5-6-2021

Does coding method matter? An examination of propensity score methods when the treatment group is larger than the comparison group

Beth A. Perkins
James Madison University

Does Coding Method Matter? An Examination of Propensity Score Methods When the Treatment Group is Larger Than the Comparison Group

Beth A. Perkins

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

May 2021

FACULTY COMMITTEE:

Committee Chair: Dr. S. Jeanne Horst

Committee Members/ Readers:

Dr. Christine DeMars

Dr. Heather Harris

Dr. Dena Pastor

Dedication

I dedicate this dissertation to my son, Avery Perkins and my daughter, Lorelei Perkins. Every day you both inspire to me to pursue my goals to their fullest extent. You motivate me to be the best mom and person I can be. I hope that you both set big goals for yourselves and pursue them with purpose and passion. I promise to cheer you on and love you every step of the way.

Acknowledgements

During the last four years I have had the privilege of having two wonderful advisors. First, I would like to express my unending gratitude to my current advisor and dissertation chair, Dr. Jeanne Horst. I could not have completed this dissertation without you. Thank you for taking me on as an advisee right before the start of my dissertation. I have learned so much from you and had a lot of fun along the way. You have helped me strengthen my professional identity by encouraging me to lean into experiences that felt out of my depths. You have cheered me on through challenges and celebrated my victories. Thank you for caring about me as a person, student, and professional.

Second, I would like to express my gratitude to my advisor for the first two and a half years of my doctoral journey, Dr. Sara Finney. You helped me lay the foundation for all of my goals and accomplishments. Under your guidance I learned how to define my professional goals and be deliberate about achieving those goals. Thank you for challenging me to step outside of my comfort zone, advocating for me, and celebrating my accomplishments with enthusiasm and pride.

I also want to thank my wonderful dissertation committee members: Dr. Christine DeMars, Dr. Heather Harris, and Dr. Dena Pastor. This dissertation would be far worse with your thoughtful feedback and support. Christine, I have learned so much from you throughout the many courses you've taught. Thank you for always being a patient educator. Heather, your questions and comments have pushed me to understand propensity score methods at a deeper level. Thank you for your unending enthusiasm and kindness. Dena, you have taught me so much about effectively explaining complex topics. Thank you for your constant encouragement, exuberance, and humor.

Thank you to all of the Assessment and Measurement faculty members for the top notch education and endless support. The sense of community that each of you foster is one thing that makes the A & M program unique. I am so grateful that I got to learn from and work closely with all of you. I look forward to catching up with you at conferences!

I would be remiss to not thank my fellow students, both past and present. Thank you for providing camaraderie through the difficult courses, late nights, and weekends spent working. I have enjoyed working with and learning from you all. To my original cohort, thank you for the endless support, amusing shenanigans, and many game nights.

I would like to express my appreciation to my family for their support and understanding through the last four years. Thank you for understanding the less frequent visits and many missed phone calls. Rose, thank you for the constant encouragement and for always believing in me. The many fun adventures we've had helped me stay sane during this entire process. I look forward to having more time to spend with you!

Last, and certainly not least, my deepest appreciation goes to my husband Brandon and my children Avery and Lorelei. Nobody has been deeper in the trenches with me over the last four years than the three of you. Brandon, thank you for being willing to uproot our lives so that I could pursue my goals. Thank you for believing in me and celebrating even my smallest of accomplishments. Thank you for patiently listening to me practice presentations and explain my work. Avery and Lorelei, my love for you knows no bounds. Though you may not have always understood why I worked so much, thank you for your patience throughout this process. If there is one thing you gain from going through this experience with me, I hope it is an eternal love for learning. No words can express my love for you all.

Table of Contents

List of Tables

List of Figures

Abstract

In educational contexts, students often self-select into specific interventions (e.g., courses, majors, extracurricular programming). When students self-select into an intervention, systematic group differences may impact the validity of inferences made regarding the effect of the intervention. Propensity score methods are commonly used to reduce selection bias in estimates of treatment effects. In educational contexts, often a larger number of students receive a treatment than not. However, recommendations regarding the application of propensity score methods when the treatment group is larger than the comparison group have not been empirically examined. The current study examined the recommendation to recode the treatment and comparison groups (i.e., two types of treatment effect coding; Ho et al., 2007).

A simulation study was conducted to examine the performance of three propensity score methods (nearest neighbor matching, nearest neighbor matching with a 0.20 *SD* caliper, and generalized boosted modeling), using two coding methods (ATT and ATC) when the treatment group was larger than the comparison group. Additionally, three treatment sample sizes (200, 600, 1,000), three treatment to comparison group ratios (2:1, 4:3, 1:4), and four true treatment effects (Cohen's *d* of 0, 0.20, 0.50, 0.80) were simulated.

For nearest neighbor matching with a 0.20 *SD* caliper, adequate group covariate balance and low bias in the estimated treatment effect were observed across both coding methods regardless of which group was larger. In contrast, for generalized boosted modeling and nearest neighbor matching, group covariate balance and bias in the estimated treatment effect differed across coding method. When the treatment group was

larger than the comparison group, ATC coding resulted in better group covariate balance and lower bias than ATT coding. However, ideal balance was not obtained on all covariates, and bias in the estimated treatment effect was high for generalized boosted modeling and nearest neighbor matching. In sum, when the treatment group was larger than the comparison group, coding method did not matter for nearest neighbor matching with a 0.20 *SD* caliper. Conversely, for generalized boosted modeling, ATC coding performed better than ATT coding. Nearest neighbor matching did not perform well regardless of coding method.

CHAPTER 1

Introduction

Although random assignment is often considered the gold standard when designing studies from which causal inferences are drawn, there are circumstances under which random assignment to groups is not justified (e.g., Austin, 2011; Shadish et al., 2002). Consider the assessment specialist at a university who wishes to understand the effectiveness of general education writing courses on student learning. If desiring the ideal research design for making causal inferences about course effectiveness, students would be randomly assigned across general education writing courses (e.g., Shadish et al., 2002). However, realistically, it is typically not feasible to randomly assign students to a particular class.

When students cannot be randomly assigned to classes, there may be systematic differences between students who choose one class over another. Perhaps students who have stronger writing skills choose a more difficult writing course than students who have weaker writing skills. Students who completed the more difficult writing class may have higher scores on an end-of-the-semester test of general education writing skills than students who completed another class, simply because the students who enroll in those classes may have stronger writing skills. Thus, the inference regarding the effect of taking one course over another may be biased due to self-selection into a particular class.

Causal Inferences

When group comparisons are made between those who have and have not received an intervention, researchers are often interested in making a *causal* claim about the effectiveness of that intervention. That is, if students who completed a general

education writing course scored higher on a writing prompt than students who did not complete a general education writing course, the general education faculty would be likely to claim that the curriculum of the writing course resulted in increased writing ability. Thus, the causal claim would be that the writing course increased students' writing ability.

To understand the causal effect of a treatment for a single individual, the outcome when treatment is implemented and when treatment is not implemented must be known (e.g., Rubin's Causal Model, potential outcomes framework; Rubin, 1974). An approximation of the *counterfactual*, or what the outcome would have been for an individual under both treatment conditions, is necessary for making causal claims (Shadish et al., 2002). However, the counterfactual is unobservable at the individual level (because an individual cannot both receive and not receive treatment); instead, groups of individuals exposed (or not exposed) to treatment can be compared on the outcome of interest (Rubin, 1974; Shadish et al., 2002; Wainer, 2016). Thus, the treatment effect is the difference between the outcome when treatment is received and not received across all individuals (Rubin, 1974; Wainer, 2016).

Random assignment to groups is essential for approximating the counterfactual (Rubin, 1974). When random assignment is employed, the treatment and comparison groups should not differ systematically. The groups are expected to be balanced on both known and unknown variables, thus there are no variables that are related to treatment selection (Pedhazur & Schmelkin, 1991b; Rubin, 1974). When random assignment is not feasible, groups may systematically differ on known and unknown variables related to treatment selection. Specifically, when individuals self-select or are not randomly

assigned to receive or not receive a treatment, there are likely to be variables related to self-selection. Systematic group differences due to self-selected or non-random participation threaten the internal validity of the inferences that are drawn regarding the treatment effect (Shadish et al., 2002).

Internal Validity

Internal validity refers to whether the observed relation between treatment and outcome can be deemed causal as a result of study design (Shadish et al., 2002). That is, can an observed effect be attributed to the treatment that was implemented given the study design? Or, are there factors or confounds related to the treatment that could have led to the observed effect? When random assignment to treatment is not feasible, selection bias is a threat to the validity of the causal claims that can be made regarding the effect of receiving treatment.

In non-randomized (quasi-experimental) studies, the estimated treatment effect could appear to be stronger or weaker than it is in the population due to the systematic group differences on variables related to treatment selection (Rosenbaum & Rubin, 1983b; Rubin, 1973a, 1973b). However, when random assignment is not feasible, there are statistical methods that can be employed to reduce systematic differences between the groups, thus approximating random assignment to groups. By reducing systematic differences across the treatment and comparison groups, the researcher can reduce the impact of selection bias on the estimated treatment effect (Austin, 2011; Cochran & Rubin, 1973; Rosenbaum & Rubin, 1983b; Rubin, 1973a, 1973b, 1974; Shadish et al., 2008). Propensity score methods are commonly used to balance treatment and

comparison groups on covariates to reduce systematic group differences (e.g., Austin, 2011, 2013; Bai, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010).

Propensity Score Methods

Propensity score methods are one way to reduce systematic group differences, thus reducing bias in the estimated treatment effect (Austin, 2011; Caliendo & Kopeinig, 2008; Rosenbaum & Rubin, 1983b, 1985; Stuart, 2010; Stuart & Rubin, 2008). Propensity for treatment is estimated (typically via logistic regression) from variables that relate to treatment selection, the outcome, or both treatment selection and the outcome (i.e., confounders; Austin, 2011; Bai, 2011; Caliendo & Kopeinig, 2008; Rosenbaum & Rubin, 1985). After propensity scores are estimated, they can be used to match comparison group members to similar treatment group members, or to weight individuals based on their propensity score (Austin, 2009, 2011; Caliendo & Kopeinig, 2008; Rosenbaum & Rubin, 1983b, 1985; Stuart, 2010). When groups are balanced on all confounding covariates, selection bias is no longer present in the estimated treatment effect (Rubin, 1973a, 1973b).

When matching on the propensity score, there are various matching algorithms that can be used. With nearest neighbor matching (a greedy algorithm), a comparison group member is matched to a treatment group member with the closest propensity score value (Austin, 2011; Caliendo & Kopeinig, 2008; Rosenbaum & Rubin, 1983b, 1985; Stuart, 2010). Using this method, all treatment group members receive a match from the comparison pool. Nearest neighbor matching can also be implemented with a distance caliper. When a caliper is specified, the researcher defines the maximum difference on the propensity score that is allowable for a comparison group member to be matched to a

treatment group member (Austin, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010). A typical caliper that is used is 0.20 standard deviations of the logit of the propensity score (Austin, 2009). When nearest neighbor matching is employed with a caliper, there is the potential for some treatment group members to not receive a match from the comparison pool (Austin, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010).

Propensity score methods work well to create balanced groups when certain key assumptions are not violated (Austin, 2011; Caliendo & Kopeinig, 2008; Ho et al., 2007). The first crucial assumption is that of strong ignorability (or no unmeasured confounders). The strong ignorability assumption means that conditional on the covariates, treatment assignment is not related to the outcome (Rosenbaum & Rubin, 1983a; Stuart, 2010). Thus, if all sources of selection bias are included in the estimation of the propensity score, treatment assignment should not be related to the outcome. The second crucial assumption is that of common support (or sufficient overlap of propensity scores across treatment and comparison groups). Common support refers to the degree to which the propensity score distributions of each group overlap (Guo & Fraser, 2015). Sufficient overlap between groups is necessary for obtaining adequately balanced matched treatment and comparison groups. In contrast, if there is little overlap between the propensity score distributions of the groups, then obtaining good comparison pool matches for the treatment group members will be difficult. The third crucial assumption is that of stable unit treatment value. Stable unit treatment value refers to independence between the treatment of one individual and another individual (and thus, independence between the outcome of individuals; Stuart, 2010). In order to estimate the effect of the

treatment, comparison group members should receive no treatment and all treatment members should receive the intended degree of treatment.

The goal of propensity score matching is to construct a comparison group that does not systematically differ from the treatment group on variables related to treatment selection. Thus, in order to attain a match for every treatment group member, it is beneficial to have a comparison group that is larger than the treatment group (Pan & Bai, 2015; Rosenbaum & Rubin, 1985; Rubin, 1979; Stuart, 2010; Stuart & Rubin, 2008). However, there may be situations in which the number of individuals who receive treatment is greater than the number of individuals in the comparison pool. For example, in higher education, it is desirable that all students complete general education courses. That is, all students should receive an intervention aimed at increasing their general education knowledge, skills, and abilities. If educational researchers are interested in comparing learning outcomes of students who have and have not completed coursework, it is possible that they might face a scenario in which there will be a larger number of students who received treatment than who did not. A larger treatment than comparison group might also be obtained if the desire is to compare learning outcomes across students who completed different general education courses within the same domain. More research is needed regarding the use of propensity score methods to reduce selection bias when there are more treatment group members than comparison group members.

In the quasi-experimental literature, propensity score methods have been used with a larger treatment than comparison group; to compare two labor success programs (Lechner, 2000), to compare results obtained with different comparison to treatment

group ratios (Holzman & Horst, 2019), to compare survival rates of those who did and did not receive a smoking-cessation intervention (Austin & Cafri, 2020), to compare results obtained from different matching and coding methods (Perkins & Horst, 2020), and to compare math performance of students who completed traditional or new math curricula (Powell et al., 2020). Additionally, some guidance regarding how to implement propensity score methods when the treatment is larger than the comparison group has been provided; however, the recommendations lack empirical support (Ho et al., 2007; Stuart, 2010). These recommendations include using subclassification, full matching, weighting by the odds (Stuart, 2010), matching with replacement, or switching the coding of the treatment and comparison group (Ho et al., 2007).

Current Study

The aim of the current study was to examine how well propensity score methods reduce systematic differences between groups when the treatment group is larger than the comparison group. Although methods for addressing a larger treatment than comparison group have been suggested in the literature, there is little evidence to support those suggestions. Thus, it is important to understand whether propensity score methods result in adequate balance between the treatment and comparison groups and accurate estimates of the treatment effect when the treatment group is larger than the comparison group. The aim of the current study was to examine bias in the estimated treatment effect, covariate balance after matching or weighting, loss of treatment group members, and the direction and magnitude of the estimated treatment effect across different propensity score methods. Therefore, a simulation study was used in order to specify the true treatment effect and examine the research questions under varying conditions (Feinberg &

Rubright, 2016). Given the lack of research on the use of propensity score methods when the treatment group is larger than the comparison group, three propensity score methods (i.e., nearest neighbor matching, nearest neighbor matching with a caliper, and generalized boosted modeling) were evaluated across different coding methods, treatment to comparison group ratios, treatment sample sizes, and true treatment effect sizes. The following research questions were specified:

Research Question 1a: When the Treatment Group is Larger Than the Comparison Group, Can Propensity Score Methods Accurately Recover the True Treatment Effect?

The first research question was crucial for providing guidance regarding the use of propensity score methods when the treatment group is larger than the comparison group. The goal of propensity score methods is to reduce bias in the estimated treatment effect that is due to systematic differences between the groups on confounding variables (Austin, 2011, 2013; Bai, 2011; Caliendo & Kopeinig, 2008; Rosenbaum & Rubin, 1983b, 1985). Thus, information regarding the accuracy of propensity score methods under various conditions when the treatment group is larger than the comparison group can provide guidance to researchers.

Research Question 1b: Does the Magnitude and Direction of the Estimated Treatment Effect Differ Across Propensity Score Methods Depending on Group Coding?

The first research question also pertained to whether the magnitude and direction of the estimated treatment effect differed depending on the whether the treatment group was coded 1 or 0. In typical applications of propensity score methods, the treatment group is coded 1 and the comparison group is coded 0. This coding aligns with the estimation of the average treatment effect for the treated (ATT). One recommendation

when the treatment group is larger than the comparison group is to switch the coding of the treatment and comparison groups. In doing so, the comparison group will be coded 1 and the treatment group coded 0. When the coding is reversed, a treatment group member match is selected for each comparison group member, with the goal of creating a treatment group that is similar to the comparison group on the covariates. This coding aligns with the estimation of the average treatment effect for the comparison group (ATC). In other words, ATC coding answers a different research question than ATT coding. There is little explanation of the ATC in the propensity score literature. Examination of the magnitude and direction of the treatment effect for the same propensity score method across different coding methods will expand the current understanding of the ATC.

Research Question 2: When the Treatment Group is Larger Than the Comparison Group, Can Propensity Score Methods Achieve Adequate Group Balance on the Covariates?

The second research question examined the extent to which balanced groups could be created when there is a comparison pool that is smaller than the treatment group. Adequate group balance on the covariates after matching or weighting indicates that selection bias due to the included covariates has been reduced or removed from the estimated treatment effect (Ho et al., 2007). Thus, the validity of the inferences made regarding an observed treatment effect is strengthened. Inadequate group balance after matching or weighting on the covariates indicates that the groups systematically differ and estimation of the treatment effect from the matched or weighted sample is not appropriate (Ho et al., 2007). Examination of group balance on the covariates after

matching or weighting will provide information regarding the quality of the matched or weighted samples.

Research Question 3: When the Treatment Group is Larger Than the Comparison Group, Does the Loss of Treatment Group Members for Nearest Neighbor Matching With a Caliper Differ Across Coding Methods, Treatment to Comparison Ratios, and Treatment Sample Sizes?

When the treatment group is larger than the comparison group and one-to-one matching is implemented, there will be treatment members who do not receive a match. Specifically, the number of treatment group members who receive a match can, at a maximum, equal the number of comparison group members. When nearest neighbor matching is used with a caliper, there can be a larger loss of treatment group members if there is no comparison group match within the specified caliper distance. Loss of treatment group members is undesirable because it leads to a loss of treatment representation and a decrease in the matched sample size (Jacovidis et al., 2017).

CHAPTER 2

Review of the Literature

Within educational research, random assignment of students to groups (e.g., courses, remediation programs, majors) is often outside of the researcher's control. When random assignment to groups is not feasible, there are statistical methods that can be employed to reduce systematic group differences. Throughout the quasi-experimental literature, covariate adjustment (e.g., Cochran, 1953; Cochran & Rubin, 1973; Lord, 1960; Pascarella et al., 2013; Pedhazur & Schmelkin, 1991a; Rubin, 1973b, 1974, 1979), stratification (Austin, 2011; Maxwell & Delaney, 2004; Rosenbaum & Rubin, 1983b; Stuart, 2010), matching (Rubin, 1973a, 1973b, 1979), and propensity score matching (Austin, 2011, 2013; Bai, 2011; Caliendo & Kopeinig, 2008; Rosenbaum & Rubin, 1983b, 1985) are among the most frequently used methods for reducing systematic group differences when random assignment is not feasible. The goal of these methods is to control for variables that relate to group selection, thus reducing systematic differences between the groups.

In this chapter, I review the literature on four common quasi-experimental methods that are used to reduce systematic group differences (covariate adjustment, stratification, matching, and propensity score matching). Next, I provide an in-depth description of the steps involved in propensity score matching. I then explain generalized boosted modeling (an extension of propensity score methods). Finally, I review the limited literature on propensity score matching when the treatment group is larger than the comparison group.

Covariate Adjustment

Researchers refer to covariate adjustment by different names throughout the literature (e.g., analysis of covariance [ANCOVA], regression adjustment for background variables). In the traditional ANCOVA approach, group differences are estimated on a continuous outcome variable after controlling for covariates that are included in the model. Instead of estimating raw group mean differences on the outcome variable, the researcher is now examining differences in the group means of the outcome variable that have been adjusted based on how each groups' means on the covariate(s) differ from the grand mean of the covariate(s). This technique is straightforward, as it is simply multiple regression with categorical and continuous predictors. However, there is conflicting empirical evidence regarding the ability of ANCOVA to accurately reduce selection bias in the estimated treatment effect. Some studies have shown that covariate adjustment did not adequately reduce selection bias (Rubin, 1973b, 1979); whereas, others have argued that covariate adjustment works just as well as, or better than matching methods (Cochran & Rubin, 1973; Pascarella et al., 2013).

In a study on liberal arts versus other 4-year universities' impacts on students' critical thinking skills, need for cognition, and positive attitudes toward literacy, Pascarella et al. (2013) examined differences between the use of covariate adjustment and propensity score matching. Twelve covariates (race, sex, parental education, ACT composite score, federal grant receipt, institutional grant receipt, precollege major intent, precollege political views, precollege purpose in life scores, precollege critical thinking scores, precollege need for cognition scores, and precollege attitude toward literacy scores) were identified as variables that were confounded with selection into attending a

liberal arts or 4-year university. For both propensity score matching and covariate adjustment, a model without the pretest as a covariate and a model with the pretest as a covariate were estimated. Baseline mean differences (unadjusted) on critical thinking skills, need for cognition, and positive attitudes toward literacy between those who attended a liberal university versus a 4-year university were statistically significant. The magnitude of each mean difference decreased from the baseline difference for all four models except for need for cognition scores using covariate adjustment with the pretest excluded. Pascarella et al. (2013) concluded that propensity score matching does not provide a substantial benefit over covariate adjustment. Additionally, Pascarella et al. (2013) concluded that both methods led to a substantial reduction in the amount of bias in the estimated mean difference between liberal and 4-year university student impact, and the models including the pretest as a covariate resulted in the largest reduction in bias.

The comparison of covariate adjustment to propensity score matching is useful to inform researchers' selection between the two methods. However, the claims made by Pascarella et al. (2013) based on their study overreach. The researchers focused on how different models including covariates (both propensity score matching and covariate adjustment) reduced the amount of bias in the estimated treatment effect. Given that they used empirical data from an applied study, it was not possible to determine whether and to what extent bias was present in the estimated treatment effect. Pascarella et al. (2013) assumed that the unadjusted treatment effect was biased, and that any change to the estimated treatment effect was due solely to a reduction in bias.

A fundamental consideration when choosing an appropriate statistical method is how well the method aligns with the research questions and the data that were collected.

That is, statistical methods were developed to answer specific types of research questions. ANCOVA was developed to allow for the comparison of randomly assigned groups' adjusted means, and, in particular, to reduce the error term, which results in a more powerful test of statistical significance (Pedhazur & Schmelkin, 1991a; 1991b). ANCOVA was intended for randomly assigned groups or groups for which there are no expected differences on a covariate, rather than quasi-experiments prone to self-selection bias. Although ANCOVA is a useful tool for understanding group differences while controlling for known covariates, it is not appropriate for accounting for self-selection into treatment unless the research question of interest is how groups differ on the outcome when both groups have equivalent means on the covariates (Pedhazur & Schmelkin, 1991b).

Indeed, in two different simulation studies, Rubin (1973b, 1979) demonstrated that ANCOVA alone did not adequately reduce selection bias. ANCOVA only reduced bias well when there were equal numbers of treatment and comparison group members and equal group covariate variances. The exception was when there was a large ratio of comparison to treatment group members (at least 4:1; Rubin, 1973b). ANCOVA can also be conducted after creating treatment and comparison groups that are matched on covariates related to selection bias. Specifically, each treatment group member receives a comparison group member match with similar covariate values. Instead of conducting the ANCOVA on scores from the original, unbalanced sample, scores from the matched sample are used for the analysis. Thus, the matched sample approximates random assignment of participants to treatment and comparison groups (if the treatment and comparison group are balanced on the covariates in the matched sample). Researchers

might use ANCOVA on matched samples to understand group differences while controlling for known covariates.

Extending his earlier work, Rubin (1979) showed that ANCOVA after matching worked consistently well because very few conditions resulted in increased bias. An alternative to samples matched on individual covariate values is to estimate propensity scores, and then conduct an ANCOVA using the propensity score as the covariate (Austin, 2011; Rubin, 1973b, 1979). ANCOVA used in conjunction with a matching procedure or with propensity scores has been shown to work reasonably well for reducing selection bias (Austin, 2011; Austin et al., 2007; Rubin, 1973b, 1979).

When using ANCOVA, group differences on the outcome are adjusted as if both groups were the same on the covariates. In other methods that use covariates to reduce systematic group differences the covariates are used as a *balancing score*.

Balancing Score

In order to reduce the bias in the estimated treatment effect that is due to treatment selection, the treatment and comparison groups must be balanced on the variables related to treatment selection (Rosenbaum & Rubin, 1983b). Rosenbaum and Rubin (1983b) described balancing scores as scores that are used to aid in the creation of balanced treatment and comparison groups. Balancing scores range from *fine* to *coarse*, with the observed values on the covariates being the finest balancing score. Using the values of the covariates results in the *finest* balancing score because a treatment and comparison group member with the same balancing score will have the same values on each covariate. Conversely, the propensity score is the *coarsest* balancing score, because two individuals with the same propensity score do not necessarily have the exact same

values on all covariates used to estimate the propensity score (Rosenbaum & Rubin, 1983b).

The balancing score is sufficient for eliminating systematic differences between the treatment and comparison groups assuming that all relevant covariates have been included. Balancing scores can be used for stratification (strata are created based on the balancing score; Austin, 2011; Austin et al., 2007; Rosenbaum & Rubin, 1983b) and matching (Austin, 2011; Austin et al., 2007; Rosenbaum & Rubin, 1983b, 1985; Rubin, 1973b, 1979).

Stratification

Another method for reducing systematic group differences is stratification. Stratifying (i.e., subclassifying, blocking) is the process of sorting the entire sample into blocks based on the covariate values. In doing so, treatment and comparison group members within each block should be relatively homogenous on the covariates that are related to treatment selection (Austin, 2011; Cochran, 1968; Guo & Fraser, 2015; Maxwell & Delaney, 2004). After creating strata, the treatment effect is estimated for each stratum, after which the individual estimates are pooled together (Austin, 2011; Cochran, 1968; Guo & Fraser, 2015). Pooling allows for each stratum to be weighted based on the number of treatment and comparison group members within each stratum (Austin, 2011; Cochran, 1968; Guo & Fraser, 2015).

Five strata are recommended (Austin, 2011; Cochran, 1968), which resulted in a reduction of 90% to 95% of the bias in the treatment effect that was due to one covariate (Cochran, 1968). Stratification can be done using the actual values of the covariates (Cochran, 1968) or using the propensity score (Austin, 2011; Caliendo & Kopeinig,

2008; Guo & Fraser, 2015). Using too many strata can result in smaller reduction in bias in the treatment effect than using fewer strata (Cochran, 1968). Whereas stratification uses the balancing score to sort the entire sample into subsets (or strata), matching methods use the balancing score to create matched treatment and comparison groups that have similar covariate distributions.

Matching

Matching methods were introduced as a means of creating balanced treatment and comparison groups by matching on variables related to treatment selection (covariates; Rubin, 1973a). Matching methods involve selecting a portion of the comparison group sample that most similarly resembles the treatment group sample on the covariates (Rubin, 1973a). This matching can occur in different ways. Pair-matching (or exact matching) works by selecting a comparison group member as a match for a treatment group member if all values of the covariates match those of the treatment group member. Nearest pair-matching works in the same way except the values of the covariates do not have to match exactly. Rather, the comparison group member with the *closest* values on the covariates to the treatment group member is selected as the match. Mean-matching works by selecting a comparison group member match for the treatment group member that will result in balance between the two group means (Rubin, 1973a).

When all appropriate covariates are selected, matching methods will remove selection bias from the treatment effect conditional upon the covariates included in the model by balancing the groups on the covariates (Rubin, 1973a, 1973b). When bias in each covariate is reduced by the same amount, matching methods result in equal percent bias reduction (EPBR; Rosenbaum & Rubin, 1985). If a method is EPBR, then the

method is appropriate for reducing systematic group differences and bias in the estimated treatment effect (Rubin, 1979; Rosenbaum & Rubin, 1985).

Although the procedure used with matching is straightforward, the process of matching becomes increasingly complicated as the number of covariates increases (Rosenbaum & Rubin, 1985). First, to ease the matching process, researchers may make the decision to categorize continuous variables. When this is done, it can be easier to match individuals on the covariate; however, there is a substantial loss of information when a continuous variable is categorized (MacCallum et al., 2002). Due to the loss of information, categorization of continuous variables is typically not recommended, unless it makes sense to do so (MacCallum et al., 2002). Second, even if all covariates are categorical variables, the potential for unmatched treatment group individuals increases as the number of covariates increase. Rosenbaum and Rubin (1985) demonstrated this problem by considering a scenario where there are 20 covariates on which the treatment and comparison groups are matched. If each of the 20 variables were binary (only two response options), there would be over one million different possible response patterns across the 20 variables. Even with a very large sample, there will likely be many treatment group members for which there is no exact match from the comparison group pool. Thus, it may not be possible to find an appropriate comparison group member match for all treatment group members (Rosenbaum & Rubin, 1985). Propensity score matching was developed to reduce the complexity of matching on a large number of covariates while still creating matched treatment and comparison groups (Rosenbaum & Rubin, 1983b, 1985; Stuart & Rubin, 2008).

Propensity Score Methods

Another method for reducing selection bias is through the estimation of propensity scores. Propensity score methods involve the estimation of the propensity for treatment for each individual in the sample, conditional upon researcher-identified covariates. The propensity score is estimated such that individuals with the same propensity score will have the same propensity for treatment, regardless of whether they belong to the treatment or comparison group (Rosenbaum & Rubin, 1983). When comparison group members are matched to treatment group members using the propensity score, the matched treatment and comparison groups should have similar covariate distributions. Thus, the propensity score is a *balancing score* conditional upon the covariates used to estimate the propensity score (Austin, 2011). After the propensity score is estimated, treatment group members can be matched to comparison group members with the same or similar propensity score (depending on the matching method implemented). Treatment assignment is said to be ignorable conditional upon the propensity score, because a treatment group member and a comparison group member with the same propensity score are, in theory, interchangeable even if their specific covariate values differ (Rosenbaum & Rubin, 1983b). By creating matched pairs of treatment and comparison group members, a comparison group is created that is similar to the treatment group on variables that are related to selection into treatment. Thus, systematic differences between the treatment and comparison group will be reduced or eliminated and the estimated treatment effect will not be biased due to self-selection.

Matching on the balancing score results in groups that are balanced on the *observed* covariates. Any covariates on which the groups are unbalanced that are not

included when estimating the propensity score (whether these covariates are measured or not) will result in systematic group differences (Rosenbaum & Rubin, 1985). Thus, the estimated propensity scores are only as good as the model used to estimate them. That is, if relevant covariates are omitted or the functional form of the relation between the covariates and the propensity score is misspecified, the use of the propensity scores to create balanced groups may not effectively reduce systematic differences between the groups (Austin, 2009; Craig, 2020; Rosenbaum & Rubin, 1985). By matching treatment group members to comparison group members on the propensity score, the researcher can create groups that have the same distribution of propensity for treatment selection, conditional upon the covariates. Achieving this balance makes the examination of the treatment effect more reasonable when randomization is not feasible (Rosenbaum & Rubin, 1983b), strengthening the validity of the inferences made regarding the treatment effect. Without balancing the groups on the covariates, researchers have little evidence that any observed treatment effect is indeed attributable to the occurrence of the treatment (Cochran & Rubin, 1973; Rosenbaum & Rubin, 1983b, 1985; Rubin, 1973a, 1979; Shadish et al., 2002). Any magnitude of estimated treatment effect could reflect systematic group differences rather than the true treatment effect in the population. Thus, without reducing or removing systematic group differences, researchers can easily conclude that an ineffective treatment is indeed effective or that an effective treatment is ineffective (Rosenbaum & Rubin, 1983b, 1985; Rubin, 1973a, 1979).

Assumptions of Propensity Score Matching

Propensity score matching has key assumptions, which if not met, can impact the credibility and generalizability of the results (Austin, 2011; Caliendo & Kopeinig, 2008;

Ho et al., 2007; Rosenbaum & Rubin, 1983b). When the assumptions of propensity score matching hold for the data that are being analyzed, propensity score matching should reduce the effect of selection bias on the estimated treatment effect (Rosenbaum & Rubin, 1983b). Some assumptions are difficult to evidence with empirical data; however, researchers can examine these assumptions to some extent, providing evidence for the appropriateness of propensity score matching for their study.

Strong Ignorability. A key assumption that is two-fold is the strong ignorability assumption (Rosenbaum & Rubin, 1983b; Stuart, 2010). Strong ignorability means that the treatment selection is independent of the measured outcome, given the covariates. For example, with random assignment, treatment assignment is strongly ignorable. This is the case because the outcome is not dependent upon treatment assignment. When using propensity score matching, we are trying to mimic a similar situation. Thus, after balancing groups on the covariates, treatment selection is strongly ignorable if assignment or selection into treatment conditions is unrelated to the observed outcome. This assumption is important because if treatment selection is related to the outcome, then treatment selection is not independent of the outcome (i.e., treatment selection is a confound). When treatment selection is not independent of the outcome, the treatment and comparison groups should not be directly compared on the outcome of interest (Rosenbaum & Rubin, 1983b). Interwoven with strong ignorability is the assumption that all variables that relate to treatment selection are included in the estimation of the propensity score (i.e., no unmeasured confounders). If variables that are related to treatment selection are omitted from the propensity score model, the estimated treatment effect will be biased due to treatment selection. Moreover, any unmeasured covariates

will threaten the independence of treatment selection and the outcome. Although this assumption is not directly testable, researchers can guard against severe violations of this assumption by consulting theory and empirical studies to determine what variables are likely related to treatment selection (Austin, 2011; Rosenbaum & Rubin, 1983a; Steiner et al., 2010; Stuart, 2010; Stuart & Rubin, 2010). A lack of careful consideration of which variables to measure during the study design phase cannot be corrected for through statistical analyses. Careful thought and thorough review of the relevant literature is the best safeguard against violation of this key assumption (Steiner et al., 2010). Although there is no method to determine whether the strong ignorability assumption has been violated, sensitivity analysis can provide an *indication* as to whether the strong ignorability assumption has been violated (Austin, 2011; Steiner et al., 2010; Stuart, 2010). If the estimated treatment effect differs with the inclusion of additional covariates, it is likely that the strong ignorability assumption has been violated (Stuart, 2010).

Common Support. Common support (or sufficient overlap of propensity scores) is related to the assumption of strong ignorability (Guo & Fraser, 2015). Common support refers to the amount of overlap between the propensity score distributions in the treatment and comparison group. If there is not sufficient overlap between the groups' propensity score distributions, the strong ignorability assumption will not be met (Caliendo & Kopeinig, 2008). That is, if the propensity score distributions differ to a large degree, sufficient treatment and comparison group matched pairs cannot be found. Thus, groups will likely not be balanced on the covariates after matching. If there are

systematic differences between the groups after matching, selection bias is still present and the treatment assignment is not independent of the outcome.

Stable Unit Treatment Value. The stable unit treatment value (SUTVA) assumption specifies that the treatment of one individual is not affected by the treatment of another individual (Stuart, 2010). The biggest threat to this assumption is interaction between treatment and comparison group members. For example, if a treatment group member shares details of the treatment with a comparison group member, the comparison group member's outcome may be influenced by this information. Thus, rather than the comparison group member's outcome score representing *no* treatment (and no treatment was intended), it is instead the outcome under receiving *some* treatment. If the intention is to understand the treatment effect by comparing treated and non-treated individuals, the inclusion of this comparison group member in the sample would threaten the validity of the estimated treatment effect. To determine whether this assumption holds, researchers should give careful consideration to potential interaction of group members during the study design phase. Additionally, during the treatment implementation, researchers should record any events that indicate that this assumption has been violated.

Propensity Score Matching Steps

When implementing propensity score matching, there are two general stages; the non-parametric, pre-processing stage and the treatment effect estimation stage (e.g., Ho et al., 2007). The first non-parametric, pre-processing stage can be further broken down into steps. Some of the steps may be combined or omitted in the propensity score matching literature (Austin, 2011; Bai, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008). Nonetheless, the steps are the same and each step requires careful

consideration from the researcher (Austin, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008). The stages can be broken down as follows:

1. Selection of covariates/estimation of propensity scores.
2. Selection of matching method(s).
3. Evaluation of common support.
4. Evaluation of matching quality.
5. Estimation of the treatment effect.
6. Evaluation of sensitivity analysis.

Steps one through four comprise the non-parametric, pre-processing stage; step five is the treatment effect estimation stage. Step six is a follow-up procedure to determine the potential impact of unmeasured covariates on the estimated treatment effect.

Step 1: Selection of Covariates and Estimation of Propensity Scores. The first step (which is most often conducted during the study design phase) is the selection of relevant variables to include in the propensity score model and specification of the propensity score model. First, I will discuss important considerations for variable selection, then specification of the propensity score model.

Variable Selection. Variable selection is important when specifying the propensity score model (Caliendo & Kopeinig, 2008; Dehejia & Wahba, 1999; Steiner et al., 2010). The goal of matching on the propensity score is to reduce systematic differences between the groups, which in turn will decrease the bias in the estimated treatment effect that is due to treatment selection. If relevant variables are omitted from the model, the estimated treatment effect may be more biased than if no matching was

implemented (Bai, 2011; Dehejia & Wahba, 1999; Rosenbaum & Rubin, 1983b; Stuart, 2010). Some researchers caution that the inclusion of variables that are not meaningful for the estimation of the propensity (i.e., including all variables that a researcher has available to use) can be detrimental (Bai & Clark, 2019; Shadish et al., 2008), whereas others indicate that the inclusion of irrelevant variables does not bias the estimate of the treatment effect (Caliendo & Kopeinig, 2008).

Due to the importance of variable selection, there is a wealth of literature that provides guidance for selecting variables to include in the propensity score model (Austin, 2011; Bai, 2011; Bai & Clark, 2019; Brookhart et al., 2006; Caliendo & Kopeinig, 2008; Ho et al., 2007; Steiner et al., 2010; Stuart, 2010; Stuart & Rubin, 2008). In general, theory and previous research should guide variable selection. A review of the relevant literature should reveal variables that are theoretically or empirically related to either treatment selection or the outcome of interest (Austin, 2011; Bai, 2011; Bai & Clark, 2019; Caliendo & Kopeinig, 2008; Ho et al., 2007; Stuart, 2010; Stuart & Rubin, 2008). Bias is best reduced when the propensity score model includes covariates that relate to both treatment selection and the outcome of interest (Austin et al., 2007; Bai & Clark, 2019; Brookhart et al., 2006; Caliendo & Kopeinig, 2008; Ho et al., 2007; Stuart & Rubin, 2008). Variables that are only related to the outcome (and not treatment selection) should also be included to reduce bias in the estimated treatment effect (Bai & Clark, 2019; Brookhart et al., 2006). Including variables that are only related to treatment selection may be inefficient and provide no benefit in terms of bias reduction because the variables are not related to the outcome (Brookhart et al., 2006; Stuart & Rubin, 2008). Others state that in most settings researchers can likely include all available covariates

without risking severe consequences in the form of bias in the estimated treatment effect (Austin, 2011; Caliendo & Kopeinig, 2008).

The covariates selected for the propensity score model should be measured prior to treatment implementation (Austin, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008). That is, covariates that are affected by the treatment should not be included in the propensity score model. If it is not possible to measure certain covariates prior to treatment implementation, then they should not be included in the propensity score model. The only exception is variables that are constant over time (e.g., demographic variables that do not change, proxies for student ability, historical variables; Caliendo & Kopeinig, 2008). Because the propensity score model is predicting the probability of receiving treatment, the inclusion of a variable that is affected by treatment would result in propensity scores that are conflated with the treatment itself.

Propensity Score Model. Once the covariates are selected and measured, the propensity score model is selected. When selecting the propensity score model, one consideration is the number of treatment options (Caliendo & Kopeinig, 2008). That is, how many different possible group membership variables are to be estimated? Typically, there are two treatment options (i.e., received treatment, did not receive treatment), which would indicate that a model that allows for the estimation of a binary outcome would be necessary. Although the logit model is used most frequently in the propensity score matching literature, any model that accommodates a binary outcome will work well (Caliendo & Kopeinig, 2008). If there are more than two treatment options (e.g.,

comparing two different treatment methods and no treatment), a multinomial probit model or a series of binomial models may be appropriate (Caliendo & Kopeinig, 2008).

Consideration must also be given to the functional form of the relation between the covariates and treatment selection (Rosenbaum & Rubin, 1985). The researcher should determine the appropriate form for the propensity score model. Rosenbaum and Rubin (1985) cautioned that estimated propensity scores from an incorrect model will not be useful for balancing the groups nor for reducing selection bias. However, propensity score methods have been shown to result in adequate balance even when the propensity score model is misspecified (Craig, 2020).

Logistic Regression. Logistic regression is most frequently used to estimate propensity scores when there are two groups (e.g., treatment group, comparison group; Austin, 2011; Bai, 2011; Caliendo & Kopeinig, 2008, Rosenbaum & Rubin, 1985). When using logistic regression, the value predicted by the model is the logit of the propensity score. Thus, the predicted outcome (probability of receiving or not receiving treatment) is on the logit metric, which allows for the estimation of a linear relation between the covariates and treatment selection via the following model:

$$ln\left[\frac{P(y_i=1|\boldsymbol{X}_i)}{1-P(y_i=1|\boldsymbol{X}_i)}\right] = \mathbf{X}_i\boldsymbol{\beta} \quad (1)$$

where the logit is the natural log of the odds of treatment. The odds of treatment is represented as the probability of treatment ($y_i = 1$) conditional upon the vector of covariates ($\mathbf{X}_i$), divided by one minus the probability of treatment conditional upon the vector of covariates. The logit for person i is equal to the product of $\mathbf{X}_i$ (a vector of covariate values for person i) and $\boldsymbol{\beta}$ (a vector of logistic regression coefficients; Guo & Fraser, 2015). The propensity score on the logit metric can also be transformed onto a

probability metric, resulting in the probability (or propensity) for treatment selection, conditional upon the covariates in the model.

Logistic regression can accommodate categorical and continuous covariates as well as interactions and polynomial terms (Tabachnick & Fidell, 2013). Although logistic regression is often used as a parametric procedure (where the model parameters are tested for statistical significance), that is not the case when logistic regression is used to estimate the propensity scores. The regression parameters and model deviance from the propensity score model are not of primary interest. The interest is predicting the propensity for treatment in order to create matched treatment and comparison groups (Pan & Bai, 2015; Stuart, 2010).

Step 2: Selection of Matching Method(s). There are a variety of matching methods from which to choose, and they can result in different matched treatment and comparison groups depending on the data. Thus, when selecting a matching method it is common practice to audition multiple matching methods (Austin, 2011; Caliendo & Kopeinig, 2008; Ho et al., 2007; Rubin, 1973a; Stuart & Rubin, 2008). After evaluating the quality of the matched sample resulting from each matching method, researchers are then able to champion one or multiple matching methods best suited to their study. Although all matching methods work well in general, this does not mean that they work well for all samples, under all circumstances (Caliendo & Kopeinig, 2008; Stuart & Rubin, 2008). The recommendation to implement multiple matching methods allows the researcher to evaluate the quality of matches for each method individually and compare the quality of matches across methods. Sometimes evaluation of the matching quality will illuminate that one method results in better group balance after matching than the other

methods. Other times, the results can be less clear. Caliendo and Kopeinig (2008) suggested that when the matching quality is similar across methods, then the choice of which method to champion matters less. That is, if all methods result in similar balance, then all methods should result in similar adjustment for selection bias in the eventual estimation of the treatment effect.

The suggestion to audition multiple matching methods may seem daunting to researchers with little propensity score matching experience. It is important to point out that the researcher is not conducting a full analysis using each method (Stuart & Rubin, 2008). Specifically, several matching methods are used to create different matched samples from the complete dataset. The quality of matches for each of the resulting matched datasets is evaluated using multiple criteria. The treatment effect is *not* estimated for each matching method (Stuart & Rubin, 2008). If a researcher estimated the treatment effect for matched samples from five different matching methods, there is the potential that different conclusions would be made regarding the treatment effect depending on which matching method the researcher championed. Indeed, when the quality of matches differed across methods, the magnitude and direction of the estimated treatment effect differed across matching methods (Austin, 2013; Austin et al., 2007; Jacovidis et al., 2017; Perkins & Horst, 2020). Therefore, researchers are cautioned against examining the treatment effect prior to the selection of which matching method results in the best balance for the study (Bai, 2011; Stuart & Rubin, 2008)

Matching algorithms can be categorized into optimal and greedy methods. Optimal methods are designed to select matches so as to optimize group balance on the covariates; whereas greedy methods are designed to select the closest match from the

remaining comparison group pool (Austin, 2011; Cochran & Rubin, 1973; Gu & Rosenbaum, 1993; Rubin, 1973a; Stuart, 2010).

Optimal matching methods minimize the average propensity score difference across all matched pairs (Austin, 2013; Gu & Rosenbaum, 1993). The optimal algorithm works to ensure that these individual matches result in the lowest average within-pair difference on the propensity score. A nice feature of optimal matching methods is that they result in every single treatment member receiving a match. That is, there is no loss of treatment group representation when optimal matching is used (Austin, 2013; Gu & Rosenbaum, 1993).

Greedy matching methods process through the treatment group members one by one and select the comparison group member with the closest propensity score to the treatment member (Austin, 2011; Gu & Rosenbaum, 1993). At first glance, greedy matching methods may sound identical to optimal matching methods, but there are certain distinctions between the two algorithms. Greedy matching methods differ from optimal matching methods in that with the greedy algorithm, there is no value that the method is trying to minimize. After a match is made, it is not re-evaluated. However, greedy matching methods tend to result in similar group balance to optimal matching methods, with generally the same comparison members selected from the comparison group pool (Gu & Rosenbaum, 1993). Although the same comparison group members were selected from the comparison pool, the matched treatment and comparison group pairs were not the same across greedy and optimal methods (Gu & Rosenbaum, 1993).

Nearest Neighbor Matching. Nearest Neighbor (NN) matching is a greedy matching method (Rubin, 1973a). After the propensity score is estimated for treatment

and comparison group members, the treatment group members are ordered based on the estimated propensity score (Rubin, 1973a). Ordering of treatment member selection can be done from high to low, from low to high, or randomly (Rubin, 1973a). Random ordering of treatment members based on the propensity score in order to select matches tends to result in better matches and lower bias than high to low or low to high ordering (Austin, 2013).

For each treatment member, the NN algorithm will select the comparison group member with the closest propensity score. After the treatment group member receives a comparison group member match, both participants are removed from the unmatched sample and are placed into the matched sample (unless matching with replacement, where the comparison group member would be returned to the comparison group pool and could be matched to additional treatment group members). As long as the comparison group is larger than the treatment group, all treatment group members will receive a comparison group member match, with only unmatched comparison group members excluded from the matched sample.

If there are large differences in the distribution of the propensity score across the treatment and comparison groups, NN matching can result in poorer balance than other methods (Stuart, 2010). Importantly, greedy matching methods do not ensure that there will be a small difference between propensity scores for each match. For example, a treatment group member with a propensity score of 0.80 could be matched with a comparison group member with a propensity score of 0.50 because this comparison group member is indeed the closest match. If matched pairs differ to a large extent on the propensity score, there may still be systematic differences between the treatment and

comparison group even after matching. However, greedy matching methods have been shown to produce similar group balance to optimal matching methods (Gu & Rosenbaum, 1993). Like optimal matching, greedy matching methods result in every treatment member receiving a match (Austin, 2013).

Nearest Neighbor Matching With Caliper. Calipers can be specified when using greedy matching methods to restrict the maximum distance between a matched treatment and comparison pair (Austin, 2011, 2013; Rosenbaum & Rubin, 1985). By employing a caliper method with propensity score matching, the researcher defines the maximum possible difference on the propensity score between a matched treatment and comparison pair (Austin, 2011; Caliendo & Kopeinig, 2008; Cochran & Rubin, 1973; Rosenbaum & Rubin, 1985; Stuart & Rubin, 2008). If there is no comparison group member with a propensity score within the specified caliper of the treatment member, no match is selected and the treatment member will not be included in the matched sample. Common calipers are 0.20 standard deviations of the logit of the propensity score or 0.02 or 0.03 standard deviations of the propensity score (Austin, 2009).

Compared to non-caliper methods, caliper matching can improve the balance on the covariates between the treatment and comparison group, but often comes at a cost to treatment group (and matched sample) size (Austin, 2009; Austin, 2011; Austin, 2013; Caliendo & Kopeinig, 2008; Jacovidis, 2017; Jacovidis et al., 2017; Stuart, 2010). Loss of treatment group members can be problematic for two reasons. First, there is a loss of treatment representation. That is, depending on the treatment and comparison group propensity score distributions, the loss of treatment members may result in decreased variance and restriction of range of the propensity score distribution (Caliendo &

Kopeinig, 2008). Second, loss of treatment members will decrease the matched sample size (Austin, 2009, 2011; Jacovidis et al., 2007). Although the original, unmatched sample size may have been adequate for the outcome analysis, loss of treatment members may lead to the matched sample no longer being of adequate size for the outcome analysis. Even if the matched sample size is sufficient for the outcome analysis, the results may be impacted by loss of power (Stuart, 2010). When using caliper matching methods, researchers need to be cognizant of the benefits and drawbacks of the method and use their best judgement as to whether the loss of treatment sample is concerning or acceptable (Jacovidis et al., 2007). If very few treatment group members are lost from the matched sample or there is a high degree of common support, the researcher may not have reason to be concerned about whether the estimated treatment effect is generalizable to the population.

