ExMatch 1:2	0.712	0.334	0.147
	ExMatch 1:1	0.630	0.290	0.131
	Group	0.298	0.133	0.061
	Regression	0.268	0.126	0.063
β_g	Wt x Odds	0.757	0.462	0.207
	GBM	0.392	0.210	0.099
	Match 1:2	0.537	0.260	0.113
	Match 1:1	0.311	0.141	0.070
	ExMatch 1:2	0.549	0.246	0.112
	ExMatch 1:1	0.336	0.138	0.069
	Group	0.154	0.068	0.034
	Regression	0.137	0.065	0.032

Table B13

MSE for Data Generating Model A, ATT Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.286	0.132	0.060
	GBM	0.251	0.129	0.053
	Match 1:2	0.299	0.139	0.059
	Match 1:1	0.262	0.124	0.054
	ExMatch 1:2	0.309	0.136	0.061
	ExMatch 1:1	0.256	0.121	0.054
	Group	0.267	0.121	0.052
	Regression	0.157	0.072	0.032
β_{tg}	Wt x Odds	0.609	0.342	0.147
	GBM	0.500	0.263	0.115
	Match 1:2	0.608	0.305	0.134
	Match 1:1	0.554	0.245	0.115
	ExMatch 1:2	0.622	0.290	0.133
	ExMatch 1:1	0.517	0.241	0.112
	Group	0.253	0.119	0.058
	Regression	0.269	0.126	0.063
β_g	Wt x Odds	0.439	0.237	0.111
	GBM	0.328	0.167	0.079
	Match 1:2	0.420	0.206	0.103
	Match 1:1	0.292	0.134	0.068
	ExMatch 1:2	0.436	0.201	0.100
	ExMatch 1:1	0.270	0.129	0.069
	Group	0.131	0.062	0.031
	Regression	0.137	0.065	0.032

Table B14

Bias for Data Generating Model A, ATT Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.014	-0.005	-0.016
	GBM	0.139	0.102	0.073
	Match 1:2	0.037	0.015	0.003
	Match 1:1	0.040	0.014	0.018
	ExMatch 1:2	0.038	0.025	0.003
	ExMatch 1:1	0.043	0.010	0.010
	Group	0.036	0.013	0.011
	Regression	0.019	0.001	0.000
β_{tg}	Wt x Odds	-0.088	-0.084	-0.048
	GBM	-0.073	-0.073	-0.045
	Match 1:2	-0.059	-0.036	0.000
	Match 1:1	-0.038	-0.014	-0.008
	ExMatch 1:2	-0.068	-0.043	-0.002
	ExMatch 1:1	-0.053	-0.011	-0.003
	Group	0.004	0.002	-0.003
	Regression	-0.021	-0.017	0.000
β_g	Wt x Odds	0.026	0.025	-0.007
	GBM	0.011	0.014	-0.011
	Match 1:2	-0.014	-0.029	-0.057
	Match 1:1	-0.056	-0.080	-0.083
	ExMatch 1:2	0.012	-0.008	-0.049
	ExMatch 1:1	-0.045	-0.079	-0.087
	Group	-0.022	-0.030	-0.025
	Regression	0.008	0.008	-0.004

Table B15

Variance for Data Generating Model A, ATT Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.285	0.132	0.060
	GBM	0.232	0.118	0.048
	Match 1:2	0.298	0.139	0.059
	Match 1:1	0.261	0.124	0.054
	ExMatch 1:2	0.307	0.135	0.061
	ExMatch 1:1	0.254	0.121	0.054
	Group	0.266	0.121	0.052
	Regression	0.157	0.072	0.032
β_{tg}	Wt x Odds	0.602	0.335	0.144
	GBM	0.495	0.258	0.113
	Match 1:2	0.604	0.304	0.134
	Match 1:1	0.552	0.245	0.115
	ExMatch 1:2	0.617	0.288	0.133
	ExMatch 1:1	0.514	0.241	0.112
	Group	0.253	0.119	0.058
	Regression	0.268	0.126	0.063
β_g	Wt x Odds	0.438	0.236	0.110
	GBM	0.328	0.166	0.079
	Match 1:2	0.420	0.206	0.099
	Match 1:1	0.289	0.128	0.061
	ExMatch 1:2	0.435	0.201	0.098
	ExMatch 1:1	0.268	0.123	0.062
	Group	0.131	0.061	0.030
	Regression	0.137	0.065	0.032

Table B16

MSE for Data Generating Model A, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.286	0.132	0.057
	GBM	0.238	0.122	0.052
	Match 1:2	0.305	0.128	0.059
	Match 1:1	0.253	0.118	0.051
	ExMatch 1:2	0.294	0.134	0.058
	ExMatch 1:1	0.260	0.112	0.050
	Group	0.256	0.113	0.051
	Regression	0.157	0.072	0.032
β_{tg}	Wt x Odds	0.634	0.358	0.146
	GBM	0.471	0.258	0.110
	Match 1:2	0.612	0.295	0.135
	Match 1:1	0.543	0.251	0.113
	ExMatch 1:2	0.610	0.291	0.129
	ExMatch 1:1	0.548	0.244	0.112
	Group	0.248	0.113	0.052
	Regression	0.269	0.126	0.063
β_g	Wt x Odds	0.462	0.252	0.111
	GBM	0.303	0.165	0.076
	Match 1:2	0.433	0.203	0.104
	Match 1:1	0.282	0.124	0.067
	ExMatch 1:2	0.423	0.205	0.099
	ExMatch 1:1	0.286	0.133	0.068
	Group	0.138	0.062	0.029
	Regression	0.137	0.065	0.032

Table B17

Bias for Data Generating Model A, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.010	-0.006	-0.016
	GBM	0.155	0.109	0.081
	Match 1:2	0.032	0.006	0.010
	Match 1:1	0.046	0.018	0.015
	ExMatch 1:2	0.030	0.023	0.004
	ExMatch 1:1	0.050	0.008	0.013
	Group	0.032	0.022	0.009
	Regression	0.019	0.001	0.000
β_{tg}	Wt x Odds	-0.094	-0.088	-0.048
	GBM	-0.077	-0.068	-0.044
	Match 1:2	-0.063	-0.029	0.000
	Match 1:1	-0.069	-0.019	-0.018
	ExMatch 1:2	-0.071	-0.041	0.002
	ExMatch 1:1	-0.070	-0.023	-0.014
	Group	-0.003	-0.008	0.007
	Regression	-0.021	-0.017	0.000
β_g	Wt x Odds	0.032	0.029	-0.008
	GBM	0.014	0.008	-0.012
	Match 1:2	-0.010	-0.035	-0.056
	Match 1:1	-0.048	-0.071	-0.078
	ExMatch 1:2	0.015	-0.011	-0.053
	ExMatch 1:1	-0.046	-0.079	-0.080
	Group	-0.013	-0.023	-0.026
	Regression	0.008	0.008	-0.004

Table B18

Variance for Data Generating Model A, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.286	0.132	0.057
	GBM	0.214	0.110	0.045
	Match 1:2	0.304	0.128	0.059
	Match 1:1	0.251	0.117	0.051
	ExMatch 1:2	0.294	0.134	0.058
	ExMatch 1:1	0.257	0.112	0.050
	Group	0.255	0.112	0.051
	Regression	0.157	0.072	0.032
β_{tg}	Wt x Odds	0.625	0.351	0.144
	GBM	0.465	0.253	0.108
	Match 1:2	0.608	0.295	0.135
	Match 1:1	0.538	0.251	0.113
	ExMatch 1:2	0.605	0.290	0.129
	ExMatch 1:1	0.543	0.243	0.111
	Group	0.248	0.112	0.052
	Regression	0.268	0.126	0.063
β_g	Wt x Odds	0.461	0.251	0.111
	GBM	0.303	0.165	0.076
	Match 1:2	0.433	0.201	0.100
	Match 1:1	0.279	0.119	0.061
	ExMatch 1:2	0.422	0.205	0.097
	ExMatch 1:1	0.284	0.127	0.061
	Group	0.138	0.061	0.029
	Regression	0.137	0.065	0.032

Appendix C: Simulation I Data Generating Model B

Table C1

MSE for Data Generating Model B, ATE Estimates, True Model B.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.249	0.151	0.085
	MNPS	0.275	0.164	0.097
	IPTW	0.251	0.136	0.065
	GBM	0.286	0.173	0.101
β_{tg}	Subclass	0.491	0.243	0.105
	MNPS	0.346	0.180	0.087
	IPTW	0.631	0.371	0.171
	GBM	0.370	0.202	0.099
β_g	Subclass	0.271	0.174	0.099
	MNPS	0.163	0.089	0.042
	IPTW	0.364	0.226	0.105
	GBM	0.183	0.103	0.048

Table C2

Bias for Data Generating Model B, ATE Estimates, True Model B.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	-0.158	-0.177	-0.176
	MNPS	0.320	0.267	0.230
	IPTW	0.086	0.061	0.044
	GBM	0.336	0.289	0.239
β_{tg}	Subclass	-0.034	-0.027	-0.031
	MNPS	-0.036	-0.049	-0.045
	IPTW	-0.152	-0.150	-0.140
	GBM	-0.058	-0.083	-0.081
β_g	Subclass	-0.168	-0.192	-0.198
	MNPS	-0.020	-0.003	-0.005
	IPTW	0.085	0.085	0.078
	GBM	-0.016	0.001	0.010

Table C3

Variance for Data Generating Model B, ATE Estimates, True Model B.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.224	0.120	0.054
	MNPS	0.173	0.093	0.044
	IPTW	0.244	0.132	0.063
	GBM	0.173	0.090	0.044
β_{tg}	Subclass	0.490	0.242	0.104
	MNPS	0.345	0.178	0.085
	IPTW	0.608	0.348	0.152
	GBM	0.367	0.195	0.092
β_g	Subclass	0.243	0.137	0.060
	MNPS	0.162	0.089	0.042
	IPTW	0.357	0.219	0.099
	GBM	0.183	0.103	0.048

Table C4

MSE for Data Generating Model B, ATE Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.290	0.185	0.115
	MNPS	0.322	0.219	0.150
	IPTW	0.267	0.173	0.106
	GBM	0.316	0.219	0.149
β_{tg}	Subclass	0.464	0.224	0.102
	MNPS	0.372	0.197	0.092
	IPTW	0.473	0.274	0.117
	GBM	0.366	0.201	0.094
β_g	Subclass	0.269	0.163	0.093
	MNPS	0.175	0.095	0.044
	IPTW	0.250	0.142	0.059
	GBM	0.182	0.101	0.045

Table C5

Bias for Data Generating Model B, ATE Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.259	0.254	0.248
	MNPS	0.367	0.346	0.324
	IPTW	0.247	0.239	0.235
	GBM	0.378	0.352	0.324
β_{tg}	Subclass	-0.028	-0.027	-0.026
	MNPS	-0.027	-0.055	-0.055
	IPTW	-0.107	-0.110	-0.106
	GBM	-0.051	-0.073	-0.073
β_g	Subclass	-0.173	-0.186	-0.198
	MNPS	-0.025	-0.005	-0.010
	IPTW	0.042	0.042	0.032
	GBM	-0.019	-0.004	-0.002

Table C6

Variance for Data Generating Model B, ATE Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.223	0.121	0.053
	MNPS	0.187	0.099	0.045
	IPTW	0.206	0.115	0.050
	GBM	0.173	0.095	0.044
β_{tg}	Subclass	0.464	0.223	0.102
	MNPS	0.372	0.194	0.089
	IPTW	0.461	0.262	0.106
	GBM	0.363	0.196	0.089
β_g	Subclass	0.239	0.128	0.054
	MNPS	0.175	0.095	0.044
	IPTW	0.249	0.141	0.058
	GBM	0.182	0.101	0.045

Table C7

MSE for Data Generating Model B, ATE Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.243	0.132	0.065
	MNPS	0.275	0.164	0.097
	IPTW	0.252	0.138	0.065
	GBM	0.286	0.173	0.101
β_{tg}	Subclass	0.485	0.245	0.105
	MNPS	0.346	0.180	0.087
	IPTW	0.638	0.370	0.171
	GBM	0.370	0.202	0.099
β_g	Subclass	0.269	0.177	0.101
	MNPS	0.163	0.089	0.042
	IPTW	0.372	0.224	0.105
	GBM	0.183	0.103	0.048

Table C8

Bias for Data Generating Model B, ATE Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.127	0.103	0.103
	MNPS	0.320	0.267	0.230
	IPTW	0.084	0.059	0.046
	GBM	0.336	0.289	0.239
β_{tg}	Subclass	-0.045	-0.028	-0.032
	MNPS	-0.036	-0.049	-0.045
	IPTW	-0.152	-0.146	-0.140
	GBM	-0.058	-0.083	-0.081
β_g	Subclass	-0.177	-0.199	-0.208
	MNPS	-0.020	-0.003	-0.005
	IPTW	0.087	0.080	0.078
	GBM	-0.016	0.001	0.010

Table C9

Variance for Data Generating Model B, ATE Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.227	0.121	0.054
	MNPS	0.173	0.093	0.044
	IPTW	0.245	0.135	0.063
	GBM	0.173	0.090	0.044
β_{tg}	Subclass	0.483	0.244	0.104
	MNPS	0.345	0.178	0.085
	IPTW	0.615	0.349	0.152
	GBM	0.367	0.195	0.092
β_g	Subclass	0.238	0.138	0.058
	MNPS	0.162	0.089	0.042
	IPTW	0.364	0.217	0.099
	GBM	0.183	0.103	0.048

Table C10

MSE for Data Generating Model B, ATT Estimates, True Model B.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.357	0.220	0.118
	GBM	0.260	0.149	0.073
	Match 1:2	0.327	0.168	0.071
	Match 1:1	0.292	0.137	0.061
	ExMatch 1:2	0.326	0.165	0.080
	ExMatch 1:1	0.296	0.141	0.061
	Group	0.272	0.127	0.058
	Regression	0.166	0.080	0.039
β_{tg}	Wt x Odds	0.903	0.557	0.263
	GBM	0.513	0.298	0.146
	Match 1:2	0.698	0.353	0.147
	Match 1:1	0.626	0.282	0.130
	ExMatch 1:2	0.654	0.340	0.154
	ExMatch 1:1	0.581	0.276	0.127
	Group	0.273	0.123	0.057
	Regression	0.284	0.128	0.067
β_g	Wt x Odds	0.707	0.448	0.215
	GBM	0.342	0.204	0.103
	Match 1:2	0.524	0.262	0.118
	Match 1:1	0.308	0.157	0.079
	ExMatch 1:2	0.509	0.256	0.123
	ExMatch 1:1	0.310	0.156	0.076
	Group	0.141	0.064	0.031
	Regression	0.135	0.068	0.030

Table C11

Bias for Data Generating Model B, ATT Estimates, True Model B.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	-0.017	-0.053	-0.080
	GBM	0.196	0.134	0.099
	Match 1:2	0.025	-0.001	-0.002
	Match 1:1	0.013	0.013	0.013
	ExMatch 1:2	0.011	0.001	-0.001
	ExMatch 1:1	0.043	0.018	0.012
	Group	0.004	0.010	-0.001
	Regression	-0.007	-0.008	-0.010
β_{tg}	Wt x Odds	-0.101	-0.107	-0.093
	GBM	-0.055	-0.074	-0.071
	Match 1:2	-0.011	-0.011	0.000
	Match 1:1	-0.001	-0.017	-0.010
	ExMatch 1:2	0.002	-0.018	-0.004
	ExMatch 1:1	-0.034	-0.023	-0.017
	Group	0.031	-0.001	0.006
	Regression	0.022	-0.005	0.001
β_g	Wt x Odds	0.034	0.025	0.008
	GBM	-0.012	-0.008	-0.014
	Match 1:2	-0.068	-0.074	-0.085
	Match 1:1	-0.102	-0.110	-0.115
	ExMatch 1:2	-0.052	-0.055	-0.077
	ExMatch 1:1	-0.094	-0.113	-0.116
	Group	-0.035	-0.023	-0.024
	Regression	-0.012	0.003	-0.004

Table C12

Variance for Data Generating Model B, ATT Estimates, True Model B.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.357	0.217	0.112
	GBM	0.221	0.131	0.063
	Match 1:2	0.327	0.168	0.071
	Match 1:1	0.292	0.137	0.061
	ExMatch 1:2	0.326	0.165	0.080
	ExMatch 1:1	0.294	0.140	0.061
	Group	0.272	0.127	0.058
	Regression	0.166	0.080	0.039
β_{tg}	Wt x Odds	0.893	0.546	0.254
	GBM	0.510	0.292	0.141
	Match 1:2	0.698	0.353	0.147
	Match 1:1	0.626	0.282	0.130
	ExMatch 1:2	0.654	0.340	0.154
	ExMatch 1:1	0.580	0.275	0.126
	Group	0.272	0.123	0.057
	Regression	0.284	0.128	0.067
β_g	Wt x Odds	0.705	0.447	0.215
	GBM	0.342	0.204	0.102
	Match 1:2	0.520	0.257	0.111
	Match 1:1	0.298	0.144	0.066
	ExMatch 1:2	0.506	0.253	0.117
	ExMatch 1:1	0.301	0.144	0.063
	Group	0.140	0.064	0.031
	Regression	0.135	0.068	0.030

