5-2016

Propensity score methods : a simulation and case study involving breast cancer patients.

John Craycroft
University of Louisville

PROPENSITY SCORE METHODS: A SIMULATION AND CASE STUDY INVOLVING BREAST CANCER PATIENTS

By

John Craycroft
B.A., B.S. The George Washington University, 1998
MBA, Emory University, 2004

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

Master of Science
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, KY

May 2016

PROPENSITY SCORE METHODS: A SIMULATION AND CASE STUDY

INVOLVING BREAST CANCER PATIENTS

By

John Craycroft
B.A., B.S. The George Washington University, 1998
MBA, Emory University, 2004

A Thesis Approved on

April 12, 2016

by the following Thesis Committee:

_____

Dr. Maiying Kong, Committee Chair

_____

Dr. Joseph Benitez

_____

Dr. Jeremy Gaskins

DEDICATION

This thesis is dedicated to my wife

Laurie Craycroft

who has provided immeasurable support and patience during this work.

ACKNOWLEDGEMENTS

I would like to thank my major advisor, Dr. Maiying Kong, for her weekly meeting with me and the guidance she provided for this project. I would also like to thank the other committee members, Dr. Joe Benitez and Dr. Jeremy Gaskins, for their time in reading this work and providing thoughtful discussion. I would also like to thank Dr. Carrie Geisberg Lenneman for providing the case study dataset for this analysis.

Finally, I would like to thank my family for their support and patience throughout this process.

ABSTRACT

PROPENSITY SCORE METHODS: A SIMULATION AND CASE STUDY INVOLVING BREAST CANCER PATIENTS

John A. Craycroft

April 12, 2016

Observational data presents unique challenges for analysis that are not encountered with experimental data resulting from carefully designed randomized controlled trials. Selection bias and unbalanced treatment assignments can obscure estimations of treatment effects, making the process of causal inference from observational data highly problematic. In 1983, Paul Rosenbaum and Donald Rubin formalized an approach for analyzing observational data that adjusts treatment effect estimates for the set of non-treatment variables that are measured at baseline. The propensity score is the conditional probability of assignment to a treatment group given the covariates. Using this score, one may balance the covariates across treatment groups and obtain unbiased estimates of treatment effects. This paper has three objectives: to explain propensity scores, their assumptions, and their applications; to illustrate their use and several considerations underlying various propensity score methods, by using simulation studies; and to use propensity score methods to estimate average treatment effect between two types of breast cancer chemotherapy treatment regimens, with respect to subsequent development of cardiotoxicity.

TABLE OF CONTENTS

PAGE
ACKNOWLEDGEMENTS……...…….……..iv
ABSTRACT……...…….……..……v
LIST OF TABLES …….……vii

INTRODUCTION……...…1

METHODS……...10

SIMULATION STUDY……...24

CASE STUDY……..35

DISCUSSION……...44

REFERENCES...……...51

APPENDIX…..……...……..54

CURRICULUM VITA……...……..63

LIST OF TABLES

TABLE PAGE

1. Covariate distributions and parameters for simulation analysis……....25

2. Propensity score model specifications……...26

3. Baseline distribution of covariates for randomly selected simulated dataset……....27

4. MSE, Bias, and Variance for IPTW method for n=100 vs n=1000……...28

5. Standardized Mean Differences, one representative simulated dataset…….30

6. PS application method vs PS model specification…….31

7. 95% confidence interval coverage rates……34

8. Baseline distribution of covariates by TREATMENT group……38

9. Baseline distribution of covariates by OUTCOME group…….39

10. Standardized Mean Diffs before and after adjustment on propensity score……41

A1. Summary of outcome vs treatment, representative simulated dataset……51

A2. Summary of outcome vs treatment, case study dataset……...51

A3. MSE, Bias, and Variance, 4 propensity score methods, n=100 vs n=1000……52

A4. Description of case study variables…….……53

CHAPTER I.

INTRODUCTION

1.1 Introduction.

Why do we engage in scientific research? Quite simply, to gain knowledge about our world. Rarely, however, is knowledge sought as nothing more than a snapshot of current conditions; rather, we want to understand how conditions change over time. In particular, we want to know what *causes* those changes, and *how* causes and effects relate to each other. We want to know *how things work.* It is probably only a slight oversimplification to say that scientific research is, fundamentally, intended to identify and describe "cause-and-effect" relationships in our world.

For our beliefs derived from scientific research to be true and justified – and thus constitute real knowledge – we must have sound logical reasoning underlying the inferential process by which we draw general conclusions from empirical data. One aspect of this logical reasoning is that different types of data support different levels of causal inference. *Experimental data* from a randomized controlled trial may be considered the "gold standard" for empirical data that enables justified causal inferences, but there exist significant limitations with obtaining and using that type of data. In this paper we are concerned with propensity score methodologies, a set of statistical

techniques that aid in the process of gleaning accurate causal inferences from *observational data*. We will explain what propensity scores are, explore assumptions underlying these methodologies, and demonstrate their application on both real world and simulated data.

The structure of the paper will be as follows: in this Introduction, we will describe the potential outcomes framework in which most empirical scientific research is conducted; describe two types of treatment effects, and explain how different types of covariates relate to treatments and responses. We will also explain the two main types of empirical data and the implications of each for scientific research. This, finally, will motivate the definition of the propensity score.

In the Methods section, we will present a more thorough exposition of propensity scores. We will define them formally in mathematical notation, explore their various properties, and the properties of effect estimates stemming from their use. We will also explain assumptions underlying the propensity score approach, discuss various considerations for estimating these scores, and explain several application methods that are commonly used after the scores are estimated.

The third section of this paper will describe the simulation study, while the fourth section will describe the case study. In these sections, we will describe the specific objectives of the studies, detail the characteristics of the data sets involved, explain how the propensity score methodologies were applied, and present the results. The final section in the body of the paper will provide discussion about those results, as well as what conclusions we may make regarding propensity score analysis in general.

The Appendix of this paper provides additional tables and charts related to the simulation and the case study analysis that were deemed noncritical for inclusion in the body of the paper, but useful nevertheless for the reader who may want even more details regarding the simulation and/or the case study analysis. Also in the Appendix, we present a highly summarized sketch of important developments in the philosophy of knowledge (epistemology); here, within the context of scientific research as a motivating paradigm, we highlight key concepts related to logical reasoning and causal inference and explain how these lead us to the need for the concept of propensity scores.

1.2 The potential outcomes framework.

Nearly all scientific research, at some level, explicitly or implicitly, makes assumptions that rely on the notions of counterfactuals. Counterfactuals are statements that invert statements about current or past conditions. In statistics, this use has been formalized in what has been termed the "potential outcomes framework." The idea of the "potential outcomes framework" is that for any unit in the population, there are two[1] potential outcomes: one outcome if the unit is exposed to the treatment of interest, and one outcome if the unit is not exposed to the treatment. If we let Y be an indicator of the outcome of interest (with a value of 1 indicating that the unit had the specific outcome, and 0 indicating that the unit did not have the outcome), and Z be an indicator of treatment status (with a value of 1 indicating that the unit received the treatment, and 0 indicating that the unit did not receive the treatment), then our two potential outcomes, in formula form, are: $Y|Z = 1,$ and $Y|Z = 0$. We shorten this notation to $Y(1)$and $Y(0)$. Then the "treatment effect" for a particular unit i is $Y_i(1) - Y_i(0)$. We call this the

[1] Actually it is more precise to say that there is one potential outcome for every possible treatment. However, throughout this paper we will focus on the binary treatment case only.