Additional Matching Considerations. When selecting the matching method, the researcher must also consider two additional matching specifications: matching with or without replacement and one-to-one or many-to-one matching (Austin, 2011; Caliendo & Kopeinig, 2008). Matching is typically implemented without replacement (Austin, 2009, 2013). When using a greedy algorithm, such as nearest neighbor, once a comparison group member is matched to a treatment group member, the comparison group member is removed from the comparison pool and cannot be matched with any other treatment group members (even if the comparison group member is a good match for a subsequent treatment group member). Matching without replacement ensures that each matched individual is included in the matched sample only once (Austin, 2011, 2013; Caliendo & Kopeinig, 2008). When matching with replacement, after a comparison group member is

matched to a treatment group member, the comparison group member is returned to the comparison pool and can be matched with other treatment group members. Matching with replacement can result in better quality of matches than matching without replacement (because one comparison group member may be a good match for multiple treatment group members). However, matching with replacement can result in the same comparison group member being included in the matched sample multiple times (Austin, 2011, 2013; Caliendo & Kopeinig, 2008). When matching with replacement is used, an outcome analysis that can account for the lack of independence within the comparison group must be used (Stuart, 2010). Moreover, matching with replacement resulted in a greater reduction in bias than other matching methods, however variance in the treatment effect and mean squared error were larger for matching with replacement than with other methods (Austin, 2013; Caliendo & Kopeinig, 2008).

One-to-one matching is used more frequently than many-to-one matching within the PSM literature (Austin, 2009, 2011; Stuart & Rubin, 2008). With one-to-one matching, matched pairs consist of one treatment group member and one comparison group member. This will result in equal group sizes after matching, which is preferred for many outcome analyses. Many-to-one matching (ratio matching) allows multiple comparison group members to be matched to one treatment group member (Stuart, 2010; Stuart & Rubin, 2008). The researcher can specify the number of comparison group members to be matched with one treatment group member (e.g., 2 to 1) or the number of comparison group members matched to one treatment group member can be allowed to vary (Austin, 2011). Many-to-one matching might be preferred when the comparison group pool is much larger than the treatment group (Stuart, 2010; Stuart & Rubin, 2008).

In this situation, there may be multiple comparison group members who are a good match for each treatment group member. However, using many-to-one matching can increase bias in the estimated treatment effect (Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008).

After matching method decisions are made and the matching methods are implemented, the quality of matches are examined. Doing so allows the researcher to determine which matching method is preferred for the study and whether the group balance after matching is adequate. Evaluation of matching quality is conducted using multiple criteria. Caliendo and Kopeinig (2008) separated this evaluation into two steps: evaluation of common support and evaluation of matching quality.

Step 3: Evaluation of Common Support. Common support refers to the extent to which the distribution of the propensity scores for the treatment group overlaps with the distribution of the propensity scores for the comparison group (Austin, 2011; Caliendo & Kopeinig, 2008; Ho et al., 2007; Stuart, 2010; Stuart & Rubin, 2008). A high degree of common support indicates that the propensity score distributions of the two groups are similar, or at least overlapping, and that the quality of matches may be favorable. A low degree of common support indicates that the propensity score distribution of the two groups differ greatly, and that the quality of matches may not be adequate (Austin, 2011; Caliendo & Kopeinig, 2008).

Common support is evaluated in two ways: visually checking the amount of overlap in the propensity score distributions and comparing the minimum and maximum propensity score across groups (Caliendo & Kopeinig, 2008; Ho et al., 2007; Stuart, 2010). Visual checks of common support are often done by examining jitter plots. Each

group member's propensity score is plotted, allowing the researcher to compare the propensity score distributions across groups (see Figure 1). Common support can also be examined by comparing the minimum and maximum propensity score values across groups (Caliendo & Kopeinig, 2008; Stuart, 2010). Any comparison group members with a propensity score that falls outside of the minimum or maximum in the treatment group can be removed, as these comparison group members are unlike the treatment group members in terms of their propensity for treatment selection (Caliendo & Kopeinig, 2008). Differences in the minimum and maximum values of the groups' propensity score distributions are not the only threat to common support. There could also be a range of propensity scores in which only treatment group members fall (and no comparison group members). Thus, lack of common support may result in a treatment group member being matched to a comparison group member who is qualitatively different on the covariates (depending on the matching method that is used; Caliendo & Kopeinig, 2008).

A jitter plot can also be produced after matching has been implemented, with the propensity scores plotted for each of four categories: unmatched comparison group members, matched comparison group members, matched treatment group members, and unmatched treatment group members (see Figure 1). Evaluation of common support sets the stage for examination of matching quality. If the matched treatment and comparison group members have a similar distribution of propensity scores and there are few unmatched treatment group members, matching quality may be favorable. If the matched treatment and comparison group members do not have similar distributions of propensity scores, or if there are many unmatched treatment group members, matching quality may not be ideal or adequate (Austin, 2011; Stuart, 2010). Assessing common support after

matching can also be useful for understanding lack of propensity score and covariate balance after matching. For example, if a researcher found that a large proportion of treatment group members were not matched using a caliper matching method, the jitter plot would be a means of diagnosing whether there were comparison group members with similar propensity scores (e.g., Ho et al., 2007).

The evaluation of common support allows the researcher to understand the propensity score distributions of the treatment and comparison groups (both before and after matching) and a researcher may be able to anticipate the matching quality they will observe based on the similarities or differences in the propensity score distributions. Direct evaluation of matching quality can provide a picture of how well each matching method worked, and which matching method worked the best for the study.

Step 4: Evaluation of Matching Quality. After creating matched groups and evaluating common support, the next step is to evaluate match quality for each implemented matching method. After matching quality is deemed to be adequate for a method, the treatment effect can be estimated (step five). If matching quality is not acceptable, the researcher must consider reasons why (e.g., an important covariate was excluded, the model for estimating the propensity scores was wrong, there was little common support, etc.) and either address those reasons (if possible) or conclude that the employed matching methods did not work well for the data. Estimation of the treatment effect from the poorly matched sample would be inappropriate.

Group Balance on the Covariates. Matching quality is evaluated by comparing the group balance on the covariates and propensity score before and after matching (Austin, 2011; Caliendo & Kopeinig, 2008). Group balance on each covariate and the

propensity score can be examined in multiple ways: examining raw and standardized mean differences on each covariate and the propensity score, examining the average standardized mean difference across all covariates, examining the ratio of the groups' propensity score variances, and visually examining the distributions of each covariate and the propensity score in each group (Austin, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010).

One of the easiest and most straightforward methods for examining matching quality is by comparing raw mean differences on each covariate before and after matching (Austin, 2011). If there were no systematic differences between the groups, the raw means should be equal. Any difference between the means needs to be considered in terms of the scale of the variable. When comparing raw means, it can be difficult to discern how large a difference indicates lack of balance. Similarly, the lack of balance cannot be compared across variables that are on different metrics. Therefore, standardized mean differences are often compared (Austin, 2011; Stuart, 2010).

The standardized mean difference is computed for each continuous covariate by dividing the mean difference by either the standard deviation of the covariate in the treatment group, comparison group, or pooled across groups. Different researchers or matching programs use different standardizers. The standardized mean difference is a standardized effect size, allowing for the covariates to be compared in terms of any remaining lack of balance. There are different views regarding what value of standardized mean difference constitutes adequate balance. Some consider a standardized mean difference less than |0.25| (Rubin, 2001), |0.10| (Austin, 2009, 2011, 2013), or |0.05|

(What Works Clearinghouse, 2017) as an indication of adequate balance on a covariate or the propensity score.

Variance Ratio. A ratio of the treatment group propensity score variance to the comparison group propensity score variance can provide information regarding the quality of balance for each method (Rubin, 2001; Stuart, 2010). A variance ratio close to 1 indicates that the propensity score variances in each group are similar (Stuart & Rubin, 2008). Equal propensity score variances in the treatment and comparison group (in tandem with adequate group covariate balance) indicate that the distributions of the propensity score overlap to a large degree, providing evidence in favor of common support (Rubin, 2001).

Group Covariate Distributions. Visually examining the group distributions of each covariate can reveal areas of imbalance that might not have been revealed using other methods of examining balance (Austin, 2011). Balancing on the covariates is intended to remove all systematic differences between the groups on those covariates. When there are still systematic differences between the groups after matching, the estimated treatment effect may still be biased due to treatment selection.

Evaluation of matching quality may reveal that all auditioned matching methods worked equally well (e.g., Gu & Rosenbaum, 1993). When this occurs, the choice of matching method to champion does not matter (Caliendo & Kopeinig, 2008). If the standardized mean difference across all covariates after matching is similar across all methods, then the variance ratio of the propensity scores, or amount of treatment sample loss (with caliper methods) could potentially guide the selection of which matching method to champion.

Evaluation of matching quality may reveal that some matching methods worked better than others (Caliendo & Kopeinig, 2008; Stuart & Rubin, 2008). When this occurs, the choice of which matching method to champion may be more easily made than when the results from all methods are similar (Caliendo & Kopeinig, 2008).

Evaluation of matching quality may reveal that none of the auditioned matching methods work well. When this occurs, additional evaluation of the included covariates or propensity score model is necessary (Bai, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010). If adequate balance is not achieved (i.e., groups are still unbalanced on the covariates), it is not appropriate to estimate the treatment effect because selection bias will still be present (Ho et al., 2007). If the groups are not balanced after matching, then the goal of matching was not achieved, and the researcher should specify a different propensity score model (Ho et al., 2007; Steiner et al., 2010; Stuart, 2010).

Step 5: Estimation of the Treatment Effect. When matching quality is good, a researcher should select the matching method that results in the best group balance on the covariates. After doing so, the researcher is able to use outcome scores from the resulting matched sample to estimate the treatment effect (Bai, 2011; Stuart & Rubin, 2008). If group means on the covariates are equal after matching, all of the selection bias will be removed from the treatment effect for the covariates that were included in the model (Rubin, 1973a; 1973b). In contrast, when matching does not work well (i.e., groups are still unbalanced on the covariates), little to no selection bias will be removed from the estimated treatment effect, or the direction of the bias may shift (Rubin, 1973a; 1973b).

When propensity score matching is conducted, there are three commonly estimated treatment effects that researchers can choose from depending on the population

to which they wish to make inferences. The choice between estimating the average treatment effect (ATE), average treatment effect for the treated (ATT), or average treatment effect for the control (ATC) is determined by which is most appropriate to answer the research question (Austin, 2011; Ho et al., 2007; Pan & Bai, 2015; Stuart, 2010). The treatment effect that is estimated (i.e., ATE, ATT, or ATC) can be specified via dummy coding of the groups, application of weights, or through research design and randomization.

Average Treatment Effect. The average treatment effect (ATE) provides an estimated treatment effect that would be observed if every individual in the population both did and did not receive treatment (Austin, 2011; Ho et al., 2007; Pan & Bai, 2015). The ATE is the treatment effect that is estimated when attempting to understand the counterfactual in a *randomized control trial* (Rosenbaum & Rubin, 1983). Thus, estimation of the ATE allows for the generalization of the treatment effect to the entire population.

Average Treatment Effect for the Treated. The average treatment effect for the treated (ATT) provides an estimated treatment effect only for those who received treatment (Austin, 2011; Ho et al., 2007; Pan & Bai, 2015). The ATT is the treatment effect that is estimated when attempting to understand the counterfactual *for those who received treatment*. Thus, estimation of the ATT allows for the generalization of the treatment effect to the treatment population. The ATT is frequently the treatment effect of

interest in quasi-experimental studies, such as propensity score matching (Austin, 2011; Cochran & Rubin, 1973; Ho et al., 2007; Rubin, 1973a, 1973b; Stuart, 2010).

Average Treatment Effect for the Control. The average treatment effect for the control (ATC) provides an estimated treatment effect only for those who did not receive treatment (comparison group members; Guo & Fraser, 2015; Pan & Bai, 2015). The ATC is the treatment effect that is estimated when attempting to understand the counterfactual *for those who did not receive treatment*. Thus, estimation of the ATC allows for the generalization of the treatment effect to the comparison population. The ATC is rarely the treatment effect of interest in quasi-experimental studies (Pan & Bai, 2015). The ATC may be of interest when there are fewer comparison group members than treatment group members; however, studies of this nature are infrequent (Pan & Bai, 2015).

Once matched treatment and comparison groups are created, the desired treatment effect can be estimated using whatever statistical analysis is most appropriate to answer the research question (Austin, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008). Estimation of treatment effects is conducted in the same way as randomized control studies or other observational studies where propensity score matching methods are not used (Austin, 2011). The only difference is that the treatment effect is estimated on the matched sample rather than the original sample. When there are two groups (e.g., treatment group, comparison group), the estimated treatment effect is often a comparison of group mean differences on the outcome of interest (Austin, 2011).

One consideration is whether the matched sample is an independent or dependent sample; the answer to which guides the selection of either an independent or dependent samples analysis. The matched sample may be considered a dependent sample because

the groups were created by matching on the covariates. Because the covariates relate to the outcome, it stands to reason that groups that are more similar on the covariates will also be more similar on the outcome (Austin, 2011). When a matched sample is not created (e.g., treatment assignment is randomized), the groups should be independent of one another. In contrast, the matched sample may be considered an independent sample because there should still be random differences between the distributions of the covariates in both groups. Thus, the outcomes of two matched individuals should not be related simply because the individuals are similar on the covariates (Schafer & Kang, 2008). Consideration must be given to whether the matched sample consists of independent or dependent groups to guide the selection of an appropriate outcome analysis model.

Many applications of matching methods in applied studies stop at this step (estimation of the treatment effect; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008). That is, the researcher estimates the treatment effect according to the research question that is being examined, then reports the results and moves on to a discussion of the results. However, there is an additional step that provides validity evidence for the claims regarding the estimated treatment effect. Sensitivity analysis is used to determine the range of possible conclusions that could be made from a quasi-experimental study (Rosenbaum & Rubin, 1983a). Specifically, sensitivity analysis is focused on how the estimated treatment effect would differ due to potential unmeasured (and thus, not included in the propensity score model) covariates.

Step 6: Evaluation of Sensitivity Analysis. Sensitivity analysis was introduced by Rosenbaum and Rubin (1983a) to determine how the treatment effect estimate is

impacted by potentially unmeasured covariates. Sensitivity analysis allows the researcher to speculate about what the estimated treatment effect would have been if there were any unmeasured covariates that should have been included in the propensity score model. If sensitivity analysis reveals that the estimated treatment effect would not differ with the addition of unmeasured covariates, then there is additional evidence that the matching method reduced the impact of selection bias (Rosenbaum & Rubin, 1983a). Although a researcher can never be certain that all selection bias was removed, favorable sensitivity analysis results strengthen the claims made regarding the outcome of interest (Caliendo & Kopeinig, 2008; Rosenbaum & Rubin, 1983a; Stuart, 2010).

Generalized Boosted Modeling

Generalized Boosted Modeling (GBM) is a propensity score method that does not require the creation of matched treatment and comparison groups. GBM is an iterative method that builds off of classification and regression trees (McCaffrey et al., 2004). The process of predicting the propensity scores is automated and data driven (Burgette et al., 2015). As statistical programming software packages have become more robust over time, GBM has become relatively simple to implement, with many models taking very little time to iterate over thousands of replications (Sinharay, 2016).

Classification and Regression Trees. The goal of classification and regression trees (CART) is to accurately predict a categorical (classification) or continuous (regression) outcome by optimizing the use of predictor variables. Specifically, when a binary variable is the outcome (e.g., treatment assignment), the variables that are most useful for predicting the outcome are selected through a series of splits on the data. Beginning with the full sample, the data are split into two subsets based on one variable.

The split occurs at a single value of the variable and maximizes the subsets' difference on the outcome (McCaffrey et al., 2004). After the first split, each subset is split again on the variable that *again* maximizes the subsets' difference on the outcome. Each split can either use a previously used variable (with the split occurring at a different value) or a variable that was not previously used. The splitting process continues until the maximum tree depth is reached (Burgette et al., 2015; McCaffrey et al., 2004). The maximum tree depth is set by the researcher and specifies the maximum number of interactions or largest polynomial terms that can be included in the model (Burgette et al., 2015; McCaffrey et al., 2004; Sinharay, 2016). For example, a tree depth of 3 will allow for the inclusion of 2-way interactions between variables, 3-way interactions between variables, quadratic effects, and cubic effects in the model predicting treatment assignment. Thus, the larger the specified tree depth, the more complex the model.

The classification tree is built using data from *all participants.* After the tree is built, the predicted treatment assignment (0/1) can be determined for each individual by starting at the top of the tree and following the splits based on the individual's values on the variables included in the tree (Burgette et al., 2015). Thus, a predicted treatment assignment value will be estimated for each participant in the sample.

Boosting. Boosting is a procedure that combines information from multiple CART models to improve prediction of the outcome variable (Burgette et al., 2015). When boosting is added to a CART model, the first tree that is fit is a poorly fitting model (i.e., classification error is only slightly smaller than chance classification). That is, the first model performs marginally better at predicting treatment assignment than if guessing was used to predict treatment assignment. After fitting the first classification

tree, another tree is built on the residuals from the previous tree, with individuals who were misclassified by the previous tree receiving a "boost" or larger weight (Burgette et al., 2015; Guo & Fraser, 2015; McCaffrey et al., 2004). By boosting misclassified individuals, there is a greater chance that misclassified individuals will be correctly classified by the next tree (Sinharay, 2016).

Trees are successively built, boosting on the misclassified observations from the previous tree until the algorithm reaches the optimal iteration (Guo & Fraser, 2015; McCaffrey et al., 2004). When generalized boosted modeling is used to estimate propensity scores, the optimal iteration is the one that results in the best covariate balance between the treatment and comparison group. Two different stopping rules can be used to determine which iteration results in the best covariate balance; the effect size stopping rule or the Kolmogorov-Smirnov stopping rule (Burgette et al., 2015; McCaffrey et al., 2004). The effect size stopping rule selects the best iteration as the one that minimizes the standardized group differences across covariates. The Kolmogorov-Smirnov stopping rule selects the best iteration as the one that maximizes the Kolmogorov-Smirnov statistic. Propensity scores are estimated for all participants using the boosted iteration that is determined to be the optimal iteration (based on the stopping rule; Burgette et al., 2015; McCaffrey et al., 2004).

In summary, when generalized boosted modeling is used to estimate propensity scores a series of regression trees are fit to the data. The residuals from each tree inform the successive tree, with misclassified individuals receiving additional weight. Trees are fit until the best covariate balance is reached between the treatment and comparison

groups. The iteration that results in the best covariate balance between groups is used to estimate the log odds of receiving treatment for each participant.

Treatment Effect Estimate Weighting. Instead of matching on the propensity scores, in generalized boosted modeling, the propensity score is used to weight each participants' outcome score in the estimation of the treatment effect. The way in which the propensity score is used to weight a participant's outcome score depends on whether the ATE or ATT is estimated. When the ATE is estimated, each comparison group member is weighted by their propensity score and each treatment group member is weighted by one divided by their propensity score (Ridgeway et al., 2015). When the ATT is estimated, each comparison group member is weighted by their propensity score divided by one minus their propensity score (their odds for treatment group assignment) and each treatment group member receives a weight of 1 (Ridgeway et al., 2015). When ATC coding is used, each treatment group member is weighted by their propensity score divided by one minus their propensity score (their odds for comparison group assignment) and each comparison group member receives a weight of 1. Thus, information from all treatment and comparison group members is included in the estimation of the treatment effect (McCaffrey et al., 2004).

Generalized boosted modeling is advantageous over traditional matching methods because generalized boosted models can support a large number of covariates and can model linear, nonlinear, and interaction effects when estimating the propensity score (Guo & Fraser, 2015; McCaffrey et al., 2004). The propensity score model is less prone to model misspecification when generalized boosted modeling is used (compared to traditional regression models) because information from many CART models is

combined to create the optimal propensity score model (McCaffrey et al., 2004). Boosting on the misclassified individuals results in better final estimated propensity scores than those estimated using a single regression model (McCaffrey et al., 2004). Generalized boosted modelling is also advantageous over traditional matching methods because the outcome analysis is conducted on the entire sample rather than on the reduced, matched sample.

The Role of Comparison Group Size in Propensity Score Matching

A common feature of propensity score matching is a comparison pool that is substantially larger than the treatment group sample. A large comparison group pool with common support can ensure adequate matches and reduce or eliminate loss of treatment group members (Pan & Bai, 2015; Rubin, 1979; Stuart, 2010; Stuart & Rubin, 2008). In one of the seminal PSM articles, Rosenbaum and Rubin (1985) stated "In many observational studies, there is a relatively small group of subjects exposed to a treatment and a much larger group of control subjects not exposed." (p. 33). When implementing a treatment (e.g., a drug to treat cancer, rehabilitation services, drug/alcohol cessation programs, etc.) the number of individuals who can receive treatment is often specified during the study design phase because the treatment requires significant resources. Thus, the researcher can at the same time determine how large a comparison pool from which to collect data. However, within educational contexts, for example, the number of students who take a class (i.e., receive treatment or an intervention) and who do not take a class (i.e., do not receive treatment or an intervention) may be outside of the researcher's control.

Specifically, in higher education, assessment practitioners and educational researchers are interested in the extent to which students achieve student learning outcomes. General education learning outcomes are of interest to multiple stakeholders (e.g., students, university administration, accrediting bodies). Often researchers are interested in comparing student knowledge in general education subject areas across different course completion conditions. In these situations, the researcher is unable to assign students to specific treatment conditions. In many instances, it would be unethical to assign certain students to complete certain courses. Thus, researchers may find themselves in the situation where they have a larger sample of individuals who received a treatment than the sample of those who did not. One might wonder, "Why would the researcher consider using propensity score methods in this scenario?" When there are systematic group differences between students who have and have not completed certain courses (which, there likely are), then statistical analyses that were designed for use with randomly assigned groups would not be appropriate. One seemingly obvious solution is to simply increase the comparison pool size by administering the assessment to students who have not completed certain courses. However, this may not always be feasible within higher education.

Lacking from the propensity score methods literature is empirically supported advice for whether propensity score methods should be used when the treatment group is larger than the comparison group. Some warnings against using propensity score methods when the treatment group is larger than the comparison group can be found. Stuart (2010) stated "If estimating the ATT and there are not (or not many) more control than treated individuals, appropriate choices are generally subclassification, full matching, and

weighting by the odds" (p. 19). Ho et al. (2007) provided two options when the treatment group is larger than the comparison group: match with replacement or recode the treatment and comparison groups.

Recoding of the groups (e.g., treatment group = 1 versus treatment group = 0, and vice versa for the comparison group) changes the effect that is estimated. When the treatment group is coded 1, the ATT is the estimated treatment effect of interest. In contrast, when recoded (treatment group = 0 and comparison group = 1), the actual estimated treatment effect will be the ATC (i.e., the average treatment effect for the comparison group; Ho et al., 2007). When the coding of the groups is switched from reflecting ATT coding to reflecting ATC coding, propensity score matching will result in a treatment group that is similar to the comparison group on the covariates. Thus, the ATC is the treatment effect for those who did *not* receive treatment relative to those who *did* receive treatment, which may not be of interest to educational researchers. Said another way, estimation of the ATC answers the question "what is the effect of receiving the treatment for those who did not receive the treatment?"

In an applied study using empirical data, Perkins and Horst (2020) compared results using ATT and ATC coding across nearest neighbor matching (with and without a 0.20 *SD* caliper) and generalized boosted modeling when the treatment group was larger than the comparison group. For each method, there were differences in the magnitude and direction of the estimated treatment effect between ATT and ATC coding. Specifically, across all methods, the magnitude of the estimated treatment effect was larger for ATT coding than for ATC coding. For nearest neighbor matching, the estimated treatment effect was negative for ATT coding (indicating that the mean outcome score of the

treatment group was less than the mean outcome score of the comparison group) and positive for ATC coding. For nearest neighbor matching with a caliper and generalized boosted modeling, the estimated treatment effect was positive for ATT coding (indicating that the mean outcome score of the comparison group was less than the mean outcome score of the treatment group) and negative for ATC coding. Moreover, there was substantial treatment group member loss with nearest neighbor matching with a 0.20 caliper and ATT coding (Perkins & Horst, 2020).

Using a simulation study, Holzman and Horst (2019) examined the ATT with varying treatment to comparison group ratios for nearest neighbor matching and nearest neighbor matching with a 0.20 caliper. When the treatment group was larger than the comparison group, there was weaker common support, greater loss of treatment members, and poorer estimation of the treatment effect than when the comparison group was larger than the treatment effect (Holzman & Horst, 2019).

The lack of supported guidance regarding how to conduct propensity score matching when the treatment group is larger than the comparison group is a gap in the quasi-experimental literature that must be filled. An understanding of whether propensity score methods result in accurate treatment effect estimates and inferences when the treatment group is larger than the comparison group is needed.

Purpose of the Current Study

As noted, there is limited research and guidance regarding what quasi-experimental methods work best for reducing selection bias when the treatment group is larger than the comparison group. Therefore, the aim of this study is to examine how quasi-experimental methods perform when the treatment group is larger than the

comparison group. Specifically, a simulation study was conducted because the population treatment effect can be defined, and the ability of each method to recover the population treatment effect can be examined. In order to provide recommendations regarding the use of propensity score methods when the treatment group is larger than the comparison group, three research questions were examined.

1a. When the treatment group is larger than the comparison group, are there differences in the bias in the estimated treatment effect for different coding methods (e.g., ATT, ATC), different treatment to comparison ratios (e.g., 2:1, 4:3, 1:4), different treatment sample sizes (e.g., 200, 600, 1,000), and different treatment effect sizes (e.g., Cohen's *d* of 0, 0.20, 0.50, 0.80) across propensity score methods (e.g., nearest neighbor matching, nearest neighbor matching with a 0.20 *SD* caliper, generalized boosted modeling)?

1b. When the treatment group is larger than the comparison group, are there differences in the estimated treatment effect inferences for different coding methods (e.g., ATT, ATC) across propensity score methods (e.g., nearest neighbor matching, nearest neighbor matching with a 0.20 *SD* caliper, generalized boosted modeling)?

2. When the treatment group is larger than the comparison group, are there differences in the covariate balance after matching or weighting for different coding methods (e.g., ATT, ATC), different treatment to comparison ratios (e.g., 2:1, 4:3, 1:4), different treatment sample sizes (e.g., 200, 600, 1,000), and different levels of initial covariate balance (e.g., *SMD* of 0, 0.20, 0.50, 0.80, 1.20) across propensity score methods (e.g., nearest neighbor matching,

nearest neighbor matching with a 0.20 *SD* caliper, generalized boosted modeling)?

3. When the treatment group is larger than the comparison group, are there differences in the loss of treatment group members for different coding methods (e.g., ATT, ATC), different treatment to comparison ratios (e.g., 2:1, 4:3, 1:4), and different treatment sample sizes (e.g., 200, 600, 1,000) for nearest neighbor matching with a 0.20 *SD* caliper? Do conditions in which there is loss of treatment group members also show greater bias in the estimated treatment effect?

CHAPTER 3

Method

The goal of the current study was to examine how well propensity score methods reduce systematic differences between groups when the treatment group is larger than the comparison group. Across various conditions, I evaluated bias in the estimated treatment effect, balance obtained after matching/weighting, treatment sample loss, and whether inferences regarding the treatment effect differed. The following conditions were varied: coding of the treatment and comparison group (i.e., ATT vs. ATC coding), magnitude of the treatment effect (i.e., no effect to large effect), treatment to comparison group ratio (i.e., 1:4, 2:1, and 4:3), and treatment group sample size (i.e., 200, 600, and 1000; see Table 1).

Conditions

Treatment Group Sample Size

Treatment group sample size varied across three levels: 200, 600, and 1,000. Treatment group sample size was manipulated to mimic realistic treatment group sample sizes that might be observed in educational research (e.g., Fan & Nowell, 2011; Jacovidis et al., 2017; Powell et al., 2020). Propensity score methods have been applied to varying-sized treatment groups in educational contexts (e.g., Fan & Nowell, 2011; Jacovidis et al., 2017; Perkins & Horst, 2020; Powell et al., 2020; Stone & Tang, 2013). The large treatment group sample sizes (i.e., 1,000) were included to mimic the situation where an educational researcher combines data collected from multiple cohorts (e.g., Perkins & Horst, 2020).

Treatment to Comparison Group Ratio

When the comparison group is larger than the treatment group, the overarching recommendation is that the comparison group be much larger than the treatment group (Bai & Clark, 2019; Rubin, 1979). By having a large ratio of comparison group members from which to select matches for treatment group members, there is a better chance of finding good matches for each treatment group member than otherwise. A commonly suggested treatment to comparison ratio is 1:4 (Bai & Clark, 2019; Rubin, 1979). There is no guidance regarding treatment to comparison group ratio when the treatment group is larger than the comparison group. However, the 2:1 treatment to comparison group ratio was examined in one simulation study (Holzman & Horst, 2019). The lack of guidance may be due to concerns about only capturing treatment effects for the overlap between the treatment and comparison groups or not capturing heterogeneous treatment effects within the range of overlap between the treatment and comparison groups. That is, if only a sample of treatment group members are examined, there may be a loss of information regarding the treatment effect.

When matching with a larger comparison group than treatment group, the maximum number of matched pairs possible for any method is the size of the treatment group (unless using one-to-many matching). For example, if there were 400 comparison group members and 100 treatment group members, then the maximum number of matched pairs would be 100 (i.e., the size of the treatment group). Conversely, when the treatment group is larger than the comparison group, the maximum number of matched pairs possible is the size of the comparison group (unless matching with replacement). For example, if there were 400 treatment group members and 100 comparison group

members, then the maximum number of matched pairs would again be 100 (i.e., the size of the comparison group). Thus, when the treatment group is larger than the comparison group and matching is done without replacement, there will always be a loss of treatment group members. In sum, with one-to-one matching without replacement, the number of matched pairs can only be as many as the smallest group sample size.

If the typically recommended treatment to comparison group ratio (i.e., 1:4) were reversed (i.e., a treatment to comparison ratio of 4:1), a maximum of 25% of the treatment group will be retained in the matched sample. Thus, when the treatment group is larger than the comparison group, there will always be a loss of treatment group members. Treatment to comparison group ratios of 2:1 and 4:3, allow for a maximum of 50% or 75% of treatment group members to be retained in the matched sample, respectively. In the current study, the treatment to comparison group ratio varied across three levels: 2:1, 4:3, and 1:4. Treatment sample size was fully crossed with treatment to comparison group ratio, which resulted in nine configurations (see Table 2). Total sample size ranged from 300 to 5,000 depending on the treatment group sample size and treatment to comparison group ratio.

Treatment Effect Size

Treatment effect size in the population varied across four levels: Cohen's *d* of 0, 0.20, 0.50, and 0.80. The selected levels align with no, small, medium, and large effects, respectively (Cohen, 1988). Although very large effect sizes (i.e., Cohen's *d* greater than 1) were not included as a condition, small to moderate effect sizes have been shown to be typical within educational research (Cheung & Slavin, 2016; Hill et al., 2008; Kraft, 2020).

Group Dummy Coding

Coding varied based on the recommendation to switch the coding of the treatment and comparison group when the treatment group is larger than the comparison group (Ho et al., 2007), which has also been used empirically (e.g., Perkins & Horst, 2020). Coding of the treatment and comparison groups varied across two levels: the first coding was treatment group coded 1 and comparison group coded 0, and the second coding was treatment group coded 0 and comparison group coded 1. When the treatment group is coded as 1, the ATT reflects the average treatment effect of those who received treatment because comparison group members are selected if they resemble treatment group members on the covariates. This results in a comparison group sample that is similar to the treatment group sample on the covariates. Calling back to the counterfactual, the ATT reflects whether comparison group members with propensity scores similar to treatment group members score the same on the outcome.

In contrast, when the comparison group is coded as 1, the comparison group is essentially being treated how we typically think of the treatment group, and vice versa. Propensity scores are now the probability of comparison group assignment, conditional upon the covariates in the model. As a result of coding the treatment group 0, there will be a larger pool from which to select matches for each comparison group member because the original treatment group is now considered to be the comparison group. Accordingly, the coding will result in estimation of the ATC, reflecting the average treatment effect of those who did not receive treatment. Thus, treatment group members are selected if they resemble comparison group members on the covariates, resulting in a treatment group sample that is similar to the comparison group on the covariates. Calling

back to the counterfactual, the ATC reflects whether treatment group members with propensity scores similar to comparison group members score the same on the outcome.

Simulation of Data

Data were simulated and analyzed using RStudio version 3.6.1 (RStudio Team, 2018). Using the mvtnorm package (Genz et al., 2019), data were simulated for nine configurations of treatment sample size and treatment to comparison group ratio (see Table 2 and Appendix) across 1,000 replications. Total simulees for which data were generated ranged from 300 (Configuration A) to 5,000 (Configuration I; see Table 2). Data for each configuration were simulated via four steps, following the simulation process used by Harris (2018),

1. Generate values for five continuous covariates from a random multivariate normal distribution with means of zero and standard deviations of one.
2. Calculate true propensity scores for each simulee from the values of the covariates. True propensity scores were first calculated on the probit metric, then centered according to the treatment to comparison group ratio, and finally converted to the probability metric.
3. Assign simulees to the treatment or comparison group based on whether the value of a random draw between zero and one was greater than or less than their true propensity score.
4. Simulate outcome scores based on a linear combination of treatment assignment and the covariates (plus a random error term). Four outcome scores were generated for each sample in order to vary the magnitude of the true treatment effect (Cohen's *d* of 0, 0.20, 0.50, and 0.80).

First, values for X1-X5 (the continuous covariates) were generated from a multivariate normal distribution (M = 0, SD = 1). The correlations between the five continuous covariates ranged from 0.10 to 0.65 (see Table 3) to reflect the relations typically seen between variables in educational psychology (Osbourne, 2002). The covariates were also differentially related to treatment assignment to reflect different levels of baseline balance between the groups on the covariates. The following correlations were specified between the propensity score and each covariate (r_{X1} = -0.02, r_{X2} = 0.15, r_{X3} = 0.40, r_{X4} = 0.70, r_{X5} = 0.90) in order to set specific group balance on the covariates prior to matching or weighting (SMD_{X1} = 0, SMD_{X2} = 0.20, SMD_{X3} = 0.50, SMD_{X4} = 0.80, SMD_{X5} = 1.20), respectively.

Second, true propensity scores were calculated for each simulee using matrix algebra. A vector of weights (***B***) was calculated via:

$$\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{Y}}_{group} \tag{2}$$

where $\boldsymbol{X}$ is a matrix of correlations between the covariates, $\boldsymbol{X}'\widehat{\boldsymbol{Y}}_{group}$ is a vector of the correlations between each covariate and propensity for treatment. The vector of weights (***B***) was then multiplied by the simulated covariate values (X1 through X5) for each simulee, which were summed to produce a predicted Y score for each simulee on the probit metric ($\widehat{\boldsymbol{Y}}_{group}$). To set each specific treatment to comparison group ratio, the predicted Y probit scores were rescaled by subtracting the intercept from each predicted Y probit scores (a linear transformation). The probit intercept was calculated via:

$$\frac{z}{\sqrt{\left(1-\left(\frac{\boldsymbol{B}'\boldsymbol{R}\boldsymbol{B}}{\boldsymbol{B}'\boldsymbol{R}\boldsymbol{B}+1}\right)\right)}} \tag{3}$$

were z is value from a standard, normal distribution corresponding to one minus the proportion of each sample that received treatment (i.e., $1 - 0.667$, $1 - 0.571$, and $1 -$

0.200 for 2:1, 4:3, and 1:4 treatment to comparison group ratios, respectively) and $\frac{\boldsymbol{B'RB}}{\boldsymbol{B'RB}+1}$ is the variance explained in propensity for treatment by the covariates (where $\boldsymbol{R}$ is the matrix of correlations between the covariates). After rescaling, the propensity scores were transformed from the probit metric to the probability metric. For each simulee, the proportion of scores that fell at or below the rescaled value of $\boldsymbol{Y'}_{group}$ on a normal curve with $variance = \boldsymbol{B'RB} + 1$ indicated the true propensity for treatment on the probability metric (0 to 1).

Third, each simulee was assigned a random value from 0 to 1 from a uniform distribution. To assign treatment membership to each simulee, their random value was compared to their true propensity score. If the random value was less than or equal to the true propensity score, the simulee was assigned to the comparison group (group = 0). Conversely, if the random value was greater than the true propensity score, the simulee was assigned to the treatment group (group = 1). Due to the rescaling of the propensity scores, treatment membership was assigned for each sample according to the specified treatment to comparison group ratio.

Fourth, in order to vary the magnitude of the treatment effect, values for four outcome variables were generated for each simulee via linear regression:

$$Y_1 = 0(group) + 0.05X1 + 0.05X2 + 0.05X3 + 0.05X4 + 0.05X5 + v \quad (4)$$

$$Y_2 = 0.11(group) + 0.05X1 + 0.05X2 + 0.05X3 + 0.05X4 + 0.05X5 + v \quad (5)$$

$$Y_3 = 0.28(group) + 0.05X1 + 0.05X2 + 0.05X3 + 0.05X4 + 0.05X5 + v \quad (6)$$

$$Y_4 = 0.45(group) + 0.05X1 + 0.05X2 + 0.05X3 + 0.05X4 + 0.05X5 + v \quad (7)$$

where the magnitude of the treatment effect varied across the four outcome variables (Y_1 through Y_4). To introduce random error into each outcome variable, a random error term

(v) was generated for each simulee from a normal distribution with a mean of 0 and standard deviation of 0.5. To specify the magnitude of the treatment effect when there are no systematic group differences on the covariates, the regression weight for the grouping variable was set as follows; $b_1 = 0$, $b_2 = 0.11$, $b_3 = 0.28$, $b_4 = 0.45$. Specifically, if every simulee were the same on the covariates (X1 through X5), the estimated treatment effect (on a Cohen's *d* metric) for the simulated Y scores should be; $d_1 = 0$, $d_2 = 0.20$, $d_3 = 0.50$, $d_4 = 0.80$. The regression weights for each covariate resulted in correlations between the covariates and each outcome variable that ranged from 0.15 to 0.45, which is typical within educational contexts (Osbourne, 2002). Each of the four true treatment effects were simulated to be homogeneous across levels of the propensity score. Thus, the magnitude of the true treatment effect was the same across ATT, ATC, and ATE coding.

Following the simulation of data, propensity score matching (nearest neighbor and nearest neighbor with a 0.20 *SD* caliper) and generalized boosted modeling were conducted for each of the 1,000 replications across 8 combinations of group coding and true treatment effect size (2 coding schemes * 4 true treatment effect sizes). All conditions were fully crossed (see Table 1), resulting in 216 unique combinations of 3 treatment sample sizes, 3 treatment to comparison group ratios, 4 true treatment effect sizes, 3 propensity score methods, and 2 group coding schemes ($3 * 3 * 4 * 3 * 2 = 216$).

Propensity Score Matching

For nearest neighbor matching and nearest neighbor matching with a 0.20 *SD* caliper, propensity scores were estimated using the MatchIt package in R (Ho et al., 2011). Propensity scores were estimated via a logistic regression model with the five covariates (X1, X2, X3, X4, and X5) as predictors of treatment assignment (0, 1). When

the treatment group is larger than the comparison group, the matchit function sorts the treatment group members by the propensity score from largest to smallest by default. However, random ordering of treatment group members by the propensity score resulted in better matches than high to low or low to high ordering (Austin, 2013). Thus, treatment group member ordering was specified to be random prior to matching. Nearest neighbor matching and nearest neighbor matching with a 0.20 *SD* caliper were selected, as these matching methods are frequently employed in the propensity score matching literature (e.g., Austin, 2011, 2013; Caliendo & Kopeinig, 2008; Cochran & Rubin, 1973; Stuart, 2010; Stuart & Rubin, 2008). Additionally, both nearest neighbor and nearest neighbor with a *SD* caliper have been used in the limited simulation (Austin & Cafri, 2020; Holzman & Horst, 2019) and empirical studies (Lechner, 2000; Perkins & Horst, 2020) where the treatment group was larger than the comparison group.

Generalized Boosted Modeling

Generalized boosted modeling was conducted using the Twang package in R (Ridgeway et al., 2020). Propensity scores were estimated via a logistic regression model with the five covariates (X1, X2, X3, X4, and X5) as predictors of treatment assignment (0, 1).

For each replication, the following tuning parameters were selected: 10,000 trees, interaction depth of 3, and shrinkage of .01. These values were selected based on the recommendations by Ridgeway et al. (2020). To ensure that the optimal iteration was not too close to the specified number of trees, 10,000 trees were specified (Ridgeway et al., 2020). The optimal iteration was identified as the iteration that resulted in the smallest mean standardized effect size across the five covariates. After generalized boosted

modeling was performed, the ATT weights were estimated using the Twang package (Ridgeway et al., 2020).

Treatment Effect Estimation

For nearest neighbor and nearest neighbor with a 0.20 *SD* caliper, the treatment effect was estimated from the matched group samples via linear regression:

$$Y_i = b_0 + b_1(group) + e_i \quad (8)$$

where b_1 indicates the treatment effect, b_0 is the comparison group mean on the outcome variable, and e_i is the error term, estimated for each of the four simulated outcome variables ($i = 1\ through\ 4$). For generalized boosted modeling, the ATT weights were applied during treatment effect estimation. The effect size (Cohen's *d*) of the mean difference between the treatment and comparison group was calculated for each replication by dividing the mean difference between groups by the *SD* pooled across both groups.

Criteria for Evaluating Research Questions

A summary of the values that were saved from the simulated data is provided in Table 4. Each research question was evaluated over a different number of combinations of the simulation conditions. Specifically research questions 1, 2, and 3 were evaluated over 216, 270, and 36 combinations of conditions, respectively. After collapsing across replications, bias in the estimated treatment effect, balance after matching or weighting, loss of treatment group members, and the estimated treatment effect inference were examined.

Research Question 1a: When the Treatment Group is Larger Than the Comparison Group, Can Propensity Score Methods Accurately Recover the True Treatment Effect?

To answer this research question, bias in the estimated treatment effect was examined. Bias in the estimated treatment effect is the extent to which the estimated treatment effect differs from the population treatment effect (Feinberg & Rubright, 2016). If the estimated treatment effect is unbiased, across repeated sampling, there should be little to no deviation from the population treatment effect (Feinberg & Rubright, 2016). Values different than zero indicate that the estimated treatment effect is biased, with large values indicating large bias. Bias was calculated as:

$$Bias = \frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_{True})}{n} \quad (9)$$

where the numerator is the sum of the deviation between the estimated treatment effect ($\hat{\theta}$) from the population treatment effect (θ_{True}) for each replication (i), averaged across n replications (Feinberg & Rubright, 2016). The summed deviations were then averaged by dividing the numerator by the number of replications (n). Bias in the estimated treatment effect (on the Cohen's *d* standardized effect size metric) was evaluated for each propensity score method across coding methods, treatment to comparison ratios, treatment sample sizes, and true treatment effect sizes via a 3x2x3x3x4 ANOVA.

Research Question 1b: Does the Magnitude and Direction of the Estimated Treatment Effect Differ Across Propensity Score Methods Depending on Group Coding?

The magnitude and direction of the estimated treatment effect using ATC coding was compared to the magnitude and direction of the estimated treatment effect using ATT coding. If the magnitude and/or direction of the treatment effect differed across conditions, the inference made regarding treatment effectiveness differed. The magnitude

and direction of the estimated treatment effect using different coding methods was examined across treatment to comparison ratios, treatment sample sizes, and true treatment effect sizes.

Research Question 2: When the Treatment Group is Larger Than the Comparison Group, Can Propensity Score Methods Achieve Adequate Group Balance on the Covariates?

Balance after matching or weighting was examined numerically by obtaining the standardized mean difference and percent in bias reduction for each covariate and the propensity score. Examining group balance on the covariates after matching or weighting provides an indication of how well each method works at reducing systematic group differences on the covariates. Additionally, the ratio of treatment group propensity score variance to comparison group propensity score variance was evaluated to examine group balance.

Standardized mean differences are an effect size for the lack of balance between groups, with a zero value indicating no mean difference between the groups on the covariate (Austin, 2011; Caliendo & Kopeinig, 2008; Stuart, 2010; Stuart & Rubin, 2008). Guidelines have been provided for what standardized mean difference value indicates acceptable group balance, ranging from 0.05 (What Works Clearinghouse, 2017) to 0.25 (Rubin, 2001). For this study, the standardized mean differences will be reported, and values $\leq |0.10|$ will indicate that the groups are adequately balanced after matching or weighting. This value was chosen as it is not too stringent or lenient and has been used as a guideline for group balance in other simulation studies (Austin 2009, 2011, 2013).

Percent in bias reduction quantifies the extent to which bias in each covariate is reduced relative to initial balance (Pan & Bai, 2015). A percent in bias reduction of 100 would indicate that group means on the covariate were equal after matching or weighting, and that 100% of the original covariate imbalance was corrected. For this study, the percent in bias reduction was reported, and values ≥ 80% indicated that the groups were adequately balanced after matching or weighting (Pan & Bai, 2015)

The ratio of the groups' propensity score variance provides an indication of the quality of balance for matching methods (Rubin, 2001; Stuart, 2010). Ratios close to 1 indicate that the variance of the propensity score is similar across the treatment and comparison group.

Visual examination of group balance on the propensity score is important for evaluating common support, or the extent to which the group distributions of the propensity score are similar. Typically, when using propensity score methods, balance is evaluated visually using jitter plots and histograms (Austin, 2011; Ho et al., 2007; Stuart, 2010). Examination of jitter plots for each replication of each condition is not feasible, thus, jitter plots were examined for validation datasets. In sum, for research question two, group balance on the covariates and the propensity score after matching or weighting was evaluated via standardized mean differences and percentage in bias reduction for each propensity score method across coding methods, treatment to comparison ratios, treatment sample sizes, and initial covariate balance between the groups.

Research Question 3: When the Treatment Group is Larger Than the Comparison Group, Does the Loss of Treatment Group Members Differ Across Conditions?

Loss of treatment group members was examined for nearest neighbor and nearest neighbor with a caliper across coding methods, treatment to comparison ratios, and treatment sample sizes. Loss of treatment group members was quantified as the percent of treatment group members not retained in the matched sample (for nearest neighbor and nearest neighbor with 0.20 *SD* caliper). For generalized boosted modeling, there is no loss of treatment group members because when estimating the treatment effect, the full treatment and comparison group samples are retained, and each individual receives a weight based on their propensity score. In addition to examination of sample size, the average propensity score for matched and unmatched treatment group members were compared to determine whether the unmatched treatment group members qualitatively differed from the matched treatment group members.

Summary

In summary, each research question was answered using a different criterion. By examining bias in the estimated treatment effect (RQ1a), the estimated treatment effect inference (RQ1b), balance after matching or weighting (RQ2), and loss of treatment group members (RQ3) across the varied conditions, this study will expand the current understanding of whether propensity score methods are appropriate to use when the treatment group is larger than the comparison group.

Validation Data Sets

To evaluate whether the data were simulated correctly, one replication was examined for each configuration (samples A through I). First, the relations among the

covariates, propensity score, and treatment assignment, and initial covariate balance between the groups were examined to evaluate whether the data were correctly simulated. Next, the treatment group sample size and treatment to comparison group ratio were examined to evaluate whether treatment assignment was correctly simulated. Finally, the true treatment effect for each of the four outcome variables was examined to evaluate the magnitude of the treatment effect if groups were not systematically different on the covariates. The validation data sets also allowed for the visual examination of the overlap between the treatment and comparison group propensity score distributions (before matching or weighting) via jitter plots.

Validation of Covariate Values

The standardized group mean differences and correlations between the covariates and latent propensity for treatment were evaluated for samples A through I. The results for each validation sample were compared to Table 3 to evaluate whether the values of the simulated data matched the specified values.

Across all nine validation samples, the standardized group mean differences and correlations between the covariates and latent propensity for treatment were consistent with the values that were specified (see Table 5). Of note, for validation sample D, the correlation between X1 and X2 was inflated. Additionally, the correlations between X1 and treatment selection and X1 and propensity scores were stronger than specified. Not surprisingly, the standardized mean difference for X1 was larger than specified (-0.21 instead of 0). These values were considered a product of sampling variability, yet maintained the patterns specified in the simulation.

Validation of Treatment Assignment and True Treatment Effect

Group sample sizes and treatment to comparison group ratio were examined for samples A through I to evaluate whether treatment assignment was correctly specified. The group sample sizes for each validation sample were compared to Table 2 to determine accuracy. The magnitude of the true treatment effect (e.g., the group difference on the outcome if there were no group imbalance on the covariates) was examined in each sample to evaluate whether the values for the outcome variables were simulated correctly.

Across all nine validation samples, the group sample sizes, the treatment to comparison group ratio, and the true treatment effect were consistent with the specified values (see Table 6). Of note, the treatment group was slightly larger than intended in validation sample A, however the true treatment effect was simulated well. For validation sample G, the true treatment effect was stronger than specified across all four outcome variables. These minor deviations were considered a product of sampling variability, yet maintained the patterns specified in the simulation.

Evaluation of Common Support

Examination of the overlap between the groups' propensity score distributions provided an indication of whether there was common support between the groups on the propensity scores. Across all nine validation samples, there appeared to be adequate common support between the treatment and comparison groups on the propensity scores (see Figure 2). For samples where the treatment group was larger than the comparison group (samples A through F), the propensity score distribution was more dense at higher propensity score values for the treatment group. Conversely, for samples where the

comparison group was larger than the treatment group (samples G through I), the propensity score distribution was denser at lower propensity score values for the comparison group.

In summary, the results from the nine validation samples indicated that the data were simulated as specified.

CHAPTER 4

Results

Evaluation of Simulated Data

Simulated covariate and true propensity score means and standard deviations for each scenario are presented in Table 7. Means and standard deviations for each covariate, collapsed across treatment and comparison groups, were approximately 0 and 1, respectively. Additionally, the small standard errors for the means and standard deviations for each covariate indicated that there was a small amount of variability in these values across the 1,000 replications for each scenario. The mean true propensity score, prior to matching or weighting, collapsed across treatment and comparison groups, matched the proportion of individuals assigned to the treatment group as specified by the treatment to comparison group ratio set for each scenario. For all scenarios, the treatment group had a higher mean propensity score than the comparison group, indicating a higher propensity for treatment.