Table C13

MSE for Data Generating Model B, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.301	0.184	0.096
	GBM	0.305	0.199	0.119
	Match 1:2	0.314	0.186	0.107
	Match 1:1	0.288	0.174	0.102
	ExMatch 1:2	0.317	0.187	0.103
	ExMatch 1:1	0.299	0.182	0.103
	Group	0.303	0.169	0.101
	Regression	0.166	0.080	0.039
β_{tg}	Wt x Odds	0.626	0.366	0.153
	GBM	0.481	0.275	0.125
	Match 1:2	0.602	0.316	0.144
	Match 1:1	0.502	0.260	0.118
	ExMatch 1:2	0.566	0.311	0.135
	ExMatch 1:1	0.525	0.260	0.127
	Group	0.286	0.151	0.089
	Regression	0.284	0.128	0.067
β_g	Wt x Odds	0.448	0.260	0.110
	GBM	0.312	0.179	0.082
	Match 1:2	0.429	0.227	0.107
	Match 1:1	0.261	0.138	0.070
	ExMatch 1:2	0.412	0.229	0.102
	ExMatch 1:1	0.273	0.142	0.071
	Group	0.140	0.076	0.043
	Regression	0.135	0.068	0.030

Table C14

Bias for Data Generating Model B, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.179	0.174	0.162
	GBM	0.299	0.273	0.248
	Match 1:2	0.200	0.187	0.183
	Match 1:1	0.200	0.212	0.210
	ExMatch 1:2	0.207	0.190	0.184
	ExMatch 1:1	0.215	0.205	0.206
	Group	0.210	0.215	0.202
	Regression	-0.007	-0.008	-0.010
β_{tg}	Wt x Odds	-0.090	-0.097	-0.082
	GBM	-0.052	-0.081	-0.069
	Match 1:2	-0.037	-0.026	-0.026
	Match 1:1	-0.006	-0.033	-0.030
	ExMatch 1:2	-0.050	-0.027	-0.029
	ExMatch 1:1	-0.024	-0.021	-0.026
	Group	0.190	0.180	0.184
	Regression	0.022	-0.005	0.001
β_g	Wt x Odds	0.023	0.015	-0.003
	GBM	-0.015	-0.001	-0.016
	Match 1:2	-0.039	-0.061	-0.061
	Match 1:1	-0.086	-0.090	-0.095
	ExMatch 1:2	-0.011	-0.050	-0.052
	ExMatch 1:1	-0.084	-0.096	-0.096
	Group	-0.117	-0.123	-0.120
	Regression	-0.012	0.003	-0.004

Table C15

Variance for Data Generating Model B, ATT Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.269	0.154	0.070
	GBM	0.215	0.125	0.058
	Match 1:2	0.274	0.151	0.074
	Match 1:1	0.248	0.129	0.058
	ExMatch 1:2	0.274	0.151	0.069
	ExMatch 1:1	0.253	0.140	0.061
	Group	0.259	0.123	0.060
	Regression	0.166	0.080	0.039
β_{tg}	Wt x Odds	0.618	0.357	0.146
	GBM	0.479	0.268	0.120
	Match 1:2	0.601	0.315	0.143
	Match 1:1	0.502	0.258	0.117
	ExMatch 1:2	0.564	0.311	0.134
	ExMatch 1:1	0.524	0.260	0.126
	Group	0.250	0.119	0.055
	Regression	0.284	0.128	0.067
β_g	Wt x Odds	0.447	0.260	0.110
	GBM	0.312	0.179	0.081
	Match 1:2	0.428	0.224	0.104
	Match 1:1	0.253	0.130	0.060
	ExMatch 1:2	0.412	0.227	0.099
	ExMatch 1:1	0.266	0.133	0.062
	Group	0.127	0.061	0.028
	Regression	0.135	0.068	0.030

Table C16

MSE for Data Generating Model B, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.361	0.222	0.118
	GBM	0.260	0.149	0.073
	Match 1:2	0.329	0.163	0.076
	Match 1:1	0.283	0.137	0.063
	ExMatch 1:2	0.331	0.168	0.076
	ExMatch 1:1	0.277	0.132	0.063
	Group	0.295	0.134	0.061
	Regression	0.166	0.080	0.039
β_{tg}	Wt x Odds	0.913	0.554	0.264
	GBM	0.513	0.298	0.146
	Match 1:2	0.674	0.357	0.156
	Match 1:1	0.567	0.278	0.129
	ExMatch 1:2	0.677	0.339	0.146
	ExMatch 1:1	0.576	0.281	0.134
	Group	0.265	0.116	0.057
	Regression	0.284	0.128	0.067
β_g	Wt x Odds	0.720	0.445	0.216
	GBM	0.342	0.204	0.103
	Match 1:2	0.515	0.272	0.121
	Match 1:1	0.295	0.151	0.081
	ExMatch 1:2	0.520	0.257	0.115
	ExMatch 1:1	0.300	0.153	0.081
	Group	0.144	0.062	0.031
	Regression	0.135	0.068	0.030

Table C17

Bias for Data Generating Model B, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	-0.015	-0.055	-0.075
	GBM	0.196	0.134	0.099
	Match 1:2	0.004	0.001	0.000
	Match 1:1	0.018	0.010	0.007
	ExMatch 1:2	0.012	-0.006	0.001
	ExMatch 1:1	0.023	0.009	0.010
	Group	0.012	0.006	0.002
	Regression	-0.007	-0.008	-0.010
β_{tg}	Wt x Odds	-0.105	-0.100	-0.095
	GBM	-0.055	-0.074	-0.071
	Match 1:2	0.009	-0.016	-0.002
	Match 1:1	0.000	-0.014	-0.003
	ExMatch 1:2	0.009	-0.013	-0.004
	ExMatch 1:1	0.000	-0.019	-0.009
	Group	0.034	0.001	0.006
	Regression	0.022	-0.005	0.001
β_g	Wt x Odds	0.038	0.018	0.010
	GBM	-0.012	-0.008	-0.014
	Match 1:2	-0.087	-0.069	-0.083
	Match 1:1	-0.118	-0.108	-0.120
	ExMatch 1:2	-0.058	-0.060	-0.077
	ExMatch 1:1	-0.108	-0.109	-0.121
	Group	-0.027	-0.027	-0.024
	Regression	-0.012	0.003	-0.004

Table C18

Variance for Data Generating Model B, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.361	0.219	0.112
	GBM	0.221	0.131	0.063
	Match 1:2	0.329	0.163	0.076
	Match 1:1	0.282	0.137	0.063
	ExMatch 1:2	0.330	0.168	0.076
	ExMatch 1:1	0.276	0.132	0.063
	Group	0.294	0.134	0.061
	Regression	0.166	0.080	0.039
β_{tg}	Wt x Odds	0.902	0.544	0.255
	GBM	0.510	0.292	0.141
	Match 1:2	0.674	0.356	0.156
	Match 1:1	0.567	0.278	0.129
	ExMatch 1:2	0.677	0.339	0.146
	ExMatch 1:1	0.576	0.281	0.134
	Group	0.264	0.116	0.057
	Regression	0.284	0.128	0.067
β_g	Wt x Odds	0.718	0.445	0.216
	GBM	0.342	0.204	0.102
	Match 1:2	0.507	0.267	0.114
	Match 1:1	0.281	0.140	0.066
	ExMatch 1:2	0.517	0.253	0.109
	ExMatch 1:1	0.288	0.141	0.066
	Group	0.143	0.061	0.031
	Regression	0.135	0.068	0.030

Appendix D: Simulation I Data Generating Model C

Table D1

MSE for Data Generating Model C, ATE Estimates, True Model C.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.281	0.161	0.096
	MNPS	0.171	0.097	0.049
	IPTW	0.261	0.147	0.084
	GBM	0.168	0.102	0.049
β_{tg}	Subclass	0.466	0.219	0.104
	MNPS	0.345	0.160	0.081
	IPTW	0.599	0.298	0.186
	GBM	0.362	0.179	0.088
β_g	Subclass	0.253	0.126	0.056
	MNPS	0.178	0.086	0.040
	IPTW	0.346	0.201	0.130
	GBM	0.184	0.090	0.046

Table D2

Bias for Data Generating Model C, ATE Estimates, True Model C.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	-0.253	-0.233	-0.227
	MNPS	0.120	0.125	0.104
	IPTW	-0.011	-0.009	-0.014
	GBM	0.128	0.130	0.109
β_{tg}	Subclass	-0.020	-0.032	-0.037
	MNPS	-0.005	-0.026	-0.019
	IPTW	-0.037	-0.048	-0.059
	GBM	-0.029	-0.051	-0.049
β_g	Subclass	-0.030	-0.029	-0.040
	MNPS	-0.007	0.007	-0.001
	IPTW	0.017	0.020	0.019
	GBM	0.006	0.013	0.013

Table D3

Variance for Data Generating Model C, ATE Estimates, True Model C.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.217	0.107	0.045
	MNPS	0.157	0.081	0.038
	IPTW	0.261	0.147	0.084
	GBM	0.151	0.085	0.037
β_{tg}	Subclass	0.465	0.218	0.102
	MNPS	0.345	0.160	0.080
	IPTW	0.597	0.296	0.182
	GBM	0.361	0.176	0.086
β_g	Subclass	0.252	0.125	0.054
	MNPS	0.178	0.086	0.040
	IPTW	0.346	0.201	0.129
	GBM	0.184	0.090	0.045

Table D4

MSE for Data Generating Model C, ATE Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.196	0.104	0.045
	MNPS	0.184	0.104	0.049
	IPTW	0.178	0.096	0.042
	GBM	0.170	0.103	0.050
β_{tg}	Subclass	0.441	0.199	0.093
	MNPS	0.380	0.178	0.088
	IPTW	0.431	0.200	0.101
	GBM	0.369	0.182	0.092
β_g	Subclass	0.238	0.120	0.061
	MNPS	0.192	0.093	0.043
	IPTW	0.214	0.101	0.049
	GBM	0.188	0.091	0.047

Table D5

Bias for Data Generating Model C, ATE Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.012	0.045	0.046
	MNPS	0.103	0.111	0.092
	IPTW	-0.005	0.022	0.021
	GBM	0.111	0.117	0.098
β_{tg}	Subclass	0.000	-0.038	-0.036
	MNPS	-0.003	-0.025	-0.018
	IPTW	-0.047	-0.071	-0.067
	GBM	-0.027	-0.050	-0.049
β_g	Subclass	-0.112	-0.094	-0.105
	MNPS	-0.004	0.005	0.001
	IPTW	0.020	0.029	0.026
	GBM	0.004	0.014	0.016

Table D6

Variance for Data Generating Model C, ATE Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.196	0.102	0.042
	MNPS	0.173	0.092	0.041
	IPTW	0.178	0.096	0.041
	GBM	0.158	0.090	0.040
β_{tg}	Subclass	0.441	0.197	0.092
	MNPS	0.380	0.178	0.088
	IPTW	0.429	0.195	0.097
	GBM	0.368	0.180	0.090
β_g	Subclass	0.226	0.111	0.050
	MNPS	0.192	0.093	0.043
	IPTW	0.214	0.100	0.049
	GBM	0.188	0.091	0.047

Table D7

MSE for Data Generating Model C, ATE Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.201	0.102	0.046
	MNPS	0.171	0.097	0.049
	IPTW	0.212	0.109	0.047
	GBM	0.168	0.102	0.049
β_{tg}	Subclass	0.443	0.205	0.100
	MNPS	0.345	0.160	0.081
	IPTW	0.504	0.235	0.122
	GBM	0.362	0.179	0.088
β_g	Subclass	0.245	0.118	0.059
	MNPS	0.178	0.086	0.040
	IPTW	0.261	0.131	0.063
	GBM	0.184	0.090	0.046

Table D8

Bias for Data Generating Model C, ATE Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.027	0.051	0.050
	MNPS	0.120	0.125	0.104
	IPTW	0.007	0.019	0.019
	GBM	0.128	0.130	0.109
β_{tg}	Subclass	-0.034	-0.040	-0.039
	MNPS	-0.005	-0.026	-0.019
	IPTW	-0.073	-0.078	-0.076
	GBM	-0.029	-0.051	-0.049
β_g	Subclass	-0.064	-0.067	-0.079
	MNPS	-0.007	0.007	-0.001
	IPTW	0.035	0.039	0.027
	GBM	0.006	0.013	0.013

Table D9

Variance for Data Generating Model C, ATE Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	1000
β_t	Subclass	0.200	0.100	0.043
	MNPS	0.157	0.081	0.038
	IPTW	0.212	0.108	0.047
	GBM	0.151	0.085	0.037
β_{tg}	Subclass	0.442	0.203	0.098
	MNPS	0.345	0.160	0.080
	IPTW	0.499	0.229	0.116
	GBM	0.361	0.176	0.086
β_g	Subclass	0.240	0.114	0.052
	MNPS	0.178	0.086	0.040
	IPTW	0.260	0.129	0.063
	GBM	0.184	0.090	0.045

Table D10

MSE for Data Generating Model C, ATT Estimates, True Model C.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.401	0.245	0.158
	GBM	0.194	0.119	0.052
	Match 1:2	0.293	0.154	0.066
	Match 1:1	0.257	0.121	0.058
	ExMatch 1:2	0.286	0.153	0.070
	ExMatch 1:1	0.251	0.123	0.057
	Group	0.245	0.122	0.054
	Regression	0.134	0.074	0.031
β_{tg}	Wt x Odds	0.868	0.472	0.315
	GBM	0.457	0.229	0.118
	Match 1:2	0.640	0.306	0.148
	Match 1:1	0.545	0.260	0.127
	ExMatch 1:2	0.632	0.316	0.152
	ExMatch 1:1	0.546	0.251	0.128
	Group	0.244	0.110	0.053
	Regression	0.279	0.133	0.061
β_g	Wt x Odds	0.703	0.410	0.275
	GBM	0.289	0.148	0.082
	Match 1:2	0.474	0.230	0.115
	Match 1:1	0.271	0.137	0.062
	ExMatch 1:2	0.454	0.231	0.114
	ExMatch 1:1	0.285	0.138	0.063
	Group	0.129	0.062	0.028
	Regression	0.141	0.070	0.030

Table D11

Bias for Data Generating Model C, ATT Estimates, True Model C.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	-0.045	-0.036	-0.062
	GBM	0.088	0.093	0.064
	Match 1:2	0.007	0.031	0.018
	Match 1:1	-0.014	0.025	0.028
	ExMatch 1:2	-0.007	0.030	0.023
	ExMatch 1:1	-0.018	0.011	0.025
	Group	-0.016	0.017	0.018
	Regression	-0.028	0.004	0.004
β_{tg}	Wt x Odds	-0.012	-0.037	-0.027
	GBM	-0.024	-0.050	-0.045
	Match 1:2	-0.014	-0.024	0.002
	Match 1:1	0.012	-0.030	-0.018
	ExMatch 1:2	0.009	-0.021	-0.001
	ExMatch 1:1	0.019	-0.009	-0.015
	Group	0.005	0.008	0.019
	Regression	0.022	-0.003	-0.001
β_g	Wt x Odds	-0.008	-0.003	-0.013
	GBM	0.004	0.011	0.006
	Match 1:2	-0.007	-0.017	-0.042
	Match 1:1	-0.039	-0.030	-0.032
	ExMatch 1:2	-0.026	-0.015	-0.037
	ExMatch 1:1	-0.049	-0.043	-0.033
	Group	-0.016	-0.030	-0.028
	Regression	-0.011	0.001	-0.004

Table D12

Variance for Data Generating Model C, ATT Estimates, True Model C.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.399	0.244	0.154
	GBM	0.186	0.110	0.048
	Match 1:2	0.293	0.153	0.066
	Match 1:1	0.257	0.121	0.057
	ExMatch 1:2	0.286	0.152	0.069
	ExMatch 1:1	0.251	0.123	0.057
	Group	0.244	0.122	0.054
	Regression	0.133	0.074	0.031
β_{tg}	Wt x Odds	0.868	0.470	0.314
	GBM	0.456	0.227	0.116
	Match 1:2	0.639	0.306	0.148
	Match 1:1	0.544	0.259	0.126
	ExMatch 1:2	0.632	0.315	0.152
	ExMatch 1:1	0.546	0.251	0.128
	Group	0.244	0.109	0.053
	Regression	0.278	0.133	0.061
β_g	Wt x Odds	0.703	0.410	0.275
	GBM	0.289	0.148	0.082
	Match 1:2	0.474	0.230	0.113
	Match 1:1	0.270	0.136	0.061
	ExMatch 1:2	0.453	0.230	0.113
	ExMatch 1:1	0.282	0.136	0.062
	Group	0.129	0.062	0.027
	Regression	0.141	0.070	0.030

Table D13

MSE for Data Generating Model C, ATT Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.213	0.114	0.049
	GBM	0.202	0.122	0.055
	Match 1:2	0.246	0.134	0.058
	Match 1:1	0.210	0.111	0.052
	ExMatch 1:2	0.241	0.133	0.057
	ExMatch 1:1	0.215	0.116	0.050
	Group	0.217	0.114	0.049
	Regression	0.134	0.074	0.031
β_{tg}	Wt x Odds	0.503	0.232	0.114
	GBM	0.476	0.230	0.121
	Match 1:2	0.542	0.272	0.125
	Match 1:1	0.505	0.221	0.112
	ExMatch 1:2	0.536	0.252	0.121
	ExMatch 1:1	0.489	0.221	0.106
	Group	0.232	0.101	0.051
	Regression	0.279	0.133	0.061
β_g	Wt x Odds	0.326	0.150	0.075
	GBM	0.303	0.147	0.083
	Match 1:2	0.358	0.182	0.090
	Match 1:1	0.239	0.118	0.056
	ExMatch 1:2	0.360	0.166	0.087
	ExMatch 1:1	0.252	0.118	0.057
	Group	0.117	0.058	0.026
	Regression	0.141	0.070	0.030