"potential outcomes framework" because, for any given unit i, we can observe only $Y_i(1)$ or $Y_i(0)$, but not both[2]. Holland refers to this situation as the "Fundamental Problem of Causal Inference" (1986).

1.3 Two types of treatment effects.

The goal for much research, particularly biomedical research, is to describe the causal effects of different treatments. To make accurate measurement and estimation possible, we need to be very precise in defining terms. We distinguish between two[3] types of causal effects for clear communication and analytical accuracy. These are the average treatment effect (ATE) and the average treatment effect among the treated (ATT). In short, the ATE is the effect if the treatment of interest were applied to the entire target population as compared to the entire population receiving the control, while the ATT is the effect of the treatment only within the treated population.

These two treatment effect measurements can be expressed in the following formulas:

$$ATE = E[Y(1) - Y(0)]$$

$$ATT = E[Y(1) - Y(0)|Z]$$

As noted by Austin (2011), with experimental data (more explanation on the difference between experimental and observational data below), these two measures of treatment effect are equal, on average, because by design and the application of randomization,

[2] Even in cases of repeated measures or crossover designs within a RCT, the different orderings of treatment and control status for a particular subject is fixed. These designs are outside the scope of this paper, but the emphasis here is that, obviously, we can only observe one actual course of events, and thus we still have a counterfactual outcome to estimate to determine subject-specific effects.

[3] A third type, the average treatment effect in the *untreated*, is mentioned by Williamson et al. (2012), but is unaddressed in most of the literature, and indeed would seem rarely to be pragmatically interesting. It is not explored in this paper.

there is no systematic difference in the treatment and control groups. Meanwhile with observational data, these two measures may differ because there may be a systematic difference in which units received the treatment and which units received the control.

In this analysis, due to the research goals for the case study analysis, which will be further explained in Section IV, we focus on the ATE.

1.4 Covariates and confounders.

Of course, things are not quite as simple as just computing the average outcome measures in the different treatment groups and then comparing them (either through a risk difference, or a relative risk, or an odds ratio). Other circumstances and characteristics may be influencing the levels of the outcome measure and the selection of sample units into the different treatment groups. We attempt to quantify these "other circumstances and characteristics" through measurements on additional variables, called covariates, taken before or at the same time as the treatment. Figure 1 below illustrates the causal relationships among confounders, treatment and outcome. The relationship between treatment and outcome (represented by the dashed arrow from Z to Y) is generally the central research question; however, that relationship is obscured by the influence of the confounders, which are covariates related to *both* the treatment and the outcome.

Figure 1. Relationships among confounders, treatment, and outcome.
(Dashed line indicates relationship of primary research interest)

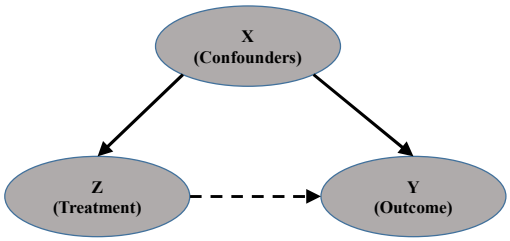
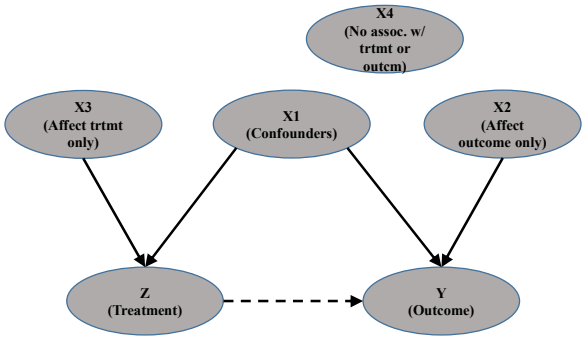


Figure 2 presents a modified causal diagram that highlights the various types of covariates and their relationships to the treatment variable and the outcome variable. The oval labeled "X1" represents confounders, i.e., those covariates related to both treatment and response. The oval labeled "X2" represents variables related to response only; "X3" represents variables related only to treatment; and "X4" represents variables related to neither treatment nor response. Although not frequently mentioned in the literature, it is entirely possible and probably common in observational data to have X4 type variables in the dataset. Determining which covariates fall into each of these four categories can be challenging, particularly in high dimensional datasets. (In later sections of this paper we will frequently refer to the different variable types simply by their label in Figure 2, e.g., "X3 variables" instead of "variables related to treatment only.")

Figure 2. Causal diagram highlighting 4 types of covariates.
(Dashed line indicates relationship of primary research interest)



1.5 Experimental vs. observational data

In biomedical research, a randomized controlled trial (RCT) is considered the "gold standard" of research methods, applying careful design decisions, sensitive measurement techniques, and sophisticated analysis methodologies in order to rigorously apply the scientific method to draw logical inferences regarding the relationships among specified treatments, covariates, and outcomes. The key feature of an RCT is the application of randomization in the assignment of units to the treatment groups.

Yet in many circumstances experimental studies are simply not an option. Cost, practicality, or ethics may imply that an RCT is either impossible or sub-optimal. In these cases, research can often still proceed by leveraging observational data.

What are the benefits of an observational study? First, data is often much easier and less expensive to come by. Second, many ethical issues are avoided. For example, if a researcher wants to be able to estimate the (detrimental) effect of some harmful behavior, such as smoking, it is not ethical to set up an experiment in which one group of human subjects would be assigned to receive the "smoking" treatment. Third, there are many cases where it is not just more difficult, but it is actually impossible to construct an RCT for the specific research question. Thus, for many scientific research questions to advance, observational data *must* be leveraged.

However, using observational data carries a significant challenge. The fundamental objective of much scientific research is to draw conclusions about *causal* relationships. An RCT supports the logical reasoning about causal relationships because that type of study carefully controls which covariates vary, and by how much, in the

experimental data. Furthermore, because of the randomization involved in the process of assigning units to treatment groups, both measured and unmeasured covariates are in theory balanced among the treatment groups. The non-randomized study enjoys no such benefits, and the logical support for causal conclusions may be tenuous. Selection bias of subjects and unmeasured covariates are very real issues. As one example: a blood pressure drug may be studied for its effect on decreasing blood pressure by interviewing a sample of physicians known to prescribe the drug. The results may show that the drug of interest appears to reduce blood pressure by more than an alternative drug. Yet that result could be due to the physicians systematically prescribing the drug to patients with higher baseline blood pressures; such patients may show more decrease in blood pressure from *any* similar drug, simply because they start off with a higher blood pressure on average. With this type of selection bias involved, we cannot really conclude anything about the performance of the drug of interest relative to the other drug.

In the above, it was stated that the treatment Z could be a control status, or some other alternative "baseline" treatment. In this, we are facing the same fine distinction lucidly delineated by Holland (1986); he states that

> [T]he effect of a cause is *always* relative to another cause. For example, the phrase 'A causes B' almost always means that A causes B relative to some other cause that includes the condition 'not A.' The terminology becomes rather tortured if we try to stick with the usual causal language, but it is straightforward if we use the language of experiments – treatment (i.e., one cause) versus control (i.e., another cause) (p. 946).