Average correlations between covariates and the true propensity score for each scenario are presented in Table 8. All values aligned with those in the simulation code, indicating that the relations between the covariates and true propensity score were simulated well across the 1,000 replications for each scenario.

Simulated treatment and comparison group outcome means and standard deviations for each scenario are presented in Table 9. For all four outcome variables (representing different true treatment effects; Cohen's *d* of Y1 = 0, Y2 = 0.20, Y3 = 0.50, and Y4 = 0.80), the treatment group had a higher mean outcome score than the comparison group. As intended, there was a difference in the average outcome between

treatment and comparison groups prior to applying propensity score methods to reduce systematic group differences, allowing for the examination of research question one across different true treatment effects.

Cohen's *d* effect sizes for each average true treatment effect are presented by scenario and coding method in Table 10. First, across all scenarios, the true treatment effect aligned with the values specified in the simulation. Specifically, on average, the simulated true treatment effect sizes approached the specified Cohen's *d* effect sizes of 0, 0.20, 0.50, and 0.80 for Y1, Y2, Y3, and Y4, respectively. As would be expected, for scenarios with smaller sample sizes (i.e., scenarios A, D, and G), the standard errors of the mean true treatment effect were larger than for scenarios with larger sample sizes (i.e., scenarios B, C, E, F, H, and I). Note that for each scenario, each true treatment effect was of the same magnitude for ATT and ATC coding. However, the direction of the effect differed across ATT and ATC coding. That is, if the true treatment effect was positive for ATT coding, the true treatment effect was negative for ATC coding. This indicated that regardless of coding method, the magnitude of the estimated treatment effect was the same, with only a directional difference. The results presented in Tables 8, 9, and 10, together indicated that the data were simulated as specified.

To evaluate convergence of the generalized boosted models, the mean, median, minimum, and maximum optimal iterations are presented in Table 11. Overall, the models reached the optimal iteration with relatively few iterations. For scenarios B through F, in which the treatment group was larger than the comparison group, generalized boosted modeling required more iterations to achieve the optimal iteration for ATT coding than for ATC coding. For scenarios G through I, in which the comparison

group was larger than the treatment group, generalized boosted modeling took more iterations to achieve the optimal iteration for ATC coding than for ATT coding. Thus, when the group coded "1" was larger than the group coded "0" (except for scenario A, which had the smallest total sample size), on average it took more iterations to achieve the optimal iteration.

Evaluation of Research Questions

Given that the data were simulated as intended and that generalized boosted models converged in fewer iterations than the maximum number of specific iterations, the research questions could be evaluated. Results are presented by research question.

Research Question 1

The first research question focused on the ability of each propensity score method to accurately recover the true treatment effect across coding methods, treatment to comparison ratios, treatment sample sizes, and true treatment effect sizes. Bias in the estimated treatment effect was evaluated for each replication for each scenario, then averaged across the 1,000 replications for each scenario. Bias values other than 0 indicated that the true treatment effect was not accurately recovered. Additionally, bias was evaluated via a 3x2x3x3x4 ANOVA. Given the large number of replications, statistical significance tests were overpowered. Thus, more weight was given to practical significance (partial eta squared) to determine whether there were significant effects for the five conditions and all interactions among conditions.

The average Cohen's *d* effect size, average bias, and standard errors are presented for each true treatment effect (Y1 through Y4) in Table 12. First, the baseline values show the amount of bias in the estimated treatment effect for the unmatched or

unweighted sample. Across all scenarios, the amount of bias in each baseline estimated treatment effect was fairly consistent, with the amount of bias increasing slightly as the true treatment effect increased. Looking at the bolded values in Table 12, three noticeable patterns emerge.

First, notice that for scenarios where the treatment group was larger than the comparison group (scenarios A through F), nearest neighbor matching always resulted in a large amount of bias in the estimated treatment effect. For ATT coding, the estimated treatment effect after nearest neighbor matching differed very little from the baseline treatment effect. Although there was a reduction in bias in the estimated treatment effect from baseline after nearest neighbor matching for ATC coding, a non-negligible amount of bias remained. Conversely, when the treatment group was smaller than the comparison group (scenarios G through I), nearest neighbor matching using ATT coding resulted in substantially less bias for Y1, Y2, and Y3 than when the treatment group was larger than the comparison group. For scenarios G through I, nearest neighbor matching using ATC coding (which, because of the 1:4 ratio, resulted in the group that was coded 0 being smaller than the group that was coded as 1) resulted in substantially more bias than for ATT coding. In fact, for scenarios G through I, the bias in the estimated treatment effect for ATC coding was very similar to the baseline bias in the estimated treatment effect. Thus, for nearest neighbor matching, using a coding scheme in which the largest group was coded "1" resulted in very little to no difference in the estimated treatment effect when compared to the baseline treatment effect.

Second, notice that for nearest neighbor matching with a 0.20 *SD* caliper, the amount of bias in estimated treatment effect was small for all scenarios, all true treatment

effects, and all coding methods. Across all scenarios and coding methods, the amount of bias in the estimated treatment effect increased as the true treatment effect increased. That is, although there was little bias in the estimated treatment effect for Y1 and Y2 (true treatment effect sizes of 0 and 0.20, respectively), there was a larger, but still relatively small, amount of bias in the estimated treatment effect for Y3 and Y4 (true treatment effect sizes of 0.50 and 0.80, respectively). These results suggested that nearest neighbor with a 0.20 caliper resulted in a generally unbiased estimate of the treatment effect when the treatment group was larger than the comparison group.

Third, for generalized boosted modeling, the amount of bias in the estimated treatment effect was relatively small for all scenarios and all coding methods for Y1 and Y2. For Y3 and Y4, bias in the estimated treatment effect after generalized boosted modeling was larger than for Y1 and Y2. Additionally for all true treatment effects, bias after generalized boosted modeling was larger than after nearest neighbor matching with a caliper, except for Y2 (scenarios A, D, and H with ATT coding, and B, C, and D with ATC coding) and Y3 (scenarios D with ATC coding, and G with ATT coding). Thus, it appeared that generalized boosted modeling using either ATT or ATC coding resulted in a larger amount of bias in the estimated treatment effect than nearest neighbor matching with a caliper and a substantially smaller amount of bias in the estimated treatment effect than nearest neighbor matching.

Additionally, across all scenarios, when comparing bias across ATT and ATC coding, the direction of the bias differed consistently. That is, if the average bias was positive for ATT coding, it was negative for ATC coding, and vice versa. This pattern was not surprising given that the true treatment effect differed in sign across ATT and

ATC coding. However, the pattern of differences in the magnitude of bias was less clear-cut than the pattern of directional differences and did not solely differ across coding method.

To better understand differences in the amount of bias in the estimated treatment effect across propensity score method, coding method, treatment to comparison ratios, treatment sample sizes, and true treatment effect sizes a 3x2x3x3x4 ANOVA was conducted[1]. Given the large sample size due to the large number of replications for each scenario, statistical significance was reported and effects that were practically significant were interpreted. Meaningful practically significant effects were identified as those that explained greater than or equal to 2% of the variance in bias. That is, partial-eta squared values greater than or equal to .02 indicated a small meaningful effect (Cohen, 1988). ANOVA results are presented in Table 13. Four effects were both statistically significant and accounted for greater than or equal to 2% of the variance in bias: propensity score method ($\eta_p^2 = 0.551$), the interaction between propensity score method and effect size ($\eta_p^2 = 0.187$), the interaction between coding method and treatment to comparison ratio ($\eta_p^2 = 0.024$), and the interaction between propensity score method, coding method, and treatment to comparison ratio ($\eta_p^2 = 0.159$). As is standard practice, main effects were not

---

[1] Given that the direction of bias differed consistently across ATT and ATC coding, the direction of bias observed for ATC coding was reversed prior to conducting the ANOVA. That is, if the ATC bias value was negative, it was made to be positive and if the ATC bias value was positive, it was made to be negative. The difference in sign across ATT and ATC coding accurately reflects that the original treatment group outcome mean was higher than the original comparison group outcome mean. Reversing the sign for only ATC coding retained information regarding the magnitude of bias in the estimated treatment effect, allowing for examination of differences in the *magnitude* of bias in the estimated treatment effect via statistical significance tests and effect sizes.

interpreted for conditions for which there was a statistically significant and meaningful interaction (Cohen, 2013). Therefore, because there was a statistically significant and meaningful three-way interaction between propensity score method, coding method, and treatment to comparison ratio, the main effect for propensity score method and the two-way interaction between coding method and treatment to comparison ratio were not interpreted.

The three-way interaction between propensity score method, coding method, and treatment to comparison ratio accounted for 15.90% of the variance in bias. Mean bias for each combination of propensity score method, coding method, and treatment to comparison group ratio are presented in Table 14. Generalized boosted modeling with ATT coding and a 1:4 ratio, generalized boosted modeling with ATC coding and a 2:1 or 4:3 ratio, and nearest neighbor matching with a caliper (for all coding methods and all ratios) had the lowest mean bias. The observation of low bias in the estimated treatment effect for generalized boosted modeling and nearest neighbor matching with a caliper with ATT coding and a 1:4 ratio were not surprising given that these conditions replicate previous research recommendations (e.g., Rubin, 1979).

The statistically significant and meaningful three-way interaction between propensity score method, coding method, and treatment to comparison ratio indicated that the interaction between coding method and treatment to comparison ratio depended on propensity score method. Said another way, there was a different pattern of interaction between coding method and treatment to comparison ratio for each propensity score method. To better understand this effect, three two-way interaction models between

coding method and treatment to comparison ratio were examined (one for each propensity score method).

The interaction between coding method and treatment to comparison ratio was statistically significant and meaningful for nearest neighbor matching ($F(2, 71994) = 22436.553$, $p < .001$, partial $\eta^2 = .384$) and generalized boosted modeling ($F(2, 71994) = 1912.556$, $p < .001$, partial $\eta^2 = .050$). However, for nearest neighbor matching with a caliper, the interaction between coding method and treatment to comparison ratio was statistically significant but not meaningful ($F(2, 71994) = 13.052$, $p = <.001$, partial $\eta^2 < .001$). The statistically significant and meaningful interactions between coding method and treatment to comparison ratio were further examined for nearest neighbor matching and generalized boosted modeling.

For both nearest neighbor matching and generalized boosted modeling, there were statistically significant and meaningful differences in bias between ATT coding and ATC coding for each treatment to comparison ratio (see Table 14 and Figure 3). For both nearest neighbor matching and generalized boosted modeling with a treatment to comparison ratio of 1:4, bias in the estimated treatment effect was larger in magnitude for ATC coding than for ATT coding. Conversely for both nearest neighbor matching and generalized boosted modeling with a treatment to comparison ratio of 2:1 or 4:3, bias in the estimated treatment effect was larger in magnitude for ATT coding than for ATC coding. Overall, generalized boosted modeling resulted in a lower magnitude of bias across all coding methods and treatment to comparison group ratios than nearest neighbor matching. Additionally, generalized boosted modeling overcorrected bias in the estimated

treatment effect as evidenced by the negative average bias values for all coding methods and treatment to comparison group ratios.

For nearest neighbor matching with a caliper, average bias was consistently low across coding method and treatment to comparison group ratio. Thus, there was no effect for coding method or treatment to comparison group ratio on bias in the estimated treatment effect after nearest neighbor matching with a 0.20 *SD* caliper.

After accounting for the three-way interaction, there was a statistically significant and meaningful two-way interaction between propensity score method and true treatment effect size. The two-way interaction between propensity score method and true treatment effect size accounted for an additional 18.70% of the variance in bias.

Further examination of the two-way interaction between propensity score method and true treatment effect size revealed that for all true treatment effect sizes, there were statistically significant and meaningful differences in bias in the estimated treatment effect across propensity score methods (see Table 15 and Figure 4). The lowest magnitude of mean bias was observed when there was no true treatment effect (i.e., Y1, SMD = 0) or when the true treatment effect was small (i.e., Y2, SMD = 0.20), for both nearest neighbor matching with a caliper and generalized boosted modeling. Nearest neighbor matching with a caliper was the only method that resulted in average bias less than |0.10| for all four true treatment effect sizes. Across all four true treatment effect sizes nearest neighbor matching (with no caliper) resulted in the highest levels of average bias compared to nearest neighbor matching and generalized boosted modeling.

For all true treatment effects, the magnitude of bias in the estimated treatment effect was largest for nearest neighbor matching, followed by generalized boosted

modeling, with the least bias observed for nearest neighbor matching with a caliper. All three propensity score methods resulted in a reduction in bias from that observed at baseline, however bias remained large after nearest neighbor matching (for all true treatment effects) and generalized boosted modeling (for treatment effects of 0.50 and 0.80).

The smallest differences in bias (although still statistically significant and meaningful) were observed between nearest neighbor matching with a caliper and generalized boosted modeling for no true treatment effect and a small true treatment effect. The magnitude of differences in bias between nearest neighbor matching with a caliper and generalized boosted modeling increased for a medium and large true treatment effect. That is, for Y3 and Y4, generalized boosted modeling resulted in a higher average magnitude of bias than nearest neighbor matching with a caliper, and there was a larger difference in bias between the two propensity score methods than was observed for Y1 and Y2.

In summary, bias in the estimated treatment effect did not differ much between generalized boosted modeling and nearest neighbor matching with a caliper when the true treatment effect was 0 or 0.20 (Y1 and Y2, respectively). In these instances, nearest neighbor matching with a caliper performed the best, followed by generalized boosted modeling. There was substantially more bias for generalized boosted modeling when the true treatment effect was 0.50 or 0.80 (Y3 and Y4, respectively). Across all true treatment effect sizes, nearest neighbor matching had the highest bias in the estimated treatment effect.

Research Question 2

The second research question focused on the extent to which group balance was achieved on the covariates across propensity score method, coding method, treatment to comparison ratio, treatment sample size, and initial covariate balance between the groups. Balance was evaluated via the standardized mean difference (values $\leq |0.10|$ indicated adequate balance; Austin, 2009, 2011, 2013), percentage in bias reduction (values $\geq$ 80.00 indicated adequate reduction in bias; Pan & Bai, 2015), and the propensity score variance ratio (values close to 1.00 indicated similar variances; Rubin, 2001).

Average (mean) standardized mean differences and median percentage in bias reduction for each covariate and the estimated propensity scores are presented by scenario in Table 16. The baseline standardized mean differences for each of the five covariates were consistent across all scenarios and coding methods. The baseline standardized mean differences aligned with the values specified of 0, 0.20, 0.50, 0.80, and 1.20 for X1, X2, X3, X4, and X5, respectively. Thus, each covariate had a different level of baseline group balance, which allowed for the examination of balance obtained across different initial levels of baseline covariate balance. Additionally, across coding method (i.e., ATT or ATC), initial covariate balance differed in sign, with small differences in magnitude for all scenarios. Examination of Table 16 revealed three main patterns for the average standardized mean difference.

First, notice that for nearest neighbor matching, the only covariate that was adequately balanced across all coding methods and scenarios was X1. This covariate was balanced in the baseline sample, so nearest neighbor matching did not improve this balance, but also did not result in a larger imbalance. For scenarios A, B, and C with

ATC coding, nearest neighbor matching also resulted in adequate group balance on X2. Conversely, for scenarios G, H, and I with ATT coding, nearest neighbor matching resulted in adequate group balance on X2 and X3. These results indicated that nearest neighbor worked well to create adequate group balance on covariates with small and medium magnitudes of initial lack of balance when the treatment to comparison ratio was 2:1 and ATC coding was used (scenarios A, B, C) or when the treatment to comparison ratio was 1:4 and ATT coding was used (scenarios G, H, I).

Second, nearest neighbor matching with a 0.20 *SD* caliper resulted in the best group balance on all five covariates and the propensity score for both ATT and ATC coding across all nine scenarios. Adequate balance was obtained on all covariates and the propensity score for nearest neighbor matching with a caliper, indicating that when the treatment group was larger than the comparison group, nearest neighbor matching with a caliper worked well at creating treatment and comparison groups that were balanced on the covariates, across five magnitudes of initial imbalance.

Third, notice that generalized boosted modeling resulted in adequate group balance on X1 and X2 (i.e., covariates with the best balance at baseline) across all coding methods and scenarios, and on X3 across all coding methods for scenarios C, E, F, H, and I. Generalized boosted modeling did not result in adequate balance on X4 or X5 (i.e., covariates with the worst balance at baseline) for any coding method or scenario, however balance was improved over baseline. These results indicated that when the treatment group was larger than comparison group, generalized boosted modeling resulted in adequate balance for covariates that had low baseline standardized mean differences. For the covariate with a medium standardized mean difference (X3, SMD =

0.50), generalized boosted modeling resulted in adequate balance in scenarios with a large treatment sample size (600 or 1,000).

To better understand the standardized mean difference across all conditions, the average absolute value of standardized mean difference was graphed for each covariate by propensity score method, coding method, and treatment to comparison ratio (see Figure 5). The average standardized mean difference was low for all covariates and the propensity score after nearest neighbor matching with a caliper. Even the covariates that were unbalanced at baseline (X3, X4, and X5) were well balanced after nearest neighbor matching with a caliper. This suggested that nearest neighbor matching with a caliper resulted in adequate group balance on the covariates regardless of the initial balance on the covariate. Additionally, for nearest neighbor matching with a caliper, there were minimal differences in average standardized mean difference for each covariate across coding methods and treatment to comparison ratio.

After generalized boosted modeling, there was a slightly different pattern. The average standardized mean difference was low for X1 and X2 after weighting, however, the average standardized mean difference was higher as the covariate's baseline standardized mean difference was higher (i.e., X3, X4, and X5). Although the average standardized mean difference was not ideal for X3, X4, and X5, generalized boosted modeling always resulted in a reduction in average standardized mean difference when compared to the baseline standardized mean difference. Additionally, for generalized boosted modeling, there were minimal differences in the average standardized mean difference for each covariate across coding methods and treatment to comparison ratio.

A noticeably different pattern emerged for nearest neighbor matching. When the baseline standardized mean difference was 0 (X1), the average standardized mean difference after nearest neighbor matching was low, and minimally differed across coding methods and treatment to comparison ratio. As the baseline standardized mean difference increased across covariates, the average standardized mean difference increased. That is, covariates with large initial standardized mean differences (large initial group imbalance) had larger standardized mean differences after nearest neighbor matching, relative to covariates with smaller initial standardized mean differences. Additionally, for nearest neighbor matching, there were noticeable differences in the average standardized mean difference for each covariate across coding method and treatment to comparison ratio. For X2, X3, X4, and X5, when the treatment group was larger than the comparison group (ATT coding with a ratio of 2:1 or 4:3 and ATC coding with a ratio of 1:4), the average standardized mean difference after nearest neighbor matching was similar to the baseline average standardized mean difference. For example, note that for X4, the average standardized mean difference after nearest neighbor matching with ATT coding for the treatment to comparison ratio of 4:3 is close to the initial standardized mean difference of 0.80. When coding was reversed (ATC coding with a ratio of 2:1 or 4:3 and ATT coding with a ratio of 1:4), the average standardized mean difference after nearest neighbor matching was smaller than the baseline standardized mean difference. In other words, when the treatment group was larger than the comparison group, balance was improved over that at baseline when nearest neighbor matching was used with ATC coding. In the typical scenario where the comparison group was larger than the treatment group (ratio of 1:4 with ATT coding), balance was improved over baseline when nearest neighbor

matching was used. Although nearest neighbor matching with a 1:4 ratio and ATT coding and nearest neighbor matching with a 2:1 or 4:3 ratio and ATC coding was more balanced than the other nearest neighbor matching conditions, balance was not ideal.

In addition to the standardized mean difference as an indicator of covariate and propensity score group balance, percentage in bias reduction was evaluated. The median percentage in bias reduction values indicated the extent to which balance was improved (or worsened, as indicated by negative values) when compared to the baseline group balance for each covariate and the propensity score (see Table 16). Thus, for covariates that were well balanced initially, there was little to no room for improvement. Although values greater than or equal to 80.00 can be used to indicate adequate reduction in initial imbalance, the percentage in bias reduction value must be evaluated in light of the initial group balance for each covariate (Pan & Bai, 2015).

The majority of the median percentage in bias reduction values were greater than 60.00, indicating that in general, balance improved after matching or weighting compared to baseline balance for initially unbalanced covariates (X2, X3, X4, X5, and the propensity score). When the treatment group was larger than the comparison group (scenarios A through F), nearest neighbor matching with ATT coding resulted in group balance on the covariates that did not differ from baseline covariate balance (i.e., low PBR values, ranging from -13.15 to 0.89). The same pattern was observed for nearest neighbor matching with ATC coding for scenarios G through I (which resulted in the group coded 0 being smaller than the group coded 1). Thus, nearest neighbor matching did not improve group balance when the treatment group was larger than the comparison group. Conversely, when ATC coding was used for scenarios A through F and ATT

coding for scenarios G through I, group balance on the covariates was improved after nearest neighbor matching compared to baseline (i.e., values ranging from 19.16 to 80.84).

The propensity score variance ratios are presented in Table 16. For all scenarios and coding methods, nearest neighbor matching with a 0.20 *SD* caliper resulted in average propensity score variance ratios close to 1.00. Nearest neighbor matching either resulted in no change relative to baseline in the propensity score variance ratio (ATT coding for scenarios A through F, and ATC coding for scenarios G through I), a small improvement relative to baseline in the propensity score variance ratio (ATT coding for scenarios G through I), or a propensity score variance ratio that was farther away from 1.00 than at baseline (ATC coding for scenarios A through F).

Research Question 3

The third research question focused on the loss of treatment group members for nearest neighbor matching and nearest neighbor matching with a caliper across coding methods, treatment to comparison ratios, and treatment sample sizes. Loss of treatment group sample size was quantified as the percentage of treatment group members (i.e., simulees) not retained in the matched sample (unmatched). Additionally, the average propensity score for the matched and unmatched treatment groups were compared to the average propensity score for the baseline treatment group to determine if the matched treatment group was similar to the baseline treatment group.

Group sample sizes for baseline, matched, and unmatched treatment and comparison groups, mean propensity score by group, and treatment loss for each scenario are presented in Table 17. The treatment loss percentages for nearest neighbor matching

aligned with the expected values given the treatment to comparison ratio. That is, when the treatment group was larger than the comparison group, nearest neighbor matching resulted in a loss of treatment sample of approximately 50% for a ratio of 2:1 (scenarios A through C) and 25% for a ratio of 4:3 (scenarios D through F). For the ratio of 1:4, there was no treatment sample loss because there were more comparison group members than treatment group members.

After nearest neighbor matching with a caliper, treatment sample loss was larger than that after nearest neighbor matching. With a ratio of 2:1 (scenarios A through C), treatment sample loss ranged from 66.76% to 67.92%, with no differences across coding method. With a ratio of 4:3 (scenarios D through F), treatment sample loss ranged from 56.78% to 58.05%, with no differences across coding method. With a ratio of 1:4 (scenarios G through H), treatment sample loss ranged from 19.43% to 21.12%, with no differences across coding method.

To better understand whether the matched treatment group had similar average propensity scores as the baseline treatment group, average propensity score was graphed by propensity score method, coding method, and treatment to comparison ratio (see Figure 6). Note that the treatment group's baseline average propensity score was higher than the comparison group's for ATT coding. For ATC coding, this pattern was reversed such that the treatment group's baseline average propensity score was lower than the comparison group's.

When the treatment group was larger than the comparison group (treatment to comparison ratios of 2:1 and 4:3), there were different patterns observed between ATT and ATC coding. Specifically, nearest neighbor matching resulted in a matched treatment

group with the same propensity score mean as the baseline treatment group (and the same for the comparison group mean). Conversely, nearest neighbor matching with a caliper resulted in a matched treatment group propensity score mean that was lower than the baseline treatment group mean, but also a matched comparison group mean that was higher than the baseline comparison group mean. In other words, the matched treatment and comparison group met in roughly the middle of the propensity score range. For ATC coding, nearest neighbor matching with a caliper also resulted in matched treatment and comparison group means that were similar to each other (with a higher matched treatment group mean than baseline treatment group mean and a lower matched comparison group mean than baseline comparison group mean). For nearest neighbor matching, the matched treatment group mean was higher than the baseline treatment group mean, however, there was no difference between the matched comparison group mean and the baseline comparison group mean.

When the comparison group was larger than the treatment group (treatment to comparison ratio of 1:4), nearest neighbor matching with ATT coding resulted in no difference between the matched treatment group propensity score mean and the baseline treatment group propensity score mean. The matched comparison group mean was higher than the baseline comparison group mean (and similar to the baseline treatment group mean). For nearest neighbor matching with a caliper with ATT coding, the matched treatment group mean was lower than the baseline treatment group mean, but more similar to the baseline treatment group mean than to the baseline comparison group mean. Additionally, the matched treatment and comparison group means were nearly identical for nearest neighbor matching with a caliper, whereas with nearest neighbor matching the

matched treatment group mean was still higher than the matched comparison group mean.

When nearest neighbor matching with ATC coding was used, there was no difference between the matched treatment group mean and baseline treatment group mean. Likewise, there was no difference between the matched comparison group mean and baseline comparison group mean. For nearest neighbor matching with a caliper with ATC coding, the matched treatment group mean was higher than the baseline treatment group mean and the matched comparison group mean was lower than the baseline comparison group mean. Additionally, the matched treatment and comparison group propensity score means were nearly identical for nearest neighbor matching with a caliper, whereas with nearest neighbor matching the matched treatment group propensity score mean was still lower than the matched comparison group propensity score mean. In sum, for nearest neighbor matching, coding method determined whether the matched treatment group resembled the original treatment or original comparison group. For nearest neighbor matching with a caliper, the matched treatment group always resembled the matched comparison group, regardless of coding method.

CHAPTER 5

Discussion

The performance of three propensity score methods when the treatment group was larger than the comparison group was evaluated in the current study. Several conditions were simulated to examine the function of each propensity score method under different situations that researchers may encounter. Specifically, two coding methods, four true treatment effects, three treatment sample sizes, three treatment to comparison group ratios, and five levels of baseline covariate balance were simulated.

Data were simulated for nine scenarios with 1,000 replications per scenario. Each scenario represented a unique combination of treatment sample size and treatment to comparison group ratio[2]. Across all scenarios, covariate baseline imbalance was varied by specifying a different amount of imbalance (SMD) for each of the five covariates (SMD for X1 = 0, X2 = 0.20, X3 = 0.50, X4 = 0.80, X5 = 1.20). Additionally, for all scenarios, four true treatment effect sizes were simulated (Cohen's *d* for Y1 = 0, Y2 = 0.20, Y3 = 0.50, and Y4 = 0.80). After data were simulated for each scenario, three propensity score methods were evaluated (nearest neighbor matching, nearest neighbor matching with a 0.20 *SD* caliper, and generalized boosted modeling) using two coding methods (ATT coding and ATC coding) for each true treatment effect.

---

[2] Each scenario represented a unique combination of treatment sample size and treatment to comparison group ratio as follows: scenario A ($N_{Treatment}$ = 200, T:C = 2:1), scenario B ($N_{Treatment}$ = 600, T:C = 2:1), scenario C ($N_{Treatment}$ = 1,000, T:C = 2:1), scenario D ($N_{Treatment}$ = 200, T:C = 4:3), scenario E ($N_{Treatment}$ = 600, T:C = 4:3), scenario F ($N_{Treatment}$ = 1,000, T:C = 4:3), scenario G ($N_{Treatment}$ = 200, T:C = 1:4), scenario H ($N_{Treatment}$ = 600, T:C = 1:4), and scenario I ($N_{Treatment}$ = 1,000, T:C = 1:4).

When the treatment group is larger than the comparison group, one recommendation is to use ATC coding (instead of ATT coding; Ho et al., 2007). Using ATC coding when the treatment group is larger than the comparison group results in coding that is equivalent to using ATT coding when the treatment group is smaller than the comparison group. That is, ATC coding when the treatment group is larger than the comparison group results in the more ideal situation where the group coded "1" is smaller than the group coded "0" (i.e., a smaller "treatment" group than "comparison" group). The current study is the first simulation study (to my knowledge) to examine the effect of coding method on bias when the treatment group is larger than the comparison group. Specifically, three research questions were examined to evaluate the performance of propensity score methods when the treatment group was larger than the comparison group.

Bias in Estimated Treatment Effect: Research Question 1

The first research question regarded the magnitude and direction of bias in the estimated treatment effect across propensity score methods, coding methods, true treatment effect size, treatment sample size, and treatment to comparison group ratio. Specifically, the first research question was two-fold: 1a. *When the treatment group is larger than the comparison group, can propensity score methods accurately recover the true treatment effect* and 1b. *Does the magnitude and direction of the estimated treatment effect differ across propensity score methods depending on group coding?* Bias in the estimated treatment effect was quantified as the average deviation from the true treatment effect across 1,000 replications. Additionally, a 3x2x3x3x4 ANOVA was conducted to determine whether there were statistically significant and practically meaningful

differences in bias across conditions (i.e., propensity score method, coding method, treatment to comparison group ratio, treatment sample size, and true treatment effect size). The magnitude and direction of the estimated treatment effect (averaged across 1,000 replications) was compared across propensity score method, treatment to comparison group ratio, treatment sample size, and true treatment effect size for both coding methods (i.e., ATT and ATC coding). In sum, bias depended on two effects: the three-way interaction between propensity score method, coding method, and treatment to comparison group ratio and the two-way interaction between propensity score method and true treatment effect size.

To facilitate understanding, the three-way interaction will be discussed in two sections: the "typical" scenario (i.e., treatment group smaller than the comparison group) and the "atypical" scenario (i.e., treatment group larger than the comparison group). Within each section, differences in bias across propensity score methods and coding methods are explored.

Treatment Group Smaller Than Comparison Group: Traditional 1:4 Ratio

Under traditional conditions where the treatment group is smaller than the comparison group, propensity score methods have been shown to result in minimally biased estimates of the treatment effect (e.g., Rubin, 1973a, 1973b, 1979). For nearest neighbor matching and generalized boosted modeling, bias was lower for ATT coding than for ATC coding when the treatment to comparison group ratio was 1:4. However, when comparing across the two methods, bias was higher for nearest neighbor matching than for generalized boosted modeling. In sum, for conditions that simulated the "typical" propensity score scenario, using ATT coding replicated previous findings of minimally

biased treatment effect estimates for generalized boosted modeling. In contrast, when using ATC coding, bias was high for both nearest neighbor matching and generalized boosted modeling. Thus, when the "typical" scenario was reversed (via the use of ATC coding) so that group coded "1" was larger than the group coded "0", nearest neighbor matching and generalized boosted modeling did not result in a minimally biased treatment effect estimate.

Unlike the other two methods, nearest neighbor matching with a caliper *did* result in minimally biased treatment effect estimates for *both* ATT and ATC coding. Moreover, for ATT coding, nearest neighbor matching with a caliper resulted in substantially lower bias than nearest neighbor matching and slightly lower bias than generalized boosted modeling. For ATC coding, however, nearest neighbor matching with a caliper resulted in substantially lower bias than both nearest neighbor matching or generalized boosted modeling. In the current study, coding did not matter for nearest neighbor matching with a caliper under the "typical" scenario where the comparison group was larger than the treatment group.

Treatment Group Larger Than Comparison Group: 2:1 and 4:3 Ratios

Under conditions where the treatment group is larger than the comparison group, the use of ATC coding is one recommendation (Ho et al., 2007). Indeed, for nearest neighbor matching and generalized boosted modeling, bias was lower for ATC coding than for ATT coding when the treatment to comparison group ratio was 2:1 or 4:3. However, like the 1:4 ratio findings, bias was higher for nearest neighbor matching than for generalized boosted. In contrast, for nearest neighbor matching and generalized boosted modeling, bias was highest for ATT coding. Like the ATC coding findings, bias

was higher for nearest neighbor matching than for generalized boosted modeling. Thus, for nearest neighbor matching and generalized boosted modeling in the "atypical" scenario, the current study findings support the Ho et al. (2007) recommendation to use ATC coding.

Unlike the other two methods, nearest neighbor matching with a caliper resulted in low bias for both ATT and ATC coding. The observation of low bias for nearest neighbor matching with a 0.20 *SD* caliper and ATT coding when the treatment group was larger than the comparison group provides new information regarding the application of propensity score methods. In the current study, coding did not matter for nearest neighbor matching with a caliper under the "atypical" scenario where the treatment group was larger than the comparison group. Thus, the recommendation to reverse the coding when the treatment group is larger than the comparison group (Ho et al., 2007) may not be necessary when nearest neighbor matching with a caliper is implemented. Therefore, researchers interested in examining the effect of the treatment on those who received treatment (ATT), may not need to alter the question of interest if the treatment group is larger than the comparison group.

Not only did bias differ by propensity score method, coding method, and treatment to comparison group ratio, bias also differed by propensity score method and true treatment effect size. Averaging across coding method and treatment to comparison group ratio, nearest neighbor matching with a caliper resulted in low bias for all four true treatment effect sizes. Generalized boosted modeling resulted in low bias when there was no true treatment effect or a small true treatment effect. Nearest neighbor matching, however, resulted in high bias for all four true treatment effects. Thus, recovery of the

true treatment effect depended on the true treatment effect size to a greater extent for generalized boosted modeling than for nearest neighbor matching with a caliper or nearest neighbor matching.

Averaging across coding method and treatment to comparison group ratio, all three propensity score methods resulted in a reduction in bias over baseline bias. This was not surprising, as propensity score methods aim to reduce bias by reducing systematic group differences (Austin, 2011; Cochran & Rubin, 1973; Ho et al., 2007; Rosenbaum & Rubin, 1983b; Rubin, 1973a, 1973b, 1974; Shadish et al., 2008). However, the three propensity score methods did not reduce bias in the estimated treatment effect to the same extent. Nearest neighbor matching resulted in bias in the estimated treatment effect that was not largely different from initial, unmatched baseline bias. Generalized boosted modeling reduced bias but resulted in an overcorrection, in which the estimated treatment effect was less than the true treatment effect. Nearest neighbor matching with a caliper reduced bias to the greatest extent and resulted in lowest bias in the estimated treatment effect.

Direction and Magnitude of Bias

For each scenario (i.e., each unique combination of treatment to comparison group ratio and treatment sample size), the direction of bias consistently differed across ATT and ATC coding. As might be expected, if the estimated treatment effect for a certain propensity score method was positive for ATT coding it was negative for the same propensity score method for ATC coding (and vice versa).

When the treatment group was larger than the comparison group, the magnitude of bias across ATT and ATC coding method differed for each propensity score method.

For nearest neighbor matching, the magnitude of bias for ATC coding was consistently smaller than for ATT coding. For generalized boosted modeling, the magnitude of bias in the estimated treatment effect for ATC coding was similar to or smaller than that for ATT coding. For nearest neighbor matching with a caliper, the magnitude of bias in the estimated treatment effect did not meaningfully differ across ATT and ATC coding. In sum, nearest neighbor matching with a caliper resulted in the least amount of bias of the three methods, and resulted in low bias for both ATT and ATC coding regardless of which group was larger (i.e., treatment or comparison group) for all true treatment effect sizes.

Covariate Balance: Research Question 2

The second research question regarded covariate balance obtained across propensity score methods, coding methods, treatment sample size, treatment to comparison group ratio, and baseline covariate balance: *When the treatment group is larger than the comparison group, can propensity score methods achieve adequate group balance on the covariates?* Balance on the covariates after the application of propensity score matching was quantified as the standardized mean difference between the treatment and comparison groups on each covariate. Additionally, the percentage in bias reduction was examined to supplement the information provided by the standardized mean difference. Each of the five covariates had a different level of baseline balance (SMD X1 = 0, X2 = 0.20, X3 = 0.50, X4 = 0.80, and X5 = 1.20), which allowed for the comparison of each propensity score method across differing levels of baseline group covariate balance. In sum, balance differed depending on propensity score method and initial

covariate balance. Balance also differed depending on coding method for nearest neighbor matching.

The three propensity score methods differed in terms of covariate balance after matching or weighting. Across all five covariates and the propensity score, nearest neighbor matching with a caliper resulted in the lowest standardized mean differences. Even for covariates with large baseline standardized mean differences (X4 and X5), nearest neighbor matching with a caliper resulted in low standardized mean differences. Recall that nearest neighbor matching with a caliper was also the method that resulted in the least bias in the treatment effect. That is, nearest neighbor matching with a caliper achieved *both* the best balance and least bias of the three propensity score methods.

Generalized boosted modeling resulted in low standardized mean differences for covariates that had medium, low, or no baseline standardized mean differences (X3, X2, and X1, respectively). For covariates with large standardized mean differences (X4 and X5), generalized boosted modeling resulted in a reduction in the standardized mean difference from that at baseline; however, systematic group differences remained on these covariates. Recall that generalized boosted modeling resulted in slightly larger bias than nearest neighbor matching with a caliper. That is, slight bias in the estimated treatment effect remained likely because ideal covariate balance was not achieved for all covariates.

Nearest neighbor matching, on the other hand, had the most imbalance, with the exception of the covariate that was balanced at baseline (X1). Recall that nearest neighbor matching consistently resulted in the highest bias. That is, nearest neighbor matching resulted in the least reduction of systematic group differences on the covariates and had the highest bias of the three propensity score methods.

For nearest neighbor matching with a caliper and generalized boosted modeling, there were no meaningful differences in covariate balance across coding method. Conversely, for nearest neighbor matching, coding method mattered. When coding resulted in a larger comparison group than treatment group, after nearest neighbor matching there was a reduction in standardized mean differences from that at baseline for X2, X3, X4, and X5. However, when coding resulted in a larger treatment group than comparison group, there was no change in the standardized mean difference from that at baseline for X2, X3, X4, and X5. That is, the covariates remained imbalanced, and selection bias was still present.

The removal of systematic group differences results in the reduction of bias in the estimated treatment effect (Austin, 2011; Cochran & Rubin, 1973; Ho et al., 2007; Rosenbaum & Rubin, 1983b; Rubin, 1973a, 1973b, 1974; Shadish et al., 2008). Thus, it was not surprising that the propensity score methods that resulted in low standardized mean differences also resulted in low bias in the estimated treatment effect.

Treatment Group Loss: Research Question 3

The third research question regarded similarity between the matched and baseline treatment groups for matching methods (i.e., nearest neighbor matching and nearest neighbor matching with a caliper): *When the treatment group is larger than the comparison group, does the loss of treatment group members differ across conditions?* Loss of treatment group members was quantified as the percentage of treatment group members not retained in the matched sample. Additionally, the average propensity scores were compared for baseline and matched treatment and comparison groups to determine similarity between the matched and baseline treatment groups. The percent of treatment

group loss differed across nearest neighbor matching and nearest neighbor matching with a caliper. Additionally, for nearest neighbor matching, whether the average matched treatment group propensity score resembled the average baseline treatment group propensity score differed across coding method.

Nearest Neighbor Matching

For nearest neighbor matching, the percentage of unmatched treatment group members aligned with expectations. That is, when using one-to-one matching with a larger treatment group than comparison group, the maximum number of retained treatment group members is the number of comparison group members. When the comparison group was larger than the treatment group, there was no loss of treatment group members. However, the matched treatment group did not always resemble the baseline treatment group with regards to the propensity score.

After nearest neighbor matching when the treatment group was larger than the comparison group with ATT coding, there was no change in the average propensity score compared to baseline. However, recall that nearest neighbor matching with ATT coding always resulted in the largest bias in the estimated treatment effect and worst covariate balance. When the treatment group was larger than the comparison group and ATC coding was used, the average matched treatment group propensity score was similar to the average baseline comparison group propensity score. Recall that nearest neighbor matching with ATC coding resulted in small improvements in covariate balance over that at baseline and slightly less bias in the estimated treatment effect than at baseline. Although there was minimal loss of treatment group members, selection bias remained

after nearest neighbor matching when the treatment group was larger than the comparison group for both ATT and ATC coding.

Conversely, after nearest neighbor matching when the comparison group was larger than the treatment group with ATT coding (i.e., the typical scenario where the comparison group is larger than the treatment group), the matched comparison group was selected to resemble the baseline (and matched) treatment group. When ATC coding was used, there was no change in the average propensity score compared to baseline. In short, matched propensity scores resembled whichever group was coded "1", yet initial selection bias remained.

Nearest Neighbor Matching With a 0.20 SD Caliper

For nearest neighbor matching, the percentage of unmatched treatment group members was slightly larger than for nearest neighbor matching. For nearest neighbor matching with a caliper, when the treatment group was larger than the comparison group, the matched treatment group and matched comparison group resembled neither the baseline treatment group nor the baseline comparison group regardless of coding method. In these cases, the propensity scores of the matched groups met midway between the baseline propensity scores of the two groups. Conversely, when the treatment group was smaller than the comparison group, there was little change in the average treatment group propensity score after matching regardless of coding method. Under the "typical" scenario, the matched comparison group was created to resemble the original treatment group.

Despite the lack of similarity between the matched and baseline treatment group propensity scores for nearest neighbor matching with a caliper, adequate group balance

was achieved on all covariates and bias in the estimated treatment effect was lowest. Loss of treatment representation and the smaller matched sample size that occurred after nearest neighbor matching with a caliper did not impact recovery of the true treatment effect, contrary to the suggestion by Jacovidis et al. (2017). Despite a loss of treatment group sample of 19-68%, nearest neighbor matching with a caliper resulted in adequate group balance on the covariates and low bias in all four estimated treatment effects.

Although loss of treatment representation did not impact recovery of the true treatment effect for nearest neighbor matching with a caliper, there may be additional concerns regarding loss of representativeness of the matched treatment group in terms of diversity and equity. Thus, if the matched treatment group differs from the baseline treatment group, researchers may have additional concerns regarding treatment group representation for the matched treatment group.

Future Research and Limitations

There are several limitations to the current study worth noting. These limitations center around the limited number of conditions examined and the well-known disadvantage that simulation studies may not represent applied practice.

First, the aim of the current study was to examine commonly used propensity score methods under varying, realistic conditions. However, without examination of additional methods it is difficult to make broad recommendations regarding the reduction of selection bias when the treatment group is larger than the comparison group. Additional recommendations when the treatment group is larger than the comparison group include subclassification, full matching, weighting by the odds (Stuart, 2010), and matching with replacement (Ho et al., 2007). The current study focused on the

recommendation to switch the coding of the treatment and comparison group (Ho et al., 2007). Although bias was the lowest for nearest neighbor matching with a caliper (for both ATT and ATC coding), bias was not always zero. That is, a minimal amount of bias remained in the estimated treatment effect for the best performing method in the current study. Examination of subclassification, full matching, weighting by the odds, and matching with replacement is necessary to provide comprehensive guidance to researchers.

Second, each of the four true treatment effects were simulated to be homogeneous across all levels of the propensity score. Thus, the magnitude of the true treatment effect was the same across ATT, ATC, and ATE coding. In applied practice, treatment effects are likely to be heterogeneous. That is, some participants may be more responsive to an intervention than others. In this case, the true treatment effect will not be the same across ATT, ATC, and ATE coding. When treatment effects are heterogeneous across levels of the propensity score, subclassification may be preferred over the methods examined in the current study.

Third, for nearest neighbor matching and nearest neighbor matching with a caliper, treatment group members were randomly ordered by propensity score before selecting matches. Although random ordering is recommended over high to low or low to high ordering (Austin, 2013) the default ordering for the MatchIt package in R is high to low. The order in which comparison group members are selected for treatment group members can impact balance between the matched groups. Although random ordering is recommended when using matching methods, ordering of treatment group members has not been examined when the treatment group is larger than the comparison group. The

current study demonstrated that adequate group balance was obtained after nearest neighbor matching with a 0.20 *SD* caliper with random ordering. However, multiple ordering methods were not examined. Future studies should examine balance obtained and bias in the estimated treatment effect after matching using different methods for ordering treatment group members when the treatment group is larger than the comparison group.

Fourth, only one caliper width was employed with nearest matching (*SD* of 0.20). Although this is a recommended caliper width (Austin, 2009), the results from this study may not extend to nearest neighbor matching using different caliper widths. That is, although the best group balance and lowest bias was observed for nearest neighbor matching with a 0.20 *SD* caliper, the same might or might not be true for nearest neighbor matching using other caliper widths. Future studies should examine whether performance differs depending on the selected caliper.

Fifth, there is a lack of sample size guidelines for the use of propensity score methods when the treatment group is larger than the comparison group. In the current study, three treatment group sample sizes were examined (treatment group sample size of 200, 600, and 1,000), however these sample sizes do not represent all possible treatment group sample sizes that may occur in applied research. In the current study, sample size did not relate to bias in the estimated treatment effect or balance obtained after matching or weighting. Thus, future research regarding the performance of propensity score matching when the treatment group is larger than the comparison group using different sample sizes would fill a gap in the literature.

Sixth, a small number of covariates (5) were simulated in the current study, all of which were continuous covariates. In applied research, more than five covariates is typical, and the covariates may be a mix of categorical and continuous variables. Future studies should examine the performance of propensity score methods when the treatment group is larger than the comparison group using different numbers of covariates and both categorical and continuous covariates.

Finally, there is always a trade-off between applied and simulation studies. Simulation studies allow for the specification of "truth"; however, it may not be clear whether that truth is realistic. Although applied studies allow for the examination of research questions under realistic circumstances, "truth" cannot be known. In the current study, conditions were selected to represent realistic scenarios; however, all possible scenarios cannot be examined in a single study.

Practical Implications

The recommendation to reverse the coding of the groups when the treatment group is larger than the comparison group (Ho et al., 2007) may not be necessary for all propensity score methods. For nearest neighbor matching with a caliper, there was little difference in the estimated treatment effect between ATT and ATC coding when the treatment group was larger than the comparison group. When nearest neighbor matching with a caliper was used, both ATT and ATC coding resulted in estimated treatment effects of similar magnitude, differing only in direction. Simply put, coding method did not matter for nearest neighbor matching with a caliper.

For generalized boosted modeling and nearest neighbor matching, coding mattered depending on the treatment to comparison group ratio. Specifically, when the

treatment to comparison group ratio was 1:4 (i.e., comparison group larger than the treatment group), bias in the estimated treatment effect was lower for ATT coding than for ATC coding. Conversely, when the treatment to comparison group ratio was 2:1 or 4:3 (i.e., treatment group larger than the comparison group), bias in the estimated treatment effect was lower for ATC coding than for ATT coding. Nearest neighbor matching differed from generalized boosted modeling in that a high amount of bias was observed for all coding methods and ratios (except for ATT coding when the treatment to comparison group ratio was 1:4). Thus, when the treatment group is larger than the comparison group, reversing the coding may not be necessary if using nearest neighbor matching with a 0.20 *SD* caliper but should be considered for other methods.

The differences in bias that were observed across propensity score methods for the simulated true treatment effect sizes also provide meaningful information for practice when the treatment group is larger than the comparison group. Although researchers never know the true treatment effect when conducting applied research, two recommendations emerge from this study. First, if prior research indicates that a small treatment effect is expected, nearest neighbor matching with a caliper and generalized boosted modeling with ATT coding may be methods to consider. Second, if prior research indicates that a medium or large treatment effect is expected, nearest neighbor matching with a caliper with ATT coding may be a method to consider. For larger true treatment effects, generalized boosted modeling overcorrected, resulting in a lower estimated treatment effect than true treatment effect. This overcorrection was most likely observed because the covariates that were most unbalanced at baseline had the strongest relation with latent propensity for treatment and with one another. For all true treatment

effects, nearest neighbor matching is not recommended when the treatment group is larger than the comparison group. Nearest neighbor matching resulted in a reduction in bias over that at baseline; however, the bias in the estimated treatment effect was still large. Thus, consulting prior research could provide guidance for which propensity score methods to audition when the treatment group is larger than the comparison group.

Propensity score methods that resulted in the best balance were those that resulted in the least bias in the estimated treatment effect. Thus, when the treatment group was larger than the comparison group, examination of group balance after matching or weighting provided useful information regarding potential reduction in bias in the estimated treatment effect. Researchers who wish to employ propensity score matching when the treatment group is larger than the comparison group should use the balance checking methods that are recommended in the propensity score literature to determine whether systematic group differences are reduced and whether treatment effects should be estimated using the matched or weighted data. However, researchers are cautioned against equating adequate covariate balance on all covariates with the complete removal of selection bias. Given the use of a simulation study, the current study was conducted under ideal conditions. That is, all covariates related to treatment selection were used for the estimation of the propensity score and creation of matched groups. Thus, the assumption of no unmeasured confounders has been met. However, in practice, there may be covariates related to selection bias that are not measured. If there are unmeasured confounders, all selection bias may not be removed from the estimated treatment effect even if adequate balance is obtained on all covariates.

As illustrated by the findings, nearest neighbor matching with a caliper *can* result in a larger loss of treatment group representation than nearest neighbor matching. Of particular interest was whether loss of treatment representation related to recovery of the true treatment effect. Loss of treatment representation (due to the inclusion of a caliper for nearest neighbor matching) did not impact the recovery of the true treatment effect. Conversely, nearest neighbor matching (which resulted in less loss of treatment representation than nearest neighbor matching with a caliper) resulted in poor recovery of the true treatment effect. Nearest neighbor matching was not able to obtain adequate group balance on the covariates (except for the covariate for which groups were balanced at baseline). Additionally, there was a large amount of bias in the estimated treatment effect despite the matched treatment group resembling the baseline treatment group. Thus, covariate balance after matching or weighting was a better indicator of bias reduction than was the similarity between the matched and baseline treatment group propensity scores.

Conclusion

To draw appropriate causal inferences from quasi-experimental studies, researchers must be cognizant of and account for selection bias. Although additional research is needed to understand how to best reduce selection bias when the treatment group is larger than the comparison group, the current study adds to the limited existing research. In educational research, if selection bias is present, practitioners have little evidence for causal claims. However, reducing selection bias strengthens the validity of inferences made regarding the treatment effect when random assignment is not feasible.