Table D14

Bias for Data Generating Model C, ATT Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.002	0.032	0.027
	GBM	0.073	0.082	0.061
	Match 1:2	0.004	0.040	0.035
	Match 1:1	-0.005	0.038	0.035
	ExMatch 1:2	0.002	0.032	0.026
	ExMatch 1:1	-0.001	0.018	0.023
	Group	-0.011	0.019	0.012
	Regression	-0.028	0.004	0.004
β_{tg}	Wt x Odds	-0.022	-0.054	-0.049
	GBM	-0.018	-0.047	-0.049
	Match 1:2	0.006	-0.041	-0.027
	Match 1:1	-0.010	-0.047	-0.042
	ExMatch 1:2	0.013	-0.016	-0.011
	ExMatch 1:1	0.002	-0.013	-0.017
	Group	-0.008	0.001	0.011
	Regression	0.022	-0.003	-0.001
β_g	Wt x Odds	0.002	0.015	0.009
	GBM	-0.002	0.008	0.009
	Match 1:2	-0.036	-0.003	-0.014
	Match 1:1	-0.035	-0.023	-0.023
	ExMatch 1:2	-0.032	-0.019	-0.026
	ExMatch 1:1	-0.042	-0.042	-0.040
	Group	-0.023	-0.035	-0.031
	Regression	-0.011	0.001	-0.004

Table D15

Variance for Data Generating Model C, ATT Estimates, Misspecified mA.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.213	0.113	0.048
	GBM	0.197	0.116	0.051
	Match 1:2	0.246	0.133	0.057
	Match 1:1	0.210	0.110	0.051
	ExMatch 1:2	0.241	0.132	0.057
	ExMatch 1:1	0.215	0.115	0.050
	Group	0.217	0.113	0.049
	Regression	0.133	0.074	0.031
β_{tg}	Wt x Odds	0.503	0.229	0.112
	GBM	0.475	0.228	0.118
	Match 1:2	0.542	0.270	0.124
	Match 1:1	0.505	0.219	0.110
	ExMatch 1:2	0.536	0.251	0.121
	ExMatch 1:1	0.489	0.221	0.105
	Group	0.232	0.101	0.051
	Regression	0.278	0.133	0.061
β_g	Wt x Odds	0.326	0.149	0.075
	GBM	0.303	0.147	0.083
	Match 1:2	0.357	0.182	0.090
	Match 1:1	0.238	0.118	0.056
	ExMatch 1:2	0.359	0.166	0.086
	ExMatch 1:1	0.250	0.116	0.056
	Group	0.116	0.057	0.025
	Regression	0.141	0.070	0.030

Table D16

MSE for Data Generating Model C, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.281	0.151	0.068
	GBM	0.194	0.119	0.052
	Match 1:2	0.274	0.137	0.058
	Match 1:1	0.223	0.115	0.053
	ExMatch 1:2	0.276	0.134	0.062
	ExMatch 1:1	0.240	0.115	0.053
	Group	0.230	0.106	0.052
	Regression	0.134	0.074	0.031
β_{tg}	Wt x Odds	0.625	0.313	0.161
	GBM	0.457	0.229	0.118
	Match 1:2	0.612	0.261	0.129
	Match 1:1	0.517	0.238	0.113
	ExMatch 1:2	0.590	0.273	0.135
	ExMatch 1:1	0.520	0.237	0.114
	Group	0.220	0.106	0.051
	Regression	0.279	0.133	0.061
β_g	Wt x Odds	0.457	0.235	0.121
	GBM	0.289	0.148	0.082
	Match 1:2	0.430	0.185	0.096
	Match 1:1	0.260	0.121	0.058
	ExMatch 1:2	0.417	0.193	0.100
	ExMatch 1:1	0.282	0.128	0.063
	Group	0.122	0.059	0.028
	Regression	0.141	0.070	0.030

Table D17

Bias for Data Generating Model C, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.001	0.022	0.008
	GBM	0.088	0.093	0.064
	Match 1:2	0.020	0.041	0.036
	Match 1:1	0.001	0.029	0.035
	ExMatch 1:2	0.002	0.023	0.029
	ExMatch 1:1	-0.013	0.022	0.025
	Group	-0.017	0.010	0.019
	Regression	-0.028	0.004	0.004
β_{tg}	Wt x Odds	-0.045	-0.067	-0.045
	GBM	-0.024	-0.050	-0.045
	Match 1:2	-0.031	-0.049	-0.032
	Match 1:1	-0.014	-0.032	-0.037
	ExMatch 1:2	0.007	-0.021	-0.014
	ExMatch 1:1	0.004	-0.024	-0.018
	Group	-0.025	0.004	0.003
	Regression	0.022	-0.003	-0.001
β_g	Wt x Odds	0.025	0.028	0.005
	GBM	0.004	0.011	0.006
	Match 1:2	0.003	0.007	-0.008
	Match 1:1	-0.039	-0.030	-0.022
	ExMatch 1:2	-0.025	-0.016	-0.023
	ExMatch 1:1	-0.051	-0.034	-0.038
	Group	0.000	-0.028	-0.023
	Regression	-0.011	0.001	-0.004

Table D18

Variance for Data Generating Model C, ATT Estimates, Misspecified mB.

		Sample Size		
Coefficient	Method	250	500	250
β_t	Wt x Odds	0.281	0.151	0.068
	GBM	0.186	0.110	0.048
	Match 1:2	0.274	0.135	0.057
	Match 1:1	0.223	0.115	0.051
	ExMatch 1:2	0.276	0.134	0.061
	ExMatch 1:1	0.240	0.115	0.052
	Group	0.230	0.106	0.052
	Regression	0.133	0.074	0.031
β_{tg}	Wt x Odds	0.623	0.308	0.159
	GBM	0.456	0.227	0.116
	Match 1:2	0.611	0.259	0.128
	Match 1:1	0.517	0.237	0.112
	ExMatch 1:2	0.590	0.272	0.135
	ExMatch 1:1	0.520	0.236	0.114
	Group	0.220	0.106	0.051
	Regression	0.278	0.133	0.061
β_g	Wt x Odds	0.456	0.234	0.121
	GBM	0.289	0.148	0.082
	Match 1:2	0.430	0.185	0.096
	Match 1:1	0.259	0.120	0.058
	ExMatch 1:2	0.417	0.192	0.100
	ExMatch 1:1	0.280	0.127	0.061
	Group	0.122	0.058	0.027
	Regression	0.141	0.070	0.030

Appendix E: Simulation II Data Generating Model D

Table E1
Metrics for Data Generating Model D, ATE Estimates, Misspecified model mA, n=500.

			Percentile Condition			
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.1027	0.1512	0.1233	0.0997
		MNPS	0.1190	0.1224	0.1134	0.0994
		IPTW	0.1148	0.1161	0.1014	0.0890
		GBM	0.1145	0.1127	0.1009	0.0900
	β_{tg}	Subclass	0.2099	0.2719	0.2491	0.2170
		MNPS	0.1680	0.1914	0.1783	0.1720
		IPTW	0.2560	0.2452	0.2542	0.2657
		GBM	0.1803	0.1986	0.1897	0.1840
	β_g	Subclass	0.1184	0.1536	0.1228	0.1212
		MNPS	0.0916	0.1059	0.1015	0.0919
		IPTW	0.1488	0.2182	0.1912	0.1923
		GBM	0.0975	0.1442	0.1290	0.1080
Bias	β_t	Subclass	0.0731	0.0737	0.0583	0.0558
		MNPS	0.1814	0.1516	0.1608	0.1515
		IPTW	0.0107	0.0267	0.0091	0.0021
		GBM	0.1701	0.1193	0.1189	0.1213
	β_{tg}	Subclass	-0.0009	-0.0008	-0.0185	-0.0009
		MNPS	-0.0115	0.0221	-0.0217	0.0020
		IPTW	-0.0081	-0.0168	-0.0295	-0.0172
		GBM	-0.0086	0.0613	0.0395	0.0630
	β_g	Subclass	-0.0091	-0.0174	0.0041	0.0208
		MNPS	0.0024	0.1036	0.0903	0.0553
		IPTW	0.0001	0.2887	0.2335	0.1691
		GBM	-0.0007	0.2131	0.1726	0.1008
Variance	β_t	Subclass	0.0974	0.1458	0.1199	0.0966
		MNPS	0.0862	0.0994	0.0876	0.0764
		IPTW	0.1147	0.1154	0.1013	0.0890
		GBM	0.0856	0.0985	0.0867	0.0753
	β_{tg}	Subclass	0.2099	0.2719	0.2488	0.2170
		MNPS	0.1679	0.1909	0.1778	0.1720

		IPTW	0.2560	0.2449	0.2534	0.2654
		GBM	0.1802	0.1948	0.1881	0.1800
	β_g	Subclass	0.1183	0.1533	0.1228	0.1208
		MNPS	0.0915	0.0952	0.0933	0.0888
		IPTW	0.1488	0.1349	0.1367	0.1637
		GBM	0.0975	0.0988	0.0992	0.0979

Table E2

Metrics for Data Generating Model D, ATE Estimates, Misspecified model mA, n=1000.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.0524	0.0969	0.0774	0.0589
		MNPS	0.0641	0.0675	0.0610	0.0640
		IPTW	0.0577	0.0584	0.0540	0.0497
		GBM	0.0610	0.0595	0.0554	0.0558
	β_{tg}	Subclass	0.1024	0.1402	0.1415	0.1102
		MNPS	0.0866	0.0908	0.0972	0.0917
		IPTW	0.1239	0.1160	0.1349	0.1244
		GBM	0.0933	0.0956	0.1058	0.0978
	β_g	Subclass	0.0579	0.0871	0.0775	0.0596
		MNPS	0.0455	0.0524	0.0490	0.0490
		IPTW	0.0679	0.1479	0.1333	0.1016
		GBM	0.0485	0.0972	0.0748	0.0604
Bias	β_t	Subclass	0.0577	0.0867	0.0772	0.0594
		MNPS	0.0455	0.0429	0.0452	0.0472
		IPTW	0.0678	0.0633	0.0796	0.0747
		GBM	0.0484	0.0460	0.0500	0.0512
	β_{tg}	Subclass	0.0183	-0.0076	0.0145	0.0100
		MNPS	0.0131	-0.0067	0.0088	-0.0083
		IPTW	0.0150	-0.0310	-0.0103	-0.0264
		GBM	0.0151	0.0381	0.0727	0.0509
	β_g	Subclass	-0.0108	-0.0198	-0.0153	0.0119
		MNPS	-0.0048	0.0972	0.0615	0.0431
		IPTW	-0.0075	0.2908	0.2316	0.1640
		GBM	-0.0084	0.2262	0.1577	0.0959

Variance	β_t	Subclass	0.0485	0.0923	0.0739	0.0542
		MNPS	0.0437	0.0492	0.0442	0.0433
		IPTW	0.0577	0.0580	0.0539	0.0493
		GBM	0.0442	0.0509	0.0477	0.0437
	β_{tg}	Subclass	0.1020	0.1402	0.1412	0.1101
		MNPS	0.0864	0.0907	0.0971	0.0917
		IPTW	0.1237	0.1151	0.1348	0.1237
		GBM	0.0931	0.0942	0.1005	0.0952
	β_g	Subclass	0.0577	0.0867	0.0772	0.0594
		MNPS	0.0455	0.0429	0.0452	0.0472
		IPTW	0.0678	0.0633	0.0796	0.0747
		GBM	0.0484	0.0460	0.0500	0.0512

Table E3
Metrics for Data Generating Model D, ATE Estimates, True Model D, n=250.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.2144	0.2486	0.2025	0.1942
		MNPS	0.2213	0.2123	0.1909	0.1847
		IPTW	0.2398	0.2252	0.1990	0.1894
		GBM	0.2136	0.1982	0.1811	0.1791
	β_{tg}	Subclass	0.4632	0.5145	0.4772	0.4702
		MNPS	0.3690	0.3656	0.3463	0.3810
		IPTW	0.5160	0.5245	0.5338	0.5273
		GBM	0.3722	0.3666	0.3499	0.3911
	β_g	Subclass	0.2129	0.2334	0.2432	0.2375
		MNPS	0.1808	0.2060	0.1971	0.2042
		IPTW	0.2827	0.3835	0.4121	0.3241
		GBM	0.1854	0.2282	0.2300	0.2142
Bias	β_t	Subclass	0.0604	0.0656	0.0722	0.0799
		MNPS	0.1884	0.1562	0.1723	0.1935
		IPTW	0.0056	0.0043	0.0223	0.0345
		GBM	0.1847	0.1239	0.1459	0.1671
	β_{tg}	Subclass	0.0045	0.0102	-0.0349	-0.0171
		MNPS	-0.0080	0.0485	0.0082	-0.0001

		IPTW	-0.0015	0.0003	-0.0717	-0.0234
		GBM	0.0049	0.0875	0.0507	0.0681
	β_g	Subclass	0.0031	0.0286	0.0598	0.0297
		MNPS	0.0119	0.1073	0.1147	0.0516
		IPTW	0.0102	0.2783	0.2874	0.1483
		GBM	0.0028	0.1933	0.1856	0.0762
Variance	β_t	Subclass	0.2107	0.2443	0.1973	0.1878
		MNPS	0.1858	0.1880	0.1612	0.1472
		IPTW	0.2397	0.2252	0.1985	0.1882
		GBM	0.1795	0.1829	0.1599	0.1512
	β_{tg}	Subclass	0.4632	0.5145	0.4772	0.4702
		MNPS	0.3690	0.3656	0.3463	0.3810
		IPTW	0.5160	0.5245	0.5338	0.5273
		GBM	0.3722	0.3666	0.3499	0.3911
	β_g	Subclass	0.2129	0.2325	0.2396	0.2366
		MNPS	0.1806	0.1945	0.1839	0.2016
		IPTW	0.2826	0.3060	0.3295	0.3021
		GBM	0.1854	0.1909	0.1956	0.2084

Table E4
Metrics for Data Generating Model D, ATE Estimates, True Model D, n=500.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.1044	0.1445	0.1222	0.1004
		MNPS	0.1190	0.1224	0.1134	0.0994
		IPTW	0.1143	0.1168	0.1017	0.0891
		GBM	0.1146	0.1123	0.1009	0.0901
	β_{tg}	Subclass	0.2143	0.2538	0.2437	0.2205
		MNPS	0.1680	0.1914	0.1783	0.1720
		IPTW	0.2544	0.2470	0.2538	0.2640
		GBM	0.1813	0.2001	0.1906	0.1855
	β_g	Subclass	0.1171	0.1402	0.1208	0.1198
		MNPS	0.0916	0.1059	0.1015	0.0919
		IPTW	0.1483	0.2167	0.1896	0.1904
		GBM	0.0980	0.1426	0.1282	0.1085

Bias	β_t	Subclass	0.0745	0.0748	0.0584	0.0563
		MNPS	0.1814	0.1516	0.1608	0.1515
		IPTW	0.0112	0.0213	0.0081	0.0016
		GBM	0.1696	0.1151	0.1171	0.1204
	β_{tg}	Subclass	-0.0049	0.0030	-0.0164	0.0016
		MNPS	-0.0115	0.0221	-0.0217	0.0020
		IPTW	-0.0072	-0.0181	-0.0300	-0.0158
		GBM	-0.0084	0.0640	0.0402	0.0629
	β_g	Subclass	-0.0051	0.0187	0.0215	0.0265
		MNPS	0.0024	0.1036	0.0903	0.0553
		IPTW	-0.0005	0.2822	0.2304	0.1664
		GBM	-0.0003	0.2071	0.1695	0.0993
Variance	β_t	Subclass	0.0989	0.1389	0.1188	0.0972
		MNPS	0.0862	0.0994	0.0876	0.0764
		IPTW	0.1142	0.1163	0.1016	0.0891
		GBM	0.0858	0.0990	0.0872	0.0756
	β_{tg}	Subclass	0.2143	0.2538	0.2435	0.2205
		MNPS	0.1679	0.1909	0.1778	0.1720
		IPTW	0.2543	0.2467	0.2529	0.2637
		GBM	0.1812	0.1960	0.1890	0.1815
	β_g	Subclass	0.0577	0.0867	0.0772	0.0594
		MNPS	0.0455	0.0429	0.0452	0.0472
		IPTW	0.0678	0.0633	0.0796	0.0747
		GBM	0.0484	0.0460	0.0500	0.0512

Table E5
Metrics for Data Generating Model D, ATE Estimates, True Model D, n=1000.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.0527	0.0909	0.0751	0.0582
		MNPS	0.0641	0.0675	0.0610	0.0640
		IPTW	0.0575	0.0591	0.0542	0.0498
		GBM	0.0610	0.0592	0.0553	0.0559
	β_{tg}	Subclass	0.1025	0.1327	0.1334	0.1119
		MNPS	0.0866	0.0908	0.0972	0.0917

		IPTW	0.1234	0.1177	0.1350	0.1240
		GBM	0.0939	0.0973	0.1074	0.0983
	β_g	Subclass	0.0564	0.0795	0.0715	0.0578
		MNPS	0.0455	0.0524	0.0490	0.0490
		IPTW	0.0678	0.1455	0.1310	0.1010
		GBM	0.0486	0.0945	0.0741	0.0604
Bias	β_t	Subclass	0.0637	0.0617	0.0611	0.0667
		MNPS	0.1427	0.1350	0.1297	0.1439
		IPTW	0.0010	0.0143	0.0057	0.0200
		GBM	0.1292	0.0875	0.0848	0.1092
	β_{tg}	Subclass	0.0174	0.0003	0.0135	0.0103
		MNPS	0.0131	-0.0067	0.0088	-0.0083
		IPTW	0.0148	-0.0322	-0.0099	-0.0256
		GBM	0.0151	0.0418	0.0749	0.0512
	β_g	Subclass	-0.0100	0.0092	0.0052	0.0142
		MNPS	-0.0048	0.0972	0.0615	0.0431
		IPTW	-0.0072	0.2846	0.2269	0.1630
		GBM	-0.0084	0.2196	0.1535	0.0944
Variance	β_t	Subclass	0.0577	0.0867	0.0772	0.0594
		MNPS	0.0455	0.0429	0.0452	0.0472
		IPTW	0.0678	0.0633	0.0796	0.0747
		GBM	0.0484	0.0460	0.0500	0.0512
	β_{tg}	Subclass	0.0486	0.0871	0.0714	0.0537
		MNPS	0.0437	0.0492	0.0442	0.0433
		IPTW	0.0575	0.0589	0.0542	0.0494
		GBM	0.0443	0.0516	0.0481	0.0440
	β_g	Subclass	0.0563	0.0794	0.0715	0.0576
		MNPS	0.0455	0.0429	0.0452	0.0472
		IPTW	0.0677	0.0645	0.0796	0.0745
		GBM	0.0485	0.0463	0.0505	0.0514