So, in our example above, the effect of the blood pressure drug under study may be relative to the absence of *any* drug, or it may be relative to the effect of an established (let us say standard) treatment. It will simplify our exposition to refer to the two treatments simply as "treatment" and "control," but it should be understood (and is in fact the case in

our case study described later) that the "control" may be simply a different treatment – the "not A" condition Holland describes.

1.6 Propensity score methodologies: Enabling causal inference from observational data.

More than 30 years ago, Rosenbaum and Rubin (1983) formalized a methodology in which the probability of a subject's treatment group is determined as a function of the measured covariates for that subject. Conditioning subsequent analysis on this probability enables unbiased estimation of the average treatment effect. Bias due to unmeasured covariates may still exist. In the Methods section we will go into greater detail about the propensity score and its associated methodologies. For now, the key idea is that this constitutes a method for handling the limitations inherent in observational data described above; multiple covariates can be effectively summarized into one scalar measure, a score. Then, conditioning properly on this score, one may obtain unbiased estimates of treatment effects. The insights articulated by Rosenbaum and Rubin in 1983 (and extended and refined in subsequent papers, both by them and by other authors) have proved immensely beneficial for the goal of gleaning causal inferences from observational data.

In the next section of this paper, we will define propensity scores in depth, explain the process for estimating them and how to evaluate those estimates, examine several ways that treatment effects are estimated using propensity scores, and highlight certain assumptions inherent in the approach.

CHAPTER II.

METHODS

2.1 Definition.

The propensity score is defined by Rosenbaum and Rubin (1983) as the probability of receiving the treatment given the covariates, which is expressed as

$$e(\boldsymbol{X}) = \Pr(Z = 1|\boldsymbol{X}).$$

It should be noted that the propensity score as defined by Rosenbaum and Rubin (1983) implies a treatment with exactly two levels, such as treatment versus control, or new therapy versus standard therapy. Imbens (2000) extended the concept to multi-level treatments; for our purposes in this paper, however, we will restrict attention to only binary treatment scenarios.

The propensity score is a balancing score. This means that, conditional on the propensity score, the distribution of all covariates is the same in the treatment group as in the control group. If $b(\boldsymbol{X})$ is a balancing score, then $\boldsymbol{X} \perp Z|b(\boldsymbol{X})$. An important result is that

> if treatment assignment is strongly ignorable [see next paragraph] given **X**,… then *the difference between treatment and control means at each value of a balancing score is an unbiased estimate of the treatment effect at that value*, and consequently pair matching, subclassification, and covariance adjustment on a

> balancing score can produce unbiased estimates of the average treatment effect (Rosenbaum and Rubin 1983) (italics added).

2.2 Assumptions.

Several assumptions are involved with the application of propensity scores. First is the assumption of "strongly ignorable treatment assignment" (SITA). There are two aspects to this. The first is the assumption that the response of a unit is conditionally independent of its treatment group, given its covariates: $(Y(1), Y(0) \perp Z|\boldsymbol{X}$. The second aspect is elsewhere referred to as "positivity" (e.g., Williamson et al., 2012; Cole and Hernan, 2008; and Funk et al., 2011) and is the notion that every unit has a positive probability of being assigned to either treatment group: $0 < \Pr(Z = 1|\boldsymbol{X}) < 1$. Austin (2011) notes that this SITA assumption is common to other statistical methodologies beyond propensity score methods, such as regression-based approaches.

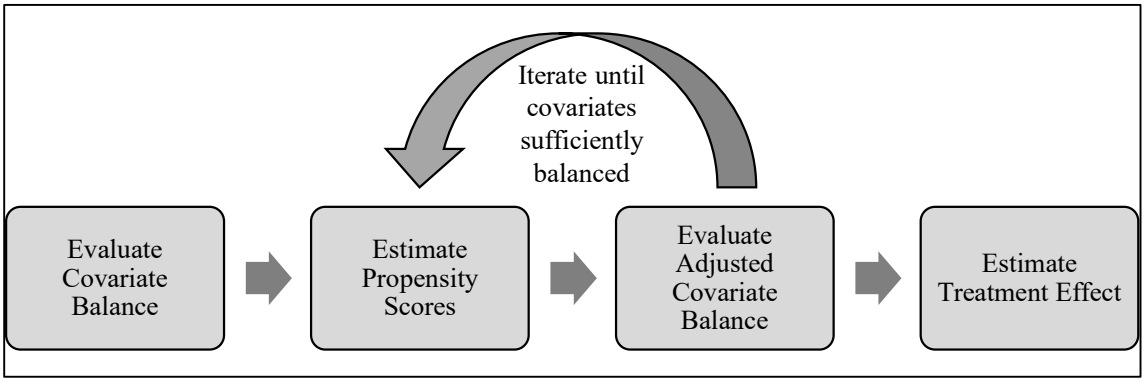
Many authors, such as Austin (2011), Funk, et al. (2011), Hirano and Imbens (2001), Lunceford and Davidian (2004), and Rosenbaum and Rubin (1984), highlight the fact that the SITA assumption is the same as an assumption of no unmeasured confounders; in other words, all covariates that are related to both the outcome and the treatment are measured and included in the propensity score model. This is a strong assumption, and it is in fact untestable, because it is an assumption about unmeasured variables (McCaffrey et al., 2013). If unmeasured confounders do exist, then estimates from propensity score methods (and other methods, for that matter) may be biased, and the more so as those nonmeasured confounders are less correlated with the other confounders (Funk et al., 2011).

Another major assumption of propensity score methodologies is the Stable Unit Treatment Value Assumption (SUTVA). This is the assumption that the treatment effect for one individual is not affected or influenced by the treatment status of another. It is a necessary assumption for limiting the potential outcomes for analysis, as without it, we would need to consider the two potential outcomes for unit *i* conditional on the treatment assignments for every other unit. Little and Rubin (2000) point out that "[w]ithout some such exclusion restrictions (to use the language of economists) which limit the range of potential outcomes, causal inference is impossible. Nothing is wrong with making assumptions…. The quality of these assumptions, not their existence, is the issue" (p. 123).

2.3 Typical propensity score analysis process.

Figure 3, below, illustrates the typical propensity score analysis process, assuming as a starting point that observational data have been collected and are ready for analysis. First, covariates are evaluated for their initial balance before adjustment. Then, the propensity scores are estimated for each unit, and the adjusted balance of the covariates is assessed. These two steps are iterated as necessary to achieve satisfactory covariate balance. Finally, treatment effects are estimated using the estimated propensity scores together with the chosen propensity score application method.

Figure 3. Typical propensity score analysis process.



2.4 Assessing balance of covariates.

It was stated in the previous section that the propensity score is a balancing score, and in Figure 3 above it is indicated that covariates should be examined for balance between the treatment and control groups both before and after adjustment for the propensity score. In this section we describe several approaches for assessing this balance; fundamentally, this step comes down to applying graphical and/or quantitative techniques for assessing the similarities of distributions in two different groups, much as might be done in an exploratory data analysis.

Within graphical approaches, the most common tool employed for assessing balance is boxplots (Austin and Stuart, 2015). Austin and Stuart advise that empirical cumulative distribution functions can also be used. Others, such as Rosenbaum and Rubin (1984), use simple bar charts to compare the means and proportions of particular covariates within subclasses, or strata, defined on the propensity score quintiles. It should be noted, however, that the covariates for treatment and control groups after balancing on the propensity score should be balanced on their entire distributions, not

solely their means or medians (Austin 2011), so bar charts may not be sufficiently informative.