Table 1

Simulation Conditions (Repeated Across Nearest Neighbor Matching, Nearest Neighbor Matching with 0.20 SD Caliper, and Generalized Boosted Modeling)

ATT Coding (Treatment = 1)			ATC Coding (Treatment = 0)		
T:C Ratio	N_T/N_C	True Treatment Effect (Cohen's *d*)	T:C Ratio	N_T/N_C	True Treatment Effect (Cohen's *d*)
2:1	200/100	0	2:1	200/100	0
		0.20			0.20
		0.50			0.50
		0.80			0.80
	600/300	0		600/300	0
		0.20			0.20
		0.50			0.50
		0.80			0.80
	1000/500	0		1000/500	0
		0.20			0.20
		0.50			0.50
		0.80			0.80
4:3	200/150	0	4:3	200/150	0
		0.20			0.20
		0.50			0.50
		0.80			0.80
	600/450	0		600/450	0
		0.20			0.20
		0.50			0.50
		0.80			0.80
	1000/750	0		1000/750	0
		0.20			0.20
		0.50			0.50
		0.80			0.80
1:4	200/800	0	1:4	200/800	0
		0.20			0.20
		0.50			0.50
		0.80			0.80
	600/2400	0		600/2400	0
		0.20			0.20
		0.50			0.50
		0.80			0.80
	1000/4000	0		1000/4000	0
		0.20			0.20
		0.50			0.50
		0.80			0.80

Table 2

Treatment Group, Comparison Group, and Total Sample Sizes for Configurations A through I

Sample	T:C Ratio	Treatment *N*	Comparison *N*	Total *N*
Configuration A	2:1	200	100	300
Configuration B	2:1	600	300	900
Configuration C	2:1	1000	500	1500
Configuration D	4:3	200	150	350
Configuration E	4:3	600	450	1050
Configuration F	4:3	1000	750	1750
Configuration G	1:4	200	800	1000
Configuration H	1:4	600	2400	3000
Configuration I	1:4	1000	4000	5000

Table 3

Specified Standardized Group Mean Differences and

Correlations between Covariates and Latent Propensity Scores

Variable	X1	X2	X3	X4	X5
X1	1.00				
X2	0.10	1.00			
X3	0.20	0.30	1.00		
X4	0.30	0.30	0.30	1.00	
X5	0.30	0.35	0.45	0.65	1.00
Latent Propensity	-0.02	0.15	0.40	0.70	0.90
SMD	0.00	0.20	0.50	0.80	1.20

Table 4

Method of Evaluation, Conditions Examined, and Values Saved from Simulated Data for Each Research Question

Research Question	Method of Evaluation	Conditions Examined	Values Saved from Simulated Data
RQ1a: When the treatment group is larger than the comparison group, can propensity score methods accurately recover the true treatment effect?	• Bias - the extent to which the estimated treatment effect differs from the population (or specified) treatment effect.	• Treatment sample size (3) • T:C ratio (3) • True treatment effect size (4) • Propensity score methods (3) • Coding methods (2)	• Intercept (mean outcome for group coded "0") • Regression coefficient for grouping variable (estimated treatment effect)
RQ1b: Does the magnitude and direction of the estimated treatment effect differ across propensity score methods depending on group coding?	• Magnitude of effect - Cohen's *d* effect size • Direction of effect - sign of the regression coefficient for grouping variable	• Treatment sample size (3) • T:C ratio (3) • True treatment effect size (4) • Propensity score methods (3) • Coding methods (2)	• *t*-value for regression coefficient (to compute Cohen's d effect size) • Regression coefficient for grouping variable (estimated treatment effect)
RQ2: When the treatment group is larger than the comparison group, can propensity score methods achieve adequate group balance on the covariates?	• Standardized mean difference (SMD) on each covariate and propensity score • Percentage in bias reduction for each covariate and propensity score • Propensity score variance ratio	• Treatment sample size (3) • T:C ratio (3) • Propensity score methods (3) • Coding methods (2) • Initial covariate balance (5)	• Matched & unmatched sample sizes • SMD for covariates and propensity score (before and after matching/weighting) • Percentage in bias reduction (after matching/weighting) • Propensity score variance ratio (after matching/weighting)
RQ3: When the treatment group is larger than the comparison group, does the loss of treatment group members differ across conditions?	• Percent of unmatched treatment group members • Average group propensity score	• Treatment sample size (3) • T:C ratio (3) • Propensity score methods (2) • Coding methods (2)	• Matched & unmatched sample sizes • Propensity score variance ratio (after matching/weighting) • Group propensity score averages (for matched and unmatched simulees)

Note. RQ3 was evaluated over 2 of the 3 propensity score methods (nearest neighbor matching and nearest neighbor matching with a 0.20 *SD* caliper). RQ3 was not evaluated for generalized boosted modeling because there would be no loss of treatment group members.

Table 5

Standardized Group Mean Differences and Correlations between Covariates and Latent Propensity Scores by Validation Sample

Variable	SMD	Latent Propensity	X1	X2	X3	X4	X5
Sample A							
X1	-0.01	-0.01	1.00				
X2	0.17	0.13	0.28	1.00			
X3	0.62	0.48	0.23	0.30	1.00		
X4	0.80	0.64	0.33	0.30	0.36	1.00	
X5	1.14	0.83	0.37	0.41	0.42	0.65	1.00
Sample B							
X1	0.07	0.06	1.00				
X2	0.23	0.19	0.12	1.00			
X3	0.54	0.43	0.22	0.31	1.00		
X4	0.76	0.62	0.32	0.28	0.30	1.00	
X5	1.18	0.89	0.36	0.37	0.42	0.63	1.00
Sample C							
X1	0.06	0.05	1.00				
X2	0.22	0.18	0.14	1.00			
X3	0.48	0.39	0.20	0.30	1.00		
X4	0.89	0.68	0.33	0.31	0.29	1.00	
X5	1.22	0.88	0.33	0.36	0.43	0.64	1.00
Sample D							
X1	-0.21	-0.18	1.00				
X2	0.17	0.13	0.23	1.00			
X3	0.46	0.36	0.26	0.31	1.00		
X4	0.75	0.55	0.36	0.28	0.37	1.00	
X5	1.13	0.79	0.38	0.37	0.43	0.64	1.00
Sample E							
X1	0.02	0.01	1.00				
X2	0.21	0.18	0.13	1.00			
X3	0.46	0.38	0.20	0.31	1.00		
X4	0.81	0.65	0.32	0.30	0.30	1.00	
X5	1.20	0.89	0.35	0.37	0.43	0.64	1.00
Sample F							
X1	0.00	0.00	1.00				
X2	0.26	0.21	0.14	1.00			
X3	0.46	0.37	0.21	0.30	1.00		
X4	0.91	0.68	0.33	0.31	0.31	1.00	
X5	1.26	0.89	0.33	0.37	0.44	0.64	1.00

Table 5 Cont.

Variable	SMD	Latent Propensity	X1	X2	X3	X4	X5
Sample G							
X1	-0.02	-0.01	1.00				
X2	0.21	0.16	0.13	1.00			
X3	0.52	0.37	0.21	0.32	1.00		
X4	0.88	0.62	0.33	0.30	0.29	1.00	
X5	1.19	0.80	0.36	0.38	0.43	0.64	1.00
Sample H							
X1	-0.05	-0.03	1.00				
X2	0.21	0.15	0.10	1.00			
X3	0.59	0.40	0.20	0.29	1.00		
X4	0.99	0.64	0.31	0.31	0.31	1.00	
X5	1.29	0.80	0.31	0.35	0.46	0.64	1.00
Sample I							
X1	-0.01	0.00	1.00				
X2	0.17	0.12	0.10	1.00			
X3	0.47	0.34	0.21	0.28	1.00		
X4	1.00	0.64	0.29	0.30	0.31	1.00	
X5	1.36	0.82	0.30	0.35	0.44	0.65	1.00

Table 6

Treatment and Comparison Group Size, Treatment to Comparison Ratio, and True Treatment Effect by Validation Sample

Sample	Treatment *N*	Comparison *N*	T:C Ratio	True Treatment Effect Y1	Y2	Y3	Y4
Sample A	212	88	0.707	-0.01	0.18	0.48	0.78
Sample B	609	291	0.677	0.00	0.19	0.49	0.79
Sample C	1004	496	0.669	0.10	0.28	0.56	0.84
Sample D	203	147	0.580	-0.02	0.16	0.44	0.73
Sample E	600	450	0.571	0.02	0.20	0.49	0.77
Sample F	995	755	0.569	-0.04	0.14	0.41	0.69
Sample G	187	813	0.187	0.15	0.34	0.63	0.91
Sample H	597	2403	0.199	0.02	0.21	0.50	0.79
Sample I	1009	3991	0.202	0.02	0.21	0.50	0.79

Note. T:C Ratio was set as follows: samples A-C, .667 (ratio of 2:1); samples D-F, .571 (ratio of 4:3); samples G-I, .200 (ratio of 1:4). For all samples, true treatment effect was set as follows: Y1, 0; Y2, 0.20; Y3, 0.50; Y4, 0.80.

Table 7

Simulated Means, Standard Deviations, and Standard Errors for Covariates and True Propensity Scores by Scenario

		X1				X2				X3				X4				X5				True Propensity Score			
Scenario	*N*	*M*	*SE*	*SD*	*SE*	*M*	*SE*	*SD*	*SE*	*M*	*SE*	*SD*	*SE*	*M*	*SE*	*SD*	*SE*	*M*	*SE*	*SD*	*SE*	*M*	*SE*	*SD*	*SE*
Scenario A																									
Treatment	200.378	-0.008	0.071	1.001	0.048	0.062	0.072	0.996	0.050	0.157	0.071	0.976	0.050	0.272	0.066	0.930	0.046	0.352	0.062	0.881	0.043	0.781	0.022	0.204	0.011
Comparison	99.622	0.019	0.103	0.999	0.075	-0.117	0.099	0.996	0.072	-0.312	0.101	0.972	0.068	-0.545	0.090	0.904	0.065	-0.698	0.084	0.837	0.060	0.441	0.038	0.257	0.015
All	300.000	0.001	0.082	1.001	0.057	0.002	0.081	0.996	0.057	0.001	0.081	0.974	0.056	0.001	0.074	0.922	0.052	0.004	0.070	0.866	0.049	0.668	0.027	0.222	0.012
Scenario B																									
Treatment	600.219	-0.009	0.041	0.998	0.029	0.056	0.041	0.997	0.028	0.152	0.038	0.979	0.027	0.271	0.039	0.932	0.027	0.347	0.036	0.884	0.025	0.776	0.013	0.203	0.007
Comparison	299.781	0.014	0.057	1.001	0.041	-0.116	0.058	0.994	0.040	-0.311	0.058	0.971	0.040	-0.543	0.052	0.908	0.037	-0.698	0.050	0.840	0.034	0.449	0.023	0.255	0.009
All	900.000	-0.001	0.046	0.999	0.033	-0.001	0.047	0.996	0.032	-0.002	0.045	0.976	0.032	0.000	0.043	0.924	0.030	-0.001	0.041	0.869	0.028	0.667	0.016	0.220	0.008
Scenario C																									
Treatment	1000.800	-0.011	0.031	1.000	0.022	0.057	0.033	0.997	0.021	0.155	0.031	0.979	0.022	0.269	0.030	0.932	0.020	0.346	0.028	0.886	0.019	0.775	0.011	0.203	0.005
Comparison	499.196	0.015	0.043	0.997	0.031	-0.116	0.043	0.995	0.031	-0.309	0.042	0.970	0.030	-0.542	0.041	0.906	0.028	-0.697	0.037	0.838	0.027	0.452	0.017	0.254	0.007
All	1500.000	-0.002	0.035	0.999	0.025	0.000	0.036	0.996	0.025	0.001	0.035	0.976	0.025	-0.001	0.033	0.923	0.023	-0.001	0.031	0.870	0.022	0.667	0.013	0.220	0.006
Scenario D																									
Treatment	199.827	-0.006	0.071	1.001	0.049	0.074	0.072	0.996	0.051	0.195	0.067	0.974	0.050	0.344	0.063	0.922	0.046	0.439	0.060	0.870	0.045	0.717	0.025	0.225	0.010
Comparison	150.173	0.014	0.081	0.997	0.058	-0.098	0.080	0.995	0.059	-0.260	0.079	0.969	0.057	-0.454	0.076	0.909	0.053	-0.584	0.069	0.850	0.050	0.377	0.031	0.247	0.012
All	350.000	0.002	0.075	0.999	0.053	0.000	0.076	0.996	0.054	0.000	0.072	0.972	0.053	0.002	0.069	0.917	0.049	0.000	0.064	0.861	0.047	0.571	0.028	0.234	0.011
Scenario E																									
Treatment	600.046	-0.009	0.042	1.000	0.029	0.071	0.040	0.995	0.027	0.194	0.041	0.976	0.029	0.341	0.040	0.923	0.027	0.439	0.036	0.868	0.025	0.715	0.014	0.224	0.006
Comparison	449.954	0.012	0.048	0.999	0.032	-0.097	0.049	0.996	0.034	-0.260	0.047	0.971	0.033	-0.455	0.045	0.913	0.030	-0.586	0.040	0.852	0.029	0.381	0.018	0.246	0.007
All	1050.000	0.000	0.044	0.999	0.030	-0.001	0.044	0.995	0.030	-0.001	0.043	0.974	0.030	0.000	0.042	0.919	0.028	0.000	0.038	0.861	0.027	0.571	0.016	0.234	0.007
Scenario F																									
Treatment	999.427	-0.008	0.032	0.999	0.022	0.074	0.031	0.995	0.023	0.196	0.031	0.975	0.021	0.343	0.030	0.922	0.021	0.441	0.028	0.868	0.019	0.714	0.011	0.224	0.005
Comparison	750.573	0.012	0.037	1.000	0.025	-0.098	0.037	0.996	0.025	-0.259	0.037	0.972	0.025	-0.454	0.035	0.913	0.023	-0.586	0.032	0.852	0.022	0.381	0.014	0.246	0.005
All	1750.000	0.000	0.034	0.999	0.024	0.000	0.033	0.996	0.024	0.001	0.034	0.973	0.023	0.001	0.032	0.918	0.022	0.001	0.030	0.862	0.020	0.571	0.012	0.234	0.005
Scenario G																									
Treatment	200.040	-0.018	0.070	0.999	0.050	0.150	0.071	0.996	0.050	0.397	0.069	0.966	0.049	0.700	0.067	0.897	0.045	0.897	0.060	0.826	0.043	0.441	0.027	0.256	0.013
Comparison	799.960	0.005	0.034	1.000	0.024	-0.039	0.035	0.997	0.025	-0.098	0.035	0.983	0.025	-0.174	0.033	0.946	0.024	-0.225	0.032	0.911	0.022	0.140	0.009	0.162	0.007
All	1000.000	0.000	0.042	0.999	0.029	-0.001	0.042	0.997	0.030	0.001	0.042	0.980	0.030	0.000	0.040	0.936	0.028	0.000	0.037	0.894	0.026	0.200	0.013	0.181	0.008
Scenario H																									
Treatment	599.374	-0.020	0.041	1.000	0.028	0.147	0.039	0.997	0.029	0.396	0.039	0.968	0.028	0.695	0.037	0.897	0.026	0.893	0.034	0.825	0.024	0.436	0.015	0.254	0.007
Comparison	2400.630	0.005	0.021	1.001	0.015	-0.036	0.020	0.997	0.014	-0.099	0.020	0.983	0.015	-0.174	0.019	0.948	0.014	-0.223	0.019	0.911	0.013	0.141	0.006	0.162	0.004
All	3000.000	0.000	0.025	1.000	0.017	0.000	0.024	0.997	0.017	0.000	0.024	0.980	0.017	0.000	0.023	0.938	0.017	-0.001	0.022	0.894	0.015	0.200	0.007	0.180	0.005
Scenario I																									
Treatment	997.619	-0.019	0.031	0.999	0.022	0.151	0.031	0.996	0.021	0.399	0.032	0.967	0.022	0.697	0.027	0.900	0.019	0.896	0.026	0.825	0.019	0.437	0.012	0.255	0.006
Comparison	4002.380	0.005	0.015	1.000	0.011	-0.038	0.016	0.997	0.011	-0.099	0.015	0.983	0.011	-0.174	0.015	0.947	0.011	-0.224	0.014	0.911	0.011	0.140	0.004	0.162	0.003
All	5000.000	0.000	0.019	0.999	0.013	0.000	0.019	0.997	0.013	0.000	0.018	0.980	0.013	0.000	0.018	0.938	0.013	0.000	0.017	0.894	0.012	0.200	0.006	0.180	0.004

Note. Covariate means and standard deviations are averaged across 1,000 replications for each scenario. Standard errors indicate the variability in each parameter across the 1,000 replications. Each scenario represents a unique combination of treatment sample size and treatment to comparison group ratio as follows: scenario A ($N_{\text{Treatment}}$ = 200, T:C = 2:1), scenario B ($N_{\text{Treatment}}$ = 600, T:C = 2:1), scenario C ($N_{\text{Treatment}}$ = 1,000, T:C = 2:1), scenario D ($N_{\text{Treatment}}$ = 200, T:C = 4:3), scenario E ($N_{\text{Treatment}}$ = 600, T:C = 4:3), scenario F ($N_{\text{Treatment}}$ = 1,000, T:C = 4:3), scenario G ($N_{\text{Treatment}}$ = 200, T:C = 1:4), scenario H ($N_{\text{Treatment}}$ = 600, T:C = 1:4), and scenario I ($N_{\text{Treatment}}$ = 1,000, T:C = 1:4).

Table 8

Average Simulated Correlations between Covariates and True Propensity Scores by Scenario

Scenario	True Propensity Scores	X1	X2	X3	X4	X5
Scenario A						
X1	-0.02	1.00				
X2	0.14	0.10	1.00			
X3	0.38	0.20	0.30	1.00		
X4	0.66	0.30	0.30	0.30	1.00	
X5	0.85	0.30	0.35	0.45	0.65	1.00
Scenario B						
X1	-0.02	1.00				
X2	0.14	0.10	1.00			
X3	0.38	0.20	0.30	1.00		
X4	0.67	0.30	0.30	0.30	1.00	
X5	0.86	0.30	0.35	0.45	0.65	1.00
Scenario C						
X1	-0.02	1.00				
X2	0.14	0.10	1.00			
X3	0.38	0.20	0.30	1.00		
X4	0.67	0.30	0.30	0.30	1.00	
X5	0.86	0.30	0.35	0.45	0.65	1.00
Scenario D						
X1	-0.02	1.00				
X2	0.15	0.10	1.00			
X3	0.39	0.20	0.30	1.00		
X4	0.68	0.30	0.30	0.30	1.00	
X5	0.87	0.30	0.35	0.45	0.65	1.00
Scenario E						
X1	-0.02	1.00				
X2	0.14	0.10	1.00			
X3	0.39	0.20	0.30	1.00		
X4	0.68	0.30	0.30	0.30	1.00	
X5	0.88	0.30	0.35	0.45	0.65	1.00
Scenario F						
X1	-0.02	1.00				
X2	0.15	0.10	1.00			
X3	0.39	0.20	0.30	1.00		
X4	0.68	0.30	0.30	0.30	1.00	
X5	0.88	0.30	0.35	0.45	0.65	1.00

Table 8 Cont.

Scenario	True Propensity Scores	X1	X2	X3	X4	X5
Scenario G						
X1	-0.02	1.00				
X2	0.14	0.10	1.00			
X3	0.36	0.20	0.30	1.00		
X4	0.63	0.30	0.30	0.30	1.00	
X5	0.81	0.30	0.35	0.45	0.65	1.00
Scenario H						
X1	-0.02	1.00				
X2	0.13	0.10	1.00			
X3	0.36	0.20	0.30	1.00		
X4	0.63	0.30	0.30	0.30	1.00	
X5	0.82	0.30	0.35	0.45	0.65	1.00
Scenario I						
X1	-0.02	1.00				
X2	0.14	0.10	1.00			
X3	0.36	0.20	0.30	1.00		
X4	0.64	0.30	0.30	0.30	1.00	
X5	0.82	0.30	0.35	0.45	0.65	1.00

Note. Correlations are averaged across 1,000 replications for each scenario. Each scenario represents a unique combination of treatment sample size and treatment to comparison group ratio as follows: scenario A ($N_{\text{Treatment}}$ = 200, T:C = 2:1), scenario B ($N_{\text{Treatment}}$ = 600, T:C = 2:1), scenario C ($N_{\text{Treatment}}$ = 1,000, T:C = 2:1), scenario D ($N_{\text{Treatment}}$ = 200, T:C = 4:3), scenario E ($N_{\text{Treatment}}$ = 600, T:C = 4:3), scenario F ($N_{\text{Treatment}}$ = 1,000, T:C = 4:3), scenario G ($N_{\text{Treatment}}$ = 200, T:C = 1:4), scenario H ($N_{\text{Treatment}}$ = 600, T:C = 1:4), and scenario I ($N_{\text{Treatment}}$ = 1,000, T:C = 1:4).

Table 9

Simulated Group Means, Standard Deviations, and Standard Errors for Outcome Variables by Scenario

		Y1				Y2				Y3				Y4			
Scenario	*N*	*M*	*SE*	*SD*	*SE*	*M*	*SE*	*SD*	*SE*	*M*	*SE*	*SD*	*SE*	*M*	*SE*	*SD*	*SE*
Scenario A																	
Treatment	200.378	0.041	0.038	0.525	0.026	0.151	0.038	0.525	0.026	0.321	0.038	0.525	0.026	0.491	0.038	0.525	0.026
Comparison	99.622	-0.082	0.054	0.522	0.038	-0.082	0.054	0.522	0.038	-0.082	0.054	0.522	0.038	-0.082	0.054	0.522	0.038
All	300.000	0.000	0.043	0.524	0.030	0.074	0.043	0.524	0.030	0.187	0.043	0.524	0.030	0.301	0.043	0.524	0.030
Scenario B																	
Treatment	600.219	0.041	0.022	0.524	0.016	0.151	0.022	0.524	0.016	0.321	0.022	0.524	0.016	0.491	0.022	0.524	0.016
Comparison	299.781	-0.083	0.030	0.523	0.021	-0.083	0.030	0.523	0.021	-0.083	0.030	0.523	0.021	-0.083	0.030	0.523	0.021
All	900.000	0.000	0.025	0.524	0.017	0.073	0.025	0.524	0.017	0.187	0.025	0.524	0.017	0.300	0.025	0.524	0.017
Scenario C																	
Treatment	1000.800	0.041	0.017	0.525	0.011	0.151	0.017	0.525	0.011	0.321	0.017	0.525	0.011	0.491	0.017	0.525	0.011
Comparison	499.196	-0.083	0.025	0.524	0.017	-0.083	0.025	0.524	0.017	-0.083	0.025	0.524	0.017	-0.083	0.025	0.524	0.017
All	1500.000	0.000	0.019	0.525	0.013	0.073	0.019	0.525	0.013	0.187	0.019	0.525	0.013	0.300	0.019	0.525	0.013
Scenario D																	
Treatment	199.827	0.051	0.037	0.525	0.026	0.161	0.037	0.525	0.026	0.331	0.037	0.525	0.026	0.501	0.037	0.525	0.026
Comparison	150.173	-0.068	0.043	0.523	0.030	-0.068	0.043	0.523	0.030	-0.068	0.043	0.523	0.030	-0.068	0.043	0.523	0.030
All	350.000	0.000	0.039	0.524	0.028	0.063	0.039	0.524	0.028	0.160	0.039	0.524	0.028	0.257	0.039	0.524	0.028
Scenario E																	
Treatment	600.046	0.052	0.021	0.524	0.016	0.162	0.021	0.524	0.016	0.332	0.021	0.524	0.016	0.502	0.021	0.524	0.016
Comparison	449.954	-0.070	0.024	0.524	0.019	-0.070	0.024	0.524	0.019	-0.070	0.024	0.524	0.019	-0.070	0.024	0.524	0.019
All	1050.000	0.000	0.022	0.524	0.017	0.063	0.022	0.524	0.017	0.160	0.022	0.524	0.017	0.257	0.022	0.524	0.017
Scenario F																	
Treatment	999.427	0.052	0.017	0.524	0.012	0.162	0.017	0.524	0.012	0.332	0.017	0.524	0.012	0.502	0.017	0.524	0.012
Comparison	750.573	-0.070	0.019	0.524	0.013	-0.070	0.019	0.524	0.013	-0.070	0.019	0.524	0.013	-0.070	0.019	0.524	0.013
All	1750.000	0.000	0.018	0.524	0.013	0.062	0.018	0.524	0.013	0.160	0.018	0.524	0.013	0.257	0.018	0.524	0.013
Scenario G																	
Treatment	200.040	0.106	0.036	0.523	0.027	0.216	0.036	0.523	0.027	0.386	0.036	0.523	0.027	0.556	0.036	0.523	0.027
Comparison	799.960	-0.027	0.019	0.526	0.013	-0.027	0.019	0.526	0.013	-0.027	0.019	0.526	0.013	-0.027	0.019	0.526	0.013
All	1000.000	0.000	0.022	0.526	0.016	0.022	0.022	0.526	0.016	0.056	0.022	0.526	0.016	0.090	0.022	0.526	0.016
Scenario H																	
Treatment	599.374	0.106	0.021	0.523	0.015	0.216	0.021	0.523	0.015	0.386	0.021	0.523	0.015	0.556	0.021	0.523	0.015
Comparison	2400.630	-0.026	0.011	0.525	0.008	-0.026	0.011	0.525	0.008	-0.026	0.011	0.525	0.008	-0.026	0.011	0.525	0.008
All	3000.000	0.000	0.013	0.525	0.009	0.022	0.013	0.525	0.009	0.056	0.013	0.525	0.009	0.090	0.013	0.525	0.009
Scenario I																	
Treatment	997.619	0.106	0.017	0.523	0.012	0.216	0.017	0.523	0.012	0.386	0.017	0.523	0.012	0.556	0.017	0.523	0.012
Comparison	4002.380	-0.026	0.008	0.526	0.006	-0.026	0.008	0.526	0.006	-0.026	0.008	0.526	0.006	-0.026	0.008	0.526	0.006
All	5000.000	0.000	0.010	0.525	0.007	0.022	0.010	0.525	0.007	0.056	0.010	0.525	0.007	0.090	0.010	0.525	0.007

Note. Outcome variable (Y1 through Y4) means and standard deviations are averaged across 1,000 replications for each scenario. Standard errors indicate the variability in each parameter across the 1,000 replications. Each scenario represents a unique combination of treatment sample size and treatment to comparison group ratio as follows: scenario A ($N_{\text{Treatment}}$ = 200, T:C = 2:1), scenario B ($N_{\text{Treatment}}$ = 600, T:C = 2:1), scenario C ($N_{\text{Treatment}}$ = 1,000, T:C = 2:1), scenario D ($N_{\text{Treatment}}$ = 200, T:C = 4:3), scenario E ($N_{\text{Treatment}}$ = 600, T:C = 4:3), scenario F ($N_{\text{Treatment}}$ = 1,000, T:C = 4:3), scenario G ($N_{\text{Treatment}}$ = 200, T:C = 1:4), scenario H ($N_{\text{Treatment}}$ = 600, T:C = 1:4), and scenario I ($N_{\text{Treatment}}$ = 1,000, T:C = 1:4).

Table 10

Average Simulated True Treatment Effect for Each Outcome Variable by Scenario and Coding Method

	Y1		Y2		Y3		Y4	
Scenario	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Scenario A								
ATT	0.00	0.12	0.18	0.12	0.46	0.13	0.75	0.13
ATC	0.00	0.12	-0.18	0.12	-0.46	0.13	-0.75	0.13
Scenario B								
ATT	0.00	0.07	0.19	0.07	0.47	0.07	0.76	0.07
ATC	0.00	0.07	-0.19	0.07	-0.47	0.07	-0.76	0.07
Scenario C								
ATT	0.00	0.06	0.18	0.06	0.47	0.06	0.75	0.06
ATC	0.00	0.06	-0.18	0.06	-0.47	0.06	-0.75	0.06
Scenario D								
ATT	0.00	0.11	0.18	0.11	0.46	0.11	0.74	0.11
ATC	0.00	0.11	-0.18	0.11	-0.46	0.11	-0.74	0.11
Scenario E								
ATT	0.00	0.06	0.18	0.06	0.47	0.06	0.75	0.06
ATC	0.00	0.06	-0.18	0.06	-0.47	0.06	-0.75	0.06
Scenario F								
ATT	0.00	0.05	0.18	0.05	0.46	0.05	0.75	0.05
ATC	0.00	0.05	-0.18	0.05	-0.46	0.05	-0.75	0.05
Scenario G								
ATT	0.00	0.08	0.19	0.08	0.49	0.08	0.78	0.08
ATC	0.00	0.08	-0.19	0.08	-0.49	0.08	-0.78	0.08
Scenario H								
ATT	0.00	0.05	0.19	0.05	0.49	0.05	0.79	0.05
ATC	0.00	0.05	-0.19	0.05	-0.49	0.05	-0.79	0.05
Scenario I								
ATT	0.00	0.04	0.19	0.04	0.49	0.04	0.78	0.04
ATC	0.00	0.04	-0.19	0.04	-0.49	0.04	-0.78	0.04

Note. The presented means for each Y outcome variable are averaged across 1,000 replications for each scenario. Standard errors indicate the variability in each parameter across the 1,000 replications. Each scenario represents a unique combination of treatment sample size and treatment to comparison group ratio as follows: scenario A ($N_{\text{Treatment}}$ = 200, T:C = 2:1), scenario B ($N_{\text{Treatment}}$ = 600, T:C = 2:1), scenario C ($N_{\text{Treatment}}$ = 1,000, T:C = 2:1), scenario D ($N_{\text{Treatment}}$ = 200, T:C = 4:3), scenario E ($N_{\text{Treatment}}$ = 600, T:C =

4:3), scenario F ($N_{\text{Treatment}}$ = 1,000, T:C = 4:3), scenario G ($N_{\text{Treatment}}$ = 200, T:C = 1:4), scenario H ($N_{\text{Treatment}}$ = 600, T:C = 1:4), and scenario I ($N_{\text{Treatment}}$ = 1,000, T:C = 1:4).

Table 11

Mean, Median, Minimum, and Maximum Optimal Iterations for Generalized Boosted Models by Scenario and Coding Method

Scenario	Mean	Median	Min	Max
Scenario A				
ATT Coding	683.59	586.00	219.00	2675.00
ATC Coding	709.93	641.00	265.00	2342.00
Scenario B				
ATT Coding	1217.34	1040.00	382.00	7595.00
ATC Coding	998.90	933.00	435.00	2757.00
Scenario C				
ATT Coding	1617.80	1360.00	476.00	6575.00
ATC Coding	1180.18	1137.50	488.00	3434.00
Scenario D				
ATT Coding	762.44	660.50	251.00	3019.00
ATC Coding	745.76	689.00	244.00	2572.00
Scenario E				
ATT Coding	1212.78	1088.00	470.00	4443.00
ATC Coding	1102.21	1037.50	449.00	5925.00
Scenario F				
ATT Coding	1500.39	1349.50	537.00	9904.00
ATC Coding	1338.18	1235.00	545.00	8261.00
Scenario G				
ATT Coding	956.14	925.00	423.00	2599.00
ATC Coding	1423.77	1154.00	310.00	6737.00
Scenario H				
ATT Coding	1347.55	1315.00	712.00	2650.00
ATC Coding	2986.81	2540.50	622.00	9997.00
Scenario I				
ATT Coding	1544.35	1508.50	851.00	2827.00
ATC Coding	4128.92	3452.00	954.00	9992.00

Note. Each summary statistic is averaged across 1,000 replications for each scenario. Each scenario represents a unique combination of treatment sample size and treatment to comparison group ratio as follows: scenario A ($N_{\text{Treatment}}$ = 200, T:C = 2:1), scenario B ($N_{\text{Treatment}}$ = 600, T:C = 2:1), scenario C ($N_{\text{Treatment}}$ = 1,000, T:C =

2:1), scenario D ($N_{Treatment}$ = 200, T:C = 4:3), scenario E ($N_{Treatment}$ = 600, T:C = 4:3), scenario F ($N_{Treatment}$ = 1,000, T:C = 4:3), scenario G ($N_{Treatment}$ = 200, T:C = 1:4), scenario H ($N_{Treatment}$ = 600, T:C = 1:4), and scenario I ($N_{Treatment}$ = 1,000, T:C = 1:4). For all scenarios, the maximum number of iterations allowed for generalized boosted modeling was 10,000.

Table 12

Average Cohen's D Estimated Treatment Effect, Average Bias in Estimated Treatment Effect, and Standard Errors by Propensity Score Method and Coding Method

	Cohen's D Y1				Cohen's D Y2				Cohen's D Y3				Cohen's D Y4			
Method	*M*	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*
							Scenario A									
ATT Coding																
Baseline	0.234	0.122	0.234	0.122	0.444	0.123	0.244	0.123	0.768	0.126	0.268	0.126	1.093	0.130	0.293	0.130
NN	0.232	0.142	0.232	0.142	0.443	0.143	0.243	0.143	0.769	0.147	0.269	0.147	1.095	0.152	0.295	0.152
NN Cal	0.000	0.176	**0.000**	0.176	0.212	0.176	**0.012**	0.176	0.540	0.179	**0.040**	0.179	0.868	0.183	**0.068**	0.183
GBM	0.045	0.133	**0.045**	0.133	0.198	0.142	**-0.002**	0.142	0.434	0.167	**-0.066**	0.167	0.671	0.200	-0.129	0.200
ATC Coding																
Baseline	-0.234	0.122	-0.234	0.122	-0.444	0.123	-0.244	0.123	-0.768	0.126	-0.268	0.126	-1.093	0.130	-0.293	0.130
NN	-0.106	0.141	-0.106	0.141	-0.317	0.142	-0.117	0.142	-0.644	0.145	-0.144	0.145	-0.971	0.151	-0.171	0.151
NN Cal	0.008	0.174	**0.008**	0.174	-0.205	0.174	**-0.005**	0.174	-0.534	0.177	**-0.034**	0.177	-0.862	0.182	**-0.062**	0.182
GBM	-0.044	0.124	**-0.044**	0.124	-0.210	0.130	**-0.010**	0.130	-0.467	0.143	**0.033**	0.143	-0.724	0.162	**0.076**	0.162
							Scenario B									
ATT Coding																
Baseline	0.238	0.071	0.238	0.071	0.448	0.072	0.248	0.072	0.772	0.074	0.272	0.074	1.097	0.077	0.297	0.077
NN	0.240	0.082	0.240	0.082	0.451	0.083	0.251	0.083	0.776	0.085	0.276	0.085	1.101	0.088	0.301	0.088
NN Cal	0.002	0.098	**0.002**	0.098	0.213	0.098	**0.013**	0.098	0.540	0.100	**0.040**	0.100	0.866	0.102	**0.066**	0.102
GBM	0.033	0.072	**0.033**	0.072	0.171	0.081	**-0.029**	0.081	0.384	0.103	-0.116	0.103	0.597	0.131	-0.203	0.131
ATC Coding																
Baseline	-0.238	0.071	-0.238	0.071	-0.448	0.072	-0.248	0.072	-0.772	0.074	-0.272	0.074	-1.097	0.077	-0.297	0.077
NN	-0.106	0.082	-0.106	0.082	-0.317	0.083	-0.117	0.083	-0.643	0.085	-0.143	0.085	-0.970	0.088	-0.170	0.088
NN Cal	0.001	0.097	**0.001**	0.097	-0.210	0.097	**-0.010**	0.097	-0.537	0.099	**-0.037**	0.099	-0.863	0.101	**-0.063**	0.101
GBM	-0.037	0.070	**-0.037**	0.070	-0.198	0.073	**0.002**	0.073	-0.447	0.082	**0.053**	0.082	-0.695	0.096	0.105	0.096

Table 12 Cont.

Method	Cohen's D Y1				Cohen's D Y2				Cohen's D Y3				Cohen's D Y4			
	M	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*
Scenario C																
ATT Coding																
Baseline	0.235	0.058	0.235	0.058	0.445	0.058	0.245	0.058	0.769	0.059	0.269	0.059	1.092	0.061	0.292	0.061
NN	0.235	0.068	0.235	0.068	0.445	0.068	0.245	0.068	0.769	0.070	0.269	0.070	1.093	0.072	0.293	0.072
NN Cal	0.007	0.076	**0.007**	0.076	0.218	0.076	**0.018**	0.076	0.544	0.077	**0.044**	0.077	0.871	0.078	**0.071**	0.078
GBM	0.031	0.055	**0.031**	0.055	0.164	0.063	**-0.036**	0.063	0.370	0.083	-0.130	0.083	0.575	0.108	-0.225	0.108
ATC Coding																
Baseline	-0.235	0.058	-0.235	0.058	-0.445	0.058	-0.245	0.058	-0.769	0.059	-0.269	0.059	-1.092	0.061	-0.292	0.061
NN	-0.102	0.066	-0.102	0.066	-0.313	0.066	-0.113	0.066	-0.639	0.067	-0.139	0.067	-0.964	0.069	-0.164	0.069
NN Cal	-0.005	0.075	**-0.005**	0.075	-0.216	0.075	**-0.016**	0.075	-0.543	0.076	**-0.043**	0.076	-0.869	0.078	**-0.069**	0.078
GBM	-0.030	0.055	**-0.030**	0.055	-0.187	0.058	**0.013**	0.058	-0.430	0.067	**0.070**	0.067	-0.673	0.079	0.127	0.079
Scenario D																
ATT Coding																
Baseline	0.227	0.108	0.227	0.108	0.438	0.109	0.238	0.109	0.762	0.111	0.262	0.111	1.087	0.115	0.287	0.115
NN	0.226	0.117	0.226	0.117	0.436	0.118	0.236	0.118	0.761	0.121	0.261	0.121	1.086	0.125	0.286	0.125
NN Cal	0.008	0.149	**0.008**	0.149	0.221	0.149	**0.021**	0.149	0.549	0.152	**0.049**	0.152	0.877	0.156	**0.077**	0.156
GBM	0.040	0.111	**0.040**	0.111	0.194	0.118	**-0.006**	0.118	0.433	0.137	**-0.067**	0.137	0.672	0.163	-0.128	0.163
ATC Coding																
Baseline	-0.227	0.108	-0.227	0.108	-0.438	0.109	-0.238	0.109	-0.762	0.111	-0.262	0.111	-1.087	0.115	-0.287	0.115
NN	-0.157	0.115	-0.157	0.115	-0.368	0.116	-0.168	0.116	-0.695	0.118	-0.195	0.118	-1.021	0.121	-0.221	0.121
NN Cal	-0.010	0.149	**-0.010**	0.149	-0.222	0.149	**-0.022**	0.149	-0.550	0.150	**-0.050**	0.150	-0.878	0.154	**-0.078**	0.154
GBM	-0.041	0.109	**-0.041**	0.109	-0.203	0.114	**-0.003**	0.114	-0.453	0.129	**0.047**	0.129	-0.703	0.150	**0.097**	0.150
Scenario E																
ATT Coding																
Baseline	0.232	0.061	0.232	0.061	0.442	0.061	0.242	0.061	0.766	0.063	0.266	0.063	1.090	0.065	0.290	0.065
NN	0.233	0.065	0.233	0.065	0.443	0.065	0.243	0.065	0.767	0.067	0.267	0.067	1.092	0.069	0.292	0.069
NN Cal	0.003	0.089	**0.003**	0.089	0.214	0.089	**0.014**	0.089	0.540	0.090	**0.040**	0.090	0.867	0.093	**0.067**	0.093
GBM	0.034	0.063	**0.034**	0.063	0.178	0.069	**-0.022**	0.069	0.399	0.085	-0.101	0.085	0.621	0.106	-0.179	0.106
ATC Coding																
Baseline	-0.232	0.061	-0.232	0.061	-0.442	0.061	-0.242	0.061	-0.766	0.063	-0.266	0.063	-1.090	0.065	-0.290	0.065
NN	-0.160	0.065	-0.160	0.065	-0.371	0.066	-0.171	0.066	-0.697	0.067	-0.197	0.067	-1.023	0.069	-0.223	0.069
NN Cal	-0.002	0.089	**-0.002**	0.089	-0.213	0.089	**-0.013**	0.089	-0.540	0.091	**-0.040**	0.091	-0.866	0.093	**-0.066**	0.093
GBM	-0.031	0.061	**-0.031**	0.061	-0.183	0.066	**0.017**	0.066	-0.418	0.079	**0.082**	0.079	-0.653	0.097	0.147	0.097

Table 12 Cont.

Method	Cohen's D Y1				Cohen's D Y2				Cohen's D Y3				Cohen's D Y4			
	M	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*
Scenario F																
ATT Coding																
Baseline	0.232	0.048	0.232	0.048	0.442	0.048	0.242	0.048	0.767	0.050	0.267	0.050	1.091	0.051	0.291	0.051
NN	0.233	0.051	0.233	0.051	0.443	0.052	0.243	0.052	0.767	0.053	0.267	0.053	1.092	0.055	0.292	0.055
NN Cal	0.007	0.066	**0.007**	0.066	0.218	0.066	**0.018**	0.066	0.544	0.067	**0.044**	0.067	0.870	0.069	**0.070**	0.069
GBM	0.029	0.049	**0.029**	0.049	0.168	0.054	**-0.032**	0.054	0.383	0.069	-0.117	0.069	0.599	0.088	-0.201	0.088
ATC Coding																
Baseline	-0.232	0.048	-0.232	0.048	-0.442	0.048	-0.242	0.048	-0.767	0.050	-0.267	0.050	-1.091	0.051	-0.291	0.051
NN	-0.162	0.050	-0.162	0.050	-0.372	0.051	-0.172	0.051	-0.698	0.052	-0.198	0.052	-1.024	0.054	-0.224	0.054
NN Cal	-0.005	0.067	**-0.005**	0.067	-0.216	0.067	**-0.016**	0.067	-0.543	0.068	**-0.043**	0.068	-0.869	0.070	**-0.069**	0.070
GBM	-0.030	0.047	**-0.030**	0.047	-0.179	0.052	**0.021**	0.052	-0.410	0.064	**0.090**	0.064	-0.641	0.080	0.159	0.080
Scenario G																
ATT Coding																
Baseline	0.252	0.077	0.252	0.077	0.462	0.077	0.262	0.077	0.785	0.078	0.285	0.078	1.109	0.080	0.309	0.080
NN	0.051	0.100	**0.051**	0.100	0.262	0.100	**0.062**	0.100	0.589	0.102	**0.089**	0.102	0.915	0.104	0.115	0.104
NN Cal	-0.002	0.113	**-0.002**	0.113	0.209	0.113	**0.009**	0.113	0.536	0.115	**0.036**	0.115	0.863	0.118	**0.063**	0.118
GBM	0.039	0.080	**0.039**	0.080	0.209	0.083	**0.009**	0.083	0.471	0.093	**-0.029**	0.093	0.734	0.106	**-0.066**	0.106
ATC Coding																
Baseline	-0.252	0.077	-0.252	0.077	-0.462	0.077	-0.262	0.077	-0.785	0.078	-0.285	0.078	-1.109	0.080	-0.309	0.080
NN	-0.256	0.096	-0.256	0.096	-0.466	0.097	-0.266	0.097	-0.790	0.100	-0.290	0.100	-1.115	0.105	-0.315	0.105
NN Cal	-0.004	0.110	**-0.004**	0.110	-0.215	0.110	**-0.015**	0.110	-0.541	0.111	**-0.041**	0.111	-0.867	0.114	**-0.067**	0.114
GBM	-0.036	0.084	**-0.036**	0.084	-0.161	0.095	**0.039**	0.095	-0.355	0.124	0.145	0.124	-0.549	0.159	0.251	0.159
Scenario H																
ATT Coding																
Baseline	0.253	0.044	0.253	0.044	0.462	0.044	0.262	0.044	0.786	0.045	0.286	0.045	1.110	0.046	0.310	0.046
NN	0.051	0.056	**0.051**	0.056	0.262	0.056	**0.062**	0.056	0.588	0.057	**0.088**	0.057	0.914	0.059	0.114	0.059
NN Cal	0.005	0.063	**0.005**	0.063	0.216	0.063	**0.016**	0.063	0.543	0.064	**0.043**	0.064	0.869	0.066	**0.069**	0.066
GBM	0.027	0.045	**0.027**	0.045	0.192	0.047	**-0.008**	0.047	0.447	0.054	**-0.053**	0.054	0.702	0.063	**-0.098**	0.063
ATC Coding																
Baseline	-0.253	0.044	-0.253	0.044	-0.462	0.044	-0.262	0.044	-0.786	0.045	-0.286	0.045	-1.110	0.046	-0.310	0.046
NN	-0.253	0.056	-0.253	0.056	-0.463	0.056	-0.263	0.056	-0.787	0.058	-0.287	0.058	-1.112	0.060	-0.312	0.060
NN Cal	-0.012	0.062	**-0.012**	0.062	-0.223	0.063	**-0.023**	0.063	-0.549	0.064	**-0.049**	0.064	-0.876	0.066	**-0.076**	0.066
GBM	-0.030	0.047	**-0.030**	0.047	-0.144	0.057	**0.056**	0.057	-0.321	0.080	0.179	0.080	-0.497	0.108	0.303	0.108

Table 12 Cont.

Method	Cohen's D Y1				Cohen's D Y2				Cohen's D Y3				Cohen's D Y4			
	M	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*	*M*	*SE*	*M* Bias	*SE*
							Scenario I									
ATT Coding																
Baseline	0.253	0.036	0.253	0.036	0.462	0.036	0.262	0.036	0.786	0.037	0.286	0.037	1.110	0.037	0.310	0.037
NN	0.049	0.046	**0.049**	0.046	0.260	0.046	**0.060**	0.046	0.586	0.047	**0.086**	0.047	0.912	0.048	0.112	0.048
NN Cal	0.001	0.048	**0.001**	0.048	0.212	0.048	**0.012**	0.048	0.538	0.049	**0.038**	0.049	0.864	0.050	**0.064**	0.050
GBM	0.021	0.036	**0.021**	0.036	0.183	0.038	**-0.017**	0.038	0.434	0.044	**-0.066**	0.044	0.684	0.052	-0.116	0.052
ATC Coding																
Baseline	-0.253	0.036	-0.253	0.036	-0.462	0.036	-0.262	0.036	-0.786	0.037	-0.286	0.037	-1.110	0.037	-0.310	0.037
NN	-0.253	0.046	-0.253	0.046	-0.463	0.046	-0.263	0.046	-0.787	0.047	-0.287	0.047	-1.112	0.048	-0.312	0.048
NN Cal	-0.007	0.048	**-0.007**	0.048	-0.218	0.048	**-0.018**	0.048	-0.544	0.049	**-0.044**	0.049	-0.870	0.050	**-0.070**	0.050
GBM	-0.025	0.037	**-0.025**	0.037	-0.134	0.046	**0.066**	0.046	-0.302	0.067	0.198	0.067	-0.470	0.093	0.330	0.093

Note. Each scenario represents a unique combination of treatment sample size and treatment to comparison group ratio as follows: scenario A ($N_{\text{Treatment}}$ = 200, T:C = 2:1), scenario B ($N_{\text{Treatment}}$ = 600, T:C = 2:1), scenario C ($N_{\text{Treatment}}$ = 1,000, T:C = 2:1), scenario D ($N_{\text{Treatment}}$ = 200, T:C = 4:3), scenario E ($N_{\text{Treatment}}$ = 600, T:C = 4:3), scenario F ($N_{\text{Treatment}}$ = 1,000, T:C = 4:3), scenario G ($N_{\text{Treatment}}$ = 200, T:C = 1:4), scenario H ($N_{\text{Treatment}}$ = 600, T:C = 1:4), and scenario I ($N_{\text{Treatment}}$ = 1,000, T:C = 1:4). Mean bias values ≤ |0.10| are bolded.

Table 13

ANOVA Results for Bias in the Estimated Treatment Effect

Source	*df*	SS	MS	F Value	*p*	Partial η^2
Method	**2**	**2491.966**	**1245.983**	**132231.000**	**<.001**	**0.551**
Coding	1	0.742	0.742	78.770	<.001	0.000
Ratio	2	18.461	9.230	979.580	<.001	0.009
Size	2	5.796	2.898	307.560	<.001	0.003
EffSize	3	16.278	5.426	575.830	<.001	0.008
Method*Coding	2	2.144	1.072	113.770	<.001	0.001
Method*Ratio	4	13.860	3.465	367.720	<.001	0.007
Method*Size	4	13.355	3.339	354.320	<.001	0.007
Method*EffSize	**6**	**466.586**	**77.764**	**8252.800**	**<.001**	**0.187**
Coding*Ratio	**2**	**48.995**	**24.497**	**2599.800**	**<.001**	**0.024**
Coding*Size	2	0.010	0.005	0.530	0.589	0.000
Coding*EffSize	3	0.641	0.214	22.680	<.001	0.000
Ratio*Size	4	0.174	0.043	4.610	0.001	0.000
Ratio*EffSize	6	1.392	0.232	24.620	<.001	0.001
Size*EffSize	6	2.259	0.377	39.960	<.001	0.001
Method*Coding*Ratio	**4**	**384.535**	**96.134**	**10202.300**	**<.001**	**0.159**
Method*Coding*Size	4	0.164	0.041	4.360	0.002	0.000
Method*Coding*EffSize	6	1.412	0.235	24.980	<.001	0.001
Method*Ratio*Size	8	0.864	0.108	11.460	<.001	0.000
Method*Ratio*EffSize	12	2.359	0.197	20.860	<.001	0.001
Method*Size*EffSize	12	3.377	0.281	29.860	<.001	0.002
Coding*Ratio*Size	4	0.247	0.062	6.550	<.001	0.000
Coding*Ratio*EffSize	6	14.391	2.399	254.550	<.001	0.007
Coding*Size*EffSize	6	0.011	0.002	0.190	0.981	0.000
Ratio*Size*EffSize	12	0.034	0.003	0.300	0.990	0.000
Method*Coding*Ratio*Size	8	0.542	0.068	7.190	<.001	0.000
Method*Coding*Ratio*EffSize	12	25.852	2.154	228.630	<.001	0.013
Method*Coding*Size*EffSize	12	0.014	0.001	0.130	1.000	0.000
Method*Ratio*Size*EffSize	24	0.013	0.001	0.060	1.000	0.000
Coding*Ratio*Size*EffSize	12	0.160	0.013	1.410	0.152	0.000
Method*Coding*Ratio*Size*EffSize	24	0.369	0.015	1.630	0.027	0.000
Method*Coding*Ratio*Size*EffSize	24	0.981	0.041	7.400	<.001	0.001

Note. Method refers to propensity score method, coding refers to ATT or ATC coding, size refers to treatment sample size, and effSize refers to true treatment effect size. Effects that were statistically significant and meaningful (partial $\eta^2 \geq 0.02$) are bolded.