Table E6

Bias and variance for Data Generating Model D, ATT Estimates, Misspecified model mA, n=250.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
Bias	β_t	Wt x Odds	-0.0347	-0.0193	-0.0199	-0.0071
		GBM	0.0949	0.0615	0.0698	0.0829
		Match 1:2	-0.0054	-0.0177	-0.0159	0.0074
		Match 1:1	0.0045	-0.0343	-0.0002	0.0146
		ExMatch 1:2	0.0009	-0.0074	0.0061	0.0074
		ExMatch 1:1	-0.0208	-0.0098	0.0073	0.0116
		Group	-0.0012	-0.0197	0.0040	0.0168
		Regression	-0.0230	-0.0148	-0.0009	0.0052
	β_{tg}	Wt x Odds	-0.0046	-0.0044	-0.0673	-0.0099
		GBM	0.0177	0.0565	0.0234	0.0714
		Match 1:2	-0.0005	0.0427	0.0040	0.0336
		Match 1:1	-0.0208	0.0525	0.0064	0.0051
		ExMatch 1:2	-0.0075	0.0249	-0.0306	0.0010
		ExMatch 1:1	0.0028	0.0333	-0.0165	0.0080
		Group	-0.0033	0.0088	0.0093	-0.0104
		Regression	-0.0015	0.0116	-0.0205	0.0005
	β_g	Wt x Odds	0.0154	0.2899	0.3219	0.1801
		GBM	-0.0069	0.2290	0.2312	0.0988
		Match 1:2	0.0106	0.2272	0.2322	0.1169
		Match 1:1	0.0042	0.1247	0.1411	0.0478
		ExMatch 1:2	0.0167	0.2747	0.2843	0.1498
		ExMatch 1:1	0.0091	0.1469	0.1410	0.0640
		Group	0.0187	0.0953	0.0947	0.0594
		Regression	0.0095	-0.0084	0.0233	-0.0001
Variance	β_t	Wt x Odds	0.3300	0.2617	0.2487	0.2335
		GBM	0.2676	0.2250	0.2087	0.2057
		Match 1:2	0.3142	0.2848	0.2721	0.2444
		Match 1:1	0.2660	0.2391	0.2391	0.2228
		ExMatch 1:2	0.3114	0.2671	0.2549	0.2328
		ExMatch 1:1	0.2644	0.2582	0.2173	0.2143
		Group	0.2446	0.2701	0.2247	0.2290

	Regression	0.1493	0.1580	0.1307	0.1287
β_{tg}	Wt x Odds	0.6992	0.6954	0.7213	0.7064
	GBM	0.5290	0.4829	0.4924	0.5429
	Match 1:2	0.6386	0.6010	0.6187	0.6664
	Match 1:1	0.5551	0.4966	0.5357	0.5531
	ExMatch 1:2	0.6346	0.5719	0.6349	0.6586
	ExMatch 1:1	0.5476	0.5504	0.5234	0.5970
	Group	0.2602	0.2454	0.2945	0.3429
	Regression	0.2658	0.2819	0.2613	0.2913
β_g	Wt x Odds	0.5433	0.5406	0.5899	0.5642
	GBM	0.3665	0.3286	0.3601	0.3952
	Match 1:2	0.4773	0.4578	0.4857	0.5356
	Match 1:1	0.2593	0.2746	0.2685	0.2908
	ExMatch 1:2	0.4503	0.4288	0.5010	0.5327
	ExMatch 1:1	0.2722	0.2837	0.2748	0.3116
	Group	0.1316	0.1388	0.1560	0.1789
	Regression	0.1316	0.1489	0.1399	0.1463

Table E7
Metrics for Data Generating Model D, ATT Estimates, Misspecified model mA, n=500.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Wt x Odds	0.1646	0.1449	0.1264	0.1237
		GBM	0.1251	0.1195	0.1138	0.1108
		Match 1:2	0.1490	0.1505	0.1324	0.1247
		Match 1:1	0.1226	0.1325	0.1144	0.1018
		ExMatch 1:2	0.1403	0.1362	0.1304	0.1150
		ExMatch 1:1	0.1248	0.1326	0.1114	0.1087
		Group	0.1183	0.1256	0.1141	0.1041
		Regression	0.0706	0.0792	0.0715	0.0658
	β_{tg}	Wt x Odds	0.3579	0.3303	0.3385	0.3991
		GBM	0.2437	0.2588	0.2613	0.2656
		Match 1:2	0.3045	0.3140	0.2968	0.3246
		Match 1:1	0.2485	0.2821	0.2558	0.2458
		ExMatch 1:2	0.2855	0.3036	0.3128	0.3201
		ExMatch 1:1	0.2442	0.2721	0.2554	0.2430

		Group	0.1156	0.1147	0.1399	0.1404
		Regression	0.1169	0.1385	0.1323	0.1253
	β_g	Wt x Odds	0.2935	0.3344	0.3132	0.3612
		GBM	0.1860	0.2316	0.2336	0.2076
		Match 1:2	0.2470	0.2809	0.2583	0.2667
		Match 1:1	0.1312	0.1656	0.1452	0.1279
		ExMatch 1:2	0.2295	0.3014	0.2873	0.2761
		ExMatch 1:1	0.1354	0.1670	0.1447	0.1334
		Group	0.0685	0.0712	0.0782	0.0794
		Regression	0.0637	0.0692	0.0704	0.0646
Bias	β_t	Wt x Odds	-0.0263	-0.0156	-0.0270	-0.0285
		GBM	0.0885	0.0433	0.0521	0.0493
		Match 1:2	0.0214	-0.0131	-0.0031	-0.0192
		Match 1:1	0.0320	-0.0073	0.0041	-0.0060
		ExMatch 1:2	0.0120	0.0052	0.0070	-0.0061
		ExMatch 1:1	0.0159	0.0087	0.0093	-0.0049
		Group	0.0236	0.0093	0.0141	-0.0097
		Regression	0.0108	-0.0017	0.0020	-0.0106
	β_{tg}	Wt x Odds	-0.0033	-0.0147	-0.0213	-0.0187
		GBM	-0.0072	0.0450	0.0193	0.0494
		Match 1:2	-0.0022	0.0423	0.0143	0.0528
		Match 1:1	-0.0124	0.0455	-0.0080	0.0123
		ExMatch 1:2	0.0026	0.0154	-0.0060	0.0134
		ExMatch 1:1	-0.0140	0.0173	-0.0278	0.0086
		Group	0.0110	0.0188	-0.0150	0.0080
		Regression	-0.0071	0.0110	-0.0213	-0.0019
	β_g	Wt x Odds	-0.0028	0.2988	0.2488	0.2028
		GBM	0.0011	0.2392	0.2082	0.1346
		Match 1:2	-0.0031	0.2352	0.2065	0.1215
		Match 1:1	0.0170	0.1385	0.1143	0.0613
		ExMatch 1:2	-0.0067	0.2812	0.2405	0.1672
		ExMatch 1:1	0.0025	0.1588	0.1220	0.0582
		Group	-0.0035	0.0849	0.0774	0.0481
		Regression	0.0025	-0.0141	0.0097	0.0073
Variance	β_t	Wt x Odds	0.1639	0.1447	0.1257	0.1229
		GBM	0.1173	0.1176	0.1111	0.1083

	Match 1:2	0.1486	0.1503	0.1324	0.1244
	Match 1:1	0.1216	0.1325	0.1144	0.1017
	ExMatch 1:2	0.1402	0.1362	0.1303	0.1150
	ExMatch 1:1	0.1245	0.1325	0.1114	0.1086
	Group	0.1177	0.1255	0.1139	0.1040
	Regression	0.0705	0.0792	0.0715	0.0657
β_{tg}	Wt x Odds	0.3579	0.3301	0.3381	0.3987
	GBM	0.2437	0.2568	0.2609	0.2632
	Match 1:2	0.3045	0.3122	0.2966	0.3218
	Match 1:1	0.2484	0.2800	0.2558	0.2457
	ExMatch 1:2	0.2855	0.3034	0.3128	0.3199
	ExMatch 1:1	0.2440	0.2718	0.2547	0.2430
	Group	0.1155	0.1144	0.1397	0.1403
	Regression	0.1168	0.1384	0.1318	0.1253
β_g	Wt x Odds	0.2935	0.2451	0.2513	0.3201
	GBM	0.1860	0.1744	0.1902	0.1895
	Match 1:2	0.2469	0.2256	0.2156	0.2519
	Match 1:1	0.1309	0.1464	0.1321	0.1242
	ExMatch 1:2	0.2294	0.2224	0.2294	0.2481
	ExMatch 1:1	0.1354	0.1418	0.1298	0.1300
	Group	0.0685	0.0640	0.0722	0.0771
	Regression	0.0637	0.0690	0.0703	0.0645

Table E8
Metrics for Data Generating Model D, ATT Estimates, Misspecified model mA, n=1000.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Wt x Odds	0.0890	0.0698	0.0662	0.0608
		GBM	0.0671	0.0619	0.0592	0.0565
		Match 1:2	0.0734	0.0709	0.0675	0.0649
		Match 1:1	0.0612	0.0669	0.0597	0.0568
		ExMatch 1:2	0.0729	0.0714	0.0638	0.0624
		ExMatch 1:1	0.0602	0.0638	0.0587	0.0586
		Group	0.0620	0.0621	0.0583	0.0551
		Regression	0.0361	0.0415	0.0373	0.0347

	β_{tg}	Wt x Odds	0.1902	0.1573	0.1879	0.1810
		GBM	0.1389	0.1234	0.1368	0.1348
		Match 1:2	0.1645	0.1454	0.1516	0.1532
		Match 1:1	0.1251	0.1304	0.1368	0.1297
		ExMatch 1:2	0.1556	0.1400	0.1471	0.1518
		ExMatch 1:1	0.1209	0.1266	0.1374	0.1307
		Group	0.0589	0.0593	0.0644	0.0710
		Regression	0.0634	0.0672	0.0698	0.0655
	β_g	Wt x Odds	0.1446	0.2088	0.2167	0.1892
		GBM	0.0971	0.1447	0.1273	0.1138
		Match 1:2	0.1230	0.1634	0.1428	0.1314
		Match 1:1	0.0643	0.0830	0.0762	0.0724
		ExMatch 1:2	0.1158	0.1840	0.1580	0.1426
		ExMatch 1:1	0.0650	0.0886	0.0783	0.0693
		Group	0.0314	0.0383	0.0365	0.0388
		Regression	0.0331	0.0317	0.0331	0.0343
Bias	β_t	Wt x Odds	-0.0391	-0.0246	-0.0250	-0.0143
		GBM	0.0628	0.0283	0.0263	0.0358
		Match 1:2	0.0074	-0.0200	-0.0201	-0.0045
		Match 1:1	0.0140	-0.0053	-0.0091	0.0069
		ExMatch 1:2	0.0024	0.0074	-0.0024	0.0056
		ExMatch 1:1	0.0123	0.0101	0.0103	0.0110
		Group	0.0189	0.0084	0.0064	0.0123
		Regression	-0.0001	-0.0012	-0.0029	0.0050
	β_{tg}	Wt x Odds	0.0141	-0.0258	-0.0177	-0.0307
		GBM	0.0139	0.0252	0.0531	0.0518
		Match 1:2	0.0154	0.0276	0.0532	0.0371
		Match 1:1	0.0137	0.0254	0.0469	0.0272
		ExMatch 1:2	0.0142	-0.0035	0.0247	0.0060
		ExMatch 1:1	0.0108	0.0007	0.0076	0.0113
		Group	0.0215	0.0089	0.0177	0.0253
		Regression	0.0172	-0.0036	0.0124	0.0025
	β_g	Wt x Odds	-0.0037	0.2997	0.2636	0.2056
		GBM	-0.0034	0.2487	0.1927	0.1231
		Match 1:2	-0.0053	0.2437	0.1904	0.1352
		Match 1:1	-0.0126	0.1407	0.0909	0.0563

		ExMatch 1:2	-0.0045	0.2873	0.2272	0.1701
		ExMatch 1:1	-0.0071	0.1551	0.1051	0.0520
		Group	0.0058	0.0914	0.0631	0.0355
		Regression	-0.0081	-0.0036	-0.0104	0.0025
Variance	β_t	Wt x Odds	0.0875	0.0692	0.0656	0.0606
		GBM	0.0631	0.0611	0.0585	0.0552
		Match 1:2	0.0734	0.0705	0.0670	0.0649
		Match 1:1	0.0610	0.0669	0.0596	0.0567
		ExMatch 1:2	0.0728	0.0713	0.0638	0.0623
		ExMatch 1:1	0.0601	0.0637	0.0586	0.0585
		Group	0.0617	0.0620	0.0583	0.0549
		Regression	0.0361	0.0415	0.0373	0.0347
	β_{tg}	Wt x Odds	0.1900	0.1566	0.1876	0.1801
		GBM	0.1387	0.1228	0.1340	0.1321
		Match 1:2	0.1643	0.1447	0.1488	0.1518
		Match 1:1	0.1249	0.1298	0.1346	0.1289
		ExMatch 1:2	0.1554	0.1400	0.1465	0.1517
		ExMatch 1:1	0.1208	0.1266	0.1373	0.1306
		Group	0.0584	0.0592	0.0641	0.0704
		Regression	0.0631	0.0672	0.0696	0.0655
	β_g	Wt x Odds	0.1446	0.1189	0.1472	0.1469
		GBM	0.0971	0.0828	0.0901	0.0987
		Match 1:2	0.1230	0.1040	0.1066	0.1131
		Match 1:1	0.0641	0.0632	0.0679	0.0692
		ExMatch 1:2	0.1158	0.1015	0.1064	0.1136
		ExMatch 1:1	0.0650	0.0646	0.0673	0.0666
		Group	0.0313	0.0299	0.0325	0.0375
		Regression	0.0330	0.0317	0.0330	0.0343

Table E9
Metrics for Data Generating Model D, ATT Estimates, True Model D, n=250.

			Percentile Condition			
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Wt x Odds	0.3293	0.2645	0.2491	0.2318
		GBM	0.2773	0.2322	0.2153	0.2118
		Match 1:2	0.3174	0.2863	0.2597	0.2453
		Match 1:1	0.2754	0.2474	0.2260	0.2234
		ExMatch 1:2	0.3091	0.2807	0.2506	0.2347
		ExMatch 1:1	0.2627	0.2614	0.2388	0.2200
		Group	0.2541	0.2608	0.2218	0.2293
		Regression	0.1498	0.1582	0.1307	0.1287
	β_{tg}	Wt x Odds	0.6936	0.7055	0.7255	0.6984
		GBM	0.5288	0.4912	0.5014	0.5509
		Match 1:2	0.6520	0.6127	0.6143	0.6528
		Match 1:1	0.5640	0.5192	0.5389	0.5411
		ExMatch 1:2	0.6171	0.5659	0.6206	0.6593
		ExMatch 1:1	0.5683	0.5291	0.5743	0.5593
		Group	0.2629	0.2427	0.3048	0.3368
		Regression	0.2658	0.2820	0.2618	0.2913
	β_g	Wt x Odds	0.5387	0.6363	0.6945	0.5878
		GBM	0.3657	0.3835	0.4175	0.4048
		Match 1:2	0.4773	0.5108	0.5445	0.5288
		Match 1:1	0.2782	0.3009	0.3004	0.2815
		ExMatch 1:2	0.4459	0.5109	0.5697	0.5529
		ExMatch 1:1	0.2704	0.3062	0.3210	0.2982
		Group	0.1349	0.1397	0.1671	0.1836
		Regression	0.1317	0.1490	0.1404	0.1463
Bias	β_t	Wt x Odds	-0.0330	-0.0225	-0.0198	-0.0059
		GBM	0.0947	0.0542	0.0682	0.0811
		Match 1:2	0.0064	-0.0337	0.0034	-0.0074
		Match 1:1	-0.0088	-0.0111	-0.0023	0.0162
		ExMatch 1:2	-0.0024	-0.0147	0.0038	0.0131
		ExMatch 1:1	-0.0071	-0.0088	0.0003	0.0294
		Group	-0.0038	-0.0072	0.0112	0.0203