Within quantitative approaches, a recommended approach is to compute the standardized mean difference (SMD) for each covariate (also called the "standardized bias," as in Harder et al. (2010)). Note that this metric may also be viewed in terms of its absolute value, i.e., the ASMD. As given in Austin (2011), the SMD is computed for continuous covariates as:

$$SMD = \frac{(\bar{x}_t - \bar{x}_c)}{\sqrt{\frac{s_t^2 + s_c^2}{2}}}$$

and for dichotomous covariates as:

$$SMD = \frac{(\hat{p}_t - \hat{p}_c)}{\sqrt{\frac{\hat{p}_t(1-\hat{p}_t) + \hat{p}_c(1-\hat{p}_c)}{2}}}$$

where the subscripts *t* and *c* stand for *treatment* and *control*, respectively. Categorical variables with more than two levels may be expressed via a series of dichotomous dummy variables, and then the balance assessed on those dummy variables. Note that the above formulae are for unweighted comparisons. If propensity scores have been estimated and the inverse probability of treatment weighting application method (see below) is used, then the propensity score-adjusted SMD may be computed by using the weighted means and variances in the above formulae (Austin and Stuart 2015).

There is no strong consensus as to what SMD threshold indicates a variable is sufficiently balanced. Austin (2011) and Austin and Stuart (2015) have suggested 0.1 is a

good threshold and indicated that other authors have advised that level as well, while Harder (2010) employed 0.25 in an empirical study, but suggested that for very important confounders (i.e., those highly related to the outcome), a lower threshold may be preferable.

Another quantitative approach is to use F statistics ((D'Agostino 1998), (Rosenbaum and Rubin 1984)) or some variation (t tests, Kolmogorov-Smirnov statistic) to assess the balance of covariates.

2.5 Estimating propensity scores and variable selection.

The second step illustrated in Figure 3 is to estimate the propensity score. The true underlying propensity score, the probability of selection into the treatment group given the covariates, is never definitely known. We must estimate it. Building a model for estimating the propensity score is not a trivial task. Several methodologies may be used here, although logistic regression and the nonparametric Generalized Boosted Model (GBM) are perhaps the two most common. Within logistic regression, which is the methodology of focus in this paper, there are multiple decisions related to which covariates to include and whether to include interactions or higher order terms of covariates.

Many authors, including among others Austin (2011), Brookhart et al. (2006), Emsley et al. (2008), Harder (2010), and Rosenbaum and Rubin (1984), have explored the question of which covariates are important to include in a logistic regression model for estimating the propensity scores. Some difference in opinion is evident in the literature (Millimet and Tchernis 2009). A few authors, such as (Emsley et al. 2008) say

that including *all* measured covariates in the propensity score model is the simplest approach and enhances the precision of the effect estimates, and while it may introduce some bias into the eventual treatment effect estimates, the bias introduced is small and so the "cost" is low. Other authors, such as (Brookhart et al. 2006), have performed simulations that illustrate that including all confounders and all variables related to outcome only (i.e., the X1 and X2 variables in Figure 2) is required for obtaining the least biased estimates of treatment effect, and as long as the confounders are included, the inclusion of X3 variables (those related to treatment only) will have only a modest effect on the MSE by increasing the variance modestly[4].

This latter view seems to be the prevailing view, although our current research did not attempt to evaluate the prevalence of each approach either in reviews, or tutorials, empirical/simulation studies, or real life data applications. Our sentiment is that part of the analyst's job is to categorize *all* measured covariates into the X1, X2, X3, or X4 type covariates defined in the Introduction section; thus, since there seems to be general agreement that including X1 and X2 variables in the propensity score model is optimal[5], only these two types of variables should be used. Nevertheless, given that it is not always straightforward to categorize covariates into these four types, particularly in high dimensional datasets, it is reassuring, although perhaps worthy of further theoretical testing and sensitivity testing in specific applications, that the inclusion of X3 type

[4] Relevantly, Brookhart et al. (2006) also indicate that the relative impact on the treatment effect estimator under various propensity score specification models depends to an extent on the propensity score application method (matching, stratification, etc.) that is used. In their simulations, they found that the stratification (subclassification) method was more sensitive to the misspecification of the propensity score model, in terms of effect on MSE, than was the cubic regression spline method.

[5] Even if not *very* optimal, due to low cost of including X3 variables as long as all X1 variables are included.

covariates seems to pose little risk to subsequent treatment effect estimation. Our simulation analysis described below appears consistent with this approach.

When building a logistic regression model to estimate the propensity scores, a further question presents itself, namely, what sorts of interaction or higher order terms should be used? In order to answer this, we feel it is important at this point to emphasize the purpose of the propensity score.

In the preceding discussion, we find it interesting (and believe it critical) to remember that the objective of the propensity score model is *not* to make the best predictive model of treatment group (Rubin 2004). In fact, one can hypothesize a propensity score model that perfectly predicts treatment group assignment based on the covariates. In this case, given a unit's vector of covariates, the probability of assignment to the treatment group for that unit would be either 0 or 1. This would be a perfectly predictive model, and, under normal regression purposes, a highly "successful" model. Yet, in the paradigm of propensity scores it would be a useless model. There would be no ability to estimate average treatment effect from the data if the treatment group were perfectly determined by the covariates. It is the same as saying that in an RCT in which only males received the test treatment, one could not make conclusions about the effect of the treatment on females, unless one makes the *assumption* that the effect of the treatment would be exactly the same on males and females. But RCTs are not designed to provide the treatment to only males, if females are also in the target population. RCTs, *by design*, ensure that covariates are balanced across treatment groups, and randomization assures us that, on average, unmeasured covariates/confounders will be balanced as well. Instead of a perfectly predictive propensity score model, we want a model that results in

balancing the covariates across the treatment groups. Rubin (2004) even points out that traditional regression diagnostics are not relevant for evaluating a propensity score model.

Having reinforced the fundamental objective of the propensity score model, namely to balance the covariates across the treatment groups, we can consider the question of variable selection for the propensity score model in a different light. A practical approach is to include in the propensity score model first all main effects for all measured covariates. Any covariates known *a priori* to be unrelated to the outcome could be omitted, but if there is uncertainty, they should be included. For any covariates that remain unbalanced after adjustment on the initial propensity score, interaction and higher order terms may be added to the propensity score model. This leads to the iterative process diagrammed in Figure 3. After each revision of the propensity score model, balance on all important covariates should be inspected, and a decision should be made whether the covariates are sufficiently balanced or not. Note that Harder et al. (2010) make the eminently reasonable suggestion that the balance required need not necessarily be the same for all covariates; covariates strongly related to outcome or treatment likely require a higher degree of balance, but covariates unrelated, or very weakly related, particularly to the outcome, may not need to be stringently balanced.

2.6 Typical approaches for applying propensity scores.

After propensity scores are estimated and covariates are deemed sufficiently balanced, the next step is to return to what is the main objective of the study: estimating average treatment effects. Several approaches can be used, and we next describe four that are highly prevalent in the literature. We will refer to these as propensity score

application methods, to distinguish this analysis step from the modeling of the propensity scores themselves. Harder et al. (2010) point out that any of the propensity score modeling methods (such as logistic regression, or generalized boosted models) may be paired with any of these application methods, leading to quite a number of potential paths that could be taken in a propensity score-based analysis.