Table 14

Bias Means, Standard deviations, and Differences for ATC Coding and ATT Coding by Treatment to Comparison Ratio for Each Propensity Score Method

	ATC Coding		ATT Coding				
Ratio	*M*	*SD*	*M*	*SD*	Difference	*p*	Cohen's *d*
NN							
1:4	0.280	0.075	0.078	0.076	0.201	<.001	1.878
2:1	0.133	0.107	0.263	0.108	-0.130	<.001	-0.855
4:3	0.187	0.087	0.257	0.088	-0.069	<.001	-0.560
GBM							
1:4	-0.123	0.153	-0.030	0.081	-0.093	<.001	-0.537
2:1	-0.030	0.115	-0.069	0.148	0.039	<.001	0.209
4:3	-0.046	0.114	-0.062	0.126	0.016	<.001	0.095
NN Cal							
1:4	0.036	0.083	0.029	0.085	0.006	<.001	0.053
2:1	0.028	0.127	0.032	0.129	-0.004	0.007	-0.021
4:3	0.035	0.112	0.035	0.112	0.000	0.919	-0.001

Note. Because the direction of bias differed consistently across ATT and ATC coding, the sign was reversed for all bias values for ATC coding prior to conducting the ANOVA.

Table 15

Bias Means, Standard deviations, and Differences for Nearest Neighbor Matching, Nearest Neighbor Matching with a Caliper, and Generalized Boosted Modeling by True Treatment Effect Size

True Treatment	NN		NN Cal		GBM		NN with NN Cal			NN with GBM			NN Cal with GBM		
Effect Size	*M*	*SD*	*M*	*SD*	*M*	*SD*	Difference	*p*	Cohen's *d*	Difference	*p*	Cohen's *d*	Difference	*p*	Cohen's *d*
None (0)	0.173	0.114	0.004	0.105	0.033	0.077	0.169	<.001	1.090	0.139	<.001	1.012	-0.030	<.001	-0.229
Small (0.20)	0.183	0.114	0.015	0.105	-0.019	0.085	0.168	<.001	1.083	0.202	<.001	1.420	0.034	<.001	0.253
Medium (0.50)	0.208	0.115	0.042	0.107	-0.091	0.109	0.167	<.001	1.062	0.300	<.001	1.890	0.133	<.001	0.873
Large (0.80)	0.234	0.117	0.069	0.109	-0.163	0.140	0.165	<.001	1.031	0.397	<.001	2.173	0.232	<.001	1.303

Note. Due to the negative average bias values for generalized boosted modeling (for small, medium, and large true treatment effect sizes), the magnitude of the differences (absolute value) in bias between generalized boosted modeling and nearest neighbor matching or nearest neighbor matching with a caliper are slightly inflated. Comparison of the magnitude (absolute value) of differences between nearest neighbor matching and generalized boosted modeling resulted in $\text{Difference}_{\text{Small}} = 0.164$, Cohen's $d_{\text{Small}} = 1.153$, $\text{Difference}_{\text{Medium}} = 0.117$, Cohen's $d_{\text{Medium}} = 0.741$, and $\text{Difference}_{\text{Large}} = 0.071$, Cohen's $d_{\text{Large}} = 0.388$. Comparison of the magnitude (absolute value) of differences between nearest neighbor matching with a caliper and generalized boosted modeling resulted in $\text{Difference}_{\text{Small}} = -0.004$, Cohen's $d_{\text{Small}} = -0.029$, $\text{Difference}_{\text{Medium}} = -0.049$, Cohen's $d_{\text{Medium}} = -0.322$, and $\text{Difference}_{\text{Large}} = -0.094$, Cohen's $d_{\text{Large}} = -0.530$.

Table 16

Standardized Mean Differences and Percentage in Bias Reduction for Covariates and Estimated Propensity Scores

	X1		X2		X3		X4		X5		Propensity Score		
Method	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	Variance Ratio
Scenario A													
ATT Coding													
Baseline	**-0.03**	-	0.18	-	0.48	-	0.88	-	1.19	-	1.66	-	0.64
NN	**-0.03**	-12.61	0.18	-6.27	0.48	0.89	0.88	-0.23	1.19	0.03	1.66	-0.07	0.64
NN Cal	**0.00**	25.60	**0.01**	64.12	**0.01**	**87.06**	**0.02**	**94.34**	**0.03**	**96.92**	**0.03**	**98.39**	1.04
GBM	**-0.09**	0.00	**-0.01**	65.31	0.13	72.05	0.28	67.81	0.40	66.25	-	-	-
ATC Coding													
Baseline	**0.03**	-	-0.18	-	-0.49	-	-0.91	-	-1.26	-	1.32	-	0.64
NN	**0.01**	32.93	**-0.08**	48.29	-0.22	55.02	-0.41	55.01	-0.56	55.29	0.74	44.40	0.57
NN Cal	**0.01**	22.20	**0.00**	64.78	**0.00**	**87.89**	**0.00**	**94.48**	**-0.01**	**97.53**	**0.01**	**99.18**	0.98
GBM	**0.06**	25.00	**0.00**	73.75	-0.12	74.41	-0.25	72.22	-0.37	70.53	-	-	-
Scenario B													
ATT Coding													
Baseline	**-0.02**	-	0.17	-	0.47	-	0.87	-	1.18	-	1.60	-	0.64
NN	**-0.03**	-13.15	0.17	0.07	0.47	0.54	0.87	0.10	1.18	-0.07	1.60	-0.01	0.64
NN Cal	**0.00**	33.86	**0.01**	79.05	**0.01**	**92.84**	**0.02**	**96.67**	**0.03**	**97.64**	**0.03**	**98.33**	1.04
GBM	**-0.06**	-1.74	**-0.01**	**82.87**	0.11	76.69	0.22	74.07	0.31	73.08	-	-	-
ATC Coding													
Baseline	**0.02**	-	-0.17	-	-0.48	-	-0.90	-	-1.25	-	1.28	-	0.64
NN	**0.01**	39.66	**-0.08**	54.94	-0.21	55.68	-0.39	56.21	-0.55	56.24	0.69	45.72	0.55
NN Cal	**0.00**	28.57	**0.00**	79.38	**0.00**	**92.85**	**-0.01**	**96.88**	**-0.01**	**98.50**	**0.01**	**99.21**	0.98
GBM	**0.04**	28.57	**0.00**	**88.06**	**-0.10**	79.12	-0.20	77.92	-0.28	77.95	-	-	-

Table 16 Cont.

	X1		X2		X3		X4		X5		Propensity Score		
Method	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	Variance Ratio
						Scenario C							
ATT Coding													
Baseline	**-0.03**	-	0.17	-	0.47	-	0.87	-	1.18	-	1.59	-	0.64
NN	**-0.03**	-8.42	0.18	-1.37	0.47	-0.34	0.87	0.10	1.18	-0.22	1.59	-0.16	0.64
NN Cal	**0.00**	30.14	**0.00**	**84.92**	**0.01**	**94.72**	**0.02**	**97.28**	**0.03**	**97.67**	**0.03**	**98.32**	1.04
GBM	**-0.05**	-15.52	**0.00**	**87.76**	**0.10**	78.44	0.20	76.23	0.28	76.03	-	-	-
ATC Coding													
Baseline	**0.03**	-	-0.17	-	-0.48	-	-0.90	-	-1.25	-	1.27	-	0.64
NN	**0.01**	38.64	**-0.08**	55.81	-0.21	56.79	-0.39	56.68	-0.54	56.68	0.68	46.35	0.55
NN Cal	**0.00**	36.45	**0.00**	**83.92**	**0.00**	**94.58**	**-0.01**	**97.74**	**-0.01**	**98.77**	**0.01**	**99.21**	0.97
GBM	**0.03**	31.37	**0.00**	**91.52**	**-0.09**	**81.84**	-0.17	**81.28**	-0.23	**81.44**	-	-	-
						Scenario D							
ATT Coding													
Baseline	**-0.02**	-	0.17	-	0.47	-	0.87	-	1.18	-	1.51	-	0.84
NN	**-0.02**	-3.87	0.17	0.38	0.46	0.61	0.87	0.41	1.18	0.29	1.50	0.31	0.84
NN Cal	**0.00**	29.75	**0.00**	66.04	**0.01**	**89.53**	**0.01**	**95.00**	**0.02**	**97.27**	**0.02**	**98.52**	1.04
GBM	**-0.07**	7.67	**-0.01**	74.20	0.12	74.35	0.25	70.83	0.36	69.36	-	-	-
ATC Coding													
Baseline	**0.02**	-	-0.17	-	-0.47	-	-0.88	-	-1.21	-	1.37	-	0.84
NN	**0.01**	19.16	-0.12	26.73	-0.33	29.74	-0.61	30.36	-0.84	30.71	1.07	22.45	0.73
NN Cal	**0.00**	26.49	**0.00**	66.93	**0.00**	**89.78**	**-0.01**	**95.08**	**-0.01**	**97.64**	**0.02**	**98.95**	0.97
GBM	**0.06**	21.57	**0.00**	74.11	-0.12	75.31	-0.25	71.52	-0.35	70.69	-	-	-

Table 16 Cont.

	X1		X2		X3		X4		X5		Propensity Score		
Method	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	Variance Ratio
						Scenario E							
ATT Coding													
Baseline	**-0.02**	-	0.17	-	0.47	-	0.86	-	1.18	-	1.49	-	0.83
NN	**-0.02**	-4.72	0.17	-0.14	0.47	-0.01	0.86	0.18	1.18	0.12	1.49	0.16	0.83
NN Cal	**0.00**	32.58	**0.00**	**80.73**	**0.01**	**93.74**	**0.02**	**96.90**	**0.02**	**98.13**	**0.02**	**98.50**	1.04
GBM	**-0.05**	2.67	**0.00**	**85.77**	**0.10**	77.92	0.20	76.32	0.29	75.77	-	-	-
ATC Coding													
Baseline	**0.02**	-	-0.17	-	-0.47	-	-0.87	-	-1.20	-	1.36	-	0.83
NN	**0.01**	21.29	-0.12	30.33	-0.32	30.89	-0.60	31.14	-0.83	31.25	1.04	23.18	0.72
NN Cal	**0.00**	30.32	**0.00**	**80.39**	**-0.01**	**93.91**	**-0.01**	**97.23**	**-0.01**	**98.47**	**0.01**	**98.90**	0.97
GBM	**0.04**	27.27	**0.00**	**89.05**	**-0.09**	**80.19**	-0.19	78.01	-0.26	77.83	-	-	-
						Scenario F							
ATT Coding													
Baseline	**-0.02**	-	0.17	-	0.47	-	0.87	-	1.18	-	1.48	-	0.83
NN	**-0.02**	-6.14	0.17	-0.17	0.47	0.05	0.86	0.09	1.18	0.04	1.48	-0.01	0.83
NN Cal	**0.00**	36.76	**0.00**	**85.68**	**0.01**	**95.13**	**0.01**	**97.61**	**0.02**	**98.22**	**0.02**	**98.50**	1.04
GBM	**-0.04**	0.00	**0.00**	**91.20**	**0.09**	**80.40**	0.18	79.06	0.25	78.65	-	-	-
ATC Coding													
Baseline	**0.02**	-	-0.17	-	-0.47	-	-0.87	-	-1.20	-	1.35	-	0.83
NN	**0.01**	22.57	-0.12	29.93	-0.32	30.77	-0.60	30.88	-0.83	30.96	1.04	22.96	0.71
NN Cal	**0.00**	40.11	**0.00**	**86.36**	**0.00**	**95.64**	**-0.01**	**97.84**	**-0.01**	**98.73**	**0.01**	**98.90**	0.97
GBM	**0.03**	26.56	**0.00**	**92.31**	**-0.08**	**81.80**	-0.17	**80.63**	-0.23	**80.82**	-	-	-

Table 16 Cont.

	X1		X2		X3		X4		X5		Propensity Score		
Method	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	Variance Ratio
						Scenario G							
ATT Coding													
Baseline	**-0.02**	-	0.19	-	0.51	-	0.98	-	1.36	-	1.17	-	2.52
NN	**-0.01**	34.88	**0.04**	72.56	**0.10**	79.66	0.19	**80.17**	0.27	**80.24**	0.36	69.67	1.90
NN Cal	**0.00**	25.55	**0.00**	75.96	**0.00**	**92.09**	**0.00**	**96.49**	**0.00**	**98.41**	**0.00**	**99.63**	1.01
GBM	**-0.04**	35.00	**0.00**	**89.40**	**0.10**	**80.07**	0.20	78.94	0.29	79.13	-	-	-
ATC Coding													
Baseline	**0.02**	-	-0.19	-	-0.50	-	-0.92	-	-1.23	-	1.86	-	2.52
NN	**0.02**	-25.74	-0.19	-2.19	-0.51	0.06	-0.93	-0.19	-1.23	0.00	1.86	-0.10	2.59
NN Cal	**0.00**	21.98	**0.00**	76.58	**-0.02**	**91.61**	**-0.03**	**95.74**	**-0.04**	**96.82**	**0.03**	**98.28**	0.97
GBM	**0.08**	-25.90	**0.02**	77.22	-0.11	77.67	-0.25	71.70	-0.36	70.26	-	-	-
						Scenario H							
ATT Coding													
Baseline	**-0.02**	-	0.18	-	0.51	-	0.97	-	1.35	-	1.16	-	2.48
NN	**0.00**	42.31	**0.03**	79.61	**0.10**	**80.84**	0.18	**80.84**	0.26	**80.84**	0.34	70.50	1.88
NN Cal	**0.00**	28.38	**0.00**	**86.87**	**0.00**	**95.42**	**0.00**	**98.06**	**0.00**	**99.06**	**0.00**	**99.64**	1.01
GBM	**-0.02**	49.18	**0.00**	**95.79**	**0.08**	**84.94**	0.14	**85.46**	0.19	**85.89**	-	-	-
ATC Coding													
Baseline	**0.02**	-	-0.18	-	-0.50	-	-0.92	-	-1.22	-	1.83	-	2.48
NN	**0.03**	-22.11	-0.18	0.19	-0.50	0.06	-0.92	0.02	-1.23	-0.02	1.83	0.02	2.49
NN Cal	**0.00**	31.82	**-0.01**	**86.47**	**-0.02**	**94.83**	**-0.03**	**96.73**	**-0.04**	**96.93**	**0.03**	**98.24**	0.96
GBM	**0.07**	-75.00	**0.01**	**86.89**	**-0.10**	79.48	-0.21	76.33	-0.30	75.03	-	-	-

Table 16 Cont.

	X1		X2		X3		X4		X5		Propensity Score		
Method	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	SMD	Median PBR	Variance Ratio
						Scenario I							
ATT Coding													
Baseline	**-0.02**	-	0.19	-	0.52	-	0.97	-	1.36	-	1.16	-	2.49
NN	**0.00**	44.95	**0.04**	**80.79**	**0.10**	**81.22**	0.19	**80.80**	0.26	**80.77**	0.34	70.40	1.89
NN Cal	**0.00**	41.28	**0.00**	**89.58**	**0.00**	**96.78**	**0.00**	**98.49**	**0.00**	**99.29**	**0.00**	**99.64**	1.01
GBM	**-0.01**	57.66	**0.01**	**96.61**	**0.07**	**87.17**	0.12	**88.01**	0.16	**88.28**	-	-	-
ATC Coding													
Baseline	**0.02**	-	-0.19	-	-0.51	-	-0.92	-	-1.23	-	1.83	-	2.49
NN	**0.02**	-23.11	-0.19	0.90	-0.51	0.03	-0.92	-0.05	-1.23	-0.08	1.83	-0.01	2.51
NN Cal	**0.00**	37.60	**-0.01**	**90.28**	**-0.02**	**95.95**	**-0.03**	**96.93**	**-0.04**	**96.97**	**0.03**	**98.24**	0.96
GBM	**0.06**	-82.66	**0.00**	**90.30**	**-0.09**	**81.03**	-0.20	77.73	-0.28	76.93	-	-	-

Note. Each scenario represents a unique combination of treatment sample size and treatment to comparison group ratio as follows: scenario A ($N_{\text{Treatment}}$ = 200, T:C = 2:1), scenario B ($N_{\text{Treatment}}$ = 600, T:C = 2:1), scenario C ($N_{\text{Treatment}}$ = 1,000, T:C = 2:1), scenario D ($N_{\text{Treatment}}$ = 200, T:C = 4:3), scenario E ($N_{\text{Treatment}}$ = 600, T:C = 4:3), scenario F ($N_{\text{Treatment}}$ = 1,000, T:C = 4:3), scenario G ($N_{\text{Treatment}}$ = 200, T:C = 1:4), scenario H ($N_{\text{Treatment}}$ = 600, T:C = 1:4), and scenario I ($N_{\text{Treatment}}$ = 1,000, T:C = 1:4). Each covariate (X1 through X5) represented a different magnitude of standardized mean difference at baseline (i.e., baseline imbalance) as follows: X1 = 0, X2 = 0.20, X3 = 0.50, X4 = 0.80, X5 = 1.20. Standardized mean differences, percentages in bias reduction, and propensity score variance ratios are not provided for generalized boosted modeling (for all scenarios) because outcome variables are weighted by the propensity score based on group membership (i.e., treatment or comparison group). Thus, there was no change

to the sample that was used in the outcome analysis, and no change in the group balance on the propensity score or propensity score variance ratio over that at baseline. SMD values ≤ |0.10| and PBR values ≥ 80.00 are bolded.

Table 17

Baseline, Matched, and Unmatched Treatment and Comparison Group Sizes and Average Propensity Scores by Matching Method and Coding Method

	Baseline Treatment		Baseline Comparison		Matched Treatment		Unmatched Treatment		Matched Comparison		Unmatched Comparison		
Method	*N*	Mean PS	*N*	Mean PS	*N*	Mean PS	*N*	Mean PS	*N*	Mean PS	*N*	Mean PS	Treatment Loss
						Scenario A							
ATT Coding													
NN	200.378	0.781	99.622	0.441	99.622	0.781	100.756	0.781	99.622	0.441	0.000	-	50.28%
NN Cal	200.378	0.781	99.622	0.441	64.290	0.582	136.088	0.873	64.290	0.576	35.332	0.203	67.92%
ATC Coding													
NN	200.378	0.219	99.622	0.559	99.622	0.368	100.756	0.070	99.622	0.559	0.000	-	50.28%
NN Cal	200.378	0.219	99.622	0.559	64.316	0.423	136.062	0.125	64.316	0.425	35.306	0.794	67.90%
						Scenario B							
ATT Coding													
NN	600.219	0.776	299.781	0.449	299.781	0.776	300.438	0.776	299.781	0.449	0.000	-	50.05%
NN Cal	600.219	0.776	299.781	0.449	199.505	0.580	400.714	0.873	199.505	0.575	100.276	0.201	66.76%
ATC Coding													
NN	600.219	0.224	299.781	0.551	299.781	0.373	300.438	0.075	299.781	0.551	0.000	-	50.05%
NN Cal	600.219	0.224	299.781	0.551	199.445	0.425	400.774	0.125	199.445	0.427	100.336	0.796	66.77%
						Scenario C							
ATT Coding													
NN	1000.800	0.775	499.196	0.452	499.196	0.775	501.604	0.775	499.196	0.452	0.000	-	50.12%
NN Cal	1000.800	0.775	499.196	0.452	333.474	0.581	667.326	0.873	333.474	0.575	165.722	0.198	66.68%
ATC Coding													
NN	1000.800	0.225	499.196	0.548	499.196	0.375	501.604	0.076	499.196	0.548	0.000	-	50.12%
NN Cal	1000.800	0.225	499.196	0.548	333.382	0.424	667.418	0.125	333.382	0.426	165.814	0.798	66.69%

Table 17 Cont.

Method	Baseline Treatment *N*	Baseline Treatment Mean PS	Baseline Comparison *N*	Baseline Comparison Mean PS	Matched Treatment *N*	Matched Treatment Mean PS	Unmatched Treatment *N*	Unmatched Treatment Mean PS	Matched Comparison *N*	Matched Comparison Mean PS	Unmatched Comparison *N*	Unmatched Comparison Mean PS	Treatment Loss
Scenario D													
ATT Coding													
NN	199.827	0.717	150.173	0.377	150.163	0.716	49.664	0.719	150.163	0.377	0.010	-	24.85%
NN Cal	199.827	0.717	150.173	0.377	83.825	0.536	116.002	0.848	83.825	0.531	66.348	0.018	58.05%
ATC Coding													
NN	199.827	0.283	150.173	0.623	150.163	0.358	49.664	0.052	150.163	0.623	0.010	-	24.85%
NN Cal	199.827	0.283	150.173	0.623	83.824	0.468	116.003	0.149	83.824	0.471	66.349	0.560	58.05%
Scenario E													
ATT Coding													
NN	600.046	0.715	449.954	0.381	449.954	0.714	150.092	0.715	449.954	0.381	0.000	-	25.01%
NN Cal	600.046	0.715	449.954	0.381	258.109	0.536	341.937	0.848	258.109	0.531	191.845	0.180	56.99%
ATC Coding													
NN	600.046	0.285	449.954	0.619	449.954	0.362	150.092	0.054	449.954	0.619	0.000	-	25.01%
NN Cal	600.046	0.285	449.954	0.619	257.973	0.468	342.073	0.149	257.973	0.472	191.981	0.817	57.01%
Scenario F													
ATT Coding													
NN	999.427	0.714	750.573	0.381	750.573	0.714	248.854	0.714	750.573	0.381	0.000	-	24.90%
NN Cal	999.427	0.714	750.573	0.381	431.941	0.536	567.486	0.849	431.941	0.531	318.632	0.179	56.78%
ATC Coding													
NN	999.427	0.286	750.573	0.619	750.573	0.363	248.854	0.054	750.573	0.619	0.000	-	24.90%
NN Cal	999.427	0.286	750.573	0.619	431.833	0.468	567.594	0.148	431.833	0.472	318.740	0.817	56.79%
Scenario G													
ATT Coding													
NN	200.040	0.441	799.960	0.140	200.040	0.441	0.000	-	200.040	0.348	599.920	0.070	0.00%
NN Cal	200.040	0.441	799.960	0.140	157.791	0.352	42.249	0.767	157.791	0.351	642.169	0.088	21.12%
ATC Coding													
NN	200.040	0.559	799.960	0.860	200.040	0.559	0.000	-	200.040	0.861	599.920	0.860	0.00%
NN Cal	200.040	0.559	799.960	0.860	157.868	0.649	42.172	0.230	157.868	0.654	642.092	0.911	21.08%

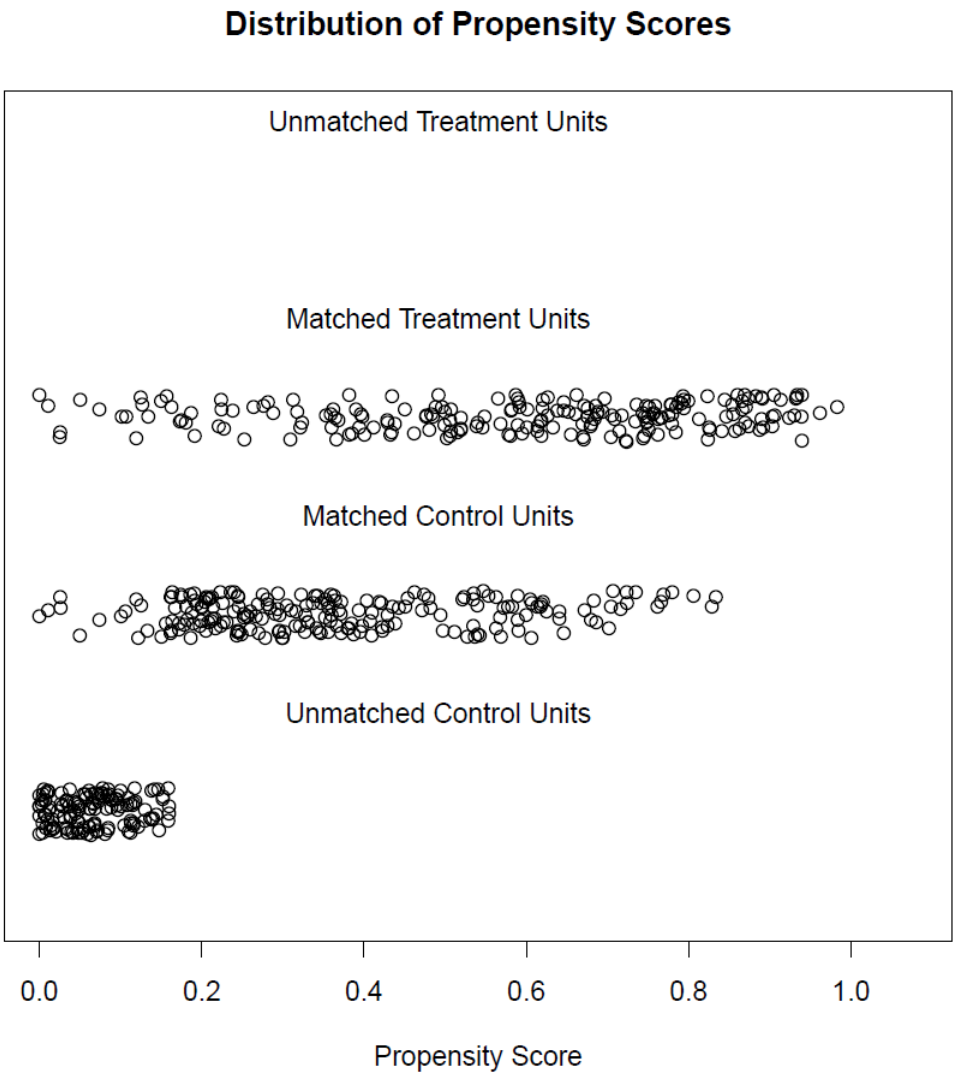
Table 17 Cont.

Method	Baseline Treatment *N*	Baseline Treatment Mean PS	Baseline Comparison *N*	Baseline Comparison Mean PS	Matched Treatment *N*	Matched Treatment Mean PS	Unmatched Treatment *N*	Unmatched Treatment Mean PS	Matched Comparison *N*	Matched Comparison Mean PS	Unmatched Comparison *N*	Unmatched Comparison Mean PS	Treatment Loss
Scenario H													
ATT Coding													
NN	599.374	0.436	2400.630	0.141	599.374	0.436	0.000	-	599.374	0.349	1801.256	0.071	0.00%
NN Cal	599.374	0.436	2400.630	0.141	481.147	0.354	118.227	0.775	481.147	0.353	1919.483	0.088	19.73%
ATC Coding													
NN	599.374	0.564	2400.630	0.859	599.374	0.564	0.000	-	599.374	0.859	1801.256	0.859	0.00%
NN Cal	599.374	0.564	2400.630	0.859	481.205	0.647	118.169	0.222	481.205	0.652	1919.425	0.911	19.72%
Scenario I													
ATT Coding													
NN	997.619	0.437	4002.380	0.140	997.619	0.437	0.000	-	997.619	0.349	3004.761	0.071	0.00%
NN Cal	997.619	0.437	4002.380	0.140	803.472	0.354	194.147	0.775	803.472	0.353	3198.908	0.087	19.46%
ATC Coding													
NN	997.619	0.563	4002.380	0.860	997.619	0.563	0.000	-	997.619	0.860	3004.761	0.860	0.00%
NN Cal	997.619	0.563	4002.380	0.860	803.735	0.647	193.884	0.221	803.735	0.652	3198.645	0.911	19.43%

Note. Each scenario represents a unique combination of treatment sample size and treatment to comparison group ratio as follows: scenario A ($N_{\text{Treatment}}$ = 200, T:C = 2:1), scenario B ($N_{\text{Treatment}}$ = 600, T:C = 2:1), scenario C ($N_{\text{Treatment}}$ = 1,000, T:C = 2:1), scenario D ($N_{\text{Treatment}}$ = 200, T:C = 4:3), scenario E ($N_{\text{Treatment}}$ = 600, T:C = 4:3), scenario F ($N_{\text{Treatment}}$ = 1,000, T:C = 4:3), scenario G ($N_{\text{Treatment}}$ = 200, T:C = 1:4), scenario H ($N_{\text{Treatment}}$ = 600, T:C = 1:4), and scenario I ($N_{\text{Treatment}}$ = 1,000, T:C = 1:4). Scenario D shows a fraction of an unmatched comparison group member after nearest neighbor matching, for both ATT and ATC coding. One replication (out of 1,000) in scenario D had a comparison group that was larger than the treatment group. Thus, for the one replication, there were a few unmatched comparison group members.

Figure 1

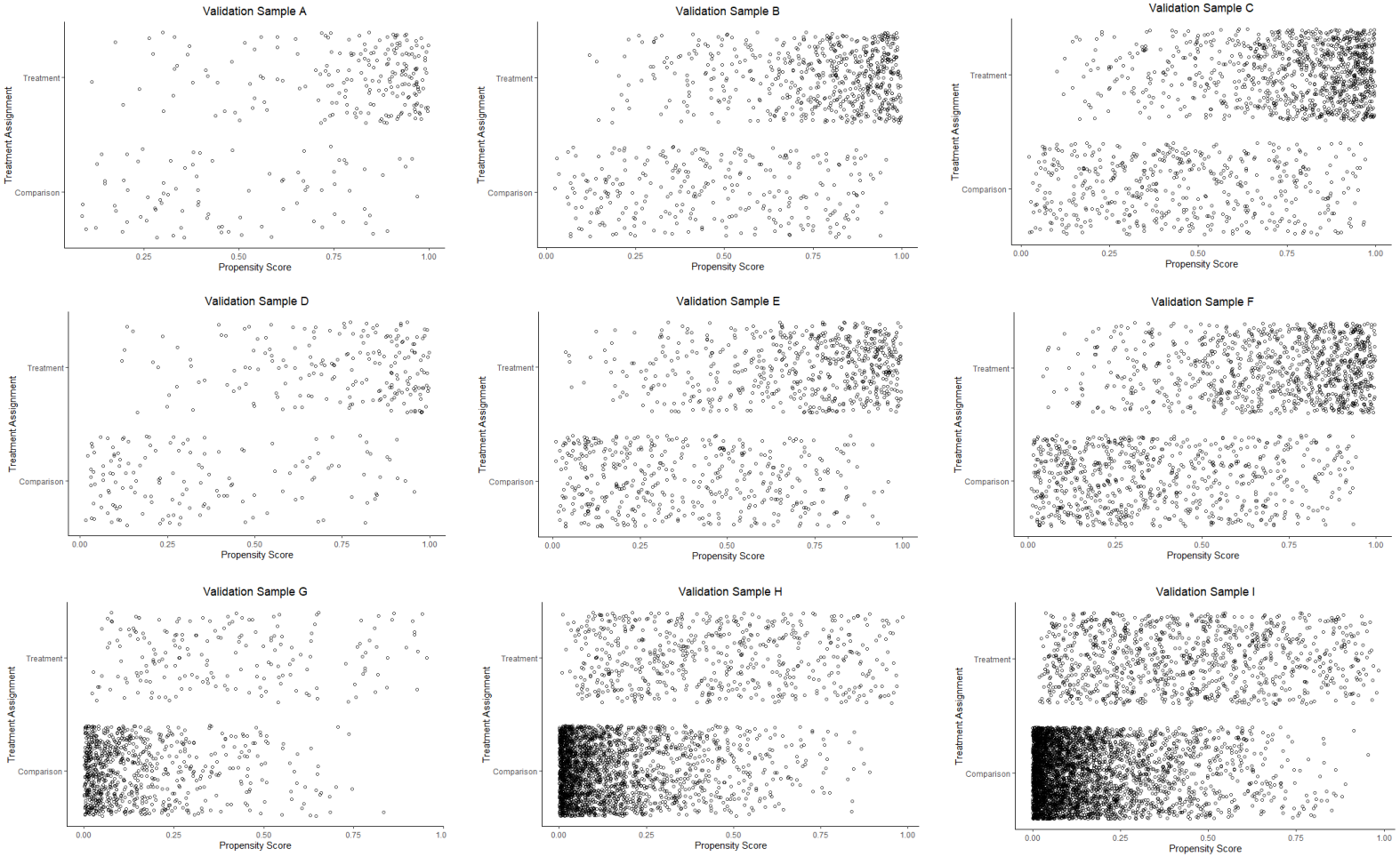
Example Jitter Plot after Matching on the Propensity Score



Note. Example jitter plot produced after matching treatment and comparison group members on the propensity score (Perkins & Horst, 2020). Jitter plots allow for the examination of the propensity score distributions of each group.

Figure 2

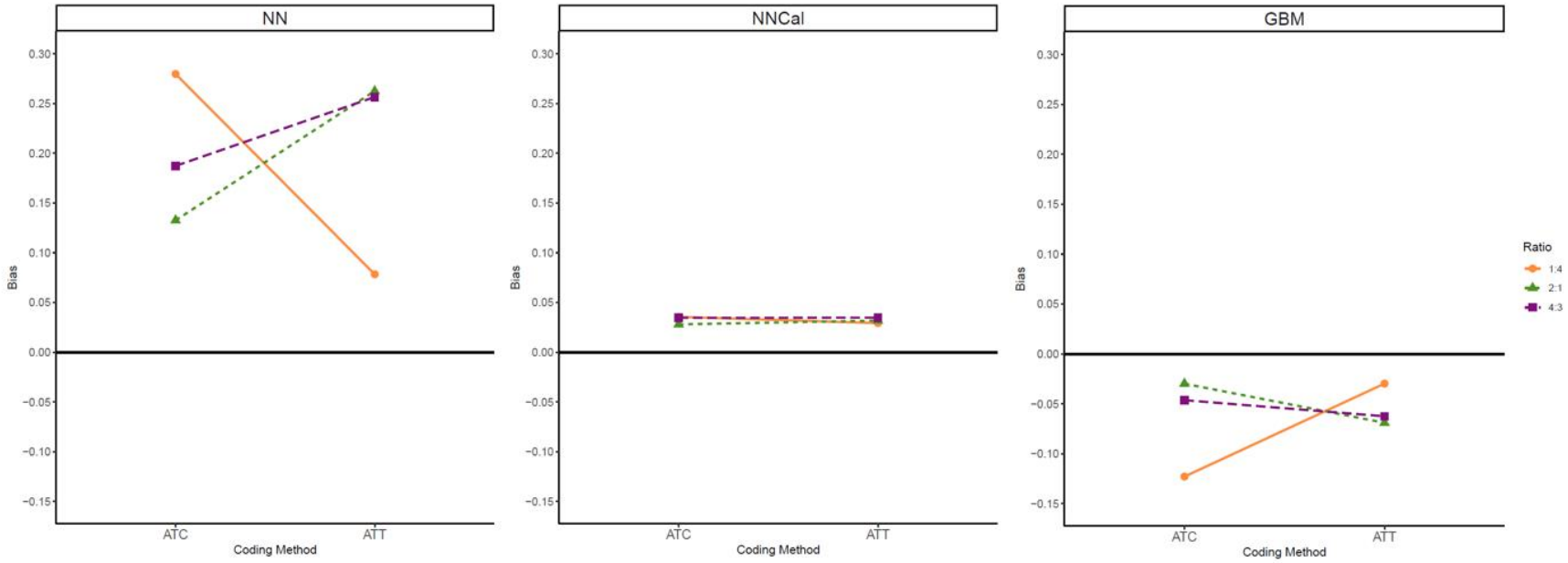
Jitter Plots Demonstrating Group Propensity Score Distributions Prior to Matching or Weighting



Note. Jitter plots demonstrating group propensity score distributions prior to matching or weighting. Samples A through C have a treatment to comparison ratio of 2:1 (treatment N = 200, 600, and 1,000 from left to right). Samples D through F have a treatment to comparison ration of 4:3 (treatment N = 200, 600, and 1,000 from left to right). Samples G through I have a treatment to comparison ratio of 1:4 (treatment N = 200, 600, and 1,000 from left to right). All validation samples have adequate common support between treatment and comparison group propensity score distributions.

Figure 3

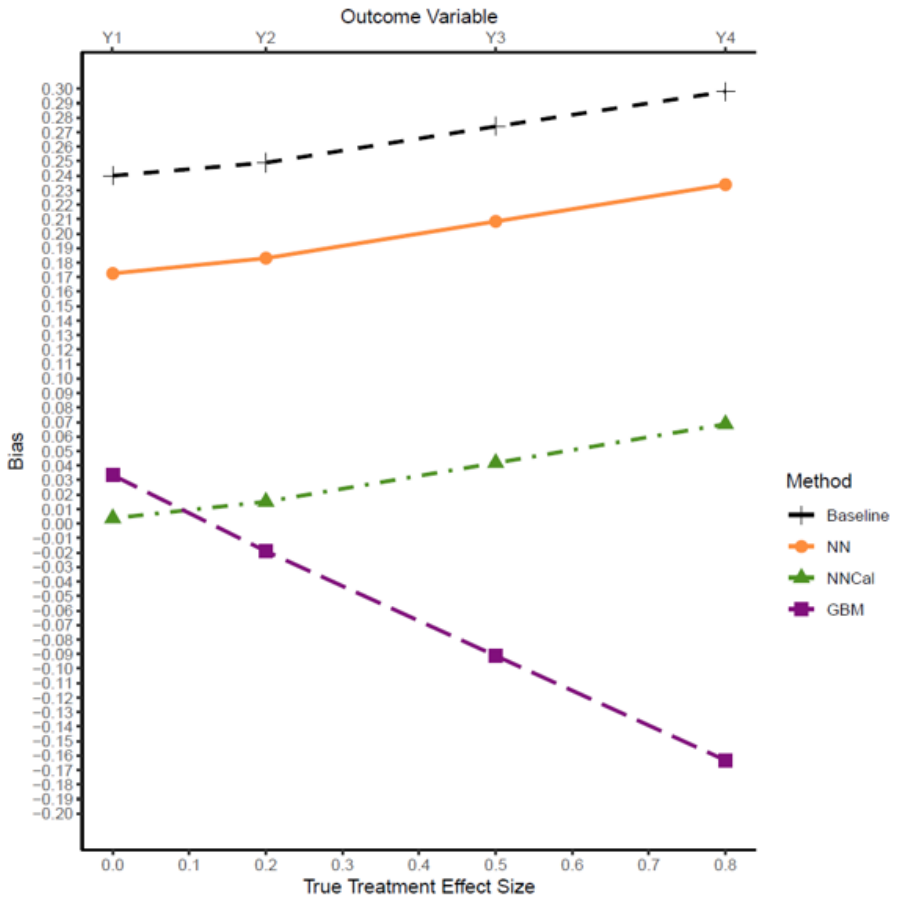
Average Bias for the Interaction between Coding Method and Treatment to Comparison Ratio for Each Propensity Score Method



Note. Line graph of the interaction between coding method and treatment to comparison ratio for each propensity score method. A 1:4 treatment to comparison ratio is indicated by orange, circles, and a solid line. A 2:1 treatment to comparison ratio is indicated by green, triangles, and a short, dashed line. A 4:3 treatment to comparison ratio is indicated by purple, squares, and a long, dashed line. The solid black line indicates bias of zero (the ideal average bias over a large number of replications).

Figure 4

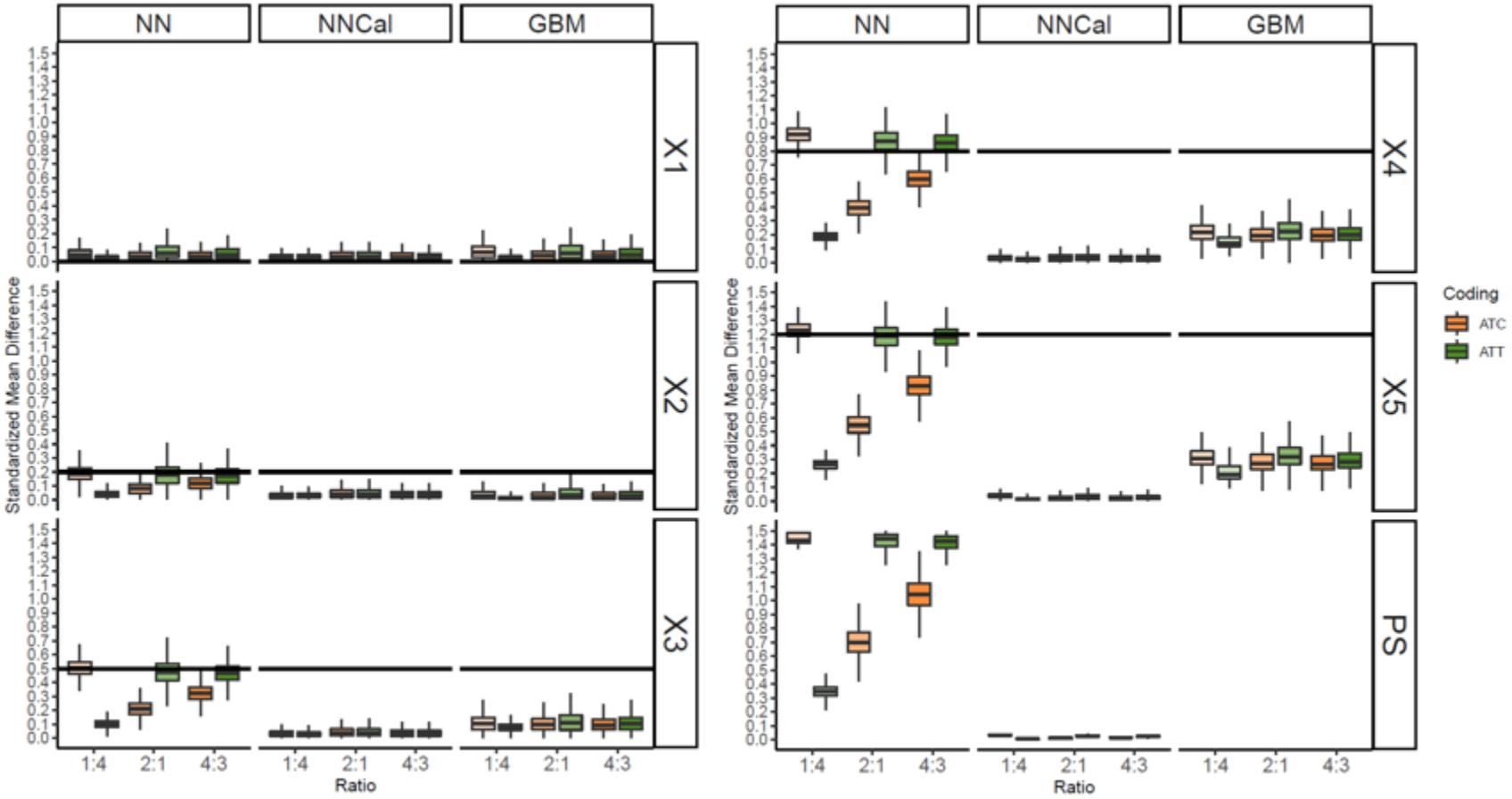
Average Bias for the Interaction between Propensity Score Method and True Treatment Effect Size



Note. Line graph of the average magnitude of bias (absolute value) in the estimated treatment effect for the two-way interaction between propensity score method and true treatment effect size. True treatment effect sizes were Y1 = 0, Y2 = 0.20, Y3 = 0.40, and Y4 = 0.80. Nearest neighbor matching is indicated by orange, circles, and a solid line. Nearest neighbor matching with a caliper is indicated by green, triangles, and a dash-dot line. Generalized boosted modeling is indicated by purple, squares, and a long, dashed line. Bias in the estimated treatment effect at baseline (prior to matching or weighting) is indicated by black, crosses, and a short, dashed line.

Figure 5

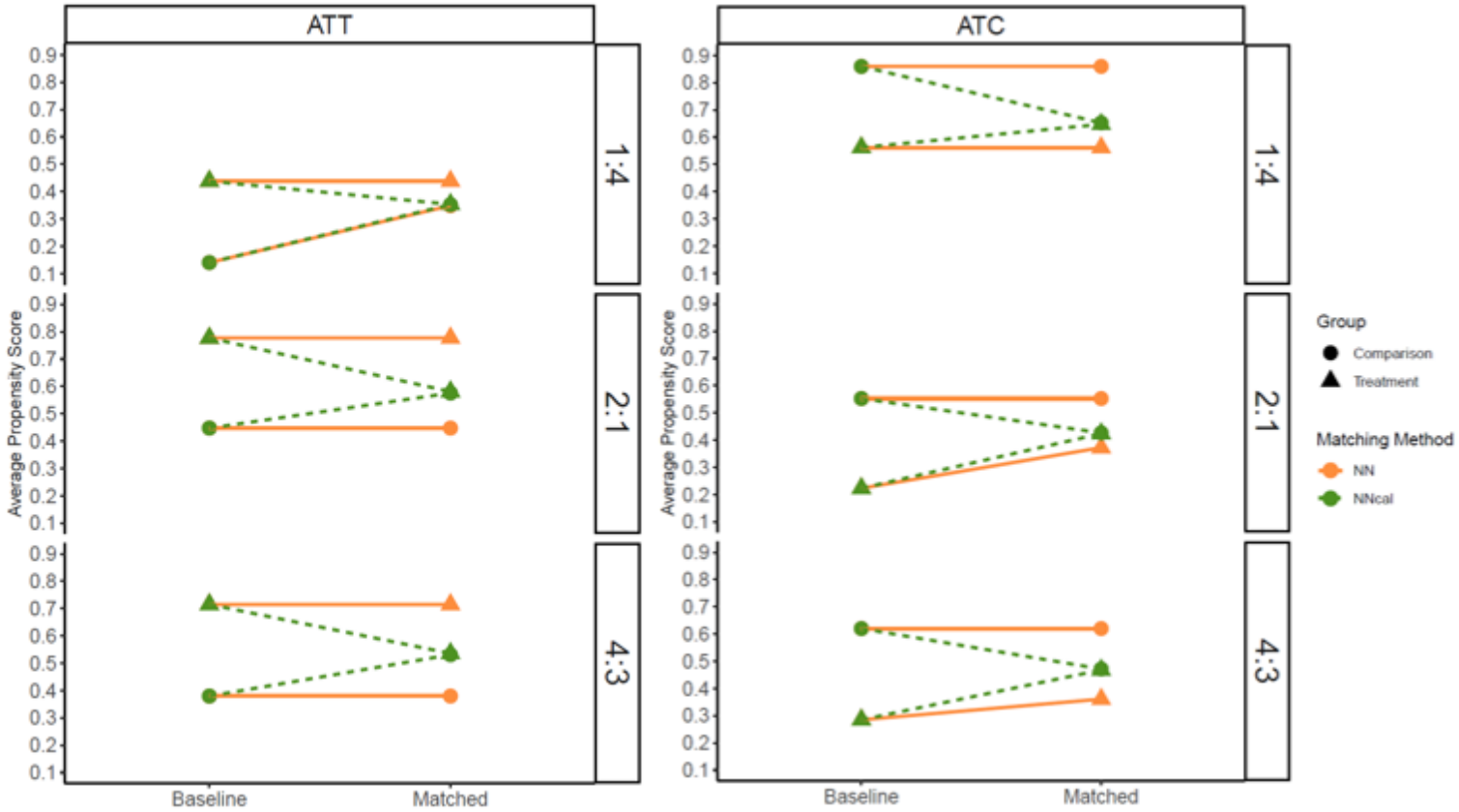
Average Standardized Mean Difference for Covariates and Propensity Score across Propensity Score Method, Coding Method, and Treatment to Comparison Ratio



Note. The standardized mean difference before matching or weighting is noted by the horizontal line for each covariate (X1 = 0, X2 = 0.20, X3 = 0.50, X4 = 0.80, and X5 = 1.20). Standardized mean difference for the propensity score is not presented for GBM because the full sample is retained, and there is no change to the distribution of the propensity score. PS stands for propensity score. ATC coding is indicated in orange and ATT coding is indicated in green. For each coding method, each treatment to comparison ratio is indicated by a lighter shading (ratio of 1:4), medium shading (ratio of 2:1), and darker shading (ratio of 4:3) of each respective color.

Figure 6

Average Propensity Score for Baseline and Matched Treatment and Comparison Groups across Matching Method, Coding Method, and Treatment to Comparison Ratio



Note. For each coding method, the baseline treatment and comparison group means did not differ across matching method (as shown by the overlapping symbols for both methods at baseline). Comparison group means are indicated by circles and treatment group means are indicated by triangle. Orange, solid lines and shapes indicate means associated with nearest neighbor matching and green, dashed lines and shapes indicate means associated with nearest neighbor matching with a caliper. Results were not included for generalized boosted modeling because there is no matched treatment or comparison group. That is, the original sample is retained and average propensity scores for each group do not differ from baseline.