		Regression	-0.0230	-0.0148	-0.0009	0.0052
	β_{tg}	Wt x Odds	-0.0044	-0.0058	-0.0680	-0.0088
		GBM	0.0199	0.0631	0.0264	0.0700
		Match 1:2	-0.0070	0.0662	-0.0232	0.0351
		Match 1:1	-0.0124	0.0271	0.0003	0.0135
		ExMatch 1:2	0.0107	0.0267	-0.0310	-0.0008
		ExMatch 1:1	-0.0302	0.0227	-0.0254	-0.0188
		Group	-0.0159	0.0141	-0.0116	0.0163
		Regression	-0.0015	0.0116	-0.0205	0.0005
	β_{g}	Wt x Odds	0.0152	0.2912	0.3226	0.1790
		GBM	-0.0092	0.2224	0.2282	0.1002
		Match 1:2	0.0179	0.2046	0.2576	0.1153
		Match 1:1	0.0103	0.1602	0.1366	0.0643
		ExMatch 1:2	-0.0007	0.2700	0.2837	0.1530
		ExMatch 1:1	0.0227	0.1523	0.1490	0.0770
		Group	0.0208	0.0947	0.1004	0.0412
		Regression	0.0095	-0.0084	0.0233	-0.0001
Variance	β_{t}	Wt x Odds	0.3282	0.2640	0.2487	0.2317
		GBM	0.2683	0.2293	0.2106	0.2052
		Match 1:2	0.3174	0.2852	0.2597	0.2452
		Match 1:1	0.2753	0.2473	0.2260	0.2232
		ExMatch 1:2	0.3090	0.2805	0.2506	0.2345
		ExMatch 1:1	0.2626	0.2613	0.2388	0.2191
		Group	0.2541	0.2608	0.2216	0.2289
		Regression	0.1493	0.1580	0.1307	0.1287
	β_{tg}	Wt x Odds	0.6936	0.7055	0.7208	0.6984
		GBM	0.5284	0.4872	0.5007	0.5460
		Match 1:2	0.6520	0.6083	0.6137	0.6515
		Match 1:1	0.5639	0.5185	0.5389	0.5409
		ExMatch 1:2	0.6170	0.5652	0.6197	0.6593
		ExMatch 1:1	0.5674	0.5286	0.5736	0.5590
		Group	0.2626	0.2425	0.3047	0.3366
		Regression	0.2658	0.2819	0.2613	0.2913
	β_{g}	Wt x Odds	0.5385	0.5514	0.5905	0.5557
		GBM	0.3656	0.3341	0.3655	0.3948
		Match 1:2	0.4770	0.4690	0.4781	0.5155

	Match 1:1	0.2781	0.2752	0.2817	0.2773
	ExMatch 1:2	0.4459	0.4380	0.4893	0.5295
	ExMatch 1:1	0.2699	0.2830	0.2988	0.2922
	Group	0.1345	0.1307	0.1570	0.1819
	Regression	0.1316	0.1489	0.1399	0.1463

Table E10
Metrics for Data Generating Model D, ATT Estimates, True Model D, n=500.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Wt x Odds	0.1637	0.1478	0.1267	0.1239
		GBM	0.1256	0.1218	0.1151	0.1122
		Match 1:2	0.1435	0.1549	0.1382	0.1274
		Match 1:1	0.1230	0.1288	0.1188	0.1066
		ExMatch 1:2	0.1417	0.1454	0.1240	0.1183
		ExMatch 1:1	0.1242	0.1343	0.1120	0.1058
		Group	0.1197	0.1238	0.1103	0.1071
		Regression	0.0706	0.0792	0.0715	0.0658
	β_{tg}	Wt x Odds	0.3557	0.3372	0.3390	0.3997
		GBM	0.2427	0.2632	0.2645	0.2679
		Match 1:2	0.3022	0.3212	0.3073	0.3116
		Match 1:1	0.2456	0.2757	0.2648	0.2633
		ExMatch 1:2	0.2817	0.3217	0.3029	0.3324
		ExMatch 1:1	0.2443	0.2717	0.2601	0.2578
		Group	0.1249	0.1197	0.1324	0.1459
		Regression	0.1169	0.1385	0.1323	0.1253
	β_g	Wt x Odds	0.2913	0.3429	0.3144	0.3616
		GBM	0.1851	0.2336	0.2362	0.2091
		Match 1:2	0.2353	0.2825	0.2621	0.2537
		Match 1:1	0.1189	0.1654	0.1448	0.1344
		ExMatch 1:2	0.2283	0.3123	0.2736	0.2943
		ExMatch 1:1	0.1294	0.1728	0.1474	0.1278
		Group	0.0645	0.0743	0.0773	0.0764
		Regression	0.0637	0.0692	0.0704	0.0646
Bias	β_t	Wt x Odds	-0.0255	-0.0199	-0.0276	-0.0285

		GBM	0.0879	0.0378	0.0511	0.0480
		Match 1:2	0.0135	-0.0197	-0.0172	-0.0143
		Match 1:1	0.0264	-0.0142	0.0000	-0.0089
		ExMatch 1:2	0.0086	0.0050	0.0018	-0.0014
		ExMatch 1:1	0.0277	0.0056	0.0101	-0.0093
		Group	0.0226	-0.0044	0.0093	-0.0117
		Regression	0.0108	-0.0017	0.0020	-0.0106
	β_{tg}	Wt x Odds	-0.0034	-0.0164	-0.0221	-0.0187
		GBM	-0.0064	0.0485	0.0208	0.0499
		Match 1:2	0.0023	0.0481	0.0194	0.0403
		Match 1:1	-0.0091	0.0376	-0.0022	0.0291
		ExMatch 1:2	0.0114	0.0168	-0.0042	0.0060
		ExMatch 1:1	-0.0191	0.0159	-0.0281	0.0095
		Group	0.0088	0.0156	-0.0220	0.0057
		Regression	-0.0071	0.0110	-0.0213	-0.0019
	β_{g}	Wt x Odds	-0.0027	0.3005	0.2496	0.2028
		GBM	0.0003	0.2356	0.2067	0.1342
		Match 1:2	-0.0080	0.2302	0.2014	0.1347
		Match 1:1	0.0051	0.1495	0.1180	0.0634
		ExMatch 1:2	-0.0152	0.2786	0.2385	0.1746
		ExMatch 1:1	0.0049	0.1532	0.1210	0.0563
		Group	0.0029	0.0904	0.0778	0.0434
		Regression	0.0025	-0.0141	0.0097	0.0073
Variance	β_{t}	Wt x Odds	0.1630	0.1474	0.1259	0.1231
		GBM	0.1179	0.1204	0.1125	0.1099
		Match 1:2	0.1433	0.1545	0.1379	0.1272
		Match 1:1	0.1223	0.1286	0.1188	0.1065
		ExMatch 1:2	0.1416	0.1453	0.1240	0.1183
		ExMatch 1:1	0.1234	0.1343	0.1119	0.1057
		Group	0.1192	0.1238	0.1102	0.1070
		Regression	0.0705	0.0792	0.0715	0.0657
	β_{tg}	Wt x Odds	0.3556	0.3370	0.3385	0.3993
		GBM	0.2427	0.2608	0.2640	0.2654
		Match 1:2	0.3022	0.3189	0.3069	0.3100
		Match 1:1	0.2455	0.2743	0.2648	0.2625
		ExMatch 1:2	0.2815	0.3214	0.3029	0.3324

	ExMatch 1:1	0.2440	0.2714	0.2593	0.2577
	Group	0.1249	0.1195	0.1319	0.1459
	Regression	0.1168	0.1384	0.1318	0.1253
β_g	Wt x Odds	0.2913	0.2525	0.2521	0.3205
	GBM	0.1851	0.1781	0.1935	0.1911
	Match 1:2	0.2352	0.2295	0.2216	0.2355
	Match 1:1	0.1189	0.1430	0.1308	0.1304
	ExMatch 1:2	0.2281	0.2347	0.2168	0.2638
	ExMatch 1:1	0.1294	0.1493	0.1328	0.1246
	Group	0.0645	0.0662	0.0713	0.0745
	Regression	0.0637	0.0690	0.0703	0.0645

Table E11
Metrics for Data Generating Model D, ATT Estimates, True Model D, n=1000.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Wt x Odds	0.0887	0.0718	0.0666	0.0608
		GBM	0.0674	0.0632	0.0600	0.0569
		Match 1:2	0.0762	0.0726	0.0684	0.0650
		Match 1:1	0.0598	0.0651	0.0595	0.0556
		ExMatch 1:2	0.0733	0.0676	0.0642	0.0612
		ExMatch 1:1	0.0614	0.0622	0.0563	0.0552
		Group	0.0580	0.0626	0.0560	0.0555
		Regression	0.0361	0.0415	0.0373	0.0347
	β_{tg}	Wt x Odds	0.1897	0.1619	0.1884	0.1802
		GBM	0.1399	0.1256	0.1391	0.1366
		Match 1:2	0.1593	0.1433	0.1539	0.1532
		Match 1:1	0.1204	0.1318	0.1362	0.1344
		ExMatch 1:2	0.1577	0.1418	0.1459	0.1470
		ExMatch 1:1	0.1219	0.1262	0.1293	0.1207
		Group	0.0600	0.0545	0.0631	0.0695
		Regression	0.0634	0.0672	0.0698	0.0655
	β_g	Wt x Odds	0.1441	0.2146	0.2173	0.1884
		GBM	0.0970	0.1446	0.1277	0.1146
		Match 1:2	0.1179	0.1667	0.1516	0.1390
		Match 1:1	0.0628	0.0890	0.0785	0.0708

		ExMatch 1:2	0.1205	0.1889	0.1595	0.1417
		ExMatch 1:1	0.0627	0.0868	0.0750	0.0625
		Group	0.0320	0.0383	0.0369	0.0384
		Regression	0.0331	0.0317	0.0331	0.0343
Bias	β_t	Wt x Odds	-0.0385	-0.0295	-0.0266	-0.0142
		GBM	0.0625	0.0217	0.0224	0.0346
		Match 1:2	0.0097	-0.0160	-0.0149	-0.0004
		Match 1:1	0.0107	-0.0025	-0.0089	0.0035
		ExMatch 1:2	0.0049	0.0099	0.0005	0.0060
		ExMatch 1:1	0.0101	0.0128	0.0029	0.0133
		Group	0.0138	0.0120	0.0064	0.0113
		Regression	-0.0001	-0.0012	-0.0029	0.0050
	β_{tg}	Wt x Odds	0.0141	-0.0278	-0.0181	-0.0306
		GBM	0.0141	0.0297	0.0561	0.0527
		Match 1:2	0.0068	0.0176	0.0368	0.0252
		Match 1:1	0.0143	0.0161	0.0386	0.0338
		ExMatch 1:2	0.0157	-0.0101	0.0169	0.0059
		ExMatch 1:1	0.0102	-0.0130	0.0177	0.0112
		Group	0.0211	0.0019	0.0202	0.0249
		Regression	0.0172	-0.0036	0.0124	0.0025
	β_g	Wt x Odds	-0.0037	0.3017	0.2639	0.2055
		GBM	-0.0037	0.2442	0.1898	0.1222
		Match 1:2	0.0033	0.2537	0.2068	0.1468
		Match 1:1	-0.0075	0.1539	0.1070	0.0533
		ExMatch 1:2	-0.0058	0.2930	0.2347	0.1699
		ExMatch 1:1	-0.0029	0.1571	0.1041	0.0564
		Group	0.0041	0.0940	0.0640	0.0361
		Regression	-0.0081	-0.0036	-0.0104	0.0025
Variance	β_t	Wt x Odds	0.0872	0.0710	0.0659	0.0606
		GBM	0.0635	0.0627	0.0595	0.0557
		Match 1:2	0.0761	0.0724	0.0682	0.0650
		Match 1:1	0.0597	0.0651	0.0594	0.0555
		ExMatch 1:2	0.0733	0.0675	0.0642	0.0612
		ExMatch 1:1	0.0613	0.0620	0.0563	0.0550
		Group	0.0578	0.0625	0.0559	0.0553

	Regression	0.0361	0.0415	0.0373	0.0347
β_{tg}	Wt x Odds	0.1895	0.1612	0.1881	0.1793
	GBM	0.1397	0.1247	0.1359	0.1338
	Match 1:2	0.1593	0.1430	0.1525	0.1526
	Match 1:1	0.1202	0.1315	0.1347	0.1332
	ExMatch 1:2	0.1575	0.1417	0.1456	0.1470
	ExMatch 1:1	0.1218	0.1261	0.1290	0.1205
	Group	0.0595	0.0545	0.0627	0.0688
	Regression	0.0631	0.0672	0.0696	0.0655
β_{g}	Wt x Odds	0.1441	0.1235	0.1476	0.1461
	GBM	0.0970	0.0850	0.0917	0.0996
	Match 1:2	0.1179	0.1023	0.1088	0.1175
	Match 1:1	0.0627	0.0653	0.0670	0.0679
	ExMatch 1:2	0.1204	0.1031	0.1045	0.1128
	ExMatch 1:1	0.0627	0.0622	0.0642	0.0593
	Group	0.0320	0.0295	0.0328	0.0371
	Regression	0.0330	0.0317	0.0330	0.0343

Appendix F: Simulation II Data Generating Model E

Table F1
Metrics for Data Generating Model E, ATE Estimates, Misspecified model mA, n=500.

			Percentile Condition			
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.1660	0.2127	0.1856	0.1534
		MNPS	0.1051	0.1015	0.0965	0.0930
		IPTW	0.1033	0.1031	0.0932	0.0903
		GBM	0.1013	0.0960	0.0889	0.0875
	β_{tg}	Subclass	0.2020	0.2948	0.2256	0.2056
		MNPS	0.1863	0.1795	0.1816	0.1769
		IPTW	0.2169	0.2215	0.2128	0.2247
		GBM	0.1900	0.1909	0.1839	0.1910
	β_g	Subclass	0.1550	0.2505	0.1925	0.1472
		MNPS	0.0946	0.0924	0.0930	0.0943
		IPTW	0.1159	0.1384	0.1424	0.1270
		GBM	0.1008	0.1159	0.1144	0.1068
Bias	β_t	Subclass	-0.2557	-0.2497	-0.2678	-0.2551
		MNPS	0.1043	0.0958	0.0944	0.1167
		IPTW	-0.0189	-0.0358	-0.0416	-0.0270
		GBM	0.0987	0.0579	0.0470	0.0831
	β_{tg}	Subclass	0.0014	-0.0147	-0.0034	0.0069
		MNPS	-0.0036	0.0045	-0.0021	-0.0195
		IPTW	0.0025	0.0448	0.0315	0.0287
		GBM	-0.0044	0.0611	0.0664	0.0422
	β_g	Subclass	-0.2004	-0.3204	-0.2536	-0.2006
		MNPS	-0.0009	0.0777	0.0650	0.0532
		IPTW	0.0015	0.2043	0.1801	0.1249
		GBM	0.0047	0.1721	0.1440	0.1066
Variance	β_t	Subclass	0.1007	0.1504	0.1138	0.0883
		MNPS	0.0942	0.0923	0.0876	0.0793
		IPTW	0.1030	0.1018	0.0915	0.0896
		GBM	0.0916	0.0927	0.0867	0.0806
	β_{tg}	Subclass	0.2020	0.2946	0.2256	0.2056
		MNPS	0.1863	0.1795	0.1816	0.1765

		IPTW	0.2169	0.2195	0.2118	0.2239
		GBM	0.1900	0.1872	0.1794	0.1892
	β_g	Subclass	0.1148	0.1479	0.1282	0.1070
		MNPS	0.0946	0.0863	0.0888	0.0915
		IPTW	0.1159	0.0966	0.1099	0.1114
		GBM	0.1008	0.0863	0.0937	0.0955

Table F2

Metrics for Data Generating Model E, ATE Estimates, Misspecified model mA, n=1000.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.1081	0.1545	0.1145	0.1032
		MNPS	0.0551	0.0603	0.0579	0.0465
		IPTW	0.0491	0.0553	0.0486	0.0399
		GBM	0.0522	0.0527	0.0492	0.0427
	β_{tg}	Subclass	0.0947	0.1646	0.1233	0.0961
		MNPS	0.0905	0.0956	0.0979	0.0814
		IPTW	0.1035	0.1129	0.1028	0.0975
		GBM	0.0948	0.1014	0.0974	0.0892
	β_g	Subclass	0.0910	0.1983	0.1200	0.0984
		MNPS	0.0444	0.0519	0.0523	0.0446
		IPTW	0.0526	0.1012	0.0893	0.0675
		GBM	0.0467	0.0878	0.0775	0.0582
Bias	β_t	Subclass	-0.2495	-0.2383	-0.2307	-0.2479
		MNPS	0.0914	0.0952	0.1069	0.1015
		IPTW	-0.0145	-0.0226	-0.0168	-0.0222
		GBM	0.0811	0.0517	0.0613	0.0680
	β_{tg}	Subclass	0.0008	-0.0352	-0.0360	0.0055
		MNPS	-0.0005	-0.0243	-0.0381	-0.0219
		IPTW	0.0032	0.0175	0.0103	0.0293
		GBM	-0.0020	0.0261	0.0269	0.0383
	β_g	Subclass	-0.2003	-0.3175	-0.2352	-0.2073
		MNPS	-0.0019	0.0830	0.0786	0.0475
		IPTW	-0.0034	0.2176	0.1919	0.1215
		GBM	-0.0029	0.1973	0.1662	0.1039

Variance	β_t	Subclass	0.0458	0.0977	0.0612	0.0418
		MNPS	0.0467	0.0512	0.0464	0.0362
		IPTW	0.0489	0.0548	0.0484	0.0394
		GBM	0.0457	0.0501	0.0455	0.0380
	β_{tg}	Subclass	0.0947	0.1634	0.1220	0.0960
		MNPS	0.0905	0.0950	0.0964	0.0809
		IPTW	0.1035	0.1126	0.1027	0.0966
		GBM	0.0947	0.1007	0.0966	0.0878
	β_g	Subclass	0.0509	0.0975	0.0646	0.0555
		MNPS	0.0444	0.0451	0.0462	0.0424
		IPTW	0.0526	0.0539	0.0525	0.0527
		GBM	0.0467	0.0488	0.0498	0.0474

Table F3
Metrics for Data Generating Model E, ATE Estimates, True Model E, n=250.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.2541	0.2877	0.2675	0.2320
		MNPS	0.1887	0.2032	0.1758	0.1686
		IPTW	0.2922	0.2473	0.2109	0.1982
		GBM	0.1868	0.1848	0.1722	0.1572
	β_{tg}	Subclass	0.4134	0.5051	0.5005	0.4395
		MNPS	0.3111	0.3224	0.3448	0.3140
		IPTW	0.6406	0.6639	0.5848	0.5881
		GBM	0.3412	0.3464	0.3761	0.2957
	β_g	Subclass	0.2440	0.2652	0.2599	0.2186
		MNPS	0.1586	0.1756	0.1801	0.1594
		IPTW	0.4120	0.4707	0.4275	0.4278
		GBM	0.1757	0.1989	0.2035	0.1636
Bias	β_t	Subclass	-0.2265	-0.2358	-0.2460	-0.2113
		MNPS	0.1568	0.1433	0.1097	0.1605
		IPTW	-0.0209	-0.0017	-0.0174	0.0370
		GBM	0.1484	0.0858	0.0739	0.1272
	β_{tg}	Subclass	-0.0268	0.0169	0.0202	-0.0629
		MNPS	-0.0121	0.0027	0.0269	-0.0417