In the following we are assuming a binary outcome, since that is the scenario for our case study data analysis and hence our simulation analysis.

Inverse probability of treatment weighting (IPTW) and Doubly Robust (DR). Each individual unit in the sample may be given a weight based on the propensity score. The weight is equal to the inverse of the probability of the treatment which that unit actually received (Rosenbaum 1987). Thus, for units in the treatment group, the weight is $1/e(\boldsymbol{X})$, while for units in the control group, the weight is $1/(1-e(\boldsymbol{X}))$. This can be expressed as a single formula for both groups as (Austin 2010):

$$\widehat{w_i} = \frac{Z_i}{\widehat{e_i}} + \frac{1-Z_i}{1-\widehat{e_i}}.$$

The IPTW estimate of average treatment effect (in terms of the risk difference) is:

$$\widehat{ATE}_{iptw} = \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i Y_i}{\hat{e}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-Z_i)Y_i}{(1-\hat{e}_i)}.$$

The doubly robust (DR) estimator, introduced by Lunceford and Davidian (2004), also involves weights based on the propensity score. In this method, the treatment effect is estimated via a combination of weights and an outcome regression model that relates outcome to the treatment variable and covariates. The name of the method refers to its

property of providing an unbiased estimate of treatment effect as long as either the propensity score model or the outcome regression model is correctly specified. To contrast the two methods, in the IPTW method above, if the propensity score model is misspecified, the treatment effect estimates will likely be biased. In the DR method, if the propensity score model is misspecified, then if the outcome regression model is correctly specified then the treatment effect estimates will still be unbiased. Lunceford and Davidian highlight that the analyst has "two chances" to specify a correct model. The DR estimate of average treatment effect is (Austin 2010):

$$\widehat{ATE}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i Y_i - (Z_i - \hat{e}_i)m_1(\boldsymbol{X_i}, \hat{\boldsymbol{\alpha}})}{\hat{e}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1 - Z_i)Y_i - (Z_i - \hat{e}_i)m_0(\boldsymbol{X_i}, \hat{\boldsymbol{\alpha}})}{(1 - \hat{e}_i)},$$

where $m_z(\boldsymbol{X_i}, \hat{\boldsymbol{\alpha}}) = E(Y|Z = z, \boldsymbol{X})$. In other words, $m_1(\boldsymbol{X_i}, \hat{\boldsymbol{\alpha}})$ and $m_0(\boldsymbol{X_i}, \hat{\boldsymbol{\alpha}})$ are the regressions of the response variable on the set of covariates $\boldsymbol{X_i}$ in the treatment and control groups, respectively.

Note that in both the IPTW and the DR application method, estimated propensity scores are used, because there is no way for us to know the true scores. Therefore we have estimated weights. This leads to some complexity in specifying the variances for these treatment effect estimators. Using bootstrapped estimates of the standard error of the estimator is recommended (Funk et al. 2011).

Stratification. In stratification (also called subclassification), observations in the sample are divided into *k* strata based upon the quantiles of the estimated propensity score. After the observations are divided into strata, the treatment effect is estimated within each stratum, and the several stratum-specific treatment effect estimates are

combined in a weighted average to estimate an overall treatment effect. The stratification estimate of average treatment effect is (Austin 2010):

$$\widehat{ATE}_{strat} = \frac{1}{k}\sum_{i=1}^{k}(p_{t,i} - p_{c,i}),$$

where $p_{t,i}$ and $p_{c,i}$ represent respectively the proportion of treated and control observations in the i^{th} stratum that experience the outcome.

The variance of this estimator is obtained by pooling the stratum-specific variances:

$$var(\widehat{ATE}_{strat}) = \left(\frac{1}{k}\right)^2 \sum_{i=1}^{k} \frac{p_{t,i}(1-p_{t,i})}{n_{t,i}} + \frac{p_{c,i}(1-p_{c,i})}{n_{c,i}}.$$

Several practical decision points and challenges do arise in the subclassification methodology. How many strata and where to draw the strata boundaries are two key decisions. The number of strata (k) can really be any number but is typically five, as it was shown by Cochran (1968) that stratification on five levels of a covariate removes 90% of the bias from that covariate. The strata boundaries are often set at the quintiles of the distribution of the estimated propensity score. This of course would imply that the total number of observations, treatment plus control, would be equal in each stratum. This in turn means that each stratum would receive equal weight in the estimation of ATE, which somewhat simplifies calculations. However, in cases where many observations in the control (treatment) group have a very low (high) propensity score, it could happen that defining the strata on the quintiles of the propensity scores of the full sample results in some strata having *only* control (treatment) observations. This situation

leads to a significant philosophical conundrum. On one hand, it seems reasonable to exclude control observations having very low propensity scores, when there are no treatment observations having similarly low propensity scores, because this is evidence that those controls are not in fact comparable to the treatment observations after adjustment for the covariates, and our reason for using propensity scores is to discern the treatment effect *after* adjustment for covariates. On the other hand, excluding a set of control observations like this would have two effects. First, the statistical power would be decreased, due to the lower sample size, making it more difficult to detect true treatment effects. Second, one could not really claim to be estimating the ATE anymore, since some types of controls were systematically excluded from comparison. Rather, assuming there remained enough controls to make treatment effect estimates with the treatment group observations, this must be interpreted now as an ATT estimate, average treatment effect among treated; and this may not be the effect that is desired from the study.

Matching. In propensity score matching, each treatment observation is matched to one or more control observations that are most like it in terms of the collective set of covariates, in other words, most similar in terms of $e(\boldsymbol{X})$. The treatment effect for every individual in the treatment group is then estimated by direct comparison against its match(es) on the outcome measure, and the multiple individual treatment effects are averaged for a population estimate.

While conceptually straightforward, propensity score matching does present challenges. The matching may be done in a number of ways. Each treatment observation may be matched to just one control or to several controls. Matching may be done without

replacement (once a control is matched to a treatment observation, it is unavailable for further matches) or with replacement. Determining which controls are "closest" can be approached in different ways, for example by using nearest neighbor matching, Mahalanobis metric matching, or nearest neighbor Mahalanobis metric matching within calipers (D'Agostino 1998).

It should be noted that the treatment effect estimate available from the matching application is the ATT, the average treatment effect among the treated (Imbens 2004). This is because an estimate is made for the treatment effect of each treated unit in the sample, but unmatched controls are not used. In this paper, we therefore do not use the matching application method, because our case study research objectives imply that it is the ATE that is of primary interest.

Covariate adjustment (i.e., regression). Another common method for employing propensity score analysis is to regress the outcome measure on the propensity score and the treatment group variable. This is a common approach in the medical literature. If the outcome measure is binary, then logistic regression would be the natural choice; however, as highlighted by Austin (2010), with logistic regression the odds ratio, not the risk difference, is the estimated effect.

Given the objectives of the case study, the analysis in this paper employs the first three of these methods only, i.e., IPTW, DR, and stratification. Now having defined propensity scores, explained assumptions and properties of the methods, described typical analysis procedures and application methods, we proceed to describe the simulation study we conducted.

CHAPTER III.