Appendix

Simulation code for Configuration A (adapted from simulation code used by Harris, 2018). The following code was adapted for each configuration by changing the proportion of the sample receiving treatment (TreatP) and the total sample size (Nexaminee). For each configuration, the following values were substituted:

Configuration	TreatP	Nexaminee
A	.667	300
B	.667	900
C	.667	1500
D	.571	350
E	.571	1050
F	.571	1750
G	.200	1000
H	.200	3000
I	.200	5000

```
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# ~~~~~~~~~~~~~~~~            Dissertation Code         ~~~~~~~~~~~~~~~~
# ~~~~~~~~~~~~~~~~               Beth Perkins           ~~~~~~~~~~~~~~~~
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

#Set working directory for saving simulated data
setwd("C:/Users/perkinba/Desktop/Dissertation/Chapter 4 - Results")
getwd()

install.packages("permute")
install.packages("mvtnorm")
install.packages("MatchIt")
install.packages("reshape2")
install.packages("twang")
install.packages("psych")
install.packages("writexl")

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#                      CREATING THE SAVE-OUT VALUES.
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
# A denotes elements of Scenario A (2:1 ratio, T=200,C=100) before
matching/weighting

AvgX1TreatA     <-  rep(NA, 1000)
AvgX2TreatA     <-  rep(NA, 1000)
AvgX3TreatA     <-  rep(NA, 1000)
AvgX4TreatA     <-  rep(NA, 1000)
AvgX5TreatA     <-  rep(NA, 1000)
AvgYA1TreatA    <-  rep(NA, 1000)
AvgYA2TreatA    <-  rep(NA, 1000)
AvgYA3TreatA    <-  rep(NA, 1000)
AvgYA4TreatA    <-  rep(NA, 1000)
AvgPSTreatA     <-  rep(NA, 1000)
AvgX1CompA      <-  rep(NA, 1000)
AvgX2CompA      <-  rep(NA, 1000)
```

```
AvgX3CompA      <-  rep(NA, 1000)
AvgX4CompA      <-  rep(NA, 1000)
AvgX5CompA      <-  rep(NA, 1000)
AvgYA1CompA     <-  rep(NA, 1000)
AvgYA2CompA     <-  rep(NA, 1000)
AvgYA3CompA     <-  rep(NA, 1000)
AvgYA4CompA     <-  rep(NA, 1000)
AvgPSCompA      <-  rep(NA, 1000)
SDX1TreatA      <-  rep(NA, 1000)
SDX2TreatA      <-  rep(NA, 1000)
SDX3TreatA      <-  rep(NA, 1000)
SDX4TreatA      <-  rep(NA, 1000)
SDX5TreatA      <-  rep(NA, 1000)
SDYA1TreatA     <-  rep(NA, 1000)
SDYA2TreatA     <-  rep(NA, 1000)
SDYA3TreatA     <-  rep(NA, 1000)
SDYA4TreatA     <-  rep(NA, 1000)
SDPSTreatA      <-  rep(NA, 1000)
SDX1CompA       <-  rep(NA, 1000)
SDX2CompA       <-  rep(NA, 1000)
SDX3CompA       <-  rep(NA, 1000)
SDX4CompA       <-  rep(NA, 1000)
SDX5CompA       <-  rep(NA, 1000)
SDYA1CompA      <-  rep(NA, 1000)
SDYA2CompA      <-  rep(NA, 1000)
SDYA3CompA      <-  rep(NA, 1000)
SDYA4CompA      <-  rep(NA, 1000)
SDPSCompA       <-  rep(NA, 1000)
SMD_X1_All      <-  rep(NA, 1000)
SMD_X2_All      <-  rep(NA, 1000)
SMD_X3_All      <-  rep(NA, 1000)
SMD_X4_All      <-  rep(NA, 1000)
SMD_X5_All      <-  rep(NA, 1000)
SMD_PS_All      <-  rep(NA, 1000)
SMD_X1_AllATC   <-  rep(NA, 1000)
SMD_X2_AllATC   <-  rep(NA, 1000)
SMD_X3_AllATC   <-  rep(NA, 1000)
SMD_X4_AllATC   <-  rep(NA, 1000)
SMD_X5_AllATC   <-  rep(NA, 1000)
SMD_PS_AllATC   <-  rep(NA, 1000)
Cor_X1.X2_A     <-  rep(NA, 1000)
Cor_X1.X3_A     <-  rep(NA, 1000)
Cor_X1.X4_A     <-  rep(NA, 1000)
Cor_X1.X5_A     <-  rep(NA, 1000)
Cor_X2.X3_A     <-  rep(NA, 1000)
Cor_X2.X4_A     <-  rep(NA, 1000)
Cor_X2.X5_A     <-  rep(NA, 1000)
Cor_X3.X4_A     <-  rep(NA, 1000)
Cor_X3.X5_A     <-  rep(NA, 1000)
Cor_X4.X5_A     <-  rep(NA, 1000)
Cor_X1.PS_A     <-  rep(NA, 1000)
Cor_X2.PS_A     <-  rep(NA, 1000)
Cor_X3.PS_A     <-  rep(NA, 1000)
Cor_X4.PS_A     <-  rep(NA, 1000)
Cor_X5.PS_A     <-  rep(NA, 1000)
Cor_X1.Y1_A     <-  rep(NA, 1000)
Cor_X2.Y1_A     <-  rep(NA, 1000)
```

```
Cor_X3.Y1_A     <-  rep(NA, 1000)
Cor_X4.Y1_A     <-  rep(NA, 1000)
Cor_X5.Y1_A     <-  rep(NA, 1000)
Cor_X1.Y2_A     <-  rep(NA, 1000)
Cor_X2.Y2_A     <-  rep(NA, 1000)
Cor_X3.Y2_A     <-  rep(NA, 1000)
Cor_X4.Y2_A     <-  rep(NA, 1000)
Cor_X5.Y2_A     <-  rep(NA, 1000)
Cor_X1.Y3_A     <-  rep(NA, 1000)
Cor_X2.Y3_A     <-  rep(NA, 1000)
Cor_X3.Y3_A     <-  rep(NA, 1000)
Cor_X4.Y3_A     <-  rep(NA, 1000)
Cor_X5.Y3_A     <-  rep(NA, 1000)
Cor_X1.Y4_A     <-  rep(NA, 1000)
Cor_X2.Y4_A     <-  rep(NA, 1000)
Cor_X3.Y4_A     <-  rep(NA, 1000)
Cor_X4.Y4_A     <-  rep(NA, 1000)
Cor_X5.Y4_A     <-  rep(NA, 1000)
Cor_G.Y1_A      <-  rep(NA, 1000)
Cor_G.Y2_A      <-  rep(NA, 1000)
Cor_G.Y3_A      <-  rep(NA, 1000)
Cor_G.Y4_A      <-  rep(NA, 1000)
PopY1A          <-  rep(NA, 1000)
PopY2A          <-  rep(NA, 1000)
PopY3A          <-  rep(NA, 1000)
PopY4A          <-  rep(NA, 1000)
tPopY1A         <-  rep(NA, 1000)
tPopY2A         <-  rep(NA, 1000)
tPopY3A         <-  rep(NA, 1000)
tPopY4A         <-  rep(NA, 1000)
treatPopNA      <-  rep(NA, 1000)
compPopNA       <-  rep(NA, 1000)
PopCohenY1A     <-  rep(NA, 1000)
PopCohenY2A     <-  rep(NA, 1000)
PopCohenY3A     <-  rep(NA, 1000)
PopCohenY4A     <-  rep(NA, 1000)
BaseY1A         <-  rep(NA, 1000)
BaseY2A         <-  rep(NA, 1000)
BaseY3A         <-  rep(NA, 1000)
BaseY4A         <-  rep(NA, 1000)
tBaseY1A        <-  rep(NA, 1000)
tBaseY2A        <-  rep(NA, 1000)
tBaseY3A        <-  rep(NA, 1000)
tBaseY4A        <-  rep(NA, 1000)
treatBaseNA     <-  rep(NA, 1000)
compBaseNA      <-  rep(NA, 1000)
BaseCohenY1A    <-  rep(NA, 1000)
BaseCohenY2A    <-  rep(NA, 1000)
BaseCohenY3A    <-  rep(NA, 1000)
BaseCohenY4A    <-  rep(NA, 1000)
PopATCY1A       <-  rep(NA, 1000)
PopATCY2A       <-  rep(NA, 1000)
PopATCY3A       <-  rep(NA, 1000)
PopATCY4A       <-  rep(NA, 1000)
tPopATCY1A      <-  rep(NA, 1000)
tPopATCY2A      <-  rep(NA, 1000)
tPopATCY3A      <-  rep(NA, 1000)
```

```
tPopATCY4A      <-  rep(NA, 1000)
treatPopNAATC   <-  rep(NA, 1000)
compPopNAATC    <-  rep(NA, 1000)
PopCohenATCY1A  <-  rep(NA, 1000)
PopCohenATCY2A  <-  rep(NA, 1000)
PopCohenATCY3A  <-  rep(NA, 1000)
PopCohenATCY4A  <-  rep(NA, 1000)
BaseATCY1A      <-  rep(NA, 1000)
BaseATCY2A      <-  rep(NA, 1000)
BaseATCY3A      <-  rep(NA, 1000)
BaseATCY4A      <-  rep(NA, 1000)
tBaseATCY1A     <-  rep(NA, 1000)
tBaseATCY2A     <-  rep(NA, 1000)
tBaseATCY3A     <-  rep(NA, 1000)
tBaseATCY4A     <-  rep(NA, 1000)
treatBaseNAATC  <-  rep(NA, 1000)
compBaseNAATC   <-  rep(NA, 1000)
BaseCohenATCY1A <-  rep(NA, 1000)
BaseCohenATCY2A <-  rep(NA, 1000)
BaseCohenATCY3A <-  rep(NA, 1000)
BaseCohenATCY4A <-  rep(NA, 1000)
VRB             <-  rep(NA, 1000)
VRBATC          <-  rep(NA, 1000)
AvgX1TreatANN   <-  rep(NA, 1000)
AvgX2TreatANN   <-  rep(NA, 1000)
AvgX3TreatANN   <-  rep(NA, 1000)
AvgX4TreatANN   <-  rep(NA, 1000)
AvgX5TreatANN   <-  rep(NA, 1000)
AvgYA1TreatANN  <-  rep(NA, 1000)
AvgYA2TreatANN  <-  rep(NA, 1000)
AvgYA3TreatANN  <-  rep(NA, 1000)
AvgYA4TreatANN  <-  rep(NA, 1000)
AvgPSTreatANN   <-  rep(NA, 1000)
AvgX1CompANN    <-  rep(NA, 1000)
AvgX2CompANN    <-  rep(NA, 1000)
AvgX3CompANN    <-  rep(NA, 1000)
AvgX4CompANN    <-  rep(NA, 1000)
AvgX5CompANN    <-  rep(NA, 1000)
AvgYA1CompANN   <-  rep(NA, 1000)
AvgYA2CompANN   <-  rep(NA, 1000)
AvgYA3CompANN   <-  rep(NA, 1000)
AvgYA4CompANN   <-  rep(NA, 1000)
AvgPSCompANN    <-  rep(NA, 1000)
SDX1TreatANN    <-  rep(NA, 1000)
SDX2TreatANN    <-  rep(NA, 1000)
SDX3TreatANN    <-  rep(NA, 1000)
SDX4TreatANN    <-  rep(NA, 1000)
SDX5TreatANN    <-  rep(NA, 1000)
SDYA1TreatANN   <-  rep(NA, 1000)
SDYA2TreatANN   <-  rep(NA, 1000)
SDYA3TreatANN   <-  rep(NA, 1000)
SDYA4TreatANN   <-  rep(NA, 1000)
SDPSTreatANN    <-  rep(NA, 1000)
SDX1CompANN     <-  rep(NA, 1000)
SDX2CompANN     <-  rep(NA, 1000)
SDX3CompANN     <-  rep(NA, 1000)
SDX4CompANN     <-  rep(NA, 1000)
```

```
SDX5CompANN     <-  rep(NA, 1000)
SDYA1CompANN    <-  rep(NA, 1000)
SDYA2CompANN    <-  rep(NA, 1000)
SDYA3CompANN    <-  rep(NA, 1000)
SDYA4CompANN    <-  rep(NA, 1000)
SDPSCompANN     <-  rep(NA, 1000)
SMD_X1_ANN      <-  rep(NA, 1000)
SMD_X2_ANN      <-  rep(NA, 1000)
SMD_X3_ANN      <-  rep(NA, 1000)
SMD_X4_ANN      <-  rep(NA, 1000)
SMD_X5_ANN      <-  rep(NA, 1000)
SMD_PS_ANN      <-  rep(NA, 1000)
PBR_X1_ANN      <-  rep(NA, 1000)
PBR_X2_ANN      <-  rep(NA, 1000)
PBR_X3_ANN      <-  rep(NA, 1000)
PBR_X4_ANN      <-  rep(NA, 1000)
PBR_X5_ANN      <-  rep(NA, 1000)
PBR_PS_ANN      <-  rep(NA, 1000)
Y1ANN           <-  rep(NA, 1000)
Y2ANN           <-  rep(NA, 1000)
Y3ANN           <-  rep(NA, 1000)
Y4ANN           <-  rep(NA, 1000)
tNNY1A          <-  rep(NA, 1000)
tNNY2A          <-  rep(NA, 1000)
tNNY3A          <-  rep(NA, 1000)
tNNY4A          <-  rep(NA, 1000)
NNtreatNA       <-  rep(NA, 1000)
NNcompNA        <-  rep(NA, 1000)
NNCohenY1A      <-  rep(NA, 1000)
NNCohenY2A      <-  rep(NA, 1000)
NNCohenY3A      <-  rep(NA, 1000)
NNCohenY4A      <-  rep(NA, 1000)
PSMeanMatchedTreatANN      <-  rep(NA, 1000)
PSMeanMatchedCompANN       <-  rep(NA, 1000)
PSMeanUnMatchedTreatANN    <-  rep(NA, 1000)
PSMeanUnMatchedCompANN     <-  rep(NA, 1000)
PSMedMatchedTreatANN       <-  rep(NA, 1000)
PSMedMatchedCompANN        <-  rep(NA, 1000)
PSMedUnMatchedTreatANN     <-  rep(NA, 1000)
PSMedUnMatchedCompANN      <-  rep(NA, 1000)
PSsdMatchedTreatANN        <-  rep(NA, 1000)
PSsdMatchedCompANN         <-  rep(NA, 1000)
PSsdUnMatchedTreatANN      <-  rep(NA, 1000)
PSsdUnMatchedCompANN       <-  rep(NA, 1000)
VRANN                <-  rep(NA, 1000)
AvgX1TreatANNATC     <-  rep(NA, 1000)
AvgX2TreatANNATC     <-  rep(NA, 1000)
AvgX3TreatANNATC     <-  rep(NA, 1000)
AvgX4TreatANNATC     <-  rep(NA, 1000)
AvgX5TreatANNATC     <-  rep(NA, 1000)
AvgYA1TreatANNATC    <-  rep(NA, 1000)
AvgYA2TreatANNATC    <-  rep(NA, 1000)
AvgYA3TreatANNATC    <-  rep(NA, 1000)
AvgYA4TreatANNATC    <-  rep(NA, 1000)
AvgPSTreatANNATC     <-  rep(NA, 1000)
AvgX1CompANNATC      <-  rep(NA, 1000)
AvgX2CompANNATC      <-  rep(NA, 1000)
```

```
AvgX3CompANNATC      <-  rep(NA, 1000)
AvgX4CompANNATC      <-  rep(NA, 1000)
AvgX5CompANNATC      <-  rep(NA, 1000)
AvgYA1CompANNATC     <-  rep(NA, 1000)
AvgYA2CompANNATC     <-  rep(NA, 1000)
AvgYA3CompANNATC     <-  rep(NA, 1000)
AvgYA4CompANNATC     <-  rep(NA, 1000)
AvgPSCompANNATC      <-  rep(NA, 1000)
SDX1TreatANNATC      <-  rep(NA, 1000)
SDX2TreatANNATC      <-  rep(NA, 1000)
SDX3TreatANNATC      <-  rep(NA, 1000)
SDX4TreatANNATC      <-  rep(NA, 1000)
SDX5TreatANNATC      <-  rep(NA, 1000)
SDYA1TreatANNATC     <-  rep(NA, 1000)
SDYA2TreatANNATC     <-  rep(NA, 1000)
SDYA3TreatANNATC     <-  rep(NA, 1000)
SDYA4TreatANNATC     <-  rep(NA, 1000)
SDPSTreatANNATC      <-  rep(NA, 1000)
SDX1CompANNATC       <-  rep(NA, 1000)
SDX2CompANNATC       <-  rep(NA, 1000)
SDX3CompANNATC       <-  rep(NA, 1000)
SDX4CompANNATC       <-  rep(NA, 1000)
SDX5CompANNATC      <-  rep(NA, 1000)
SDYA1CompANNATC     <-  rep(NA, 1000)
SDYA2CompANNATC     <-  rep(NA, 1000)
SDYA3CompANNATC     <-  rep(NA, 1000)
SDYA4CompANNATC     <-  rep(NA, 1000)
SDPSCompANNATC      <-  rep(NA, 1000)
SMD_X1_ANNATC       <-  rep(NA, 1000)
SMD_X2_ANNATC       <-  rep(NA, 1000)
SMD_X3_ANNATC       <-  rep(NA, 1000)
SMD_X4_ANNATC       <-  rep(NA, 1000)
SMD_X5_ANNATC       <-  rep(NA, 1000)
SMD_PS_ANNATC       <-  rep(NA, 1000)
PBR_X1_ANNATC       <-  rep(NA, 1000)
PBR_X2_ANNATC       <-  rep(NA, 1000)
PBR_X3_ANNATC       <-  rep(NA, 1000)
PBR_X4_ANNATC       <-  rep(NA, 1000)
PBR_X5_ANNATC       <-  rep(NA, 1000)
PBR_PS_ANNATC       <-  rep(NA, 1000)
Y1ANNATC            <-  rep(NA, 1000)
Y2ANNATC            <-  rep(NA, 1000)
Y3ANNATC            <-  rep(NA, 1000)
Y4ANNATC            <-  rep(NA, 1000)
tNNATCY1A           <-  rep(NA, 1000)
tNNATCY2A           <-  rep(NA, 1000)
tNNATCY3A           <-  rep(NA, 1000)
tNNATCY4A           <-  rep(NA, 1000)
NNATCtreatNA        <-  rep(NA, 1000)
NNATCcompNA         <-  rep(NA, 1000)
NNATCCohenY1A       <-  rep(NA, 1000)
NNATCCohenY2A       <-  rep(NA, 1000)
NNATCCohenY3A       <-  rep(NA, 1000)
NNATCCohenY4A       <-  rep(NA, 1000)
PSMeanMatchedTreatANNATC       <-  rep(NA, 1000)
PSMeanMatchedCompANNATC        <-  rep(NA, 1000)
PSMeanUnMatchedTreatANNATC     <-  rep(NA, 1000)
```

```
PSMeanUnMatchedCompANNATC      <-  rep(NA, 1000)
PSMedMatchedTreatANNATC        <-  rep(NA, 1000)
PSMedMatchedCompANNATC         <-  rep(NA, 1000)
PSMedUnMatchedTreatANNATC      <-  rep(NA, 1000)
PSMedUnMatchedCompANNATC       <-  rep(NA, 1000)
PSsdMatchedTreatANNATC         <-  rep(NA, 1000)
PSsdMatchedCompANNATC          <-  rep(NA, 1000)
PSsdUnMatchedTreatANNATC       <-  rep(NA, 1000)
PSsdUnMatchedCompANNATC        <-  rep(NA, 1000)
VRANNATC                  <-  rep(NA, 1000)
AvgX1TreatANNCal          <-  rep(NA, 1000)
AvgX2TreatANNCal          <-  rep(NA, 1000)
AvgX3TreatANNCal          <-  rep(NA, 1000)
AvgX4TreatANNCal          <-  rep(NA, 1000)
AvgX5TreatANNCal          <-  rep(NA, 1000)
AvgYA1TreatANNCal         <-  rep(NA, 1000)
AvgYA2TreatANNCal         <-  rep(NA, 1000)
AvgYA3TreatANNCal         <-  rep(NA, 1000)
AvgYA4TreatANNCal         <-  rep(NA, 1000)
AvgPSTreatANNCal          <-  rep(NA, 1000)
AvgX1CompANNCal           <-  rep(NA, 1000)
AvgX2CompANNCal           <-  rep(NA, 1000)
AvgX3CompANNCal       <-  rep(NA, 1000)
AvgX4CompANNCal       <-  rep(NA, 1000)
AvgX5CompANNCal       <-  rep(NA, 1000)
AvgYA1CompANNCal      <-  rep(NA, 1000)
AvgYA2CompANNCal      <-  rep(NA, 1000)
AvgYA3CompANNCal      <-  rep(NA, 1000)
AvgYA4CompANNCal      <-  rep(NA, 1000)
AvgPSCompANNCal       <-  rep(NA, 1000)
SDX1TreatANNCal       <-  rep(NA, 1000)
SDX2TreatANNCal       <-  rep(NA, 1000)
SDX3TreatANNCal       <-  rep(NA, 1000)
SDX4TreatANNCal       <-  rep(NA, 1000)
SDX5TreatANNCal       <-  rep(NA, 1000)
SDYA1TreatANNCal      <-  rep(NA, 1000)
SDYA2TreatANNCal      <-  rep(NA, 1000)
SDYA3TreatANNCal      <-  rep(NA, 1000)
SDYA4TreatANNCal      <-  rep(NA, 1000)
SDPSTreatANNCal       <-  rep(NA, 1000)
SDX1CompANNCal        <-  rep(NA, 1000)
SDX2CompANNCal        <-  rep(NA, 1000)
SDX3CompANNCal        <-  rep(NA, 1000)
SDX4CompANNCal        <-  rep(NA, 1000)
SDX5CompANNCal        <-  rep(NA, 1000)
SDYA1CompANNCal       <-  rep(NA, 1000)
SDYA2CompANNCal       <-  rep(NA, 1000)
SDYA3CompANNCal       <-  rep(NA, 1000)
SDYA4CompANNCal       <-  rep(NA, 1000)
SDPSCompANNCal        <-  rep(NA, 1000)
SMD_X1_ANNCal         <-  rep(NA, 1000)
SMD_X2_ANNCal         <-  rep(NA, 1000)
SMD_X3_ANNCal         <-  rep(NA, 1000)
SMD_X4_ANNCal         <-  rep(NA, 1000)
SMD_X5_ANNCal         <-  rep(NA, 1000)
SMD_PS_ANNCal         <-  rep(NA, 1000)
PBR_X1_ANNCal         <-  rep(NA, 1000)
```

```
PBR_X2_ANNCal        <-  rep(NA, 1000)
PBR_X3_ANNCal        <-  rep(NA, 1000)
PBR_X4_ANNCal        <-  rep(NA, 1000)
PBR_X5_ANNCal        <-  rep(NA, 1000)
PBR_PS_ANNCal        <-  rep(NA, 1000)
Y1ANNCal             <-  rep(NA, 1000)
Y2ANNCal             <-  rep(NA, 1000)
Y3ANNCal             <-  rep(NA, 1000)
Y4ANNCal             <-  rep(NA, 1000)
tNNCalY1A            <-  rep(NA, 1000)
tNNCalY2A            <-  rep(NA, 1000)
tNNCalY3A            <-  rep(NA, 1000)
tNNCalY4A            <-  rep(NA, 1000)
NNCaltreatNA         <-  rep(NA, 1000)
NNCalcompNA          <-  rep(NA, 1000)
NNCalCohenY1A        <-  rep(NA, 1000)
NNCalCohenY2A        <-  rep(NA, 1000)
NNCalCohenY3A        <-  rep(NA, 1000)
NNCalCohenY4A        <-  rep(NA, 1000)
PSMeanMatchedTreatANNCal      <-  rep(NA, 1000)
PSMeanMatchedCompANNCal       <-  rep(NA, 1000)
PSMeanUnMatchedTreatANNCal    <-  rep(NA, 1000)
PSMeanUnMatchedCompANNCal     <-  rep(NA, 1000)
PSMedMatchedTreatANNCal       <-  rep(NA, 1000)
PSMedMatchedCompANNCal        <-  rep(NA, 1000)
PSMedUnMatchedTreatANNCal     <-  rep(NA, 1000)
PSMedUnMatchedCompANNCal      <-  rep(NA, 1000)
PSsdMatchedTreatANNCal        <-  rep(NA, 1000)
PSsdMatchedCompANNCal         <-  rep(NA, 1000)
PSsdUnMatchedTreatANNCal      <-  rep(NA, 1000)
PSsdUnMatchedCompANNCal       <-  rep(NA, 1000)
VRANNCal                <-  rep(NA, 1000)
AvgX1TreatANNCalATC     <-  rep(NA, 1000)
AvgX2TreatANNCalATC     <-  rep(NA, 1000)
AvgX3TreatANNCalATC     <-  rep(NA, 1000)
AvgX4TreatANNCalATC     <-  rep(NA, 1000)
AvgX5TreatANNCalATC     <-  rep(NA, 1000)
AvgYA1TreatANNCalATC    <-  rep(NA, 1000)
AvgYA2TreatANNCalATC    <-  rep(NA, 1000)
AvgYA3TreatANNCalATC    <-  rep(NA, 1000)
AvgYA4TreatANNCalATC    <-  rep(NA, 1000)
AvgPSTreatANNCalATC     <-  rep(NA, 1000)
AvgX1CompANNCalATC      <-  rep(NA, 1000)
AvgX2CompANNCalATC      <-  rep(NA, 1000)
AvgX3CompANNCalATC      <-  rep(NA, 1000)
AvgX4CompANNCalATC      <-  rep(NA, 1000)
AvgX5CompANNCalATC      <-  rep(NA, 1000)
AvgYA1CompANNCalATC     <-  rep(NA, 1000)
AvgYA2CompANNCalATC     <-  rep(NA, 1000)
AvgYA3CompANNCalATC     <-  rep(NA, 1000)
AvgYA4CompANNCalATC     <-  rep(NA, 1000)
AvgPSCompANNCalATC      <-  rep(NA, 1000)
SDX1TreatANNCalATC      <-  rep(NA, 1000)
SDX2TreatANNCalATC      <-  rep(NA, 1000)
SDX3TreatANNCalATC      <-  rep(NA, 1000)
SDX4TreatANNCalATC      <-  rep(NA, 1000)
SDX5TreatANNCalATC      <-  rep(NA, 1000)
```

```
SDYA1TreatANNCalATC     <-  rep(NA, 1000)
SDYA2TreatANNCalATC     <-  rep(NA, 1000)
SDYA3TreatANNCalATC     <-  rep(NA, 1000)
SDYA4TreatANNCalATC     <-  rep(NA, 1000)
SDPSTreatANNCalATC      <-  rep(NA, 1000)
SDX1CompANNCalATC       <-  rep(NA, 1000)
SDX2CompANNCalATC       <-  rep(NA, 1000)
SDX3CompANNCalATC       <-  rep(NA, 1000)
SDX4CompANNCalATC       <-  rep(NA, 1000)
SDX5CompANNCalATC       <-  rep(NA, 1000)
SDYA1CompANNCalATC      <-  rep(NA, 1000)
SDYA2CompANNCalATC      <-  rep(NA, 1000)
SDYA3CompANNCalATC      <-  rep(NA, 1000)
SDYA4CompANNCalATC      <-  rep(NA, 1000)
SDPSCompANNCalATC       <-  rep(NA, 1000)
SMD_X1_ANNCalATC        <-  rep(NA, 1000)
SMD_X2_ANNCalATC        <-  rep(NA, 1000)
SMD_X3_ANNCalATC        <-  rep(NA, 1000)
SMD_X4_ANNCalATC        <-  rep(NA, 1000)
SMD_X5_ANNCalATC        <-  rep(NA, 1000)
SMD_PS_ANNCalATC        <-  rep(NA, 1000)
PBR_X1_ANNCalATC        <-  rep(NA, 1000)
PBR_X2_ANNCalATC     <-  rep(NA, 1000)
PBR_X3_ANNCalATC     <-  rep(NA, 1000)
PBR_X4_ANNCalATC     <-  rep(NA, 1000)
PBR_X5_ANNCalATC     <-  rep(NA, 1000)
PBR_PS_ANNCalATC     <-  rep(NA, 1000)
Y1ANNCalATC          <-  rep(NA, 1000)
Y2ANNCalATC          <-  rep(NA, 1000)
Y3ANNCalATC          <-  rep(NA, 1000)
Y4ANNCalATC          <-  rep(NA, 1000)
tNNCalATCY1A         <-  rep(NA, 1000)
tNNCalATCY2A         <-  rep(NA, 1000)
tNNCalATCY3A         <-  rep(NA, 1000)
tNNCalATCY4A         <-  rep(NA, 1000)
NNCalATCtreatNA      <-  rep(NA, 1000)
NNCalATCcompNA       <-  rep(NA, 1000)
NNCalATCCohenY1A     <-  rep(NA, 1000)
NNCalATCCohenY2A     <-  rep(NA, 1000)
NNCalATCCohenY3A     <-  rep(NA, 1000)
NNCalATCCohenY4A     <-  rep(NA, 1000)
PSMeanMatchedTreatANNCalATC      <-  rep(NA, 1000)
PSMeanMatchedCompANNCalATC       <-  rep(NA, 1000)
PSMeanUnMatchedTreatANNCalATC    <-  rep(NA, 1000)
PSMeanUnMatchedCompANNCalATC     <-  rep(NA, 1000)
PSMedMatchedTreatANNCalATC       <-  rep(NA, 1000)
PSMedMatchedCompANNCalATC        <-  rep(NA, 1000)
PSMedUnMatchedTreatANNCalATC     <-  rep(NA, 1000)
PSMedUnMatchedCompANNCalATC      <-  rep(NA, 1000)
PSsdMatchedTreatANNCalATC        <-  rep(NA, 1000)
PSsdMatchedCompANNCalATC         <-  rep(NA, 1000)
PSsdUnMatchedTreatANNCalATC      <-  rep(NA, 1000)
PSsdUnMatchedCompANNCalATC       <-  rep(NA, 1000)
VRANNCalATC          <-  rep(NA, 1000)
AvgX1TreatAGBM       <-  rep(NA, 1000)
AvgX2TreatAGBM       <-  rep(NA, 1000)
AvgX3TreatAGBM       <-  rep(NA, 1000)
```

```
AvgX4TreatAGBM       <-  rep(NA, 1000)
AvgX5TreatAGBM       <-  rep(NA, 1000)
AvgX1CompAGBM        <-  rep(NA, 1000)
AvgX2CompAGBM        <-  rep(NA, 1000)
AvgX3CompAGBM        <-  rep(NA, 1000)
AvgX4CompAGBM        <-  rep(NA, 1000)
AvgX5CompAGBM        <-  rep(NA, 1000)
SDX1TreatAGBM        <-  rep(NA, 1000)
SDX2TreatAGBM        <-  rep(NA, 1000)
SDX3TreatAGBM        <-  rep(NA, 1000)
SDX4TreatAGBM        <-  rep(NA, 1000)
SDX5TreatAGBM        <-  rep(NA, 1000)
SDX1CompAGBM         <-  rep(NA, 1000)
SDX2CompAGBM         <-  rep(NA, 1000)
SDX3CompAGBM         <-  rep(NA, 1000)
SDX4CompAGBM         <-  rep(NA, 1000)
SDX5CompAGBM         <-  rep(NA, 1000)
SMD_X1_AGBM          <-  rep(NA, 1000)
SMD_X2_AGBM          <-  rep(NA, 1000)
SMD_X3_AGBM          <-  rep(NA, 1000)
SMD_X4_AGBM          <-  rep(NA, 1000)
SMD_X5_AGBM          <-  rep(NA, 1000)
PBR_X1_AGBM          <-  rep(NA, 1000)
PBR_X2_AGBM          <-  rep(NA, 1000)
PBR_X3_AGBM          <-  rep(NA, 1000)
PBR_X4_AGBM          <-  rep(NA, 1000)
PBR_X5_AGBM          <-  rep(NA, 1000)
Y1AGBM               <-  rep(NA, 1000)
Y2AGBM               <-  rep(NA, 1000)
Y3AGBM               <-  rep(NA, 1000)
Y4AGBM               <-  rep(NA, 1000)
tGBMY1A              <-  rep(NA, 1000)
tGBMY2A              <-  rep(NA, 1000)
tGBMY3A              <-  rep(NA, 1000)
tGBMY4A              <-  rep(NA, 1000)
GBMtreatNA           <-  rep(NA, 1000)
GBMcompNA            <-  rep(NA, 1000)
GBMCohenY1A          <-  rep(NA, 1000)
GBMCohenY2A          <-  rep(NA, 1000)
GBMCohenY3A          <-  rep(NA, 1000)
GBMCohenY4A          <-  rep(NA, 1000)
AvgX1TreatAGBMATC    <-  rep(NA, 1000)
AvgX2TreatAGBMATC    <-  rep(NA, 1000)
AvgX3TreatAGBMATC    <-  rep(NA, 1000)
AvgX4TreatAGBMATC    <-  rep(NA, 1000)
AvgX5TreatAGBMATC    <-  rep(NA, 1000)
AvgX1CompAGBMATC     <-  rep(NA, 1000)
AvgX2CompAGBMATC     <-  rep(NA, 1000)
AvgX3CompAGBMATC     <-  rep(NA, 1000)
AvgX4CompAGBMATC     <-  rep(NA, 1000)
AvgX5CompAGBMATC     <-  rep(NA, 1000)
SDX1TreatAGBMATC     <-  rep(NA, 1000)
SDX2TreatAGBMATC     <-  rep(NA, 1000)
SDX3TreatAGBMATC     <-  rep(NA, 1000)
SDX4TreatAGBMATC     <-  rep(NA, 1000)
SDX5TreatAGBMATC     <-  rep(NA, 1000)
SDX1CompAGBMATC      <-  rep(NA, 1000)
```

```
SDX2CompAGBMATC      <-  rep(NA, 1000)
SDX3CompAGBMATC      <-  rep(NA, 1000)
SDX4CompAGBMATC      <-  rep(NA, 1000)
SDX5CompAGBMATC      <-  rep(NA, 1000)
SMD_X1_AGBMATC       <-  rep(NA, 1000)
SMD_X2_AGBMATC       <-  rep(NA, 1000)
SMD_X3_AGBMATC       <-  rep(NA, 1000)
SMD_X4_AGBMATC       <-  rep(NA, 1000)
SMD_X5_AGBMATC       <-  rep(NA, 1000)
PBR_X1_AGBMATC       <-  rep(NA, 1000)
PBR_X2_AGBMATC       <-  rep(NA, 1000)
PBR_X3_AGBMATC       <-  rep(NA, 1000)
PBR_X4_AGBMATC       <-  rep(NA, 1000)
PBR_X5_AGBMATC       <-  rep(NA, 1000)
Y1AGBMATC            <-  rep(NA, 1000)
Y2AGBMATC            <-  rep(NA, 1000)
Y3AGBMATC            <-  rep(NA, 1000)
Y4AGBMATC            <-  rep(NA, 1000)
tGBMATCY1A           <-  rep(NA, 1000)
tGBMATCY2A           <-  rep(NA, 1000)
tGBMATCY3A           <-  rep(NA, 1000)
tGBMATCY4A           <-  rep(NA, 1000)
GBMATCtreatNA        <-  rep(NA, 1000)
GBMATCcompNA         <-  rep(NA, 1000)
GBMATCCohenY1A       <-  rep(NA, 1000)
GBMATCCohenY2A       <-  rep(NA, 1000)
GBMATCCohenY3A       <-  rep(NA, 1000)
GBMATCCohenY4A       <-  rep(NA, 1000)
ESS_CompGBM          <-  rep(NA, 1000)
mean.esGBM           <-  rep(NA, 1000)
iterGBM              <-  rep(NA, 1000)
ESS_CompGBMATC       <-  rep(NA, 1000)
mean.esGBMATC        <-  rep(NA, 1000)
iterGBMATC           <-  rep(NA, 1000)

# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#                               BEGIN LOOP
# ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
set.seed(27)
for(i in 1:1000){

##Set treatment/comparison group ratio
#******SET RATIO
TreatP=.667
mycut=qnorm(1-TreatP) # threshold
#mycut

##initial correlation matrix
corrX=matrix(c(1,.1,.2,.3,.3,
               .1,1,.3,.3,.35,
               .2,.3,1,.3,.45,
               .3,.3,.3,1,.65,
               .3,.35,.45,.65,1),5,5)

##call these programs for simulating data (mvtnorm) & screening (psych)
library(mvtnorm)
library(psych)
```

```
library(permute)

##Set number of examinees
Nexaminee=300    #*******TOTAL SAMPLE SIZE
Nrep=1000   ##used for grouping using the PS

##Simulating the five original covariates
X=rmvnorm(Nexaminee, rep(0,5), corrX, method="chol")
##Coerce into a data frame, so can use it more easily
X <-as.data.frame(X)

#Correlation between covariates and latent propensity
#Note  - we may want to fiddle with the strength of these
relationships.
#Will specify group balance on covariates. Tried transforming cohens d
to correlation, but did not result in large
#enough baseline SMD for X4 and X5. Played with values till SMD were
consistently what I wanted them to be.
corrXp = c(-.02,.15,.40,.70,.90)

##calculate regression coefficients
#install.packages("reshape2")
library(reshape2)

##Use the above correlation matrix to calculate the regression
coefficients
##These are coefficients for the PS model (not the outcome model)
PcoefA=solve(corrX) %*% corrXp

##Variance explained for model
##Creating a temporary matrix of squared values (jh)
##This will be used to create overall R-squared for model
##this is the variance in the latent propensity explained by covariates
tempA=PcoefA %*% t(PcoefA)*corrX

varExpPA=sum(tempA) #Summing everyting in temp
#varExpPA

xnew=as.matrix.data.frame(X)

################################Propensity
scores#########################
##This is like the sum of bx for each person  (i.e., their y' predicted
scores)
noErrA=as.vector(xnew %*% PcoefA)
#describe(noErrA)

#plot(noErrA)

##Variance explained in propensity score by X1-X5
RsqA=varExpPA/(1+varExpPA) #because the error variance of a probit is 1

##Rescale around the threshold to ensure the correct treatment to
comparison group ratio
distA=noErrA-(mycut/sqrt(1-RsqA))

##Finding the probability density
```

```
##(help page says 'vector of probabilities')
##Assumes a normal distribution
truepropA=pnorm(distA)

##Randomly assign a random draw to each person Nrep times
randraw=matrix(runif(Nexaminee*Nrep),nrow=Nexaminee,ncol=Nrep)
options(scipen=999)
str(randraw)

##If propensity score is greater than the random draw,then assign to
treatment
##Otherwise, assign to comparison group
groupA=ifelse(truepropA>randraw,1,0)

##Creating a data frame with X1-X5, grouping variable, and true
propensity score
dataA=data.frame(X,groupA[,1])

##Assign variable names
library(reshape2)
names(dataA) <- c("x1","x2","x3","x4","x5","group")

##Calculate true propensity scores from logistic regression to obtain
logistic regression coefficients
dataA <- as.data.frame(dataA)
PbA=glm(formula= group ~ x1+x2+x3+x4+x5, data=dataA,  family=binomial)
#PbA

##True propensity scores
dataA$TRUEprop<-predict(PbA, type="response")
#plot(dataA$TRUEprop)
#plot(dataA$TRUEprop, dataA$group)

##Outcome model
##Random error in the model
v <- rnorm(Nexaminee, mean=0, sd=0.5)
v <- as.data.frame(v)  #coerce to data frame

##Each person's Y for the outcome model - would I create 4 different
models, one for each treatment effect size?

YA1 = 0 + 0*dataA$group + .05*dataA$x1 + .05*dataA$x2 + .05*dataA$x3 +
.05*dataA$x4 + .05*dataA$x5 + v
YA2 = 0 + .11*dataA$group + .05*dataA$x1 + .05*dataA$x2 + .05*dataA$x3
+ .05*dataA$x4 + .05*dataA$x5 + v
YA3 = 0 + .28*dataA$group + .05*dataA$x1 + .05*dataA$x2 + .05*dataA$x3
+ .05*dataA$x4 + .05*dataA$x5 + v
YA4 = 0 + .45*dataA$group + .05*dataA$x1 + .05*dataA$x2 + .05*dataA$x3
+ .05*dataA$x4 + .05*dataA$x5 + v

#describe(YA1)
#describe(YA2)
#describe(YA3)
#describe(YA4)
describe(dataA$group)

ATCgroup<-ifelse(dataA$group==0, 1, ifelse(dataA$group==1, 0, -1 ))
```

```
ATCgroup


finalDataA<-cbind(dataA,YA1,YA2,YA3,YA4,ATCgroup)
finalDataA<-as.data.frame(finalDataA)

#head(finalDataA)
describe(finalDataA)
#describeBy(finalDataA, finalDataA$group)
#cor(finalDataA)

library(reshape2)
names(finalDataA) <- c("x1","x2","x3","x4","x5","group", "PS",
"YA1","YA2","YA3","YA4", "ATCgroup")

#What should the coefficient be - if treatment and comparison group
members differ on the outcome by the specified size of treatment effect
#We want to recover these values after matching/weighting
require(MatchIt)

PopY1=lm(YA1~group+x1+x2+x3+x4+x5, data=finalDataA)
#PopY1
PopY2=lm(YA2~group+x1+x2+x3+x4+x5, data=finalDataA)
#PopY2
PopY3=lm(YA3~group+x1+x2+x3+x4+x5, data=finalDataA)
#PopY3
PopY4=lm(YA4~group+x1+x2+x3+x4+x5, data=finalDataA)
#PopY4

ATCPopY1=lm(YA1~ATCgroup+x1+x2+x3+x4+x5, data=finalDataA)
#ATCPopY1
ATCPopY2=lm(YA2~ATCgroup+x1+x2+x3+x4+x5, data=finalDataA)
#ATCPopY2
ATCPopY3=lm(YA3~ATCgroup+x1+x2+x3+x4+x5, data=finalDataA)
#ATCPopY3
ATCPopY4=lm(YA4~ATCgroup+x1+x2+x3+x4+x5, data=finalDataA)
#ATCPopY4

#Use these values to transform t to d to make sure the treatment effect
is what we want it to be
#summary(PopY1)
#summary(PopY2)
#summary(PopY3)
#summary(PopY4)
#summary(ATCPopY1)
#summary(ATCPopY2)
#summary(ATCPopY3)
#summary(ATCPopY4)

#Computing Cohen's D for Population Treatment effect
tPopY1<-summary(PopY1)$coef[2, 3]
tPopY2<-summary(PopY2)$coef[2, 3]
tPopY3<-summary(PopY3)$coef[2, 3]
tPopY4<-summary(PopY4)$coef[2, 3]
treatn<-nobs(finalDataA$group[finalDataA$group==1])
compn<-nobs(finalDataA$group[finalDataA$group==0])
```

```
cohenPopY1 <-(tPopY1)*(((1/(treatn))+(1/(compn)))^0.5)
cohenPopY2 <-(tPopY2)*(((1/(treatn))+(1/(compn)))^0.5)
cohenPopY3 <-(tPopY3)*(((1/(treatn))+(1/(compn)))^0.5)
cohenPopY4 <-(tPopY4)*(((1/(treatn))+(1/(compn)))^0.5)

tPopATCY1<-summary(ATCPopY1)$coef[2, 3]
tPopATCY2<-summary(ATCPopY2)$coef[2, 3]
tPopATCY3<-summary(ATCPopY3)$coef[2, 3]
tPopATCY4<-summary(ATCPopY4)$coef[2, 3]
treatnATC<-nobs(finalDataA$ATCgroup[finalDataA$ATCgroup==0])
compnATC<-nobs(finalDataA$ATCgroup[finalDataA$ATCgroup==1])

cohenPopATCY1 <-(tPopATCY1)*(((1/(treatnATC))+(1/(compnATC)))^0.5)
cohenPopATCY2 <-(tPopATCY2)*(((1/(treatnATC))+(1/(compnATC)))^0.5)
cohenPopATCY3 <-(tPopATCY3)*(((1/(treatnATC))+(1/(compnATC)))^0.5)
cohenPopATCY4 <-(tPopATCY4)*(((1/(treatnATC))+(1/(compnATC)))^0.5)

#Baseline group differences on outcome - deviation from PopY1-PopY4
indicate bias in the estimated treatment effect
baseout1=lm(YA1~group, data=finalDataA)
baseout2=lm(YA2~group, data=finalDataA)
baseout3=lm(YA3~group, data=finalDataA)
baseout4=lm(YA4~group, data=finalDataA)

ATCbaseout1=lm(YA1~ATCgroup, data=finalDataA)
ATCbaseout2=lm(YA2~ATCgroup, data=finalDataA)
ATCbaseout3=lm(YA3~ATCgroup, data=finalDataA)
ATCbaseout4=lm(YA4~ATCgroup, data=finalDataA)

#Computing Cohen's D for Baseline Treatment effect
tBaseY1<-summary(baseout1)$coef[2, 3]
tBaseY2<-summary(baseout2)$coef[2, 3]
tBaseY3<-summary(baseout3)$coef[2, 3]
tBaseY4<-summary(baseout4)$coef[2, 3]

cohenBaseY1 <-(tBaseY1)*(((1/(treatn))+(1/(compn)))^0.5)
cohenBaseY2 <-(tBaseY2)*(((1/(treatn))+(1/(compn)))^0.5)
cohenBaseY3 <-(tBaseY3)*(((1/(treatn))+(1/(compn)))^0.5)
cohenBaseY4 <-(tBaseY4)*(((1/(treatn))+(1/(compn)))^0.5)

#Computing Cohen's D for Baseline Treatment effect, ATC
tBaseATCY1<-summary(ATCbaseout1)$coef[2, 3]
tBaseATCY2<-summary(ATCbaseout2)$coef[2, 3]
tBaseATCY3<-summary(ATCbaseout3)$coef[2, 3]
tBaseATCY4<-summary(ATCbaseout4)$coef[2, 3]

cohenBaseATCY1 <-(tBaseATCY1)*(((1/(treatnATC))+(1/(compnATC)))^0.5)
cohenBaseATCY2 <-(tBaseATCY2)*(((1/(treatnATC))+(1/(compnATC)))^0.5)
cohenBaseATCY3 <-(tBaseATCY3)*(((1/(treatnATC))+(1/(compnATC)))^0.5)
cohenBaseATCY4 <-(tBaseATCY4)*(((1/(treatnATC))+(1/(compnATC)))^0.5)

#Baseline group differences on outcome - deviation from PopY1-PopY4
indicate bias in the estimated treatment effect
#baseout1
#baseout2
#baseout3
#baseout4
```

```
#summary(baseout1)
#summary(baseout2)
#summary(baseout3)
#summary(baseout4)
#summary(ATCbaseout1)
#summary(ATCbaseout2)
#summary(ATCbaseout3)
#summary(ATCbaseout4)

#describeBy(finalDataA, finalDataA$group)

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#~~~~~~~~~~~~~~~~~~        Matching/Weighting:       ~~~~~~~~~~~~~~~~
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#~~~~~~                      Nearest Neighbor                      ~~~~~
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#NN - ATT Coding
require(MatchIt)
m.outANN =
matchit(finalDataA$group~finalDataA$x1+finalDataA$x2+finalDataA$x3+fina
lDataA$x4+finalDataA$x5,data=finalDataA, method="nearest",
m.order="random", ratio=1)

m.outANN

MANN <- summary(m.outANN, standardize = TRUE)

ANN <- match.data(m.outANN, group="all")

FullANN<-match.data(m.outANN, group="all", drop.unmatched = FALSE)

#NN - ATC Coding
require(MatchIt)
m.outANNATC =
matchit(finalDataA$ATCgroup~finalDataA$x1+finalDataA$x2+finalDataA$x3+f
inalDataA$x4+finalDataA$x5, data=finalDataA, method="nearest",
m.order="random", ratio=1)

m.outANNATC

MANNATC <- summary(m.outANNATC, standardize = TRUE)

ANNATC <- match.data(m.outANNATC, group="all")

FullANNATC<-match.data(m.outANNATC, group="all", drop.unmatched =
FALSE)

#~~~~~~~~~~~ESTIMATED TREATMENT EFFECT~~~~~~~~~~~~~~~~~~
OutcomeA.1NN <- lm(YA1 ~ group, data = ANN)
OutcomeA.2NN <- lm(YA2 ~ group, data = ANN)
OutcomeA.3NN <- lm(YA3 ~ group, data = ANN)
OutcomeA.4NN <- lm(YA4 ~ group, data = ANN)
OutcomeA.1NNATC <- lm(YA1 ~ ATCgroup, data = ANNATC)
OutcomeA.2NNATC <- lm(YA2 ~ ATCgroup, data = ANNATC)
```

```
OutcomeA.3NNATC <- lm(YA3 ~ ATCgroup, data = ANNATC)
OutcomeA.4NNATC <- lm(YA4 ~ ATCgroup, data = ANNATC)

#Cohen's D for estimated treatment effect
tANNY1<-summary(OutcomeA.1NN)$coef[2, 3]
tANNY2<-summary(OutcomeA.2NN)$coef[2, 3]
tANNY3<-summary(OutcomeA.3NN)$coef[2, 3]
tANNY4<-summary(OutcomeA.4NN)$coef[2, 3]
treatnANN<-nobs(ANN$group[ANN$group==1])
compnANN<-nobs(ANN$group[ANN$group==0])

cohenANNY1 <-(tANNY1)*(((1/(treatnANN))+(1/(compnANN)))^0.5)
cohenANNY2 <-(tANNY2)*(((1/(treatnANN))+(1/(compnANN)))^0.5)
cohenANNY3 <-(tANNY3)*(((1/(treatnANN))+(1/(compnANN)))^0.5)
cohenANNY4 <-(tANNY4)*(((1/(treatnANN))+(1/(compnANN)))^0.5)

#Cohen's D for estimated treatment effect, ATC coding
tANNATCY1<-summary(OutcomeA.1NNATC)$coef[2, 3]
tANNATCY2<-summary(OutcomeA.2NNATC)$coef[2, 3]
tANNATCY3<-summary(OutcomeA.3NNATC)$coef[2, 3]
tANNATCY4<-summary(OutcomeA.4NNATC)$coef[2, 3]
treatnANNATC<-nobs(ANN$ATCgroup[ANN$ATCgroup==0])
compnANNATC<-nobs(ANN$ATCgroup[ANN$ATCgroup==1])

cohenANNATCY1 <-
(tANNATCY1)*(((1/(treatnANNATC))+(1/(compnANNATC)))^0.5)
cohenANNATCY2 <-
(tANNATCY2)*(((1/(treatnANNATC))+(1/(compnANNATC)))^0.5)
cohenANNATCY3 <-
(tANNATCY3)*(((1/(treatnANNATC))+(1/(compnANNATC)))^0.5)
cohenANNATCY4 <-
(tANNATCY4)*(((1/(treatnANNATC))+(1/(compnANNATC)))^0.5)

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#~~~~~~                    NN, 0.20 SD Caliper                       ~~~~~
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#NN with 0.20 Caliper - ATT Coding
require(MatchIt)
m.outANNCal =
matchit(finalDataA$group~finalDataA$x1+finalDataA$x2+finalDataA$x3+fina
lDataA$x4+finalDataA$x5, data=finalDataA, method="nearest",
caliper=0.20, m.order="random", ratio=1)

m.outANNCal

MANNCal <- summary(m.outANNCal, standardize = TRUE)

ANNCal <- match.data(m.outANNCal, group="all")

FullANNCal<-match.data(m.outANNCal, group="all", drop.unmatched =
FALSE)

#NN with 0.20 Caliper - ATC Coding
require(MatchIt)
m.outANNCalATC =
matchit(finalDataA$ATCgroup~finalDataA$x1+finalDataA$x2+finalDataA$x3+f
```

```
inalDataA$x4+finalDataA$x5, data=finalDataA, method="nearest",
caliper=0.20, m.order="random", ratio=1)

m.outANNCalATC

MANNCalATC <- summary(m.outANNCalATC, standardize = TRUE)

ANNCalATC <- match.data(m.outANNCalATC, group="all")

FullANNCalATC<-match.data(m.outANNCalATC, group="all", drop.unmatched =
FALSE)

#~~~~~~~~~~~ESTIMATED TREATMENT EFFECT~~~~~~~~~~~~~~~~~~
OutcomeA.1NNCal <- lm(YA1 ~ group, data = ANNCal)
OutcomeA.2NNCal <- lm(YA2 ~ group, data = ANNCal)
OutcomeA.3NNCal <- lm(YA3 ~ group, data = ANNCal)
OutcomeA.4NNCal <- lm(YA4 ~ group, data = ANNCal)
OutcomeA.1NNCalATC <- lm(YA1 ~ ATCgroup, data = ANNCalATC)
OutcomeA.2NNCalATC <- lm(YA2 ~ ATCgroup, data = ANNCalATC)
OutcomeA.3NNCalATC <- lm(YA3 ~ ATCgroup, data = ANNCalATC)
OutcomeA.4NNCalATC <- lm(YA4 ~ ATCgroup, data = ANNCalATC)

#Cohen's D for estimated treatment effect
tANNCalY1<-summary(OutcomeA.1NNCal)$coef[2, 3]
tANNCalY2<-summary(OutcomeA.2NNCal)$coef[2, 3]
tANNCalY3<-summary(OutcomeA.3NNCal)$coef[2, 3]
tANNCalY4<-summary(OutcomeA.4NNCal)$coef[2, 3]
treatnANNCal<-nobs(ANNCal$group[ANNCal$group==1])
compnANNCal<-nobs(ANNCal$group[ANNCal$group==0])

cohenANNCalY1 <-
(tANNCalY1)*(((1/(treatnANNCal))+(1/(compnANNCal)))^0.5)
cohenANNCalY2 <-
(tANNCalY2)*(((1/(treatnANNCal))+(1/(compnANNCal)))^0.5)
cohenANNCalY3 <-
(tANNCalY3)*(((1/(treatnANNCal))+(1/(compnANNCal)))^0.5)
cohenANNCalY4 <-
(tANNCalY4)*(((1/(treatnANNCal))+(1/(compnANNCal)))^0.5)

#Cohen's D for estimated treatment effect, ATC coding
tANNCalATCY1<-summary(OutcomeA.1NNCalATC)$coef[2, 3]
tANNCalATCY2<-summary(OutcomeA.2NNCalATC)$coef[2, 3]
tANNCalATCY3<-summary(OutcomeA.3NNCalATC)$coef[2, 3]
tANNCalATCY4<-summary(OutcomeA.4NNCalATC)$coef[2, 3]
treatnANNCalATC<-nobs(ANNCalATC$ATCgroup[ANNCalATC$ATCgroup==0])
compnANNCalATC<-nobs(ANNCalATC$ATCgroup[ANNCalATC$ATCgroup==1])

cohenANNCalATCY1 <-
(tANNCalATCY1)*(((1/(treatnANNCalATC))+(1/(compnANNCalATC)))^0.5)
cohenANNCalATCY2 <-
(tANNCalATCY2)*(((1/(treatnANNCalATC))+(1/(compnANNCalATC)))^0.5)
cohenANNCalATCY3 <-
(tANNCalATCY3)*(((1/(treatnANNCalATC))+(1/(compnANNCalATC)))^0.5)
cohenANNCalATCY4 <-
(tANNCalATCY4)*(((1/(treatnANNCalATC))+(1/(compnANNCalATC)))^0.5)

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