		IPTW	-0.0187	-0.0424	-0.0004	-0.1321
		GBM	-0.0117	0.0719	0.1023	0.0252
	β_g	Subclass	-0.1803	-0.1393	-0.1340	-0.1133
		MNPS	0.0200	0.0762	0.0615	0.0663
		IPTW	0.0083	0.2472	0.2031	0.2332
		GBM	0.0133	0.1404	0.1102	0.0997
Variance	β_t	Subclass	0.2028	0.2321	0.2070	0.1873
		MNPS	0.1641	0.1826	0.1637	0.1429
		IPTW	0.2918	0.2473	0.2106	0.1969
		GBM	0.1648	0.1775	0.1668	0.1410
	β_{tg}	Subclass	0.4127	0.5048	0.5001	0.4356
		MNPS	0.3109	0.3224	0.3441	0.3123
		IPTW	0.6402	0.6621	0.5848	0.5706
		GBM	0.3411	0.3412	0.3656	0.2951
	β_g	Subclass	0.2114	0.2458	0.2420	0.2058
		MNPS	0.1582	0.1698	0.1763	0.1550
		IPTW	0.4119	0.4095	0.3862	0.3734
		GBM	0.1756	0.1791	0.1913	0.1537

Table F4
Metrics for Data Generating Model E, ATE Estimates, True Model E, n=500.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.1715	0.2078	0.1765	0.1521
		MNPS	0.1038	0.1008	0.0965	0.0893
		IPTW	0.1649	0.1142	0.1119	0.1194
		GBM	0.0990	0.0940	0.0871	0.0860
	β_{tg}	Subclass	0.2334	0.2716	0.2464	0.2304
		MNPS	0.1731	0.1688	0.1691	0.1627
		IPTW	0.3891	0.3387	0.3423	0.3759
		GBM	0.1872	0.1926	0.1824	0.1889
	β_g	Subclass	0.1703	0.1660	0.1469	0.1487
		MNPS	0.0909	0.0891	0.0874	0.0901
		IPTW	0.2810	0.2730	0.2947	0.2945
		GBM	0.1002	0.1132	0.1121	0.1039

Bias	β_t	Subclass	-0.2502	-0.2467	-0.2549	-0.2340
		MNPS	0.1203	0.1060	0.1055	0.1295
		IPTW	-0.0490	-0.0077	-0.0207	-0.0082
		GBM	0.1108	0.0613	0.0564	0.0929
	β_{tg}	Subclass	0.0117	0.0220	0.0088	-0.0060
		MNPS	-0.0050	0.0169	0.0003	-0.0197
		IPTW	-0.0013	-0.0556	-0.0743	-0.0649
		GBM	-0.0023	0.0728	0.0712	0.0466
	β_g	Subclass	-0.2058	-0.1461	-0.1334	-0.1377
		MNPS	-0.0002	0.0822	0.0685	0.0534
		IPTW	0.0030	0.2751	0.2666	0.2056
		GBM	0.0045	0.1605	0.1375	0.1003
Variance	β_t	Subclass	0.1089	0.1470	0.1116	0.0974
		MNPS	0.0893	0.0896	0.0853	0.0726
		IPTW	0.1625	0.1141	0.1115	0.1194
		GBM	0.0867	0.0902	0.0839	0.0773
	β_{tg}	Subclass	0.2332	0.2711	0.2463	0.2304
		MNPS	0.1730	0.1685	0.1691	0.1623
		IPTW	0.3891	0.3356	0.3368	0.3717
		GBM	0.1872	0.1873	0.1774	0.1868
	β_g	Subclass	0.1279	0.1447	0.1291	0.1298
		MNPS	0.0909	0.0823	0.0827	0.0873
		IPTW	0.2810	0.1973	0.2236	0.2522
		GBM	0.1002	0.0875	0.0932	0.0939

Table F5

Metrics for Data Generating Model E, ATE Estimates, True Model E, n=1000.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Subclass	0.1058	0.1358	0.1230	0.0968
		MNPS	0.0537	0.0577	0.0562	0.0470
		IPTW	0.1117	0.0665	0.0584	0.0499
		GBM	0.0524	0.0513	0.0474	0.0422
	β_{tg}	Subclass	0.1043	0.1472	0.1365	0.1049
		MNPS	0.0836	0.0897	0.0898	0.0759

		IPTW	0.2382	0.1886	0.2135	0.2240
		GBM	0.0943	0.1015	0.0963	0.0880
	β_g	Subclass	0.0959	0.1006	0.0911	0.0796
		MNPS	0.0417	0.0502	0.0493	0.0428
		IPTW	0.1833	0.2017	0.2200	0.2196
		GBM	0.0471	0.0842	0.0753	0.0585
Bias	β_t	Subclass	-0.2349	-0.2269	-0.2301	-0.2290
		MNPS	0.1030	0.1006	0.1165	0.1150
		IPTW	-0.0415	-0.0009	0.0061	0.0082
		GBM	0.0934	0.0529	0.0703	0.0796
	β_{tg}	Subclass	-0.0039	-0.0085	-0.0119	0.0061
		MNPS	0.0031	-0.0065	-0.0327	-0.0203
		IPTW	-0.0086	-0.0886	-0.0841	-0.0960
		GBM	-0.0023	0.0336	0.0322	0.0401
	β_g	Subclass	-0.2003	-0.1342	-0.1366	-0.1458
		MNPS	-0.0028	0.0834	0.0809	0.0496
		IPTW	0.0072	0.3009	0.2647	0.2277
		GBM	-0.0013	0.1878	0.1583	0.1026
Variance	β_t	Subclass	0.0506	0.0843	0.0700	0.0444
		MNPS	0.0431	0.0476	0.0426	0.0338
		IPTW	0.1100	0.0665	0.0584	0.0499
		GBM	0.0436	0.0485	0.0424	0.0358
	β_{tg}	Subclass	0.1043	0.1472	0.1364	0.1049
		MNPS	0.0836	0.0896	0.0888	0.0755
		IPTW	0.2381	0.1808	0.2064	0.2148
		GBM	0.0943	0.1004	0.0953	0.0864
	β_g	Subclass	0.0558	0.0826	0.0724	0.0583
		MNPS	0.0417	0.0432	0.0428	0.0404
		IPTW	0.1832	0.1112	0.1499	0.1677
		GBM	0.0471	0.0489	0.0502	0.0480

Table F6
Bias and variance for Data Generating Model E, ATT Estimates, Misspecified model mA, n=250.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
Bias	β_t	Wt x Odds	0.0062	-0.0082	-0.0307	0.0121
		GBM	0.0864	0.0460	0.0324	0.0780
		Match 1:2	0.0169	-0.0317	0.0001	0.0333
		Match 1:1	0.0213	-0.0006	-0.0287	0.0319
		ExMatch 1:2	0.0331	-0.0102	-0.0100	0.0346
		ExMatch 1:1	0.0108	0.0019	-0.0127	0.0118
		Group	0.0100	0.0090	-0.0129	0.0230
		Regression	0.0069	0.0149	-0.0050	0.0193
	β_{tg}	Wt x Odds	-0.0144	0.0427	0.0585	-0.0260
		GBM	-0.0157	0.0554	0.0872	0.0039
		Match 1:2	-0.0069	0.0853	0.0586	-0.0186
		Match 1:1	-0.0285	0.0295	0.0666	-0.0259
		ExMatch 1:2	-0.0169	0.0612	0.0465	-0.0313
		ExMatch 1:1	-0.0033	0.0279	0.0325	-0.0153
		Group	-0.0040	0.0082	0.0006	-0.0110
		Regression	-0.0041	-0.0174	0.0080	-0.0356
	β_g	Wt x Odds	0.0157	0.1788	0.1700	0.1751
		GBM	0.0169	0.1661	0.1414	0.1452
		Match 1:2	0.0076	0.1194	0.1535	0.1516
		Match 1:1	0.0107	0.0919	0.0758	0.0787
		ExMatch 1:2	0.0177	0.1684	0.1794	0.1615
		ExMatch 1:1	0.0118	0.0966	0.0944	0.0746
		Group	-0.0072	0.0507	0.0435	0.0360
		Regression	0.0113	-0.0063	-0.0033	0.0188
Variance	β_t	Wt x Odds	0.2459	0.2216	0.2153	0.1880
		GBM	0.2305	0.2237	0.2170	0.1819
		Match 1:2	0.2523	0.2635	0.2520	0.2155
		Match 1:1	0.2255	0.2329	0.2172	0.1988
		ExMatch 1:2	0.2556	0.2507	0.2387	0.2124
		ExMatch 1:1	0.2439	0.2527	0.2237	0.1860

	Group	0.2315	0.2400	0.2344	0.1835
	Regression	0.1387	0.1467	0.1453	0.1290
β_{tg}	Wt x Odds	0.5223	0.4654	0.5124	0.4421
	GBM	0.4521	0.4432	0.4884	0.4019
	Match 1:2	0.5392	0.5169	0.5339	0.4954
	Match 1:1	0.4826	0.4555	0.4894	0.4341
	ExMatch 1:2	0.5348	0.5116	0.5302	0.5047
	ExMatch 1:1	0.4861	0.4746	0.5006	0.4242
	Group	0.2298	0.2119	0.2570	0.2742
	Regression	0.2575	0.2515	0.2811	0.2573
β_g	Wt x Odds	0.3680	0.3108	0.3586	0.3192
	GBM	0.3139	0.3051	0.3149	0.2814
	Match 1:2	0.3898	0.3531	0.3826	0.3650
	Match 1:1	0.2377	0.2280	0.2513	0.2116
	ExMatch 1:2	0.3684	0.3604	0.3729	0.3570
	ExMatch 1:1	0.2470	0.2556	0.2586	0.2161
	Group	0.1156	0.1135	0.1326	0.1424
	Regression	0.1326	0.1346	0.1462	0.1322

Table F7
Metrics for Data Generating Model E, ATT Estimates, Misspecified model mA, n=500.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Wt x Odds	0.1239	0.1048	0.0998	0.0986
		GBM	0.1239	0.1090	0.1030	0.1030
		Match 1:2	0.1300	0.1322	0.1172	0.1165
		Match 1:1	0.1160	0.1159	0.1105	0.0985
		ExMatch 1:2	0.1257	0.1234	0.1138	0.1042
		ExMatch 1:1	0.1149	0.1146	0.1083	0.0995
		Group	0.1084	0.1143	0.1086	0.0925
		Regression	0.0737	0.0714	0.0746	0.0637
	β_{tg}	Wt x Odds	0.2665	0.2374	0.2456	0.2648
		GBM	0.2513	0.2313	0.2236	0.2566
		Match 1:2	0.2782	0.2750	0.2549	0.2901
		Match 1:1	0.2476	0.2573	0.2561	0.2364
		ExMatch 1:2	0.2688	0.2536	0.2571	0.2691

		ExMatch 1:1	0.2434	0.2493	0.2376	0.2358
		Group	0.1142	0.1073	0.1237	0.1295
		Regression	0.1343	0.1370	0.1417	0.1313
	β_g	Wt x Odds	0.1842	0.1874	0.2075	0.2065
		GBM	0.1721	0.1817	0.1843	0.1957
		Match 1:2	0.1850	0.2016	0.1940	0.2076
		Match 1:1	0.1272	0.1234	0.1274	0.1172
		ExMatch 1:2	0.1931	0.1980	0.2101	0.2053
		ExMatch 1:1	0.1237	0.1311	0.1278	0.1205
		Group	0.0622	0.0580	0.0671	0.0688
		Regression	0.0700	0.0670	0.0670	0.0678
Bias	β_t	Wt x Odds	-0.0035	-0.0263	-0.0330	-0.0099
		GBM	0.0599	0.0233	0.0131	0.0528
		Match 1:2	0.0201	-0.0311	-0.0283	-0.0059
		Match 1:1	0.0060	-0.0215	-0.0361	-0.0012
		ExMatch 1:2	0.0089	-0.0115	-0.0176	0.0126
		ExMatch 1:1	-0.0017	-0.0088	-0.0199	0.0115
		Group	0.0036	0.0127	-0.0149	0.0136
		Regression	-0.0071	-0.0049	-0.0078	0.0038
	β_{tg}	Wt x Odds	0.0052	0.0397	0.0292	0.0290
		GBM	-0.0002	0.0457	0.0470	0.0226
		Match 1:2	0.0055	0.0799	0.0774	0.0682
		Match 1:1	-0.0062	0.0473	0.0698	0.0370
		ExMatch 1:2	0.0113	0.0501	0.0527	0.0360
		ExMatch 1:1	0.0026	0.0276	0.0319	0.0010
		Group	-0.0020	0.0182	0.0043	0.0080
		Regression	0.0026	0.0131	-0.0038	-0.0083
	β_g	Wt x Odds	-0.0034	0.2088	0.1981	0.1450
		GBM	0.0020	0.2027	0.1803	0.1515
		Match 1:2	-0.0028	0.1599	0.1426	0.0970
		Match 1:1	0.0054	0.0895	0.0745	0.0526
		ExMatch 1:2	-0.0081	0.2020	0.1759	0.1305
		ExMatch 1:1	0.0038	0.1047	0.0860	0.0752
		Group	-0.0066	0.0472	0.0423	0.0263
		Regression	-0.0023	-0.0053	0.0041	0.0137
Variance	β_t	Wt x Odds	0.1239	0.1041	0.0987	0.0985

	GBM	0.1203	0.1085	0.1028	0.1002
	Match 1:2	0.1296	0.1312	0.1164	0.1164
	Match 1:1	0.1160	0.1155	0.1092	0.0985
	ExMatch 1:2	0.1256	0.1233	0.1135	0.1041
	ExMatch 1:1	0.1149	0.1146	0.1079	0.0994
	Group	0.1084	0.1141	0.1084	0.0923
	Regression	0.0736	0.0714	0.0745	0.0637
β_{tg}	Wt x Odds	0.2665	0.2358	0.2448	0.2640
	GBM	0.2513	0.2292	0.2214	0.2561
	Match 1:2	0.2782	0.2686	0.2489	0.2854
	Match 1:1	0.2475	0.2551	0.2512	0.2350
	ExMatch 1:2	0.2686	0.2511	0.2543	0.2678
	ExMatch 1:1	0.2434	0.2486	0.2366	0.2358
	Group	0.1142	0.1070	0.1236	0.1294
	Regression	0.1343	0.1368	0.1417	0.1313
β_g	Wt x Odds	0.1842	0.1438	0.1683	0.1855
	GBM	0.1721	0.1406	0.1518	0.1728
	Match 1:2	0.1850	0.1760	0.1737	0.1982
	Match 1:1	0.1272	0.1154	0.1218	0.1145
	ExMatch 1:2	0.1931	0.1572	0.1792	0.1883
	ExMatch 1:1	0.1237	0.1201	0.1204	0.1149
	Group	0.0622	0.0558	0.0654	0.0681
	Regression	0.0700	0.0670	0.0670	0.0676

Table F8
Metrics for Data Generating Model E, ATT Estimates, Misspecified model mA, n=1000.