SIMULATION STUDY

3.1 Objectives.

We had three main objectives with this simulation. First, because our sample size in our case study was small (less than 100 total observations), we wanted to understand the impact of sample size on propensity score-based methods with respect to bias and variance of estimators. Second, we wanted to test which of three propensity score-based methods appeared to be best under varying conditions of misspecifying the propensity score model. Third, we wanted to assess confidence interval coverage for the different methods using the bootstrap estimate of the standard error. These objectives would all be instructive with respect to how we might want to analyze our case study data.

3.2 Data generating process.

In the nomenclature of Figure 2, we had three sets of simulated covariates, {X1}, {X2}, and {X3}. The set {X1}, the confounders, contained two variables, one a continuous variable with a standard normal distribution and the other a binomial variable with $p = 0.5$. The set {X2}, the prognosticators (related to outcome but not to treatment), also contained two variables, with same distributions as those in {X1}. Finally, {X3} contained a single variable, distributed as standard normal. This was related to treatment

but not to outcome. These five covariates and their distributions, along with their relationships to Z and to Y are all summarized in Table 1.

For each observation in the simulated datasets, we first generated the five covariates described above. {X1} and {X3} were then used together with specified beta parameters (see below) to compute the logit(Pr(Z=1)) for each observation. From this value, Pr(Z = 1) was computed, and then Z was generated as a binomial value with p = Pr(Z = 1).

Having generated the treatment variable, the outcome variable Y for each observation was generated in a similar fashion, but this time, {X1} and {X2} variables were used, along with Z and specified alpha parameters (see below) to compute the logit(Pr(Y=1|X1, X2, Z)) for each observation. From this value, Pr(Y = 1|X1, X2, Z) was computed, and then Y was generated as a binomial value with p = Pr(Y = 1|X1, X2, Z).

Table 1. Covariate distributions and parameters for simulation analysis.

	Covariate Set	Variable	Distributed As:	Related to:
Covariates	{X1}	x1.1	N(0, 1)	Treatment & Outcome
		x1.2	Binomial(0.5)	
	{X2}	x2.1	N(0, 1)	Outcome only
		x2.2	Binomial(0.5)	
	{X3}	x3	N(0, 1)	Treatment only
Treatment		Z	$logit(Z\|\boldsymbol{X}) = \beta_0 + \beta_{1.1}x_{1.1} + \beta_{1.2}x_{1.2} + \beta_3 x_3$ $where\ (\beta_0, \beta_{1.1}, \beta_{1.2}, \beta_3) = (-0.5, 1, 1, 1)$	
Outcome		Y	$logit(Y\|\boldsymbol{X}, Z) = \alpha_0 + \alpha_{1.1}x_{1.1} + \alpha_{1.2}x_{1.2} + \alpha_{2.1}x_{2.1} + \alpha_{2.2}x_{2.2} + \alpha_3 Z$ $where\ (\alpha_0, \alpha_{1.1}, \alpha_{1.2}, \alpha_{2.1}, \alpha_{2.2}, \alpha_3) = (-1, 1, 1, 1, 1, 1.25)$	

Note: The parameter value of 1.25 for α_3 represents an odds ratio of approximately 3.5 for the treatment effect. Together with the other parameters and covariates, this equates to a *true risk difference* of approximately 0.2.

Two sample size settings were tested, n = 100 and n = 1000. For each setting, 1000 simulated datasets were generated. For each dataset, eight propensity score models were tested. In all models, logistic regression was used to estimate the propensity scores;

the models varied according to which covariates were included in the logistic regression model. Table 2 summarizes the eight models. Note that Model 4 is the "correct" model in that it contains exactly {X1} and {X3}; Models 2, 3, 4, and 5 all contain {X1}, i.e., the confounders. Models 1, 6, 7, and 8 do not contain the confounders, although Models 7 and 8 do contain {X3}.

Table 2. Propensity score model specifications.

Model #:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Model Type:	Intercept only	Confounders Included				No Confounders		
Covariates included:	NONE	X1	X1X2	X1X3	X1X2X3	X2	X3	X2X3

For each simulated dataset, and under each of the eight propensity score model specifications, we applied four propensity score application methods: IPTW, DR, DR.MIS, and Stratification. DR is the doubly robust method with the outcome regression model correctly specified as:

$$logit(Y|\boldsymbol{X}, Z) = \alpha_0 + \alpha_{1.1}x_{1.1} + \alpha_{1.2}x_{1.2} + \alpha_{2.1}x_{2.1} + \alpha_{2.2}x_{2.2} + \alpha_3 Z$$

DR.MIS is the doubly robust method with the outcome regression model *misspecified* as:

$$logit(Y|\boldsymbol{X}, Z) = \alpha_0 + \alpha_{1.1}x_{1.1} + \alpha_{2.2}x_{2.2} + \alpha_3 Z$$

In DR.MIS, one continuous covariate and one dichotomous covariate have been left out of the outcome regression model.

3.3 Results.

The results from our simulation study will inform some of our decisions related to how we perform the analysis of the case study data (see Section IV). The specific details

of the data generating process were presented in the previous section. To give an idea of the type of dataset that was generated after all those steps were followed, we selected one of the randomly simulated datasets. Table 3, below, compares distributional statistics for all five covariates within levels of the treatment variable and within levels of the outcome variable, providing a baseline assessment of how similar the treatment groups and outcome groups are with respect to each covariate.

Table 3. Baseline distribution of covariates for randomly selected simulated dataset.

		Stratified by Treatment (Z)				Stratified by Outcome (Y)			
		0	1	p	SMD	0	1	p	SMD
	n	512	488			376	624		
CONTINUOUS VARIABLES	x1.1 (mean (sd))	-0.30 (0.90)	0.36 (0.94)	<0.001	0.718	-0.46 (0.88)	0.32 (0.91)	<0.001	0.86
	x2.1 (mean (sd))	0.03 (1.01)	0.05 (0.97)	0.805	0.016	-0.34 (0.96)	0.26 (0.93)	<0.001	0.637
	x3 (mean (sd))	-0.42 (0.91)	0.38 (0.91)	<0.001	0.877	-0.16 (0.94)	0.05 (1.02)	0.001	0.211
CATEGORICAL VARIABLES	x1.2 = 1 (%)	216 (42.2)	292 (59.8)	<0.001	0.359	126 (33.5)	382 (61.2)	<0.001	0.578
	x2.2 = 1 (%)	262 (51.2)	239 (49.0)	0.528	0.044	150 (39.9)	351 (56.2)	<0.001	0.332
	Z = 1 (%)					104 (27.7)	384 (61.5)	<0.001	0.725

We observe that for a typical dataset simulated through the process described above, we have five covariates (three continuous and two binary), one binary treatment variable, and one binary outcome variable. From the left half of Table 3, we see that x1.1, x1.2, and x3 seem associated with treatment (low p-values and high standardized mean differences). Meanwhile, x2.1 and 2.2 do not appear associated with treatment. These observations accurately reflect how the covariates, treatment, and response variables were generated. From the right half of Table 3, we see that all covariates and treatment appear associated with outcome, due to low p-values and high SMDs. We note that in the true model, x3 is not associated with outcome. We also note that, while its p-value is very low at 0.001, its SMD shown in Table 3 is the lowest of all the covariates. As one final note with respect to this representative simulated dataset, we point out that

the crude, unadjusted estimate for the average treatment effect in terms of the risk difference is 0.318; the treatment group has a probability of a positive outcome that is 0.318 higher than that of the control group (see Appendix Table A1).