```
#~~~~~~               Genearlized Boosted Modeling                ~~~~~
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#GBM - ATT Coding
require(twang)
ps.AGBM <- ps (group~x1+x2+x3+x4+x5,
               data = finalDataA,
               n.trees=10000,          ## Max #iterations (from 1 to n)
               interaction.depth=3,    ## Level of interactions (default
= 3)
               shrinkage=0.01,         ## Allowable level of shrinkage
               stop.method=c("es.mean"),   ## Other options are es.max
or ks.mean
               estimand="ATT",         ## Other option is ATE
               verbose=FALSE)    ## Do you want a ton of information?
(TRUE)

MAGBM <- summary(ps.AGBM)
MAGBM
BalAGBM <- bal.table(ps.AGBM)
BalAGBM

finalDataA$w <- get.weights(ps.AGBM, stop.method="es.mean")
options(scipen=999)

designA.ps <- svydesign (ids = ~1, weights = ~w, data = finalDataA)

#GBM - ATC Coding
require(twang)
ps.AGBMATC <- ps (ATCgroup~x1+x2+x3+x4+x5,
               data = finalDataA,
               n.trees=10000,          ## Max #iterations (from 1 to n)
               interaction.depth=3,    ## Level of interactions (default
= 3)
               shrinkage=0.01,         ## Allowable level of shrinkage
               stop.method=c("es.mean"),  ## Other options are es.max
or ks.mean
               estimand="ATT",         ##Technically ATC b/c of the
grouping variable that is being used
               verbose=FALSE)    ## Do you want a ton of information?
(TRUE)

MAGBMATC <- summary(ps.AGBMATC)

BalAGBMATC <- bal.table(ps.AGBMATC)

finalDataA$wATC <- get.weights(ps.AGBMATC, stop.method="es.mean")
options(scipen=999)

designAATC.ps <- svydesign (ids = ~1, weights = ~wATC, data =
finalDataA)

#~~~~~~~~~~~ESTIMATED TREATMENT EFFECT~~~~~~~~~~~~~~~~~
OutcomeA.1GBM <- svyglm(YA1 ~ group, design = designA.ps)
OutcomeA.2GBM <- svyglm(YA2 ~ group, design = designA.ps)
OutcomeA.3GBM <- svyglm(YA3 ~ group, design = designA.ps)
OutcomeA.4GBM <- svyglm(YA4 ~ group, design = designA.ps)
OutcomeA.1GBMATC <- svyglm(YA1 ~ ATCgroup, design = designAATC.ps)
```

```
OutcomeA.2GBMATC <- svyglm(YA2 ~ ATCgroup, design = designAATC.ps)
OutcomeA.3GBMATC <- svyglm(YA3 ~ ATCgroup, design = designAATC.ps)
OutcomeA.4GBMATC <- svyglm(YA4 ~ ATCgroup, design = designAATC.ps)

#Cohen's D for estimated treatment effect
tAGBMY1<-summary(OutcomeA.1GBM)$coef[2, 3]
tAGBMY2<-summary(OutcomeA.2GBM)$coef[2, 3]
tAGBMY3<-summary(OutcomeA.3GBM)$coef[2, 3]
tAGBMY4<-summary(OutcomeA.4GBM)$coef[2, 3]
treatnAGBM<-nobs(finalDataA$group[finalDataA$group==1])
compnAGBM<-nobs(finalDataA$group[finalDataA$group==0])

cohenAGBMY1 <-(tAGBMY1)*(((1/(treatnAGBM))+(1/(compnAGBM)))^0.5)
cohenAGBMY2 <-(tAGBMY2)*(((1/(treatnAGBM))+(1/(compnAGBM)))^0.5)
cohenAGBMY3 <-(tAGBMY3)*(((1/(treatnAGBM))+(1/(compnAGBM)))^0.5)
cohenAGBMY4 <-(tAGBMY4)*(((1/(treatnAGBM))+(1/(compnAGBM)))^0.5)

#Cohen's D for estimated treatment effect, ATC coding
tAGBMATCY1<-summary(OutcomeA.1GBMATC)$coef[2, 3]
tAGBMATCY2<-summary(OutcomeA.2GBMATC)$coef[2, 3]
tAGBMATCY3<-summary(OutcomeA.3GBMATC)$coef[2, 3]
tAGBMATCY4<-summary(OutcomeA.4GBMATC)$coef[2, 3]
treatnAGBMATC<-nobs(finalDataA$ATCgroup[finalDataA$ATCgroup==0])
compnAGBMATC<-nobs(finalDataA$ATCgroup[finalDataA$ATCgroup==1])

cohenAGBMATCY1 <-
(tAGBMATCY1)*(((1/(treatnAGBMATC))+(1/(compnAGBMATC)))^0.5)
cohenAGBMATCY2 <-
(tAGBMATCY2)*(((1/(treatnAGBMATC))+(1/(compnAGBMATC)))^0.5)
cohenAGBMATCY3 <-
(tAGBMATCY3)*(((1/(treatnAGBMATC))+(1/(compnAGBMATC)))^0.5)
cohenAGBMATCY4 <-
(tAGBMATCY4)*(((1/(treatnAGBMATC))+(1/(compnAGBMATC)))^0.5)

#~~~~~~~~~~~~~~~~~COMPUTING PBR FOR GBM~~~~~~~~~~~~~~~~~~
X1TreatABef   <-BalAGBM$unw$`tx.mn`[1]
X2TreatABef   <-BalAGBM$unw$`tx.mn`[2]
X3TreatABef   <-BalAGBM$unw$`tx.mn`[3]
X4TreatABef   <-BalAGBM$unw$`tx.mn`[4]
X5TreatABef   <-BalAGBM$unw$`tx.mn`[5]

X1CompABef   <-BalAGBM$unw$`ct.mn`[1]
X2CompABef   <-BalAGBM$unw$`ct.mn`[2]
X3CompABef   <-BalAGBM$unw$`ct.mn`[3]
X4CompABef   <-BalAGBM$unw$`ct.mn`[4]
X5CompABef   <-BalAGBM$unw$`ct.mn`[5]

X1TreatAAft   <-BalAGBM$es.mean.ATT$`tx.mn`[1]
X2TreatAAft   <-BalAGBM$es.mean.ATT$`tx.mn`[2]
X3TreatAAft   <-BalAGBM$es.mean.ATT$`tx.mn`[3]
X4TreatAAft   <-BalAGBM$es.mean.ATT$`tx.mn`[4]
X5TreatAAft   <-BalAGBM$es.mean.ATT$`tx.mn`[5]

X1CompAAft   <-BalAGBM$es.mean.ATT$`ct.mn`[1]
X2CompAAft   <-BalAGBM$es.mean.ATT$`ct.mn`[2]
X3CompAAft   <-BalAGBM$es.mean.ATT$`ct.mn`[3]
X4CompAAft   <-BalAGBM$es.mean.ATT$`ct.mn`[4]
```

```
X5CompAAft   <-BalAGBM$es.mean.ATT$`ct.mn`[5]

PBRX1GBM   <-100*((abs(X1TreatABef-X1CompABef)-abs(X1TreatAAft-
X1CompAAft))/abs(X1TreatABef-X1CompABef))
PBRX2GBM   <-100*((abs(X2TreatABef-X2CompABef)-abs(X2TreatAAft-
X2CompAAft))/abs(X2TreatABef-X2CompABef))
PBRX3GBM   <-100*((abs(X3TreatABef-X3CompABef)-abs(X3TreatAAft-
X3CompAAft))/abs(X3TreatABef-X3CompABef))
PBRX4GBM   <-100*((abs(X4TreatABef-X4CompABef)-abs(X4TreatAAft-
X4CompAAft))/abs(X4TreatABef-X4CompABef))
PBRX5GBM   <-100*((abs(X5TreatABef-X5CompABef)-abs(X5TreatAAft-
X5CompAAft))/abs(X5TreatABef-X5CompABef))

#ATC Coding
X1TreatABefATC   <-BalAGBMATC$unw$`tx.mn`[1]
X2TreatABefATC   <-BalAGBMATC$unw$`tx.mn`[2]
X3TreatABefATC   <-BalAGBMATC$unw$`tx.mn`[3]
X4TreatABefATC   <-BalAGBMATC$unw$`tx.mn`[4]
X5TreatABefATC   <-BalAGBMATC$unw$`tx.mn`[5]

X1CompABefATC   <-BalAGBMATC$unw$`ct.mn`[1]
X2CompABefATC   <-BalAGBMATC$unw$`ct.mn`[2]
X3CompABefATC   <-BalAGBMATC$unw$`ct.mn`[3]
X4CompABefATC   <-BalAGBMATC$unw$`ct.mn`[4]
X5CompABefATC   <-BalAGBMATC$unw$`ct.mn`[5]

X1TreatAAftATC   <-BalAGBMATC$es.mean.ATT$`tx.mn`[1]
X2TreatAAftATC   <-BalAGBMATC$es.mean.ATT$`tx.mn`[2]
X3TreatAAftATC   <-BalAGBMATC$es.mean.ATT$`tx.mn`[3]
X4TreatAAftATC   <-BalAGBMATC$es.mean.ATT$`tx.mn`[4]
X5TreatAAftATC   <-BalAGBMATC$es.mean.ATT$`tx.mn`[5]

X1CompAAftATC   <-BalAGBMATC$es.mean.ATT$`ct.mn`[1]
X2CompAAftATC   <-BalAGBMATC$es.mean.ATT$`ct.mn`[2]
X3CompAAftATC   <-BalAGBMATC$es.mean.ATT$`ct.mn`[3]
X4CompAAftATC   <-BalAGBMATC$es.mean.ATT$`ct.mn`[4]
X5CompAAftATC   <-BalAGBMATC$es.mean.ATT$`ct.mn`[5]

PBRX1GBMATC   <-100*((abs(X1TreatABefATC-X1CompABefATC)-
abs(X1TreatAAftATC-X1CompAAftATC))/abs(X1TreatABefATC-X1CompABefATC))
PBRX2GBMATC   <-100*((abs(X2TreatABefATC-X2CompABefATC)-
abs(X2TreatAAftATC-X2CompAAftATC))/abs(X2TreatABefATC-X2CompABefATC))
PBRX3GBMATC   <-100*((abs(X3TreatABefATC-X3CompABefATC)-
abs(X3TreatAAftATC-X3CompAAftATC))/abs(X3TreatABefATC-X3CompABefATC))
PBRX4GBMATC   <-100*((abs(X4TreatABefATC-X4CompABefATC)-
abs(X4TreatAAftATC-X4CompAAftATC))/abs(X4TreatABefATC-X4CompABefATC))
PBRX5GBMATC   <-100*((abs(X5TreatABefATC-X5CompABefATC)-
abs(X5TreatAAftATC-X5CompAAftATC))/abs(X5TreatABefATC-X5CompABefATC))

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#~~~~~~~~~~~~~~~~~         Saving Out Diagnostics       ~~~~~~~~~~~~~~~~
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

#All variables BEFORE matching/weighting
AvgX1TreatA[i]   <-  mean(finalDataA$x1[finalDataA$group==1])
AvgX2TreatA[i]   <-  mean(finalDataA$x2[finalDataA$group==1])
AvgX3TreatA[i]   <-  mean(finalDataA$x3[finalDataA$group==1])
```

```
AvgX4TreatA[i]   <-  mean(finalDataA$x4[finalDataA$group==1])
AvgX5TreatA[i]   <-  mean(finalDataA$x5[finalDataA$group==1])
AvgYA1TreatA[i]    <-  mean(finalDataA$YA1[finalDataA$group==1])
AvgYA2TreatA[i]    <-  mean(finalDataA$YA2[finalDataA$group==1])
AvgYA3TreatA[i]    <-  mean(finalDataA$YA3[finalDataA$group==1])
AvgYA4TreatA[i]    <-  mean(finalDataA$YA4[finalDataA$group==1])
AvgPSTreatA[i]     <-  mean(finalDataA$PS[finalDataA$group==1])

AvgX1CompA[i]   <-  mean(finalDataA$x1[finalDataA$group==0])
AvgX2CompA[i]   <-  mean(finalDataA$x2[finalDataA$group==0])
AvgX3CompA[i]   <-  mean(finalDataA$x3[finalDataA$group==0])
AvgX4CompA[i]   <-  mean(finalDataA$x4[finalDataA$group==0])
AvgX5CompA[i]   <-  mean(finalDataA$x5[finalDataA$group==0])
AvgYA1CompA[i]    <-  mean(finalDataA$YA1[finalDataA$group==0])
AvgYA2CompA[i]    <-  mean(finalDataA$YA2[finalDataA$group==0])
AvgYA3CompA[i]    <-  mean(finalDataA$YA3[finalDataA$group==0])
AvgYA4CompA[i]    <-  mean(finalDataA$YA4[finalDataA$group==0])
AvgPSCompA[i]     <-  mean(finalDataA$PS[finalDataA$group==0])

SDX1TreatA[i]   <-  sd(finalDataA$x1[finalDataA$group==1])
SDX2TreatA[i]   <-  sd(finalDataA$x2[finalDataA$group==1])
SDX3TreatA[i]   <-  sd(finalDataA$x3[finalDataA$group==1])
SDX4TreatA[i]   <-  sd(finalDataA$x4[finalDataA$group==1])
SDX5TreatA[i]   <-  sd(finalDataA$x5[finalDataA$group==1])
SDYA1TreatA[i]    <-  sd(finalDataA$YA1[finalDataA$group==1])
SDYA2TreatA[i]    <-  sd(finalDataA$YA2[finalDataA$group==1])
SDYA3TreatA[i]    <-  sd(finalDataA$YA3[finalDataA$group==1])
SDYA4TreatA[i]    <-  sd(finalDataA$YA4[finalDataA$group==1])
SDPSTreatA[i]     <-  sd(finalDataA$PS[finalDataA$group==1])

SDX1CompA[i]   <-  sd(finalDataA$x1[finalDataA$group==0])
SDX2CompA[i]   <-  sd(finalDataA$x2[finalDataA$group==0])
SDX3CompA[i]   <-  sd(finalDataA$x3[finalDataA$group==0])
SDX4CompA[i]   <-  sd(finalDataA$x4[finalDataA$group==0])
SDX5CompA[i]   <-  sd(finalDataA$x5[finalDataA$group==0])
SDYA1CompA[i]    <-  sd(finalDataA$YA1[finalDataA$group==0])
SDYA2CompA[i]    <-  sd(finalDataA$YA2[finalDataA$group==0])
SDYA3CompA[i]    <-  sd(finalDataA$YA3[finalDataA$group==0])
SDYA4CompA[i]    <-  sd(finalDataA$YA4[finalDataA$group==0])
SDPSCompA[i]     <-  sd(finalDataA$PS[finalDataA$group==0])

#Standardized mean differences before matching, ATT Coding
SMD_X1_All[i] <-  MANN$sum.all[c(14)]
SMD_X2_All[i] <-  MANN$sum.all[c(15)]
SMD_X3_All[i] <-  MANN$sum.all[c(16)]
SMD_X4_All[i] <-  MANN$sum.all[c(17)]
SMD_X5_All[i] <-  MANN$sum.all[c(18)]
SMD_PS_All[i] <-  MANN$sum.all[c(13)]

#Standardized mean differences before matching, ATC Coding
SMD_X1_AllATC[i] <-  MANNATC$sum.all[c(14)]
SMD_X2_AllATC[i] <-  MANNATC$sum.all[c(15)]
SMD_X3_AllATC[i] <-  MANNATC$sum.all[c(16)]
SMD_X4_AllATC[i] <-  MANNATC$sum.all[c(17)]
SMD_X5_AllATC[i] <-  MANNATC$sum.all[c(18)]
SMD_PS_AllATC[i] <-  MANNATC$sum.all[c(13)]
```

```
#Correlations
Cor_X1.X2_A[i]     <-  cor(finalDataA$x1, finalDataA$x2)
Cor_X1.X3_A[i]     <-  cor(finalDataA$x1, finalDataA$x3)
Cor_X1.X4_A[i]     <-  cor(finalDataA$x1, finalDataA$x4)
Cor_X1.X5_A[i]     <-  cor(finalDataA$x1, finalDataA$x5)
Cor_X2.X3_A[i]     <-  cor(finalDataA$x2, finalDataA$x3)
Cor_X2.X4_A[i]     <-  cor(finalDataA$x2, finalDataA$x4)
Cor_X2.X5_A[i]     <-  cor(finalDataA$x2, finalDataA$x5)
Cor_X3.X4_A[i]     <-  cor(finalDataA$x3, finalDataA$x4)
Cor_X3.X5_A[i]     <-  cor(finalDataA$x3, finalDataA$x5)
Cor_X4.X5_A[i]     <-  cor(finalDataA$x4, finalDataA$x5)
Cor_X1.PS_A[i]     <-  cor(finalDataA$x1, finalDataA$PS)
Cor_X2.PS_A[i]     <-  cor(finalDataA$x2, finalDataA$PS)
Cor_X3.PS_A[i]     <-  cor(finalDataA$x3, finalDataA$PS)
Cor_X4.PS_A[i]     <-  cor(finalDataA$x4, finalDataA$PS)
Cor_X5.PS_A[i]     <-  cor(finalDataA$x5, finalDataA$PS)
Cor_X1.Y1_A[i]     <-  cor(finalDataA$x1, finalDataA$YA1)
Cor_X2.Y1_A[i]     <-  cor(finalDataA$x2, finalDataA$YA1)
Cor_X3.Y1_A[i]     <-  cor(finalDataA$x3, finalDataA$YA1)
Cor_X4.Y1_A[i]     <-  cor(finalDataA$x4, finalDataA$YA1)
Cor_X5.Y1_A[i]     <-  cor(finalDataA$x5, finalDataA$YA1)
Cor_X1.Y2_A[i]     <-  cor(finalDataA$x1, finalDataA$YA2)
Cor_X2.Y2_A[i]     <-  cor(finalDataA$x2, finalDataA$YA2)
Cor_X3.Y2_A[i]     <-  cor(finalDataA$x3, finalDataA$YA2)
Cor_X4.Y2_A[i]     <-  cor(finalDataA$x4, finalDataA$YA2)
Cor_X5.Y2_A[i]     <-  cor(finalDataA$x5, finalDataA$YA2)
Cor_X1.Y3_A[i]     <-  cor(finalDataA$x1, finalDataA$YA3)
Cor_X2.Y3_A[i]     <-  cor(finalDataA$x2, finalDataA$YA3)
Cor_X3.Y3_A[i]     <-  cor(finalDataA$x3, finalDataA$YA3)
Cor_X4.Y3_A[i]     <-  cor(finalDataA$x4, finalDataA$YA3)
Cor_X5.Y3_A[i]     <-  cor(finalDataA$x5, finalDataA$YA3)
Cor_X1.Y4_A[i]     <-  cor(finalDataA$x1, finalDataA$YA4)
Cor_X2.Y4_A[i]     <-  cor(finalDataA$x2, finalDataA$YA4)
Cor_X3.Y4_A[i]     <-  cor(finalDataA$x3, finalDataA$YA4)
Cor_X4.Y4_A[i]     <-  cor(finalDataA$x4, finalDataA$YA4)
Cor_X5.Y4_A[i]     <-  cor(finalDataA$x5, finalDataA$YA4)
Cor_G.Y1_A[i]      <-  cor(finalDataA$group, finalDataA$YA1)
Cor_G.Y2_A[i]      <-  cor(finalDataA$group, finalDataA$YA2)
Cor_G.Y3_A[i]      <-  cor(finalDataA$group, finalDataA$YA3)
Cor_G.Y4_A[i]      <-  cor(finalDataA$group, finalDataA$YA4)
#Population Group Regression Coefficient
PopY1A[i]          <- as.numeric(PopY1$coef[2])
PopY2A[i]          <- as.numeric(PopY2$coef[2])
PopY3A[i]          <- as.numeric(PopY3$coef[2])
PopY4A[i]          <- as.numeric(PopY4$coef[2])
#Population Group t-value for each regression coefficient
tPopY1A[i]             <-summary(PopY1)$coef[2, 3]
tPopY2A[i]             <-summary(PopY2)$coef[2, 3]
tPopY3A[i]             <-summary(PopY3)$coef[2, 3]
tPopY4A[i]             <-summary(PopY4)$coef[2, 3]
#Treatment Group N (before matching)
treatPopNA[i]          <-nobs(finalDataA$group[finalDataA$group==1])
#Comparison Group N (before matching)
compPopNA[i]           <-nobs(finalDataA$group[finalDataA$group==0])
#Population Cohen's D for Treatment Effect
PopCohenY1A[i]          <-cohenPopY1
PopCohenY2A[i]          <-cohenPopY2
```

```
PopCohenY3A[i]          <-cohenPopY3
PopCohenY4A[i]          <-cohenPopY4
#Baseline Group Regression Coefficient
BaseY1A[i]          <- as.numeric(baseout1$coef[2])
BaseY2A[i]          <- as.numeric(baseout2$coef[2])
BaseY3A[i]          <- as.numeric(baseout3$coef[2])
BaseY4A[i]          <- as.numeric(baseout4$coef[2])
#Baseline Group t-value for each regression coefficient
tBaseY1A[i]             <-summary(baseout1)$coef[2, 3]
tBaseY2A[i]             <-summary(baseout2)$coef[2, 3]
tBaseY3A[i]             <-summary(baseout3)$coef[2, 3]
tBaseY4A[i]             <-summary(baseout4)$coef[2, 3]
#Treatment Group N (before matching)
treatBaseNA[i]          <-nobs(finalDataA$group[finalDataA$group==1])
#Comparison Group N (before matching)
compBaseNA[i]           <-nobs(finalDataA$group[finalDataA$group==0])
#Baseline Cohen's D for Treatment Effect
BaseCohenY1A[i]          <-cohenBaseY1
BaseCohenY2A[i]          <-cohenBaseY2
BaseCohenY3A[i]          <-cohenBaseY3
BaseCohenY4A[i]          <-cohenBaseY4
#Population Group Regression Coefficient, ATC coding
PopATCY1A[i]          <- as.numeric(ATCPopY1$coef[2])
PopATCY2A[i]          <- as.numeric(ATCPopY2$coef[2])
PopATCY3A[i]          <- as.numeric(ATCPopY3$coef[2])
PopATCY4A[i]          <- as.numeric(ATCPopY4$coef[2])
#Population Group t-value for each regression coefficient, ATC coding
tPopATCY1A[i]             <-summary(ATCPopY1)$coef[2, 3]
tPopATCY2A[i]             <-summary(ATCPopY2)$coef[2, 3]
tPopATCY3A[i]             <-summary(ATCPopY3)$coef[2, 3]
tPopATCY4A[i]             <-summary(ATCPopY4)$coef[2, 3]
#Treatment Group N (before matching), ATC coding
treatPopNAATC[i]          <-
nobs(finalDataA$ATCgroup[finalDataA$ATCgroup==0])
#Comparison Group N (before matching), ATC coding
compPopNAATC[i]           <-
nobs(finalDataA$ATCgroup[finalDataA$ATCgroup==1])
#Population Cohen's D for Treatment Effect, ATC coding
PopCohenATCY1A[i]          <-cohenPopATCY1
PopCohenATCY2A[i]          <-cohenPopATCY2
PopCohenATCY3A[i]          <-cohenPopATCY3
PopCohenATCY4A[i]          <-cohenPopATCY4
#Baseline Group Regression Coefficient, ATC coding
BaseATCY1A[i]          <- as.numeric(ATCbaseout1$coef[2])
BaseATCY2A[i]          <- as.numeric(ATCbaseout2$coef[2])
BaseATCY3A[i]          <- as.numeric(ATCbaseout3$coef[2])
BaseATCY4A[i]          <- as.numeric(ATCbaseout4$coef[2])
#Baseline Group t-value for each regression coefficient, ATC coding
tBaseATCY1A[i]             <-summary(ATCbaseout1)$coef[2, 3]
tBaseATCY2A[i]             <-summary(ATCbaseout2)$coef[2, 3]
tBaseATCY3A[i]             <-summary(ATCbaseout3)$coef[2, 3]
tBaseATCY4A[i]             <-summary(ATCbaseout4)$coef[2, 3]
#Treatment Group N (before matching), ATC coding
treatBaseNAATC[i]          <-
nobs(finalDataA$ATCgroup[finalDataA$ATCgroup==0])
#Comparison Group N (before matching), ATC coding
```

```
compBaseNAATC[i]           <-
nobs(finalDataA$ATCgroup[finalDataA$ATCgroup==1])
#Baseline Cohen's D for Treatment Effect, ATC coding
BaseCohenATCY1A[i]          <-cohenBaseATCY1
BaseCohenATCY2A[i]          <-cohenBaseATCY2
BaseCohenATCY3A[i]          <-cohenBaseATCY3
BaseCohenATCY4A[i]          <-cohenBaseATCY4
#Variance Ratio for unmatched groups (baseline data
VRB[i] <-
var(finalDataA$PS[finalDataA$group==1])/var(finalDataA$PS[finalDataA$gr
oup==0])
VRBATC[i] <-
var(finalDataA$PS[finalDataA$ATCgroup==0])/var(finalDataA$PS[finalDataA
$ATCgroup==1])

#All variables AFTER Nearest Neighbor Matching, ATT Coding
AvgX1TreatANN[i]   <-  mean(ANN$x1[ANN$group==1])
AvgX2TreatANN[i]   <-  mean(ANN$x2[ANN$group==1])
AvgX3TreatANN[i]   <-  mean(ANN$x3[ANN$group==1])
AvgX4TreatANN[i]   <-  mean(ANN$x4[ANN$group==1])
AvgX5TreatANN[i]   <-  mean(ANN$x5[ANN$group==1])
AvgYA1TreatANN[i]   <-  mean(ANN$YA1[ANN$group==1])
AvgYA2TreatANN[i]   <-  mean(ANN$YA2[ANN$group==1])
AvgYA3TreatANN[i]   <-  mean(ANN$YA3[ANN$group==1])
AvgYA4TreatANN[i]   <-  mean(ANN$YA4[ANN$group==1])
AvgPSTreatANN[i]    <-  mean(ANN$PS[ANN$group==1])

AvgX1CompANN[i]   <-  mean(ANN$x1[ANN$group==0])
AvgX2CompANN[i]   <-  mean(ANN$x2[ANN$group==0])
AvgX3CompANN[i]   <-  mean(ANN$x3[ANN$group==0])
AvgX4CompANN[i]   <-  mean(ANN$x4[ANN$group==0])
AvgX5CompANN[i]   <-  mean(ANN$x5[ANN$group==0])
AvgYA1CompANN[i]   <-  mean(ANN$YA1[ANN$group==0])
AvgYA2CompANN[i]   <-  mean(ANN$YA2[ANN$group==0])
AvgYA3CompANN[i]   <-  mean(ANN$YA3[ANN$group==0])
AvgYA4CompANN[i]   <-  mean(ANN$YA4[ANN$group==0])
AvgPSCompANN[i]    <-  mean(ANN$PS[ANN$group==0])

SDX1TreatANN[i]   <-  sd(ANN$x1[ANN$group==1])
SDX2TreatANN[i]   <-  sd(ANN$x2[ANN$group==1])
SDX3TreatANN[i]   <-  sd(ANN$x3[ANN$group==1])
SDX4TreatANN[i]   <-  sd(ANN$x4[ANN$group==1])
SDX5TreatANN[i]   <-  sd(ANN$x5[ANN$group==1])
SDYA1TreatANN[i]   <-  sd(ANN$YA1[ANN$group==1])
SDYA2TreatANN[i]   <-  sd(ANN$YA2[ANN$group==1])
SDYA3TreatANN[i]   <-  sd(ANN$YA3[ANN$group==1])
SDYA4TreatANN[i]   <-  sd(ANN$YA4[ANN$group==1])
SDPSTreatANN[i]    <-  sd(ANN$PS[ANN$group==1])

SDX1CompANN[i]   <-  sd(ANN$x1[ANN$group==0])
SDX2CompANN[i]   <-  sd(ANN$x2[ANN$group==0])
SDX3CompANN[i]   <-  sd(ANN$x3[ANN$group==0])
SDX4CompANN[i]   <-  sd(ANN$x4[ANN$group==0])
SDX5CompANN[i]   <-  sd(ANN$x5[ANN$group==0])
SDYA1CompANN[i]   <-  sd(ANN$YA1[ANN$group==0])
SDYA2CompANN[i]   <-  sd(ANN$YA2[ANN$group==0])
SDYA3CompANN[i]   <-  sd(ANN$YA3[ANN$group==0])
```

```
SDYA4CompANN[i]   <-  sd(ANN$YA4[ANN$group==0])
SDPSCompANN[i]    <-  sd(ANN$PS[ANN$group==0])

#Standardized Mean Difference after NN matching
SMD_X1_ANN[i] <-  MANN$sum.matched[c(14)]
SMD_X2_ANN[i] <-  MANN$sum.matched[c(15)]
SMD_X3_ANN[i] <-  MANN$sum.matched[c(16)]
SMD_X4_ANN[i] <-  MANN$sum.matched[c(17)]
SMD_X5_ANN[i] <-  MANN$sum.matched[c(18)]
SMD_PS_ANN[i] <-  MANN$sum.matched[c(13)]

#Percent Bias Reduction after NN matching
PBR_X1_ANN[i]  <- MANN$reduction[c(2)]
PBR_X2_ANN[i]  <- MANN$reduction[c(3)]
PBR_X3_ANN[i]  <- MANN$reduction[c(4)]
PBR_X4_ANN[i]  <- MANN$reduction[c(5)]
PBR_X5_ANN[i]  <- MANN$reduction[c(6)]
PBR_PS_ANN[i]  <- MANN$reduction[c(1)]

#Group Regression Coefficient after NN matching
Y1ANN[i]          <- as.numeric(OutcomeA.1NN$coef[2])
Y2ANN[i]          <- as.numeric(OutcomeA.2NN$coef[2])
Y3ANN[i]          <- as.numeric(OutcomeA.3NN$coef[2])
Y4ANN[i]          <- as.numeric(OutcomeA.4NN$coef[2])
#Group t-value for each regression coefficient after NN matching
tNNY1A[i]             <-summary(OutcomeA.1NN)$coef[2, 3]
tNNY2A[i]             <-summary(OutcomeA.2NN)$coef[2, 3]
tNNY3A[i]             <-summary(OutcomeA.3NN)$coef[2, 3]
tNNY4A[i]             <-summary(OutcomeA.4NN)$coef[2, 3]
#Treatment Group N (after NN matching)
NNtreatNA[i]          <-nobs(ANN$group[ANN$group==1])
#Comparison Group N (after NN matching)
NNcompNA[i]           <-nobs(ANN$group[ANN$group==0])
#Cohen's D for Treatment Effect after NN matching
NNCohenY1A[i]          <-cohenANNY1
NNCohenY2A[i]          <-cohenANNY2
NNCohenY3A[i]          <-cohenANNY3
NNCohenY4A[i]          <-cohenANNY4

#Matched & Unmatched PS mean, median, and sd after NN matching
PSMeanMatchedTreatANN[i]   <-  mean(FullANN$distance[FullANN$group==1 &
FullANN$weights==1])
PSMeanMatchedCompANN[i]   <-  mean(FullANN$distance[FullANN$group==0 &
FullANN$weights==1])
PSMeanUnMatchedTreatANN[i]   <-  mean(FullANN$distance[FullANN$group==1
& FullANN$weights==0])
PSMeanUnMatchedCompANN[i]   <-  mean(FullANN$distance[FullANN$group==0
& FullANN$weights==0])

PSMedMatchedTreatANN[i]   <-  median(FullANN$distance[FullANN$group==1
& FullANN$weights==1])
PSMedMatchedCompANN[i]   <-  median(FullANN$distance[FullANN$group==0 &
FullANN$weights==1])
PSMedUnMatchedTreatANN[i]   <-
median(FullANN$distance[FullANN$group==1 & FullANN$weights==0])
PSMedUnMatchedCompANN[i]   <-  median(FullANN$distance[FullANN$group==0
& FullANN$weights==0])
```

```
PSsdMatchedTreatANN[i]   <-  sd(FullANN$distance[FullANN$group==1 &
FullANN$weights==1])
PSsdMatchedCompANN[i]   <-  sd(FullANN$distance[FullANN$group==0 &
FullANN$weights==1])
PSsdUnMatchedTreatANN[i]   <-  sd(FullANN$distance[FullANN$group==1 &
FullANN$weights==0])
PSsdUnMatchedCompANN[i]   <-  sd(FullANN$distance[FullANN$group==0 &
FullANN$weights==0])

#Variance Ratio for matched groups, after NN Matching
VRANN[i] <-
var(ANN$distance[ANN$group==1])/var(ANN$distance[ANN$group==0])

#All variables AFTER Nearest Neighbor Matching, ATC Coding
AvgX1TreatANNATC[i]   <-  mean(ANNATC$x1[ANNATC$ATCgroup==0])
AvgX2TreatANNATC[i]   <-  mean(ANNATC$x2[ANNATC$ATCgroup==0])
AvgX3TreatANNATC[i]   <-  mean(ANNATC$x3[ANNATC$ATCgroup==0])
AvgX4TreatANNATC[i]   <-  mean(ANNATC$x4[ANNATC$ATCgroup==0])
AvgX5TreatANNATC[i]   <-  mean(ANNATC$x5[ANNATC$ATCgroup==0])
AvgYA1TreatANNATC[i]   <-  mean(ANNATC$YA1[ANNATC$ATCgroup==0])
AvgYA2TreatANNATC[i]   <-  mean(ANNATC$YA2[ANNATC$ATCgroup==0])
AvgYA3TreatANNATC[i]   <-  mean(ANNATC$YA3[ANNATC$ATCgroup==0])
AvgYA4TreatANNATC[i]   <-  mean(ANNATC$YA4[ANNATC$ATCgroup==0])
AvgPSTreatANNATC[i]    <-  mean(ANNATC$PS[ANNATC$ATCgroup==0])

AvgX1CompANNATC[i]   <-  mean(ANNATC$x1[ANNATC$ATCgroup==1])
AvgX2CompANNATC[i]   <-  mean(ANNATC$x2[ANNATC$ATCgroup==1])
AvgX3CompANNATC[i]   <-  mean(ANNATC$x3[ANNATC$ATCgroup==1])
AvgX4CompANNATC[i]   <-  mean(ANNATC$x4[ANNATC$ATCgroup==1])
AvgX5CompANNATC[i]   <-  mean(ANNATC$x5[ANNATC$ATCgroup==1])
AvgYA1CompANNATC[i]   <-  mean(ANNATC$YA1[ANNATC$ATCgroup==1])
AvgYA2CompANNATC[i]   <-  mean(ANNATC$YA2[ANNATC$ATCgroup==1])
AvgYA3CompANNATC[i]   <-  mean(ANNATC$YA3[ANNATC$ATCgroup==1])
AvgYA4CompANNATC[i]   <-  mean(ANNATC$YA4[ANNATC$ATCgroup==1])
AvgPSCompANNATC[i]    <-  mean(ANNATC$PS[ANNATC$ATCgroup==1])

SDX1TreatANNATC[i]   <-  sd(ANNATC$x1[ANNATC$ATCgroup==0])
SDX2TreatANNATC[i]   <-  sd(ANNATC$x2[ANNATC$ATCgroup==0])
SDX3TreatANNATC[i]   <-  sd(ANNATC$x3[ANNATC$ATCgroup==0])
SDX4TreatANNATC[i]   <-  sd(ANNATC$x4[ANNATC$ATCgroup==0])
SDX5TreatANNATC[i]   <-  sd(ANNATC$x5[ANNATC$ATCgroup==0])
SDYA1TreatANNATC[i]   <-  sd(ANNATC$YA1[ANNATC$ATCgroup==0])
SDYA2TreatANNATC[i]   <-  sd(ANNATC$YA2[ANNATC$ATCgroup==0])
SDYA3TreatANNATC[i]   <-  sd(ANNATC$YA3[ANNATC$ATCgroup==0])
SDYA4TreatANNATC[i]   <-  sd(ANNATC$YA4[ANNATC$ATCgroup==0])
SDPSTreatANNATC[i]    <-  sd(ANNATC$PS[ANNATC$ATCgroup==0])

SDX1CompANNATC[i]   <-  sd(ANNATC$x1[ANNATC$ATCgroup==1])
SDX2CompANNATC[i]   <-  sd(ANNATC$x2[ANNATC$ATCgroup==1])
SDX3CompANNATC[i]   <-  sd(ANNATC$x3[ANNATC$ATCgroup==1])
SDX4CompANNATC[i]   <-  sd(ANNATC$x4[ANNATC$ATCgroup==1])
SDX5CompANNATC[i]   <-  sd(ANNATC$x5[ANNATC$ATCgroup==1])
SDYA1CompANNATC[i]   <-  sd(ANNATC$YA1[ANNATC$ATCgroup==1])
SDYA2CompANNATC[i]   <-  sd(ANNATC$YA2[ANNATC$ATCgroup==1])
SDYA3CompANNATC[i]   <-  sd(ANNATC$YA3[ANNATC$ATCgroup==1])
SDYA4CompANNATC[i]   <-  sd(ANNATC$YA4[ANNATC$ATCgroup==1])
```

```
SDPSCompANNATC[i]    <-  sd(ANNATC$PS[ANNATC$ATCgroup==1])

#Standardized Mean Difference after NN matching, ATC coding
SMD_X1_ANNATC[i] <-  MANNATC$sum.matched[c(14)]
SMD_X2_ANNATC[i] <-  MANNATC$sum.matched[c(15)]
SMD_X3_ANNATC[i] <-  MANNATC$sum.matched[c(16)]
SMD_X4_ANNATC[i] <-  MANNATC$sum.matched[c(17)]
SMD_X5_ANNATC[i] <-  MANNATC$sum.matched[c(18)]
SMD_PS_ANNATC[i] <-  MANNATC$sum.matched[c(13)]

#Percent Bias Reduction after NN matching, ATC coding
PBR_X1_ANNATC[i]  <- MANNATC$reduction[c(2)]
PBR_X2_ANNATC[i]  <- MANNATC$reduction[c(3)]
PBR_X3_ANNATC[i]  <- MANNATC$reduction[c(4)]
PBR_X4_ANNATC[i]  <- MANNATC$reduction[c(5)]
PBR_X5_ANNATC[i]  <- MANNATC$reduction[c(6)]
PBR_PS_ANNATC[i]  <- MANNATC$reduction[c(1)]

#Group Regression Coefficient after NN matching, ATC coding
Y1ANNATC[i]          <- as.numeric(OutcomeA.1NNATC$coef[2])
Y2ANNATC[i]          <- as.numeric(OutcomeA.2NNATC$coef[2])
Y3ANNATC[i]          <- as.numeric(OutcomeA.3NNATC$coef[2])
Y4ANNATC[i]          <- as.numeric(OutcomeA.4NNATC$coef[2])
#Group t-value for each regression coefficient after NN matching, ATC
coding
tNNATCY1A[i]             <-summary(OutcomeA.1NNATC)$coef[2, 3]
tNNATCY2A[i]             <-summary(OutcomeA.2NNATC)$coef[2, 3]
tNNATCY3A[i]             <-summary(OutcomeA.3NNATC)$coef[2, 3]
tNNATCY4A[i]             <-summary(OutcomeA.4NNATC)$coef[2, 3]
#Treatment Group N (after NN matching), ATC coding
NNATCtreatNA[i]          <-nobs(ANNATC$group[ANNATC$group==0])
#Comparison Group N (after NN matching)
NNATCcompNA[i]           <-nobs(ANNATC$group[ANNATC$group==1])
#Cohen's D for Treatment Effect after NN matching, ATC coding
NNATCCohenY1A[i]          <-cohenANNATCY1
NNATCCohenY2A[i]          <-cohenANNATCY2
NNATCCohenY3A[i]          <-cohenANNATCY3
NNATCCohenY4A[i]          <-cohenANNATCY4

#Matched & Unmatched PS mean, median, and sd after NN matching, ATC
Coding
PSMeanMatchedTreatANNATC[i]   <-
mean(FullANNATC$distance[FullANNATC$group==0 & FullANNATC$weights==1])
PSMeanMatchedCompANNATC[i]   <-
mean(FullANNATC$distance[FullANNATC$group==1 & FullANNATC$weights==1])
PSMeanUnMatchedTreatANNATC[i]   <-
mean(FullANNATC$distance[FullANNATC$group==0 & FullANNATC$weights==0])
PSMeanUnMatchedCompANNATC[i]   <-
mean(FullANNATC$distance[FullANNATC$group==1 & FullANNATC$weights==0])

PSMedMatchedTreatANNATC[i]   <-
median(FullANNATC$distance[FullANNATC$group==0 &
FullANNATC$weights==1])
PSMedMatchedCompANNATC[i]   <-
median(FullANNATC$distance[FullANNATC$group==1 &
FullANNATC$weights==1])
```

```
PSMedUnMatchedTreatANNATC[i]   <-
median(FullANNATC$distance[FullANNATC$group==0 &
FullANNATC$weights==0])
PSMedUnMatchedCompANNATC[i]   <-
median(FullANNATC$distance[FullANNATC$group==1 &
FullANNATC$weights==0])

PSsdMatchedTreatANNATC[i]   <-
sd(FullANNATC$distance[FullANNATC$group==0 & FullANNATC$weights==1])
PSsdMatchedCompANNATC[i]   <-
sd(FullANNATC$distance[FullANNATC$group==1 & FullANNATC$weights==1])
PSsdUnMatchedTreatANNATC[i]   <-
sd(FullANNATC$distance[FullANNATC$group==0 & FullANNATC$weights==0])
PSsdUnMatchedCompANNATC[i]   <-
sd(FullANNATC$distance[FullANNATC$group==1 & FullANNATC$weights==0])

#Variance Ratio for matched groups, after NN Matching, ATC coding
VRANNATC[i] <-
var(ANNATC$distance[ANNATC$ATCgroup==0])/var(ANNATC$distance[ANNATC$ATC
group==1])

#All variables AFTER Nearest Neighbor Matching with 0.20 Caliper, ATT
Coding
AvgX1TreatANNCal[i]   <-  mean(ANNCal$x1[ANNCal$group==1])
AvgX2TreatANNCal[i]   <-  mean(ANNCal$x2[ANNCal$group==1])
AvgX3TreatANNCal[i]   <-  mean(ANNCal$x3[ANNCal$group==1])
AvgX4TreatANNCal[i]   <-  mean(ANNCal$x4[ANNCal$group==1])
AvgX5TreatANNCal[i]   <-  mean(ANNCal$x5[ANNCal$group==1])
AvgYA1TreatANNCal[i]   <-  mean(ANNCal$YA1[ANNCal$group==1])
AvgYA2TreatANNCal[i]   <-  mean(ANNCal$YA2[ANNCal$group==1])
AvgYA3TreatANNCal[i]   <-  mean(ANNCal$YA3[ANNCal$group==1])
AvgYA4TreatANNCal[i]   <-  mean(ANNCal$YA4[ANNCal$group==1])
AvgPSTreatANNCal[i]    <-  mean(ANNCal$PS[ANNCal$group==1])

AvgX1CompANNCal[i]   <-  mean(ANNCal$x1[ANNCal$group==0])
AvgX2CompANNCal[i]   <-  mean(ANNCal$x2[ANNCal$group==0])
AvgX3CompANNCal[i]   <-  mean(ANNCal$x3[ANNCal$group==0])
AvgX4CompANNCal[i]   <-  mean(ANNCal$x4[ANNCal$group==0])
AvgX5CompANNCal[i]   <-  mean(ANNCal$x5[ANNCal$group==0])
AvgYA1CompANNCal[i]   <-  mean(ANNCal$YA1[ANNCal$group==0])
AvgYA2CompANNCal[i]   <-  mean(ANNCal$YA2[ANNCal$group==0])
AvgYA3CompANNCal[i]   <-  mean(ANNCal$YA3[ANNCal$group==0])
AvgYA4CompANNCal[i]   <-  mean(ANNCal$YA4[ANNCal$group==0])
AvgPSCompANNCal[i]    <-  mean(ANNCal$PS[ANNCal$group==0])