			Percentile Condition			
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Wt x Odds	0.0588	0.0548	0.0531	0.0465
		GBM	0.0605	0.0567	0.0547	0.0488
		Match 1:2	0.0645	0.0663	0.0625	0.0540
		Match 1:1	0.0581	0.0610	0.0584	0.0465
		ExMatch 1:2	0.0636	0.0636	0.0589	0.0537
		ExMatch 1:1	0.0563	0.0596	0.0577	0.0473
		Group	0.0551	0.0565	0.0557	0.0467

		Regression	0.0370	0.0388	0.0369	0.0298
	β_{tg}	Wt x Odds	0.1278	0.1260	0.1208	0.1171
		GBM	0.1210	0.1248	0.1225	0.1174
		Match 1:2	0.1322	0.1375	0.1294	0.1200
		Match 1:1	0.1149	0.1286	0.1211	0.1100
		ExMatch 1:2	0.1306	0.1332	0.1269	0.1284
		ExMatch 1:1	0.1090	0.1205	0.1233	0.1119
		Group	0.0538	0.0519	0.0582	0.0648
		Regression	0.0660	0.0704	0.0700	0.0629
	β_{g}	Wt x Odds	0.0855	0.1304	0.1253	0.1041
		GBM	0.0781	0.1361	0.1224	0.1004
		Match 1:2	0.0887	0.1213	0.1173	0.0974
		Match 1:1	0.0578	0.0710	0.0705	0.0590
		ExMatch 1:2	0.0886	0.1336	0.1342	0.1125
		ExMatch 1:1	0.0569	0.0724	0.0732	0.0613
		Group	0.0298	0.0306	0.0338	0.0351
		Regression	0.0345	0.0359	0.0350	0.0335
Bias	β_{t}	Wt x Odds	0.0043	-0.0084	-0.0001	-0.0054
		GBM	0.0435	0.0276	0.0293	0.0358
		Match 1:2	0.0288	-0.0090	-0.0033	0.0025
		Match 1:1	0.0096	-0.0096	-0.0038	-0.0001
		ExMatch 1:2	0.0229	0.0080	0.0217	0.0192
		ExMatch 1:1	0.0120	0.0056	0.0167	0.0067
		Group	0.0093	0.0179	0.0248	0.0187
		Regression	-0.0005	0.0003	0.0163	0.0103
	β_{tg}	Wt x Odds	-0.0018	0.0086	-0.0016	0.0216
		GBM	-0.0025	-0.0028	0.0080	0.0250
		Match 1:2	-0.0004	0.0476	0.0361	0.0519
		Match 1:1	0.0038	0.0434	0.0320	0.0399
		ExMatch 1:2	0.0046	0.0188	-0.0076	0.0274
		ExMatch 1:1	-0.0014	0.0126	-0.0037	0.0176
		Group	0.0180	0.0056	-0.0023	0.0139
		Regression	0.0032	-0.0015	-0.0288	-0.0041
	β_{g}	Wt x Odds	-0.0013	0.2278	0.2149	0.1487
		GBM	-0.0006	0.2392	0.2053	0.1453
		Match 1:2	-0.0027	0.1852	0.1733	0.1143

		Match 1:1	-0.0047	0.0882	0.0890	0.0543
		ExMatch 1:2	-0.0067	0.2216	0.2222	0.1407
		ExMatch 1:1	-0.0007	0.1090	0.1015	0.0603
		Group	-0.0148	0.0502	0.0476	0.0268
		Regression	-0.0048	0.0015	0.0162	0.0054
Variance	β_t	Wt x Odds	0.0588	0.0548	0.0531	0.0465
		GBM	0.0586	0.0559	0.0538	0.0475
		Match 1:2	0.0637	0.0663	0.0625	0.0540
		Match 1:1	0.0580	0.0609	0.0584	0.0465
		ExMatch 1:2	0.0631	0.0635	0.0584	0.0534
		ExMatch 1:1	0.0562	0.0595	0.0574	0.0473
		Group	0.0550	0.0562	0.0551	0.0463
		Regression	0.0370	0.0388	0.0366	0.0297
	β_{tg}	Wt x Odds	0.1278	0.1259	0.1208	0.1166
		GBM	0.1210	0.1248	0.1225	0.1168
		Match 1:2	0.1322	0.1353	0.1281	0.1173
		Match 1:1	0.1148	0.1268	0.1201	0.1084
		ExMatch 1:2	0.1306	0.1328	0.1269	0.1276
		ExMatch 1:1	0.1090	0.1204	0.1233	0.1116
		Group	0.0535	0.0518	0.0582	0.0646
		Regression	0.0660	0.0704	0.0691	0.0629
	β_g	Wt x Odds	0.0855	0.0785	0.0792	0.0820
		GBM	0.0781	0.0789	0.0803	0.0793
		Match 1:2	0.0886	0.0870	0.0872	0.0844
		Match 1:1	0.0578	0.0632	0.0626	0.0561
		ExMatch 1:2	0.0885	0.0845	0.0848	0.0927
		ExMatch 1:1	0.0569	0.0606	0.0629	0.0576
		Group	0.0296	0.0281	0.0316	0.0344
		Regression	0.0345	0.0359	0.0347	0.0335

Table F9
Metrics for Data Generating Model E, ATT Estimates, True Model E, n=250.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Wt x Odds	0.4209	0.3169	0.2815	0.2702

		GBM	0.2225	0.2212	0.2036	0.1855
		Match 1:2	0.2995	0.3128	0.2577	0.2561
		Match 1:1	0.2609	0.2681	0.2470	0.2288
		ExMatch 1:2	0.2888	0.2998	0.2508	0.2618
		ExMatch 1:1	0.2664	0.2659	0.2376	0.2270
		Group	0.2719	0.2853	0.2428	0.2302
		Regression	0.1388	0.1470	0.1453	0.1294
	β_{tg}	Wt x Odds	0.9658	0.9461	0.8672	0.8897
		GBM	0.4489	0.4451	0.4821	0.4016
		Match 1:2	0.6564	0.6462	0.6500	0.6140
		Match 1:1	0.5398	0.5294	0.5539	0.5169
		ExMatch 1:2	0.6181	0.6166	0.6276	0.6414
		ExMatch 1:1	0.5706	0.5501	0.5717	0.5428
		Group	0.2535	0.2349	0.2676	0.2990
		Regression	0.2575	0.2518	0.2812	0.2586
	β_{g}	Wt x Odds	0.8004	0.8637	0.7825	0.8325
		GBM	0.3008	0.3270	0.3333	0.2998
		Match 1:2	0.5108	0.5358	0.5382	0.5032
		Match 1:1	0.2765	0.2949	0.3044	0.2704
		ExMatch 1:2	0.4835	0.5419	0.5305	0.5272
		ExMatch 1:1	0.2945	0.3158	0.3091	0.2785
		Group	0.1418	0.1384	0.1570	0.1726
		Regression	0.1327	0.1346	0.1462	0.1326
Bias	β_{t}	Wt x Odds	-0.0574	-0.0083	-0.0111	0.0249
		GBM	0.1048	0.0459	0.0433	0.0858
		Match 1:2	0.0076	0.0138	0.0118	0.0250
		Match 1:1	0.0424	0.0035	-0.0099	0.0386
		ExMatch 1:2	0.0097	0.0014	0.0129	0.0443
		ExMatch 1:1	0.0419	0.0145	-0.0023	0.0429
		Group	0.0271	0.0092	-0.0007	0.0362
		Regression	0.0069	0.0149	-0.0050	0.0193
	β_{tg}	Wt x Odds	-0.0029	-0.0593	-0.0134	-0.1546
		GBM	-0.0192	0.0597	0.0904	0.0086
		Match 1:2	-0.0146	0.0073	0.0163	-0.0418
		Match 1:1	-0.0405	0.0131	0.0419	-0.0455
		ExMatch 1:2	-0.0141	0.0155	0.0149	-0.0798

		ExMatch 1:1	-0.0607	0.0049	0.0362	-0.0599
		Group	0.0095	-0.0028	0.0100	-0.0183
		Regression	-0.0041	-0.0174	0.0080	-0.0356
	β_g	Wt x Odds	0.0041	0.2808	0.2419	0.3037
		GBM	0.0205	0.1618	0.1381	0.1405
		Match 1:2	0.0186	0.2020	0.1960	0.1743
		Match 1:1	0.0291	0.1476	0.1032	0.1050
		ExMatch 1:2	0.0121	0.2156	0.2060	0.2073
		ExMatch 1:1	0.0231	0.1380	0.0934	0.1064
		Group	0.0041	0.0701	0.0527	0.0486
		Regression	0.0113	-0.0063	-0.0033	0.0188
Variance	β_t	Wt x Odds	0.4176	0.3168	0.2814	0.2696
		GBM	0.2116	0.2191	0.2017	0.1781
		Match 1:2	0.2995	0.3126	0.2576	0.2555
		Match 1:1	0.2591	0.2681	0.2469	0.2273
		ExMatch 1:2	0.2887	0.2998	0.2507	0.2598
		ExMatch 1:1	0.2647	0.2657	0.2376	0.2252
		Group	0.2712	0.2852	0.2428	0.2289
		Regression	0.1387	0.1467	0.1453	0.1290
	β_{tg}	Wt x Odds	0.9658	0.9426	0.8670	0.8658
		GBM	0.4485	0.4416	0.4740	0.4016
		Match 1:2	0.6561	0.6462	0.6497	0.6122
		Match 1:1	0.5381	0.5293	0.5521	0.5148
		ExMatch 1:2	0.6179	0.6164	0.6274	0.6350
		ExMatch 1:1	0.5669	0.5501	0.5704	0.5392
		Group	0.2534	0.2349	0.2675	0.2987
		Regression	0.2575	0.2515	0.2811	0.2573
	β_g	Wt x Odds	0.8004	0.7849	0.7240	0.7402
		GBM	0.3003	0.3008	0.3142	0.2801
		Match 1:2	0.5104	0.4950	0.4997	0.4728
		Match 1:1	0.2756	0.2731	0.2937	0.2593
		ExMatch 1:2	0.4834	0.4954	0.4881	0.4842
		ExMatch 1:1	0.2939	0.2968	0.3003	0.2672
		Group	0.1417	0.1335	0.1542	0.1703
		Regression	0.1326	0.1346	0.1462	0.1322

Table F10
Metrics for Data Generating Model E, ATT Estimates, True Model E, n=500.

			Percentile Condition			
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Wt x Odds	0.2800	0.1591	0.1435	0.1655
		GBM	0.1227	0.1091	0.1044	0.1002
		Match 1:2	0.1498	0.1466	0.1297	0.1366
		Match 1:1	0.1282	0.1220	0.1197	0.1074
		ExMatch 1:2	0.1522	0.1455	0.1279	0.1322
		ExMatch 1:1	0.1192	0.1284	0.1237	0.1088
		Group	0.1221	0.1250	0.1163	0.1085
		Regression	0.0737	0.0714	0.0746	0.0637
	β_{tg}	Wt x Odds	0.6231	0.4889	0.4935	0.5901
		GBM	0.2544	0.2394	0.2275	0.2554
		Match 1:2	0.3427	0.3349	0.3079	0.3644
		Match 1:1	0.2626	0.2822	0.2858	0.2587
		ExMatch 1:2	0.3296	0.3167	0.3045	0.3612
		ExMatch 1:1	0.2566	0.2837	0.2847	0.2719
		Group	0.1214	0.1185	0.1258	0.1338
		Regression	0.1343	0.1370	0.1417	0.1313
	β_g	Wt x Odds	0.5462	0.4812	0.5211	0.5679
		GBM	0.1719	0.1893	0.1913	0.1929
		Match 1:2	0.2618	0.2775	0.2725	0.3049
		Match 1:1	0.1436	0.1613	0.1558	0.1525
		ExMatch 1:2	0.2505	0.2641	0.2832	0.3031
		ExMatch 1:1	0.1438	0.1580	0.1472	0.1415
		Group	0.0658	0.0662	0.0766	0.0763
		Regression	0.0700	0.0670	0.0670	0.0678
Bias	β_t	Wt x Odds	-0.0850	-0.0262	-0.0322	-0.0130
		GBM	0.0671	0.0262	0.0234	0.0561
		Match 1:2	0.0015	-0.0194	-0.0206	0.0178
		Match 1:1	0.0130	0.0058	-0.0163	0.0119
		ExMatch 1:2	0.0106	-0.0087	-0.0016	0.0136
		ExMatch 1:1	0.0046	0.0002	-0.0121	0.0139
		Group	0.0081	0.0039	-0.0095	0.0119

		Regression	-0.0071	-0.0049	-0.0078	0.0038
	β_{tg}	Wt x Odds	0.0086	-0.0656	-0.0978	-0.0840
		GBM	0.0001	0.0490	0.0457	0.0299
		Match 1:2	0.0235	0.0390	0.0170	0.0104
		Match 1:1	-0.0062	0.0151	0.0272	0.0086
		ExMatch 1:2	0.0169	0.0372	-0.0052	0.0095
		ExMatch 1:1	0.0055	0.0172	0.0233	0.0075
		Group	0.0115	0.0276	-0.0037	0.0162
		Regression	0.0026	0.0131	-0.0038	-0.0083
	β_{g}	Wt x Odds	-0.0068	0.3140	0.3251	0.2580
		GBM	0.0017	0.1994	0.1816	0.1442
		Match 1:2	-0.0208	0.2047	0.2058	0.1576
		Match 1:1	0.0118	0.1483	0.1130	0.0829
		ExMatch 1:2	-0.0145	0.2214	0.2366	0.1614
		ExMatch 1:1	0.0005	0.1367	0.1126	0.0774
		Group	-0.0036	0.0594	0.0595	0.0362
		Regression	-0.0023	-0.0053	0.0041	0.0137
Variance	β_{t}	Wt x Odds	0.2728	0.1584	0.1424	0.1654
		GBM	0.1182	0.1085	0.1039	0.0971
		Match 1:2	0.1498	0.1462	0.1293	0.1363
		Match 1:1	0.1281	0.1220	0.1195	0.1073
		ExMatch 1:2	0.1521	0.1454	0.1279	0.1320
		ExMatch 1:1	0.1192	0.1284	0.1235	0.1086
		Group	0.1220	0.1250	0.1162	0.1084
		Regression	0.0736	0.0714	0.0745	0.0637
	β_{tg}	Wt x Odds	0.6230	0.4846	0.4839	0.5830
		GBM	0.2544	0.2370	0.2254	0.2545
		Match 1:2	0.3422	0.3333	0.3076	0.3643
		Match 1:1	0.2625	0.2819	0.2851	0.2586
		ExMatch 1:2	0.3293	0.3153	0.3044	0.3611
		ExMatch 1:1	0.2566	0.2834	0.2841	0.2718
		Group	0.1212	0.1178	0.1258	0.1335
		Regression	0.1343	0.1368	0.1417	0.1313
	β_{g}	Wt x Odds	0.5462	0.3826	0.4154	0.5013
		GBM	0.1719	0.1495	0.1584	0.1722
		Match 1:2	0.2613	0.2356	0.2301	0.2800

	Match 1:1	0.1434	0.1393	0.1430	0.1456
	ExMatch 1:2	0.2503	0.2151	0.2272	0.2770
	ExMatch 1:1	0.1438	0.1394	0.1345	0.1355
	Group	0.0658	0.0627	0.0731	0.0750
	Regression	0.0700	0.0670	0.0670	0.0676

Table F11
Metrics for Data Generating Model E, ATT Estimates, True Model E, n=1000.

				Percentile Condition		
Metric	Coefficient	Method	Baseline	70	80	90
MSE	β_t	Wt x Odds	0.2103	0.0879	0.0843	0.0741
		GBM	0.0586	0.0568	0.0540	0.0483
		Match 1:2	0.0761	0.0720	0.0681	0.0584
		Match 1:1	0.0611	0.0629	0.0615	0.0516
		ExMatch 1:2	0.0740	0.0694	0.0672	0.0577
		ExMatch 1:1	0.0642	0.0617	0.0592	0.0491
		Group	0.0607	0.0587	0.0567	0.0471
		Regression	0.0370	0.0388	0.0369	0.0298
	β_{tg}	Wt x Odds	0.4345	0.2954	0.3473	0.3712
		GBM	0.1256	0.1285	0.1299	0.1206
		Match 1:2	0.1633	0.1646	0.1619	0.1468
		Match 1:1	0.1292	0.1341	0.1394	0.1311
		ExMatch 1:2	0.1549	0.1554	0.1599	0.1507
		ExMatch 1:1	0.1293	0.1306	0.1444	0.1251
		Group	0.0585	0.0585	0.0663	0.0674
		Regression	0.0660	0.0704	0.0700	0.0629
	β_g	Wt x Odds	0.3902	0.3530	0.3933	0.4167
		GBM	0.0822	0.1390	0.1295	0.1077
		Match 1:2	0.1206	0.1747	0.1601	0.1398
		Match 1:1	0.0659	0.0886	0.0838	0.0731
		ExMatch 1:2	0.1126	0.1746	0.1666	0.1449
		ExMatch 1:1	0.0699	0.0865	0.0879	0.0731
		Group	0.0317	0.0360	0.0395	0.0378
		Regression	0.0345	0.0359	0.0350	0.0335
Bias	β_t	Wt x Odds	-0.0693	-0.0039	-0.0066	-0.0005
		GBM	0.0455	0.0262	0.0324	0.0433

		Match 1:2	0.0233	0.0096	0.0113	0.0223
		Match 1:1	0.0120	0.0028	0.0160	0.0176
		ExMatch 1:2	0.0253	0.0219	0.0180	0.0197
		ExMatch 1:1	0.0073	0.0065	0.0157	0.0202
		Group	0.0145	0.0073	0.0193	0.0299
		Regression	-0.0005	0.0003	0.0163	0.0103
	β_{tg}	Wt x Odds	-0.0203	-0.1221	-0.1003	-0.1309
		GBM	-0.0041	-0.0025	0.0081	0.0177
		Match 1:2	-0.0127	-0.0051	0.0021	0.0049
		Match 1:1	-0.0012	0.0042	-0.0038	0.0138
		ExMatch 1:2	-0.0098	-0.0112	-0.0066	0.0031
		ExMatch 1:1	0.0004	0.0137	-0.0053	0.0023
		Group	0.0133	0.0035	0.0016	0.0305
		Regression	0.0032	-0.0015	-0.0288	-0.0041
	β_{g}	Wt x Odds	0.0171	0.3584	0.3135	0.3012
		GBM	0.0010	0.2388	0.2052	0.1526
		Match 1:2	0.0093	0.2408	0.2100	0.1644
		Match 1:1	-0.0037	0.1465	0.1321	0.0787
		ExMatch 1:2	0.0074	0.2577	0.2259	0.1692
		ExMatch 1:1	0.0004	0.1347	0.1248	0.0799
		Group	-0.0038	0.0690	0.0627	0.0355
		Regression	-0.0048	0.0015	0.0162	0.0054
Variance	β_{t}	Wt x Odds	0.2055	0.0879	0.0843	0.0741
		GBM	0.0566	0.0561	0.0530	0.0464
		Match 1:2	0.0755	0.0719	0.0679	0.0579
		Match 1:1	0.0609	0.0629	0.0612	0.0513
		ExMatch 1:2	0.0734	0.0689	0.0669	0.0573
		ExMatch 1:1	0.0641	0.0616	0.0589	0.0487
		Group	0.0605	0.0586	0.0563	0.0462
		Regression	0.0370	0.0388	0.0366	0.0297
	β_{tg}	Wt x Odds	0.4341	0.2805	0.3372	0.3541
		GBM	0.1256	0.1285	0.1298	0.1203
		Match 1:2	0.1631	0.1646	0.1619	0.1467
		Match 1:1	0.1292	0.1341	0.1394	0.1310
		ExMatch 1:2	0.1548	0.1553	0.1598	0.1507
		ExMatch 1:1	0.1293	0.1304	0.1444	0.1251