Next, we explore results for the full set of 1000 iterations. Our first objective with this simulation study was to understand the impact of sample size on propensity score-based methods with respect to bias and variance of treatment effect estimators. Table 4 presents the simulation results in terms of MSE, Bias, and Variance, for the IPTW propensity score application method. By comparing adjacent rows, we can make the following observations: (1) Increasing the sample size 10-fold (from 100 to 1000) had no notable impact on the bias of the estimates (differences in bias seen in the table for the same propensity score model specification are due solely to random sampling variations).

Table 4. MSE, Bias, and Variance for IPTW method for n=100 vs n=1000.

			Propensity Score Model							
			Int. only	Confounders Included				No Confounders		
Propensity Score Application Method			**NULL**	**X1**	**X1X2**	**X1X3**	**X1X2X3**	**X2**	**X3**	**X2X3**
IPTW	**MSE**	**n100**	0.2829	0.1237	0.1192	0.2810	0.3069	0.2746	0.3564	0.3593
		n1000	0.0216	0.0011	0.0009	0.0026	0.0025	0.0214	0.0278	0.0276
	BIAS	**n100**	0.1409	-0.0004	0.0001	-0.0017	0.0004	0.1416	0.1573	0.1604
		n1000	0.1441	0.0011	0.0009	0.0008	0.0005	0.1438	0.1637	0.1633
	VAR	**n100**	0.2630	0.1237	0.1192	0.2810	0.3069	0.2546	0.3316	0.3336
		n1000	0.0008	0.0011	0.0009	0.0026	0.0025	0.0007	0.0010	0.0009

(2) Increasing the sample size 10-fold decreased the variance by a factor of a little over 100 for propensity score models including X1, while for models excluding X1 it decreased the variance by a factor of over 300. (3) Consequently, increasing the sample size 10-fold decreased the MSE by a factor of a little over 100 for propensity score models including X1, while for models excluding X1 it decreased the MSE by a factor of

around 13. The lower decrease in MSE when moving from a sample size of 100 to a sample size of 1000 for propensity score models excluding X1 is explained by the fact that these models have substantially more bias (which remains unchanged by increasing sample size). Since $MSE(\hat{\theta}) = Var(\hat{\theta}) + \left(Bias(\hat{\theta})\right)^2$, for biased estimators the bias, being a quadratic term, has a relatively large effect on MSE as compared to the variance. Therefore, even though the variance decreased substantially with the increase in sample size, the bias component prevented the MSE from decreasing to a similar degree.

Table 4 above presents the results for the IPTW propensity score application method only. Table A3 in the Appendix presents the results for all four application methods tested in this simulation. The pattern described above is completely consistent for each of the four propensity score application methods tested[6]; thus, for brevity and clarity, we confine our subsequent analysis solely to the scenarios with sample size 1000.

Before comparing the four propensity score application methods, we demonstrate how the balance of the covariates between the treatment groups may be reassessed after adjusting for the propensity score. Table 3 earlier showed the unadjusted SMD for each of the covariates. Variables x1.1, x3, and x1.2 were significantly unbalanced with respect to the treatment, Z. Table 5, below, shows the SMDs after weighting by the propensity score. Note that the SMDs shown for Model 1, the intercept only or null propensity score model, match the SMDs shown in the 4th column of Table 3 above. In every scenario where the X1 variables were included in the propensity score model specification, the adjusted SMDs for x1.1 and x1.2 are below 0.2 – and they are usually well below 0.10.

[6] With an apparent exception, that yet actually follows the description above. See Appendix Table A3 footnote.

Table 5. Standardized Mean Differences, w/ respect to treatment group (Z), weighted by propensity score (for one representative simulated dataset).

		Propensity Score Model							
		Int. only	Confounders Included				No Confounders		
		1	2	3	4	5	6	7	8
		NULL	X1	X1X2	X1X3	X1X2X3	X2	X3	X2X3
CONTINUOUS VARIABLES	**x1.1**	0.718	0.009	0.018	0.142	0.163	0.718	0.833	0.833
	x2.1	0.016	0.022	0.021	0.009	0.086	0.000	0.020	0.000
	x3	0.877	0.981	0.979	0.029	0.014	0.874	0.022	0.021
CATEGORICAL VARIABLES	**x1.2**	0.359	0.014	0.010	0.018	0.023	0.358	0.432	0.431
	x2.2	0.044	0.054	0.020	0.028	0.065	0.000	0.004	0.012

Likewise, in every scenario where X3 is included in the propensity score model specification, the adjusted SMD for that variable is very low, always below 0.03. Finally, in Model 5, which includes all three types of covariates, all five of the variables have an adjusted SMD less than 0.20. These points demonstrate that, for this relatively noncomplex simulated dataset, summarizing multiple covariates into one scalar score can effectively serve to balance those covariates in the treatment groups.

Next, we turn to our second objective of the simulation analysis, comparing the three different propensity score application methods in terms of MSE, Bias and Variance for the different specifications of the propensity score model. Table 6 below has an identical structure to Table 3 above, but includes all propensity score application methods and suppresses the results for the smaller sample size datasets (see Table A3 in the Appendix for those results). Note that for the second propensity score application method, the doubly robust (DR) method, we actually tested two different scenarios. In the first scenario, the outcome regression model was correctly specified. In the second scenario, the outcome regression model was misspecified by omitting two covariates that were truly related to the outcome. This may be a better representation of real-life

analysis in that we usually do not know with certainty if our regression model is correctly specified.

Table 6. Varying propensity score application method and propensity score model specification.

			Propensity Score Model							
			Int. only	Confounders Included				No Confounders		
Propensity Score Application Method			**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
			NULL	**X1**	**X1X2**	**X1X3**	**X1X2X3**	**X2**	**X3**	**X2X3**
IPTW	**MSE**	**n1000**	0.0216	0.0011	0.0009	0.0026	0.0025	0.0214	0.0278	0.0276
	BIAS	**n1000**	0.1441	0.0011	0.0009	0.0008	0.0005	0.1438	0.1637	0.1633
	VAR	**n1000**	0.0008	0.0011	0.0009	0.0026	0.0025	0.0007	0.0010	0.0009
DR	**MSE**	**n1000**	0.0007	0.0008	0.0008	0.0012	0.0013	0.0007	0.0009	0.0009
	BIAS	**n1000**	-0.0003	-0.0001	0.0000	-0.0009	-0.0009	-0.0003	-0.0006	-0.0006
	VAR	**n1000**	0.0007	0.0008	0.0008	0.0012	0.0013	0.0007	0.0009	0.0009
DR.MIS	**MSE**	**n1000**	0.0021	0.0010	0.0008	0.0015	0.0013	0.0020	0.0026	0.0025
	BIAS	**n1000**	0.0352	0.0003	0.0001	-0.0002	-0.0004	0.0350	0.0398	0.0395
	VAR	**n1000**	0.0009	0.0010	0.0008	0.0015	0.0013	0.0008	0.0010	0.0009
STRAT	**MSE**	**n1000**	0.0216	0.0012	0.0010	0.0016	0.0014	0.0214	0.0271	0.0269
	BIAS	**n1000**	0.1441	0.0141	0.0141	0.0163	0.0159	0.1439	0.1619	0.1617
	VAR	**n1000**	0.0008	0.0010	0.0008	0.0013	0.0011	0.0007	0.0009	0.0008

Our first observation from Table 6 is that the Doubly Robust (DR) method consistently has the lowest MSE, and the misspecified DR model, DR.MIS, consistently has the second lowest MSE, both regardless of propensity score model specification. For scenarios where the propensity score model includes X1 variables (Models 2, 3, 4, and 5), the differences in MSE across the methods is quite small, but for the models without the confounders (Models 1, 6, 7, and 8) the differences in MSE are substantial between IPTW/STRAT and DR/DR.MIS methods.