SDX1TreatANNCal[i]   <-  sd(ANNCal$x1[ANNCal$group==1])
SDX2TreatANNCal[i]   <-  sd(ANNCal$x2[ANNCal$group==1])
SDX3TreatANNCal[i]   <-  sd(ANNCal$x3[ANNCal$group==1])
SDX4TreatANNCal[i]   <-  sd(ANNCal$x4[ANNCal$group==1])
SDX5TreatANNCal[i]   <-  sd(ANNCal$x5[ANNCal$group==1])
SDYA1TreatANNCal[i]   <-  sd(ANNCal$YA1[ANNCal$group==1])
SDYA2TreatANNCal[i]   <-  sd(ANNCal$YA2[ANNCal$group==1])
SDYA3TreatANNCal[i]   <-  sd(ANNCal$YA3[ANNCal$group==1])
SDYA4TreatANNCal[i]   <-  sd(ANNCal$YA4[ANNCal$group==1])
SDPSTreatANNCal[i]    <-  sd(ANNCal$PS[ANNCal$group==1])

SDX1CompANNCal[i]   <-  sd(ANNCal$x1[ANNCal$group==0])
```

```
SDX2CompANNCal[i]   <-  sd(ANNCal$x2[ANNCal$group==0])
SDX3CompANNCal[i]   <-  sd(ANNCal$x3[ANNCal$group==0])
SDX4CompANNCal[i]   <-  sd(ANNCal$x4[ANNCal$group==0])
SDX5CompANNCal[i]   <-  sd(ANNCal$x5[ANNCal$group==0])
SDYA1CompANNCal[i]   <-  sd(ANNCal$YA1[ANNCal$group==0])
SDYA2CompANNCal[i]   <-  sd(ANNCal$YA2[ANNCal$group==0])
SDYA3CompANNCal[i]   <-  sd(ANNCal$YA3[ANNCal$group==0])
SDYA4CompANNCal[i]   <-  sd(ANNCal$YA4[ANNCal$group==0])
SDPSCompANNCal[i]    <-  sd(ANNCal$PS[ANNCal$group==0])

#Standardized Mean Difference after NN matching with caliper
SMD_X1_ANNCal[i] <-  MANNCal$sum.matched[c(14)]
SMD_X2_ANNCal[i] <-  MANNCal$sum.matched[c(15)]
SMD_X3_ANNCal[i] <-  MANNCal$sum.matched[c(16)]
SMD_X4_ANNCal[i] <-  MANNCal$sum.matched[c(17)]
SMD_X5_ANNCal[i] <-  MANNCal$sum.matched[c(18)]
SMD_PS_ANNCal[i] <-  MANNCal$sum.matched[c(13)]

#Percent Bias Reduction after NN matching with caliper
PBR_X1_ANNCal[i]  <- MANNCal$reduction[c(2)]
PBR_X2_ANNCal[i]  <- MANNCal$reduction[c(3)]
PBR_X3_ANNCal[i]  <- MANNCal$reduction[c(4)]
PBR_X4_ANNCal[i]  <- MANNCal$reduction[c(5)]
PBR_X5_ANNCal[i]  <- MANNCal$reduction[c(6)]
PBR_PS_ANNCal[i]  <- MANNCal$reduction[c(1)]

#Group Regression Coefficient after NN matching with caliper
Y1ANNCal[i]          <- as.numeric(OutcomeA.1NNCal$coef[2])
Y2ANNCal[i]          <- as.numeric(OutcomeA.2NNCal$coef[2])
Y3ANNCal[i]          <- as.numeric(OutcomeA.3NNCal$coef[2])
Y4ANNCal[i]          <- as.numeric(OutcomeA.4NNCal$coef[2])
#Group t-value for each regression coefficient after NN matching with
caliper
tNNCalY1A[i]             <-summary(OutcomeA.1NNCal)$coef[2, 3]
tNNCalY2A[i]             <-summary(OutcomeA.2NNCal)$coef[2, 3]
tNNCalY3A[i]             <-summary(OutcomeA.3NNCal)$coef[2, 3]
tNNCalY4A[i]             <-summary(OutcomeA.4NNCal)$coef[2, 3]
#Treatment Group N (after NN matching with caliper)
NNCaltreatNA[i]          <-nobs(ANNCal$group[ANNCal$group==1])
#Comparison Group N (after NN matching)
NNCalcompNA[i]           <-nobs(ANNCal$group[ANNCal$group==0])
#Cohen's D for Treatment Effect after NN matching with caliper
NNCalCohenY1A[i]          <-cohenANNCalY1
NNCalCohenY2A[i]          <-cohenANNCalY2
NNCalCohenY3A[i]          <-cohenANNCalY3
NNCalCohenY4A[i]          <-cohenANNCalY4

#Matched & Unmatched PS mean, median, and sd after NN matching with
Caliper
PSMeanMatchedTreatANNCal[i]   <-
mean(FullANNCal$distance[FullANNCal$group==1 & FullANNCal$weights==1])
PSMeanMatchedCompANNCal[i]   <-
mean(FullANNCal$distance[FullANNCal$group==0 & FullANNCal$weights==1])
PSMeanUnMatchedTreatANNCal[i]   <-
mean(FullANNCal$distance[FullANNCal$group==1 & FullANNCal$weights==0])
PSMeanUnMatchedCompANNCal[i]   <-
mean(FullANNCal$distance[FullANNCal$group==0 & FullANNCal$weights==0])
```

```
PSMedMatchedTreatANNCal[i]   <-
median(FullANNCal$distance[FullANNCal$group==1 &
FullANNCal$weights==1])
PSMedMatchedCompANNCal[i]   <-
median(FullANNCal$distance[FullANNCal$group==0 &
FullANNCal$weights==1])
PSMedUnMatchedTreatANNCal[i]   <-
median(FullANNCal$distance[FullANNCal$group==1 &
FullANNCal$weights==0])
PSMedUnMatchedCompANNCal[i]   <-
median(FullANNCal$distance[FullANNCal$group==0 &
FullANNCal$weights==0])

PSsdMatchedTreatANNCal[i]   <-
sd(FullANNCal$distance[FullANNCal$group==1 & FullANNCal$weights==1])
PSsdMatchedCompANNCal[i]   <-
sd(FullANNCal$distance[FullANNCal$group==0 & FullANNCal$weights==1])
PSsdUnMatchedTreatANNCal[i]   <-
sd(FullANNCal$distance[FullANNCal$group==1 & FullANNCal$weights==0])
PSsdUnMatchedCompANNCal[i]   <-
sd(FullANNCal$distance[FullANNCal$group==0 & FullANNCal$weights==0])

#Variance Ratio for matched groups, after NN Matching with Caliper
VRANNCal[i] <-
var(ANNCal$distance[ANNCal$group==1])/var(ANNCal$distance[ANNCal$group=
=0])

#All variables AFTER Nearest Neighbor Matching with 0.20 Caliper, ATC
Coding
AvgX1TreatANNCalATC[i]   <-  mean(ANNCalATC$x1[ANNCalATC$ATCgroup==0])
AvgX2TreatANNCalATC[i]   <-  mean(ANNCalATC$x2[ANNCalATC$ATCgroup==0])
AvgX3TreatANNCalATC[i]   <-  mean(ANNCalATC$x3[ANNCalATC$ATCgroup==0])
AvgX4TreatANNCalATC[i]   <-  mean(ANNCalATC$x4[ANNCalATC$ATCgroup==0])
AvgX5TreatANNCalATC[i]   <-  mean(ANNCalATC$x5[ANNCalATC$ATCgroup==0])
AvgYA1TreatANNCalATC[i]   <-
mean(ANNCalATC$YA1[ANNCalATC$ATCgroup==0])
AvgYA2TreatANNCalATC[i]   <-
mean(ANNCalATC$YA2[ANNCalATC$ATCgroup==0])
AvgYA3TreatANNCalATC[i]   <-
mean(ANNCalATC$YA3[ANNCalATC$ATCgroup==0])
AvgYA4TreatANNCalATC[i]   <-
mean(ANNCalATC$YA4[ANNCalATC$ATCgroup==0])
AvgPSTreatANNCalATC[i]    <-  mean(ANNCalATC$PS[ANNCalATC$ATCgroup==0])

AvgX1CompANNCalATC[i]   <-  mean(ANNCalATC$x1[ANNCalATC$ATCgroup==1])
AvgX2CompANNCalATC[i]   <-  mean(ANNCalATC$x2[ANNCalATC$ATCgroup==1])
AvgX3CompANNCalATC[i]   <-  mean(ANNCalATC$x3[ANNCalATC$ATCgroup==1])
AvgX4CompANNCalATC[i]   <-  mean(ANNCalATC$x4[ANNCalATC$ATCgroup==1])
AvgX5CompANNCalATC[i]   <-  mean(ANNCalATC$x5[ANNCalATC$ATCgroup==1])
AvgYA1CompANNCalATC[i]   <-  mean(ANNCalATC$YA1[ANNCalATC$ATCgroup==1])
AvgYA2CompANNCalATC[i]   <-  mean(ANNCalATC$YA2[ANNCalATC$ATCgroup==1])
AvgYA3CompANNCalATC[i]   <-  mean(ANNCalATC$YA3[ANNCalATC$ATCgroup==1])
AvgYA4CompANNCalATC[i]   <-  mean(ANNCalATC$YA4[ANNCalATC$ATCgroup==1])
AvgPSCompANNCalATC[i]    <-  mean(ANNCalATC$PS[ANNCalATC$ATCgroup==1])

SDX1TreatANNCalATC[i]   <-  sd(ANNCalATC$x1[ANNCalATC$ATCgroup==0])
```

```
SDX2TreatANNCalATC[i]   <-  sd(ANNCalATC$x2[ANNCalATC$ATCgroup==0])
SDX3TreatANNCalATC[i]   <-  sd(ANNCalATC$x3[ANNCalATC$ATCgroup==0])
SDX4TreatANNCalATC[i]   <-  sd(ANNCalATC$x4[ANNCalATC$ATCgroup==0])
SDX5TreatANNCalATC[i]   <-  sd(ANNCalATC$x5[ANNCalATC$ATCgroup==0])
SDYA1TreatANNCalATC[i]   <-  sd(ANNCalATC$YA1[ANNCalATC$ATCgroup==0])
SDYA2TreatANNCalATC[i]   <-  sd(ANNCalATC$YA2[ANNCalATC$ATCgroup==0])
SDYA3TreatANNCalATC[i]   <-  sd(ANNCalATC$YA3[ANNCalATC$ATCgroup==0])
SDYA4TreatANNCalATC[i]   <-  sd(ANNCalATC$YA4[ANNCalATC$ATCgroup==0])
SDPSTreatANNCalATC[i]    <-  sd(ANNCalATC$PS[ANNCalATC$ATCgroup==0])

SDX1CompANNCalATC[i]   <-  sd(ANNCalATC$x1[ANNCalATC$ATCgroup==1])
SDX2CompANNCalATC[i]   <-  sd(ANNCalATC$x2[ANNCalATC$ATCgroup==1])
SDX3CompANNCalATC[i]   <-  sd(ANNCalATC$x3[ANNCalATC$ATCgroup==1])
SDX4CompANNCalATC[i]   <-  sd(ANNCalATC$x4[ANNCalATC$ATCgroup==1])
SDX5CompANNCalATC[i]   <-  sd(ANNCalATC$x5[ANNCalATC$ATCgroup==1])
SDYA1CompANNCalATC[i]   <-  sd(ANNCalATC$YA1[ANNCalATC$ATCgroup==1])
SDYA2CompANNCalATC[i]   <-  sd(ANNCalATC$YA2[ANNCalATC$ATCgroup==1])
SDYA3CompANNCalATC[i]   <-  sd(ANNCalATC$YA3[ANNCalATC$ATCgroup==1])
SDYA4CompANNCalATC[i]   <-  sd(ANNCalATC$YA4[ANNCalATC$ATCgroup==1])
SDPSCompANNCalATC[i]    <-  sd(ANNCalATC$PS[ANNCalATC$ATCgroup==1])

#Standardized Mean Difference after NN matching with caliper, ATC
coding
SMD_X1_ANNCalATC[i] <-  MANNCalATC$sum.matched[c(14)]
SMD_X2_ANNCalATC[i] <-  MANNCalATC$sum.matched[c(15)]
SMD_X3_ANNCalATC[i] <-  MANNCalATC$sum.matched[c(16)]
SMD_X4_ANNCalATC[i] <-  MANNCalATC$sum.matched[c(17)]
SMD_X5_ANNCalATC[i] <-  MANNCalATC$sum.matched[c(18)]
SMD_PS_ANNCalATC[i] <-  MANNCalATC$sum.matched[c(13)]

#Percent Bias Reduction after NN matching with caliper, ATC coding
PBR_X1_ANNCalATC[i]  <- MANNCalATC$reduction[c(2)]
PBR_X2_ANNCalATC[i]  <- MANNCalATC$reduction[c(3)]
PBR_X3_ANNCalATC[i]  <- MANNCalATC$reduction[c(4)]
PBR_X4_ANNCalATC[i]  <- MANNCalATC$reduction[c(5)]
PBR_X5_ANNCalATC[i]  <- MANNCalATC$reduction[c(6)]
PBR_PS_ANNCalATC[i]  <- MANNCalATC$reduction[c(1)]

#Group Regression Coefficient after NN matching with caliper, ATC
coding
Y1ANNCalATC[i]          <- as.numeric(OutcomeA.1NNCalATC$coef[2])
Y2ANNCalATC[i]          <- as.numeric(OutcomeA.2NNCalATC$coef[2])
Y3ANNCalATC[i]          <- as.numeric(OutcomeA.3NNCalATC$coef[2])
Y4ANNCalATC[i]          <- as.numeric(OutcomeA.4NNCalATC$coef[2])
#Group t-value for each regression coefficient after NN matching with
caliper, ATC coding
tNNCalATCY1A[i]             <-summary(OutcomeA.1NNCalATC)$coef[2, 3]
tNNCalATCY2A[i]             <-summary(OutcomeA.2NNCalATC)$coef[2, 3]
tNNCalATCY3A[i]             <-summary(OutcomeA.3NNCalATC)$coef[2, 3]
tNNCalATCY4A[i]             <-summary(OutcomeA.4NNCalATC)$coef[2, 3]
#Treatment Group N (after NN matching with caliper), ATC coding
NNCalATCtreatNA[i]          <-nobs(ANNCalATC$group[ANNCalATC$group==0])
#Comparison Group N (after NN matching with caliper)
NNCalATCcompNA[i]           <-nobs(ANNCalATC$group[ANNCalATC$group==1])
#Cohen's D for Treatment Effect after NN matching with caliper, ATC
coding
NNCalATCCohenY1A[i]          <-cohenANNCalATCY1
```

```
NNCalATCCohenY2A[i]          <-cohenANNCalATCY2
NNCalATCCohenY3A[i]          <-cohenANNCalATCY3
NNCalATCCohenY4A[i]          <-cohenANNCalATCY4

#Matched & Unmatched PS mean, median, and sd after NN matching with
Caliper, ATC Coding
PSMeanMatchedTreatANNCalATC[i]   <-
mean(FullANNCalATC$distance[FullANNCalATC$group==0 &
FullANNCalATC$weights==1])
PSMeanMatchedCompANNCalATC[i]   <-
mean(FullANNCalATC$distance[FullANNCalATC$group==1 &
FullANNCalATC$weights==1])
PSMeanUnMatchedTreatANNCalATC[i]   <-
mean(FullANNCalATC$distance[FullANNCalATC$group==0 &
FullANNCalATC$weights==0])
PSMeanUnMatchedCompANNCalATC[i]   <-
mean(FullANNCalATC$distance[FullANNCalATC$group==1 &
FullANNCalATC$weights==0])

PSMedMatchedTreatANNCalATC[i]   <-
median(FullANNCalATC$distance[FullANNCalATC$group==0 &
FullANNCalATC$weights==1])
PSMedMatchedCompANNCalATC[i]   <-
median(FullANNCalATC$distance[FullANNCalATC$group==1 &
FullANNCalATC$weights==1])
PSMedUnMatchedTreatANNCalATC[i]   <-
median(FullANNCalATC$distance[FullANNCalATC$group==0 &
FullANNCalATC$weights==0])
PSMedUnMatchedCompANNCalATC[i]   <-
median(FullANNCalATC$distance[FullANNCalATC$group==1 &
FullANNCalATC$weights==0])

PSsdMatchedTreatANNCalATC[i]   <-
sd(FullANNCalATC$distance[FullANNCalATC$group==0 &
FullANNCalATC$weights==1])
PSsdMatchedCompANNCalATC[i]   <-
sd(FullANNCalATC$distance[FullANNCalATC$group==1 &
FullANNCalATC$weights==1])
PSsdUnMatchedTreatANNCalATC[i]   <-
sd(FullANNCalATC$distance[FullANNCalATC$group==0 &
FullANNCalATC$weights==0])
PSsdUnMatchedCompANNCalATC[i]   <-
sd(FullANNCalATC$distance[FullANNCalATC$group==1 &
FullANNCalATC$weights==0])

#Variance Ratio for matched groups, after NN Matching with Caliper, ATC
coding
VRANNCalATC[i] <-
var(ANNCalATC$distance[ANNCalATC$ATCgroup==0])/var(ANNCalATC$distance[A
NNCalATC$ATCgroup==1])

#All variables AFTER Generalized Boosted Modeling, ATT Coding
AvgX1TreatAGBM[i]   <-  BalAGBM$es.mean.ATT$`tx.mn`[1]
AvgX2TreatAGBM[i]   <-  BalAGBM$es.mean.ATT$`tx.mn`[2]
AvgX3TreatAGBM[i]   <-  BalAGBM$es.mean.ATT$`tx.mn`[3]
AvgX4TreatAGBM[i]   <-  BalAGBM$es.mean.ATT$`tx.mn`[4]
AvgX5TreatAGBM[i]   <-  BalAGBM$es.mean.ATT$`tx.mn`[5]
```

```
AvgX1CompAGBM[i]   <-  BalAGBM$es.mean.ATT$`ct.mn`[1]
AvgX2CompAGBM[i]   <-  BalAGBM$es.mean.ATT$`ct.mn`[2]
AvgX3CompAGBM[i]   <-  BalAGBM$es.mean.ATT$`ct.mn`[3]
AvgX4CompAGBM[i]   <-  BalAGBM$es.mean.ATT$`ct.mn`[4]
AvgX5CompAGBM[i]   <-  BalAGBM$es.mean.ATT$`ct.mn`[1]

SDX1TreatAGBM[i]   <-  BalAGBM$es.mean.ATT$`tx.sd`[1]
SDX2TreatAGBM[i]   <-  BalAGBM$es.mean.ATT$`tx.sd`[2]
SDX3TreatAGBM[i]   <-  BalAGBM$es.mean.ATT$`tx.sd`[3]
SDX4TreatAGBM[i]   <-  BalAGBM$es.mean.ATT$`tx.sd`[4]
SDX5TreatAGBM[i]   <-  BalAGBM$es.mean.ATT$`tx.sd`[5]

SDX1CompAGBM[i]   <-  BalAGBM$es.mean.ATT$`ct.sd`[1]
SDX2CompAGBM[i]   <-  BalAGBM$es.mean.ATT$`ct.sd`[2]
SDX3CompAGBM[i]   <-  BalAGBM$es.mean.ATT$`ct.sd`[3]
SDX4CompAGBM[i]   <-  BalAGBM$es.mean.ATT$`ct.sd`[4]
SDX5CompAGBM[i]   <-  BalAGBM$es.mean.ATT$`ct.sd`[5]

#Standardized Mean Difference after GBM
SMD_X1_AGBM[i] <-  BalAGBM$es.mean.ATT$`std.eff.sz`[1]
SMD_X2_AGBM[i] <-  BalAGBM$es.mean.ATT$`std.eff.sz`[2]
SMD_X3_AGBM[i] <-  BalAGBM$es.mean.ATT$`std.eff.sz`[3]
SMD_X4_AGBM[i] <-  BalAGBM$es.mean.ATT$`std.eff.sz`[4]
SMD_X5_AGBM[i] <-  BalAGBM$es.mean.ATT$`std.eff.sz`[5]

#####Percent Bias Reduction after GBM
PBR_X1_AGBM[i]  <- PBRX1GBM
PBR_X2_AGBM[i]  <- PBRX2GBM
PBR_X3_AGBM[i]  <- PBRX3GBM
PBR_X4_AGBM[i]  <- PBRX4GBM
PBR_X5_AGBM[i]  <- PBRX5GBM

#Group Regression Coefficient after GBM
Y1AGBM[i]          <- as.numeric(OutcomeA.1GBM$coef[2])
Y2AGBM[i]          <- as.numeric(OutcomeA.2GBM$coef[2])
Y3AGBM[i]          <- as.numeric(OutcomeA.3GBM$coef[2])
Y4AGBM[i]          <- as.numeric(OutcomeA.4GBM$coef[2])
#Group t-value for each regression coefficient after GBM
tGBMY1A[i]             <-summary(OutcomeA.1GBM)$coef[2, 3]
tGBMY2A[i]             <-summary(OutcomeA.2GBM)$coef[2, 3]
tGBMY3A[i]             <-summary(OutcomeA.3GBM)$coef[2, 3]
tGBMY4A[i]             <-summary(OutcomeA.4GBM)$coef[2, 3]
#Treatment Group N (after GBM)
GBMtreatNA[i]          <-nobs(finalDataA$group[finalDataA$group==1])
#Comparison Group N (after GBM)
GBMcompNA[i]           <-nobs(finalDataA$group[finalDataA$group==0])
#Cohen's D for Treatment Effect after GBM
GBMCohenY1A[i]          <-cohenAGBMY1
GBMCohenY2A[i]          <-cohenAGBMY2
GBMCohenY3A[i]          <-cohenAGBMY3
GBMCohenY4A[i]          <-cohenAGBMY4

#All variables AFTER Nearest Neighbor Matching with 0.20 Caliper, ATC
Coding
AvgX1TreatAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`tx.mn`[1]
AvgX2TreatAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`tx.mn`[2]
```

```
AvgX3TreatAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`tx.mn`[3]
AvgX4TreatAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`tx.mn`[4]
AvgX5TreatAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`tx.mn`[5]

AvgX1CompAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`ct.mn`[1]
AvgX2CompAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`ct.mn`[2]
AvgX3CompAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`ct.mn`[3]
AvgX4CompAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`ct.mn`[4]
AvgX5CompAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`ct.mn`[1]

SDX1TreatAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`tx.sd`[1]
SDX2TreatAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`tx.sd`[2]
SDX3TreatAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`tx.sd`[3]
SDX4TreatAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`tx.sd`[4]
SDX5TreatAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`tx.sd`[5]

SDX1CompAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`ct.sd`[1]
SDX2CompAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`ct.sd`[2]
SDX3CompAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`ct.sd`[3]
SDX4CompAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`ct.sd`[4]
SDX5CompAGBMATC[i]   <-  BalAGBMATC$es.mean.ATT$`ct.sd`[5]

#Standardized Mean Difference after GBM, ATC Coding
SMD_X1_AGBMATC[i] <-  BalAGBMATC$es.mean.ATT$`std.eff.sz`[1]
SMD_X2_AGBMATC[i] <-  BalAGBMATC$es.mean.ATT$`std.eff.sz`[2]
SMD_X3_AGBMATC[i] <-  BalAGBMATC$es.mean.ATT$`std.eff.sz`[3]
SMD_X4_AGBMATC[i] <-  BalAGBMATC$es.mean.ATT$`std.eff.sz`[4]
SMD_X5_AGBMATC[i] <-  BalAGBMATC$es.mean.ATT$`std.eff.sz`[5]

#####Percent Bias Reduction after GBM, ATC Coding
PBR_X1_AGBMATC[i]  <- PBRX1GBMATC
PBR_X2_AGBMATC[i]  <- PBRX2GBMATC
PBR_X3_AGBMATC[i]  <- PBRX3GBMATC
PBR_X4_AGBMATC[i]  <- PBRX4GBMATC
PBR_X5_AGBMATC[i]  <- PBRX5GBMATC

#Group Regression Coefficient after GBM, ATC Coding
Y1AGBMATC[i]          <- as.numeric(OutcomeA.1GBMATC$coef[2])
Y2AGBMATC[i]          <- as.numeric(OutcomeA.2GBMATC$coef[2])
Y3AGBMATC[i]          <- as.numeric(OutcomeA.3GBMATC$coef[2])
Y4AGBMATC[i]          <- as.numeric(OutcomeA.4GBMATC$coef[2])
#Group t-value for each regression coefficient after GBM, ATC Coding
tGBMATCY1A[i]             <-summary(OutcomeA.1GBMATC)$coef[2, 3]
tGBMATCY2A[i]             <-summary(OutcomeA.2GBMATC)$coef[2, 3]
tGBMATCY3A[i]             <-summary(OutcomeA.3GBMATC)$coef[2, 3]
tGBMATCY4A[i]             <-summary(OutcomeA.4GBMATC)$coef[2, 3]
#Treatment Group N (after GBM), ATC Coding
GBMATCtreatNA[i]          <-
nobs(finalDataA$ATCgroup[finalDataA$ATCgroup==0])
#Comparison Group N (after GBM), ATC Coding
GBMATCcompNA[i]           <-
nobs(finalDataA$ATCgroup[finalDataA$ATCgroup==1])
#Cohen's D for Treatment Effect after GBM, ATC Coding
GBMATCCohenY1A[i]          <-cohenAGBMATCY1
GBMATCCohenY2A[i]          <-cohenAGBMATCY2
GBMATCCohenY3A[i]          <-cohenAGBMATCY3
GBMATCCohenY4A[i]          <-cohenAGBMATCY4
```

```
#Other GBM Values
ESS_CompGBM[i]           <-ps.AGBM$desc$es.mean.ATT$ess.ctrl
mean.esGBM[i]            <-ps.AGBM$desc$es.mean.ATT$mean.es
iterGBM[i]               <-ps.AGBM$desc$es.mean.ATT$n.trees

ESS_CompGBMATC[i]           <-ps.AGBMATC$desc$es.mean.ATT$ess.ctrl
mean.esGBMATC[i]            <-ps.AGBMATC$desc$es.mean.ATT$mean.es
iterGBMATC[i]               <-ps.AGBMATC$desc$es.mean.ATT$n.trees}

#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#~~~~~~~~~~~~~~~~~~~        Creating Excel File       ~~~~~~~~~~~~~~~~~
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Final.Sim.Data.A.BeforeMatching<-cbind(

  #All,variables,BEFORE,matching/weighting
  #Averages
AvgX1TreatA,AvgX2TreatA,AvgX3TreatA,AvgX4TreatA,AvgX5TreatA,AvgYA1Treat
A, AvgYA2TreatA,AvgYA3TreatA,

AvgYA4TreatA,AvgX1CompA,AvgX2CompA,AvgX3CompA,AvgX4CompA,AvgX5CompA,Avg
YA1CompA,AvgYA2CompA,AvgYA3CompA,AvgYA4CompA,
  #Standard Deviations

SDX1TreatA,SDX2TreatA,SDX3TreatA,SDX4TreatA,SDX5TreatA,SDYA1TreatA,SDYA
2TreatA,SDYA3TreatA,SDYA4TreatA,

SDX1CompA,SDX2CompA,SDX3CompA,SDX4CompA,SDX5CompA,SDYA1CompA,SDYA2CompA
,SDYA3CompA,SDYA4CompA,
  #Correlations

Cor_X1.X2_A,Cor_X1.X3_A,Cor_X1.X4_A,Cor_X1.X5_A,Cor_X2.X3_A,Cor_X2.X4_A
,Cor_X2.X5_A,Cor_X3.X4_A,

Cor_X3.X5_A,Cor_X4.X5_A,Cor_X1.PS_A,Cor_X2.PS_A,Cor_X3.PS_A,Cor_X4.PS_A
,Cor_X5.PS_A,Cor_X1.Y1_A,

Cor_X2.Y1_A,Cor_X3.Y1_A,Cor_X4.Y1_A,Cor_X5.Y1_A,Cor_X1.Y2_A,Cor_X2.Y2_A
,Cor_X3.Y2_A,Cor_X4.Y2_A,

Cor_X5.Y2_A,Cor_X1.Y3_A,Cor_X2.Y3_A,Cor_X3.Y3_A,Cor_X4.Y3_A,Cor_X5.Y3_A
,Cor_X1.Y4_A,Cor_X2.Y4_A,

Cor_X3.Y4_A,Cor_X4.Y4_A,Cor_X5.Y4_A,Cor_G.Y1_A,Cor_G.Y2_A,Cor_G.Y3_A,Co
r_G.Y4_A,
  #Standardized Mean Differences

SMD_X1_All,SMD_X2_All,SMD_X3_All,SMD_X4_All,SMD_X5_All,SMD_PS_All,SMD_X
1_AllATC,SMD_X2_AllATC,SMD_X3_AllATC,SMD_X4_AllATC,SMD_X5_AllATC,SMD_PS
_AllATC,

  #Outcome Variables
  #Population Regression Coefficients
PopY1A,PopY2A,PopY3A,PopY4A,PopATCY1A,PopATCY2A,PopATCY3A,PopATCY4A,
  #Population t values
```

```
tPopY1A,tPopY2A,tPopY3A,tPopY4A,tPopATCY1A,tPopATCY2A,tPopATCY3A,tPopAT
CY4A,
  #Population Cohen's d

PopCohenY1A,PopCohenY2A,PopCohenY3A,PopCohenY4A,PopCohenATCY1A,PopCohen
ATCY2A,PopCohenATCY3A,PopCohenATCY4A,
  #Ns

treatPopNA,compPopNA,treatBaseNA,compBaseNA,treatPopNAATC,compPopNAATC,
treatBaseNAATC,compBaseNAATC,
  #Baseline Regression Coefficients

BaseY1A,BaseY2A,BaseY3A,BaseY4A,BaseATCY1A,BaseATCY2A,BaseATCY3A,BaseAT
CY4A,
  #Baseline t values

tBaseY1A,tBaseY2A,tBaseY3A,tBaseY4A,tBaseATCY1A,tBaseATCY2A,tBaseATCY3A
,tBaseATCY4A,
  #Baseline Cohen's d

BaseCohenY1A,BaseCohenY2A,BaseCohenY3A,BaseCohenY4A,BaseCohenATCY1A,Bas
eCohenATCY2A,BaseCohenATCY3A,BaseCohenATCY4A,
  #Propensity Score mean, sd by group (before matching)
AvgPSTreatA,AvgPSCompA,SDPSTreatA,SDPSCompA,
  #VRs
VRB,VRBATC
)

Final.Sim.Data.A.AfterNNMatching<-cbind(

  #All,variables,AFTER,NN matching
  #Averages

AvgX1TreatANN,AvgX2TreatANN,AvgX3TreatANN,AvgX4TreatANN,AvgX5TreatANN,A
vgYA1TreatANN,AvgYA2TreatANN,

AvgYA3TreatANN,AvgYA4TreatANN,AvgPSTreatANN,AvgX1CompANN,AvgX2CompANN,A
vgX3CompANN,AvgX4CompANN,

AvgX5CompANN,AvgYA1CompANN,AvgYA2CompANN,AvgYA3CompANN,AvgYA4CompANN,Av
gPSCompANN,AvgX1TreatANNATC,

AvgX2TreatANNATC,AvgX3TreatANNATC,AvgX4TreatANNATC,AvgX5TreatANNATC,Avg
YA1TreatANNATC,AvgYA2TreatANNATC,

AvgYA3TreatANNATC,AvgYA4TreatANNATC,AvgPSTreatANNATC,AvgX1CompANNATC,Av
gX2CompANNATC,AvgX3CompANNATC,

AvgX4CompANNATC,AvgX5CompANNATC,AvgYA1CompANNATC,AvgYA2CompANNATC,AvgYA
3CompANNATC,AvgYA4CompANNATC,AvgPSCompANNATC,
  #Standard Deviations

SDX1TreatANN,SDX2TreatANN,SDX3TreatANN,SDX4TreatANN,SDX5TreatANN,SDYA1T
reatANN,SDYA2TreatANN,
```

```
SDYA3TreatANN,SDYA4TreatANN,SDPSTreatANN,SDX1CompANN,SDX2CompANN,SDX3Co
mpANN,SDX4CompANN,

SDX5CompANN,SDYA1CompANN,SDYA2CompANN,SDYA3CompANN,SDYA4CompANN,SDPSCom
pANN,SDX1TreatANNATC,

SDX2TreatANNATC,SDX3TreatANNATC,SDX4TreatANNATC,SDX5TreatANNATC,SDYA1Tr
eatANNATC,SDYA2TreatANNATC,

SDYA3TreatANNATC,SDYA4TreatANNATC,SDPSTreatANNATC,SDX1CompANNATC,SDX2Co
mpANNATC,SDX3CompANNATC,

SDX4CompANNATC,SDX5CompANNATC,SDYA1CompANNATC,SDYA2CompANNATC,SDYA3Comp
ANNATC,SDYA4CompANNATC,SDPSCompANNATC,
  #Standardized Mean Differences

SMD_X1_ANN,SMD_X2_ANN,SMD_X3_ANN,SMD_X4_ANN,SMD_X5_ANN,SMD_PS_ANN,SMD_X
1_ANNATC,SMD_X2_ANNATC,SMD_X3_ANNATC,SMD_X4_ANNATC,SMD_X5_ANNATC,SMD_PS
_ANNATC,
  #PBRs

PBR_X1_ANN,PBR_X2_ANN,PBR_X3_ANN,PBR_X4_ANN,PBR_X5_ANN,PBR_PS_ANN,PBR_X
1_ANNATC,PBR_X2_ANNATC,PBR_X3_ANNATC,PBR_X4_ANNATC,PBR_X5_ANNATC,PBR_PS
_ANNATC,
  #Outcome Variables
  #Regression Coefficients
Y1ANN,Y2ANN,Y3ANN,Y4ANN,Y1ANNATC,Y2ANNATC,Y3ANNATC,Y4ANNATC,
  #t values
tNNY1A,tNNY2A,tNNY3A,tNNY4A,tNNATCY1A,tNNATCY2A,tNNATCY3A,tNNATCY4A,
  #Cohen's d

NNCohenY1A,NNCohenY2A,NNCohenY3A,NNCohenY4A,NNATCCohenY1A,NNATCCohenY2A
,NNATCCohenY3A,NNATCCohenY4A,
  #Ns
NNtreatNA,NNcompNA,NNATCtreatNA,NNATCcompNA,
  #Propensity Score mean, median, sd by group/matching

PSMeanMatchedTreatANN,PSMeanMatchedCompANN,PSMeanUnMatchedTreatANN,PSMe
anUnMatchedCompANN,

PSMeanMatchedTreatANNATC,PSMeanMatchedCompANNATC,PSMeanUnMatchedTreatAN
NATC,PSMeanUnMatchedCompANNATC,

PSMedMatchedTreatANN,PSMedMatchedCompANN,PSMedUnMatchedTreatANN,PSMedUn
MatchedCompANN,

PSMedMatchedTreatANNATC,PSMedMatchedCompANNATC,PSMedUnMatchedTreatANNAT
C,PSMedUnMatchedCompANNATC,PSsdMatchedTreatANN,PSsdMatchedCompANN,PSsdU
nMatchedTreatANN,PSsdUnMatchedCompANN,

PSsdMatchedTreatANNATC,PSsdMatchedCompANNATC,PSsdUnMatchedTreatANNATC,P
SsdUnMatchedCompANNATC,
  #VRs
VRANN,VRANNATC
)
```

```
Final.Sim.Data.A.AfterNNCaliperMatching<-cbind(

  #All,variables,AFTER,NN matching with caliper
  #Averages

AvgX1TreatANNCal,AvgX2TreatANNCal,AvgX3TreatANNCal,AvgX4TreatANNCal,Avg
X5TreatANNCal,AvgYA1TreatANNCal,

AvgYA2TreatANNCal,AvgYA3TreatANNCal,AvgYA4TreatANNCal,AvgPSTreatANNCal,
AvgX1CompANNCal,AvgX2CompANNCal,

AvgX3CompANNCal,AvgX4CompANNCal,AvgX5CompANNCal,AvgYA1CompANNCal,AvgYA2
CompANNCal,AvgYA3CompANNCal,

AvgYA4CompANNCal,AvgPSCompANNCal,AvgX1TreatANNCalATC,AvgX2TreatANNCalAT
C,AvgX3TreatANNCalATC,

AvgX4TreatANNCalATC,AvgX5TreatANNCalATC,AvgYA1TreatANNCalATC,AvgYA2Trea
tANNCalATC,AvgYA3TreatANNCalATC,

AvgYA4TreatANNCalATC,AvgPSTreatANNCalATC,AvgX1CompANNCalATC,AvgX2CompAN
NCalATC,AvgX3CompANNCalATC,

AvgX4CompANNCalATC,AvgX5CompANNCalATC,AvgYA1CompANNCalATC,AvgYA2CompANN
CalATC,AvgYA3CompANNCalATC,AvgYA4CompANNCalATC,AvgPSCompANNCalATC,
  #Standard Deviations

SDX1TreatANNCal,SDX2TreatANNCal,SDX3TreatANNCal,SDX4TreatANNCal,SDX5Tre
atANNCal,SDYA1TreatANNCal,

SDYA2TreatANNCal,SDYA3TreatANNCal,SDYA4TreatANNCal,SDPSTreatANNCal,SDX1
CompANNCal,SDX2CompANNCal,

SDX3CompANNCal,SDX4CompANNCal,SDX5CompANNCal,SDYA1CompANNCal,SDYA2CompA
NNCal,SDYA3CompANNCal,

SDYA4CompANNCal,SDPSCompANNCal,SDX1TreatANNCalATC,SDX2TreatANNCalATC,SD
X3TreatANNCalATC,

SDX4TreatANNCalATC,SDX5TreatANNCalATC,SDYA1TreatANNCalATC,SDYA2TreatANN
CalATC,SDYA3TreatANNCalATC,

SDYA4TreatANNCalATC,SDPSTreatANNCalATC,SDX1CompANNCalATC,SDX2CompANNCal
ATC,SDX3CompANNCalATC,

SDX4CompANNCalATC,SDX5CompANNCalATC,SDYA1CompANNCalATC,SDYA2CompANNCalA
TC,SDYA3CompANNCalATC,SDYA4CompANNCalATC,SDPSCompANNCalATC,
  #Standardized Mean Differences

SMD_X1_ANNCal,SMD_X2_ANNCal,SMD_X3_ANNCal,SMD_X4_ANNCal,SMD_X5_ANNCal,S
MD_PS_ANNCal,SMD_X1_ANNCalATC,

SMD_X2_ANNCalATC,SMD_X3_ANNCalATC,SMD_X4_ANNCalATC,SMD_X5_ANNCalATC,SMD
_PS_ANNCalATC,
  #PBRs

PBR_X1_ANNCal,PBR_X2_ANNCal,PBR_X3_ANNCal,PBR_X4_ANNCal,PBR_X5_ANNCal,P
```

```
BR_PS_ANNCal,PBR_X1_ANNCalATC,PBR_X2_ANNCalATC,PBR_X3_ANNCalATC,PBR_X4_
ANNCalATC,PBR_X5_ANNCalATC,PBR_PS_ANNCalATC,
  #Outcome Variables
  #Regression Coefficients

Y1ANNCal,Y2ANNCal,Y3ANNCal,Y4ANNCal,Y1ANNCalATC,Y2ANNCalATC,Y3ANNCalATC
,Y4ANNCalATC,
  #t values

tNNCalY1A,tNNCalY2A,tNNCalY3A,tNNCalY4A,tNNCalATCY1A,tNNCalATCY2A,tNNCa
lATCY3A,tNNCalATCY4A,
  #Cohen's d

NNCalCohenY1A,NNCalCohenY2A,NNCalCohenY3A,NNCalCohenY4A,NNCalATCCohenY1
A,NNCalATCCohenY2A,NNCalATCCohenY3A,NNCalATCCohenY4A,
  #Ns
NNCaltreatNA,NNCalcompNA,NNCalATCtreatNA,NNCalATCcompNA,
  #Propensity Score mean, median, sd by group/matching

PSMeanMatchedTreatANNCal,PSMeanMatchedCompANNCal,PSMeanUnMatchedTreatAN
NCal,PSMeanUnMatchedCompANNCal,

PSMeanMatchedTreatANNCalATC,PSMeanMatchedCompANNCalATC,PSMeanUnMatchedT
reatANNCalATC,PSMeanUnMatchedCompANNCalATC,

PSMedMatchedTreatANNCal,PSMedMatchedCompANNCal,PSMedUnMatchedTreatANNCa
l,PSMedUnMatchedCompANNCal,

PSMedMatchedTreatANNCalATC,PSMedMatchedCompANNCalATC,PSMedUnMatchedTrea
tANNCalATC,PSMedUnMatchedCompANNCalATC,

PSsdMatchedTreatANNCal,PSsdMatchedCompANNCal,PSsdUnMatchedTreatANNCal,P
SsdUnMatchedCompANNCal,

PSsdMatchedTreatANNCalATC,PSsdMatchedCompANNCalATC,PSsdUnMatchedTreatAN
NCalATC,PSsdUnMatchedCompANNCalATC,
  #VRs
VRANNCal,VRANNCalATC
)

Final.Sim.Data.A.AfterGBM<-cbind(

  #All,variables,AFTER,GBM
  #Averages

AvgX1TreatAGBM,AvgX2TreatAGBM,AvgX3TreatAGBM,AvgX4TreatAGBM,AvgX5TreatA
GBM,AvgX1CompAGBM,AvgX2CompAGBM,

AvgX3CompAGBM,AvgX4CompAGBM,AvgX5CompAGBM,AvgX1TreatAGBMATC,AvgX2TreatA
GBMATC,AvgX3TreatAGBMATC,

AvgX4TreatAGBMATC,AvgX5TreatAGBMATC,AvgX1CompAGBMATC,AvgX2CompAGBMATC,A
vgX3CompAGBMATC,AvgX4CompAGBMATC,AvgX5CompAGBMATC,
  #Standard Deviations

SDX1TreatAGBM,SDX2TreatAGBM,SDX3TreatAGBM,SDX4TreatAGBM,SDX5TreatAGBM,S
DX1CompAGBM,SDX2CompAGBM,
```

```
SDX3CompAGBM,SDX4CompAGBM,SDX5CompAGBM,SDX1TreatAGBMATC,SDX2TreatAGBMAT
C,SDX3TreatAGBMATC,SDX4TreatAGBMATC,

SDX5TreatAGBMATC,SDX1CompAGBMATC,SDX2CompAGBMATC,SDX3CompAGBMATC,SDX4Co
mpAGBMATC,SDX5CompAGBMATC,
  #Standardized Mean Differences

SMD_X1_AGBM,SMD_X2_AGBM,SMD_X3_AGBM,SMD_X4_AGBM,SMD_X5_AGBM,SMD_X1_AGBM
ATC,SMD_X2_AGBMATC,SMD_X3_AGBMATC,SMD_X4_AGBMATC,SMD_X5_AGBMATC,
  #PBRs

PBR_X1_AGBM,PBR_X2_AGBM,PBR_X3_AGBM,PBR_X4_AGBM,PBR_X5_AGBM,PBR_X1_AGBM
ATC,PBR_X2_AGBMATC,PBR_X3_AGBMATC,PBR_X4_AGBMATC,PBR_X5_AGBMATC,
  #Outcome Variables
  #Regression Coefficients
Y1AGBM,Y2AGBM,Y3AGBM,Y4AGBM,Y1AGBMATC,Y2AGBMATC,Y3AGBMATC,Y4AGBMATC,
  #t values

tGBMY1A,tGBMY2A,tGBMY3A,tGBMY4A,tGBMATCY1A,tGBMATCY2A,tGBMATCY3A,tGBMAT
CY4A,
  #Cohen's d

GBMCohenY1A,GBMCohenY2A,GBMCohenY3A,GBMCohenY4A,GBMATCCohenY1A,GBMATCCo
henY2A,GBMATCCohenY3A,GBMATCCohenY4A,
  #Ns
GBMtreatNA,GBMcompNA,GBMATCtreatNA,GBMATCcompNA,
  #Additional
ESS_CompGBM,mean.esGBM,iterGBM,ESS_CompGBMATC,mean.esGBMATC,iterGBMATC
)

BeforeMatchingWeighting<-as.data.frame(Final.Sim.Data.A.BeforeMatching)
AfterNNMatching<-as.data.frame(Final.Sim.Data.A.AfterNNMatching)
AfterNNMatchingCaliper<-
as.data.frame(Final.Sim.Data.A.AfterNNCaliperMatching)
AfterGBM<-as.data.frame(Final.Sim.Data.A.AfterGBM)

library(writexl)

write_xlsx(list(BeforeMatchingWeighting = BeforeMatchingWeighting,
AfterNNMatching = AfterNNMatching, AfterNNMatchingCaliper =
AfterNNMatchingCaliper, AfterGBM = AfterGBM), path="ScenarioA.xlsx")
```

References

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

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

Austin, P. C. (2013). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine, 33,* 1057-1069.

Austin, P. C. & Cafri, G. (2020). Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes. *Statistics in Medicine, 39,* 1623-1640.

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

Bai, H. (2011). Using propensity score analysis for making causal claims in research articles. *Educational Psychology Review, 23,* 273-278.

Bai, H. & Clark, M. H. (2019). *Propensity score methods and applications* (B. Entwisle, Ed.). SAGE Publications, Inc.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163,* 1149-1156.

Burgette, L. F., McCaffrey, D. F., & Griffin, B. A. (2015). Propensity score estimation with boosted regression. In, W. Pan & H. Bai (Eds.) *Propensity score analysis: Fundamentals and developments* (pp. 49-73). Guilford Publications, Inc.

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

Cheung, A. C. K. & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45,* 283-292.

Cochran, W. G. (1953). Matching in analytical studies. *American Journal of Public Health, 43*, 684-691.

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

Cochran, W. G. & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics*, *35*, 417-446.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: L. Erlbaum Associates.

Craig, B. G. (2020). *Propensity score matching and generalized boosted modeling in the context of model misspecification: A simulation study*. [Master's Thesis, James Madison University]. Retrieved from https://commons.lib.jmu.edu/masters202029/58.

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

Fan, X. & Nowell, D. L. (2011). Using propensity score matching in educational research. *Gifted Child Quarterly, 55*, 74-79.

Feinberg, R. A. & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*, 36-49.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2019). mvtnorm: Multivariate normal and t distributions. R package version 1.0-11. https://CRAN.R-project.org/package=mvtnorm

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. W. (2015). *Propensity score analysis: Statistical methods and applications* (2nd ed.). SAGE Publications, Inc.

Harris, H. D. (2018). *The influence of covariate measurement error on treatment effect estimates and numeric balance diagnostics following several common methods of propensity score matching: A simulation study* [Doctoral dissertation, James Madison University]. Retrieved from https://commons.lib.jmu.edu/diss201019/173.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric 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*, 1-28.

Holzman, M. A., & Horst, S. J. (2019, April). Treatment-comparison group ratio and accuracy of treatment effect estimates after propensity score matching. Paper presented at the annual meeting of the American Educational Research Association, Toronto, CA.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher, 49*, 241-253.

Lechner, M. (2000). A note on the common support problem in applied evaluation studies. *Econometric Evaluation of Public Policies: Methods and Applications, 91/92*, 217-235.

Lord, F. M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association, 55*, 307-321.

Jacovidis, J. N. (2017) *Evaluating the performance of propensity score matching methods: A simulation study.* [Doctoral dissertation, James Madison University]. Retrieved from https://commons.lib.jmu.edu/diss201019/149.

Jacovidis, J. N., Foelber, K. J., & Horst, S. J. (2017) The effect of propensity score matching method on the quantity and quality of matches. *The Journal of Experimental Education, 85*, 535-558.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19-40.

Maxwell, S. E. & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective.* New York, NY: Routledge.

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.

Osbourne, J. W. (2002). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research, and Evaluation, 8,* 11.

Pedhazur, E. J. & Schmelkin, L. P. (1991a). Attribute-treatments-interactions; Analysis of covariance. In *Measurement, design, and analysis: An integrated approach* (1st ed., pp. 545-586). Psychology Press.

Pedhazur, E. J. & Schmelkin, L. P. (1991b). Quasi-experimental designs. In *Measurement, design, and analysis: An integrated approach* (1st ed., pp. 277-302). Psychology Press.

Pan, W., & Bai, H. (2015). Propensity score analysis: Concepts and issues. In, W. Pan & H. Bai (Eds.) *Propensity score analysis: Fundamentals and developments* (pp. 3-19). Guilford Publications, Inc.

Pascarella, E. T., Salisbury, M. H., & Blaich, C. (2013). Design and analysis in college impact research: Which counts more? *Journal of College Student Development, 54,* 329-335.

Perkins, B. A. & Horst, S. J. (2020, Apr). *Propensity score analysis when treatment group is larger than comparison group: An applied assessment example.* Paper presented at the Annual Meeting of the American Educational Research

Association, San Francisco, CA. https://convention2.allacademic.com/one/aera/aera20/ (Conference cancelled)

Powell, M. G., Hull, D. M., & Beaujean, A. A. (2020). Propensity score matching for education data: Worked examples. *The Journal of Experimental Education, 88,* 145-164.

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., Burgette, L., & Cefalu, M. (2020). Twang: Toolkit for weighting and analysis of nonequivalent groups. R package version 1.6. https://CRAN.R-project.org/package=twang

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, 45,* 212-218.

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

Rosenbaum, P. R., & Rubin, D. B. (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39,* 33-38.

RStudio Team (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics, 29,* 159-183.

Rubin, D. B. (1973b). 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 nonrandomized 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.

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

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*, 1334-1344.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York, NY: Houghton Mifflin Company.

Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice, 35*, 38-54.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*, 250-267.

Stone, C. A. & Tang, Y. (2013). Comparing propensity score methods in balancing covariates and recovering impact in small sample educational program evaluations. *Practical Assessment, Research, and Evaluation, 18*(13), 1-12.

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

Stuart, E. A. & Rubin, D. B. (2008). Best practices in quasi-experimental designs: Matching methods for causal inferences. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 155-176). Los Angeles, CA: SAGE Publications.

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

Wainer, H. (2016). *Truth or truthiness: Distinguishing fact from fiction by learning to think like a data scientist.* New York, NY: Cambridge University Press.

What Works Clearinghouse. (2017). *Standards handbook* (Version 4.0). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.