	Group	0.0584	0.0585	0.0663	0.0665
	Regression	0.0660	0.0704	0.0691	0.0629
β_g	Wt x Odds	0.3899	0.2245	0.2950	0.3260
	GBM	0.0822	0.0820	0.0874	0.0844
	Match 1:2	0.1206	0.1167	0.1160	0.1128
	Match 1:1	0.0659	0.0671	0.0663	0.0669
	ExMatch 1:2	0.1126	0.1081	0.1156	0.1163
	ExMatch 1:1	0.0699	0.0683	0.0723	0.0667
	Group	0.0317	0.0312	0.0356	0.0365
	Regression	0.0345	0.0359	0.0347	0.0335

Appendix G: Applied Example Descriptive Statistics

Table G1
Descriptive Statistics for Applied Example

		Public School		Private School	
		SES			
Covariate	Category	Low	High	Low	High
Perceived Competance in Math	1-1.75	780	386	77	174
	2-2.75	1053	810	122	303
	3-3.75	846	819	101	260
	4	330	269	28	95
Like Math	1=not at all	674	371	65	138
	2=a little bit	782	596	99	238
	3=mostly true	779	714	80	250
	4=very true	774	603	84	206
Enjoy Math	1=not at all	783	535	79	205
	2=a little bit	943	708	8	267
	3=mostly true	707	644	89	220
	4=very true	576	397	52	140
Disability	1=yes	517	302	37	99
	2=no	2492	1982	291	733
Race Ethnicity	1=white, nonhispanic	1684	1800	225	657
	2=black or african american	337	111	16	26
	3=hispanic, race specified	321	114	24	52
	4=hispanic, race not specified	380	56	37	28
	5=asian	125	130	12	39
	6=native hawaiian, other pacific islander	42	8	3	5
	7=american indian or alaska native	57	13	6	6
	8=more than one race, non hispanic	63	52	5	19
Region	1=Northeast	452	523	74	157
	2=Midwest	855	742	121	271
	3=South	1051	652	83	223
	4=West	651	367	50	181
Urban	1=large and mid-size city	910	553	165	452
	2=large and mid-size suburb and large town	1112	1200	98	272

	3=small town and rural	987	531	65	108
Tutor	1=yes	684	352	76	167
	2=no	2325	1932	252	665
Number of Siblings	0	533	300	54	101
	1	1183	1140	161	379
	2	823	592	66	225
	3	323	190	29	85
	4	85	42	15	26
	5	33	11	1	9
	6	14	5	0	1
	7	7	0	1	4
	8	6	3	0	1
	9	2	1	1	0
	10	0	0	0	0
	11	0	0	0	0
	12	0	0	0	1
Gender	1=male	1485	1145	155	424
	2=female	1524	1139	173	408
Total		3009	2284	328	832

Appendix H: IRB Determination

UNIVERSITY OF MARYLAND
INSTITUTIONAL REVIEW BOARD

1204 Marie Mount Hall
College Park, MD 20742-5125
TEL 301.405.4212
FAX 301.314.1475
irb@umd.edu
www.umresearch.umd.edu/IRB

DATE:	March 10, 2016
TO:	Kathy Stepien
FROM:	University of Maryland College Park (UMCP) IRB
PROJECT TITLE:	[863375-1] Applied Example for Dissertation
SUBMISSION TYPE:	New Project
ACTION:	DETERMINATION OF NOT HUMAN SUBJECT RESEARCH
DECISION DATE:	March 10, 2016

Thank you for your submission of New Project materials for this project. The University of Maryland College Park (UMCP) IRB has determined this project does not meet the definition of human subject research under the purview of the IRB according to federal regulations.

We will retain a copy of this correspondence within our records.

If you have any questions, please contact the IRB Office at 301-405-4212 or irb@umd.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within University of Maryland College Park (UMCP) IRB's records.

References

Abrahamowicz, M., Beauchamp, M.E., Fournier, P., & Dumont, A. (2013). Evidence of subgroup-specific treatment effect in the absence of an overall effect: is there really a contradiction? *Pharmacoepidemiology and Drug Safety, 22*, 1178-1188.

Angrist, J.D. (1998). Estimating the labor market impact of voluntary military service using Social Security data on military applicants. *Econometrica, 66*, 249-288.

Angrist, J.D. (2004). Treatment effect heterogeneity in theory and practice. *The Economic Journal, 114*, C52-C83.

Augurzky, B., & Kluve, J. (2007). Assessing the performance of matching algorithms when selection into treatment is strong. *Journal of Applied Econometrics, 22*, 533-557.

Austin, P.C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine, 27*, 2037-2049.

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

Austin, P.C. (2009b). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine, 29*, 2137-2148.

Austin, P.C. (2009c). The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making, 29*, 661-77.

Austin, P.C. (2009d). Some methods of propensity-score matching had superior performance to others: Results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal, 51*, 171-184.

Austin, P.C. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology, 172*, 1092-1097.

Austin, P.C. (2011a). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*, 399-424.

Austin, P.C. (2011b). 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. (2012). Using ensemble-based methods for directly estimating causal effects: An investigation of tree-based G-computation. *Multivariate Behavioral Research, 47*, 115-135.

Austin, P.C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine, 33*, 1057-1069.

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*, 734-753.

Austin, P.C., Manca, A., Zwarenstin, M., Juurlink, D.N., & Stanbrook, M.B. (2010). A substantial and confusing variation exists in handling of baseline covariates in

randomized controlled trials: a review of trials published in leading medical journals, *Journal of Clinical Epidemiology, 63*, 142-153.

Berk, R.A. (2006). An introduction to ensemble methods for data analysis. *Sociological Methods & Research, 34*, 263-295.

Brand, J.E., & Xie, Y. (2010). Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. *American Sociological Review, 75*, 273-302.

Breen, R., Choi, S., & Holm, A. (2015). Hetergeneous causal effects and sample selection bias. Sociological Science, 2, 351-369.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*, 199-215.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn R.J., Avorn J., & Sturmer T. (2006). Variable selection in propensity score models: some insights from a simulation study. *American Journal of Epidemiology, 163*, 1149–1156.

Brooks, J.M., & Fang, G. (2009). Interpreting treatment-effect estimates with heterogeneity and choice: Simulation model results. *Clinical Therapeutics, 31*, 902-919.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*, 31-72.

Chipman, H.A., George, E.I., & McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics, 4*, 266-298.

Cochran, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295-313.

Crump, R.K., Hotz, V.J., Imbens, G.W., & Mitnik, O.A. (2008). Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics, 90*, 389-405.

Cuong, N.V. (2013). Which covariates should be controlled in propensity score matching? Evidence from a simulation study. *Statistica Neerlandica, 67*, 169-180.

D'Agostino, 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*, 2265-2281.

Diamond, A., & Sekhon, J.S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *The Review of Economics and Statistics, 95*, 932-945.

Dong, N. (2015). Using propensity score methods to approximate factorial experimental designs to analyze the relationship between two variables and an outcome. American Journal of Evaluation, 36, 42-66.

Eeren, H.V., Spreeuwenberg, M.D., Bartak, A., de Rooij, M., & Busschbach, J.V. (2015). Estimating subgroup effects using the propensity score method. Medical Care, 53, 366-373.

Ellis, A.R., Dusetzina, S.B., Hansen, R.A., Gaynes, B.N., Farley, J.F., & Sturmer, T. (2013). Investigating differences in treatment effect estimates between propensity score matching and weighting: a demonstration using STAR*D trial data. *Pharmacoepidemiology and Drug Safety, 22*, 138–144.

Feng, P., Zhou, X.H., Zou, Q., Fan, M., & Li, X. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine, 31*, 681-697.

Fink, G., McConnell, M., & Vollmer, S. (2014). Testing for heterogeneous treatment effects in experimental data: False discovery risks and correction procedures. *Journal of Development Effectiveness, 6*, 44-57.

Galindo, C., & Sonnenschein, S. (2015). Decreasing the SES math achievement gap: Initial math proficiency and home learning environments. *Contemporary Educational Psychology*, 43, 35-38.

Geneletti, S., & Dawid, A. P. (2011). *Defining and identifying the effect of treatment on the treated* (pp. 728-749). Oxford University Press.

Green, K.M., & Stuart, E.A. (2014). Examining moderation analyses in propensity score methods: application to depression and substance use. *Journal of Consulting and Clinical Psychology, 82*, 773-783.

Greene, W.H. (2003). *Econometric Analysis, 5th Edition.* Upper Saddle River, NJ: Prentice Hall.

Gu, X.S., & Rosenbaum, P.R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*, 405-420.

Guo, S., & Fraser, M. (2010). *Propensity score analysis: Statistical methods and applications.* Thousand Oaks, Calif.: Sage Publications.

Hansen, B.B. (2008). The prognostic analogue of the propensity score. *Biometrika, 95*, 481-488.

Harder, V.S., Stuart, E.A., & Anthony, J.C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods, 15*, 234-249.

Hahs-Vaughn, D.L. (2005). A Primer for using and understanding weights with national datasets. *The Journal of Experimental Education*, 3, 221-248.

Hayward, R.A., Kent, D.M., Vijan, S., & Hofer, T.P. (2005). Reporting clinical trail results to inform providers and consumers. *Health Affairs, 24*, 1571-1581.

Heckman, J.J., Urzua, S., & Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics, 88*, 389-432.

Hill, J.L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics, 20*, 217-240.

Hill, J.L., Weiss, C., & Zhai, F. (2011) Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research, 46*, 477-513.

Hirano, K., Imbens, G.W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*, 1161-1189.

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2007). Matching as monparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199-236.

Ho, D.E., Imai, K., King, G., & Stuart, E.A. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, *42*(8), 1-28.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*, 945–60.

Imai, K., King, G., & Stuart, E.A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A (Statistics in Society), 171*, 481-502.

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

Imai, K., & Van Dyk, D.A. (2004) Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99*, 854-866.

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

Imbens, G.W., & Wooldridge, J.M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*, 5-86.

Joffe, M.M., & Rosenbaum, P.R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology, 150* , 327-333.

Kang, J.D., & Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22*, 523-539.

Kelcey, B. (2011). Assessing the effects of teachers' reading knowledge on students' achievement using multilevel propensity score stratification. *Educational Evaluation and Policy Analysis, 33*, 458-482.

King, M.W., & Resick, P.A. (2014). Data mining in psychological treatment research: A primer on classification and regression trees. *Journal of Consulting and Clinical Psychology, 82*, 895-905.

King, G., & Nielsen, R. (2016). Why propensity scores should not be used for matching. (February 28, 2016). Retrieved from: http://j.mp/1FQhySn.

Kreif, N., Grieve, R., Radice, R., Sadique, Z., Ramsahai, R., & Sekhon, J.S. (2012). Methods for estimating subgroup effects in cost-effectiveness analyses that use observational data. *Medical Decision Making, 32*, 750-763.

Kurth, T., Walker, A.M., Glynn, R.J., Chan, K.A., Gaziano, J.M., Berger, K., & Robins, J.M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology, 163*, 262-270.

Leacy, F.P., & Stuart, E.A. (2013). On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in Medicine, 33*, 3488-3508.

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

Luellen, J.K., Shadish, W.R., & Clark, M.H. (2005). Propensity Scores: An introduction and experimental test. *Evaluation Review , 29*, 530-558.

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*, 2937-2960.

Lunt, M., Solomon, D., Rothman, K., Glynn, R., Hyrich, K., Symmons, D.P.M., & Sturmer, T. (2009). Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *American Journal of Epidemiology, 169*, 909-917.

Luo, Z. (2011). Heterogeneity in treatment effect and comparative effectiveness research. *China Health Review, 2*(3), 2-7.

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.

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

Morgan, S.L., & Harding, D.J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research, 35*, 3-60.

Morgan, S.L., & Todd, J.J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology, 38*, 231-81.

Morgan, S.L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research.* New York: Cambridge University Press.

Murnane, R.J., & Willett, J.B. (2011). *Improving causal inference in educational and social science research.* New York: Oxford University Press.

Muthen, L.K. & Muthen, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, 4, 599-620.

Neyman, J. (1923:1990). On the application of probability theory to agricultural experiments: Essay on principles, Section 9. Translated in *Statistical Science, 5*, 465-480.

Pan, W., & Bai, H. (2015). *Propensity score analysis: Fundamentals and developments*. New York, NY: The Guilford Press.

Pocock, S.J., Assmann, S.E., Enos, L.E., & Kasten, L.E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine, 21*, 2917-2930.

Powell, J.L. (1994). Estimation of semi-parametric models. In R.F. Engle & D.L. McFadden (Eds.), *Handbook of Econometrics IV* (2443–2521). Elsevier Science.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramsahai, R.R., Grieve, R., & Sekhon, J.S. (2011). Extending iterative matching methods: an approach to improving covariate balance that allows prioritization. *Health Services and Outcomes Research Methodology, 11*, 95-114.

Rassen, J.A., Glynn, R.J., Rothmas, Setoguchi, S., & Schneeweiss, S. (2012). Applying propensity scores estimated in a full cohort to adjust for confounding in subgroup analyses. *Pharmacoepidemiology and Drug Safety, 21*, 697-709.

Rhodes, W. (2010). Heterogeneous treatment effects: What does a regression estimate? *Evaluation Review, 34*, 334-361.

Ridgeway, G., McCaffrey, D., Morral, B.A., & Burgette, L. (2015). twang: Toolkit for weighting and analysis of nonequivalent groups. R package version 3.2.2. http://CRAN.R-project.org/package-twang.

Rosenbaum, P.R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*, 387-394.

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

Rosenbaum, P.R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516-524.

Rothwell, P.M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet, 365*, 176-186.

Rubin, D.B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics, 29*, 185-203.

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

Rubin, D.B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association, 74*, 318-328.

Rubin, D.B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*, 169-188.

Rubin, D.B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics, 31*, 161-170.

Rubin, D.B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*, 20-36.

Schafer, J.L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods, 13*, 279-313.

Sekhon, J.S. (2011). Multivariate and propensity score matching software with automated balance optimization: The Matching package for R. *Journal of Statistical Software, 42*(7), 1-52.

Setoguchi, S., Schneeweiss, S., Brookhart, M., Glynn, R., & Cook, E. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and Drug Safety, 17*, 546-555.

Shadish, W.R. (2013). Propensity score analysis: promise, reality and irrational exuberance. *Journal of Experimental Criminology, 9*, 129-144.

Sloczynski, T. (2014). New evidence on linear regression and treatment effect heterogeneity. (November 14, 2014). Retrieved from: http://akson.sgh.waw.pl/~ts37864/Sloczynski_paper_regression.pdf.

Steiner, P.M., Cook, T.D., & Shadish W.R. (2011). On the importance of reliable covariate measurement in selection bias adjustments using propensity scores. *Journal of Educational and Behavioral Statistics, 36*, 213-236.

Steiner, P.M., & Cook, T.D. (2013). Matching and Propensity Scores. In T.D. Little (Ed), *The Oxford Handbook of Quantitative Methods, Volume 1: Foundations* (pp. 237-259). New York, NY: Oxford University Press.

Stuart, E.A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1-21.

Stuart, E.A., & Rubin, D.B. (2008). Best practices in quasi-experimental designs. Matching methods for causal inference. In J. Osborne (Ed), *Best practices in quantitative designs* (pp. 155-176). Thousand Oaks, CA: Sage Publications.

Sturmer, 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*, 843-854.

Sturmer, T., Rothman, K.J., & Glynn, R.J. (2006). Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiology and Drug Safety, 15*, 698-709.

Sturmer, T., Wyss, R., Glynn, R.J., & Brookhart, M.A. (2014). Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. *Journal of Internal Medicine, 275*, 570-580.

Tabachnick, B.G., & Fidell, L.S. (2013). *Using Multivariate Statistics, 6th edition.* Boston: Pearson.

Thoemmes, F. (2012). Propensity score matching in SPSS. http://arxiv.org/pdf/1201.6385.pdf

Thoemmes, F., & Kim, E.S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*, 90-118.

Tourangeau, K., Nord, C., Le, T., Sorongon, A.G., and Najarian, M. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K), Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks (NCES 2009-004)*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Tsai, S.L., & Xie, Y. (2008). Changes in Earnings Returns to Higher Education in Taiwan since the 1990s. *Population Review, 47*, 1-20.

Tsai, S.L., & Xie, Y. (2011). Heterogeneity in returns to college education: Selection bias in contemporary Taiwan. *Social Science Research, 40*, 796-810.

Varadhan, R., Segal, J.B., Boyd, C.M., Wu, A.W., & Weiss, C.O. (2013). A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology, 66*, 818-825.

Wang, R., Lagakos, S.W., Ware, J.H., Hunter, D.J., & Drazen, J.M. (2007). Statistics in Medicine - Reporting of Subgroup analyses in clinical trials. *The New England Journal of Medicine, 357*, 2189-2194.

West, S.G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology, 82*, 906-919.

Westreich, D., Lessler, J., & Funk, M.J. (2010) Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology, 63*, 826–833.

Willke, R.J., Zheng, Z., Subedi, P., Althin, R., & Mullins, C.D. (2012). From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *Medical Research Methodology, 12*(185), 1-12.

Winship, C., & Morgan, S.L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25*, 659-706.

Wooldridge, J.M. (2002). *Econometric analysis of cross section and panel data.* Cambridge, Massachusetts: The MIT Press.

Wooldridge, J.M. (2009). *Should instrumental variables be used as matching variables?* Unpublished working paper, Department of Economics, Michigan State University, East Lansing.

Wu, S., Ding, Y., Wu, F., & Hou, J. (2015). Application of propensity score matching in four leading medical journals. *Epidemiology, 26*, e19-e20.

Xie, Y., Brand, J.E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological Methodology, 42*, 314-347.

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *The Review of Economics and Statistics, 86*, 91-107.

Zubizarreta, J.R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association, 107*, 1360-1371.