The fact that the bias for the DR method is effectively zero for *all* propensity score model scenarios is a demonstration of the doubly robust property of this method. Even for propensity score models that were misspecified (i.e., all except for Model 4), the estimators are unbiased. This is because in this method, the outcome regression model is correctly specified. In the DR.MIS method, the outcome regression model is

misspecified. Here, we see that as long as the confounders, X1, are included, the bias in the estimator is negligible. If, however, the confounders are not included, then the estimators have a larger degree of bias.

Note that, compared to IPTW and STRAT methods, the DR.MIS bias is still *relatively* low. This is due in part to the particular misspecification that was used. One can imagine that as the outcome model departs further from the true model – say, by introducing unrelated variables, or in a more complex relationship, leaving off more covariates – the bias of $\widehat{ATE}$ may become very significant. Of course, this is entirely true outside of the paradigm of propensity score applications as well.

As discussed in the Methods section, some literature points out that introducing X3 type variables in the propensity score model increases the variance of $\widehat{ATE}$. This is seen in Table 6 most prominently for the IPTW application method, by comparing Model 3 against Model 5. The variance of the estimator increases from 0.0009 to 0.0025. The pattern of increasing variance holds true for the other propensity score methods as well. Here again, it should be noted that the particular conditions of our data generating process likely inherently limit the increase in variance that the introduction of X3 variables may cause. First, in our case there is only one X3 type variable; second, the distribution of that variable is standard normal. If several X3 variables were introduced to the propensity score model, or if the X3 variables differed more substantially in location or scale, then it is likely that there would be a larger impact on the variance of $\widehat{ATE}$.

Finally, just comparing IPTW and STRAT results in Table 6, the STRAT method consistently had the most bias, while the IPTW method consistently had the highest variance. The two are effectively equivalent when evaluated on MSE.

We turn now to assessing the standard errors of the $\widehat{ATE}$ estimators from the simulation. We used bootstrap sampling to estimate the standard errors and calculate the coverage rates of the resulting 95% confidence intervals under each of the propensity score application methods. We confined our bootstrap sampling to only those propensity score models that contained the confounders X1, since the previous step demonstrated that including X1 type variables is necessary for achieving unbiased treatment effect estimates.

Table 7 below summarizes the coverage rates for two types of confidence intervals, the 95% percentile interval and the 95% normal confidence interval. We can see that, using the Percentile CI approach, under most propensity score model specifications, the actual coverage rate lags the nominal 95% rate by .5 to 3 percentage points. The DR method has a very consistent actual coverage rate, regardless of propensity score model. The IPTW method usually has the coverage rate closest to nominal, although it is usually pretty close to the others. Meanwhile, in the lower panel, we can see the IPTW method consistently has a coverage rate within 1 percentage point of the nominal 95% rate. Recall from Table 6 above that the IPTW estimator always has the highest variance under our simulation conditions. Finally, the stratification method in general has the greatest departure from the nominal coverage rate.

Table 7. 95% confidence interval coverage rates.

	Propensity Score Application Method	Propensity Score Model X1	X1X2	X1X3	X1X2X3
Percentile CI	IPTW	0.934	0.944	0.935	0.929
	DR	0.923	0.924	0.929	0.926
	DR.MIS	0.945	0.935	0.932	0.927
	STRAT	0.925	0.919	0.935	0.927
Normal CI	IPTW	0.944	0.941	0.954	0.951
	DR	0.939	0.938	0.939	0.942
	DR.MIS	0.953	0.940	0.941	0.940
	STRAT	0.923	0.926	0.932	0.929

NOTE: Shaded cells indicate coverage rates differing from .95 by < 0.01.

As one final demonstration from our simulation study, we return to the representative simulated dataset summarized in Table 3. In that discussion we pointed out that the crude, unadjusted estimate of the risk difference was 0.318. Estimating the propensity scores for that dataset and applying the IPTW and DR methods (using correct propensity score model specification and correct outcome model) results in adjusted treatment effect estimates of 0.168 and 0.167, respectively. These adjusted estimates are well within two standard errors of the actual treatment effect for the simulation setup, which was approximately 0.2. This reinforces that crude treatment effect estimates can be misleading; it is critical to adjust for confounders.

CHAPTER IV.

CASE STUDY

4.1 Background and objectives.

As cancer treatment strategies are improving, more patients are living longer after diagnosis and treatment (Geisberg and Sawyer, 2010). While that is undoubtedly a positive outcome, a concern is whether the cancer treatment therapies increase the risk of cardiotoxicity in cancer survivors. Cardiotoxicity is long term damage to the heart, and impaired heart function. It manifests itself over a long period of time, as late as 20 years after initial exposure. One particular cancer treatment that has been very effective in increasing the 5-year survival rate is a class of agents known as anthracyclines. Unfortunately, an estimated 5% to 23% of patients who undergo such cancer treatment later developing late-onset heart failure secondary to anthracycline-induced cardiotoxicity (Geisberg and Sawyer, 2010).

The data for this study come from 95 breast cancer patients at Vanderbilt University. The subjects had received one of two cancer treatment regimens, either anthracycline or herceptin, an alternative treatment type. The fundamental research question of interest is whether there is a difference in the risk of cardiotoxicity depending on which type of cancer treatment was received. In this study, the binary outcome of

cardiotoxicity is determined based upon measurement of the "ejection fraction" at multiple time points during the study. The ejection fraction is a measurement of how much blood is pumped out of the heart with each beat. Two conditions must hold for cardiotoxicity to be evaluated as occurring:

1. The lowest measured ejection fraction must be less than 55.
2. The lowest measured ejection fraction must be less than 90% of the baseline ejection fraction.

A variety of demographic and biometric variables were recorded at baseline. The case study dataset contained the binary outcome variable, the binary treatment variable, and 19 covariates. Of these, 10 were on the continuous scale, and 9 were categorical (mostly dichotomous). The covariates on the continuous scale included demographic data, such as age, and biomedical baseline data, such as heart rate and blood pressure. The covariates on the categorical scale included mainly indicators related to previous health events, such as whether the subject had a history of hypertension, and current medications, such as whether the patient was on beta blockers at the time of the baseline assessment. See Table A4 in the Appendix for a full description of all the covariates in the dataset.

The goal of this study, then, is to determine whether the risk of cardiotoxicity differs between patients receiving the anthracycline treatment versus those receiving herceptin. As an observational study, the patients were not randomly assigned to receive one of these treatments or the other. No demographic or biomedical factors were specifically controlled for. The desire, however, is to see if the data can tell us anything

about the relative cardiotoxicity risk for patients under the two different treatments, after controlling for these many other factors. Given that this is observational data with high dimensionality, and the objective is to estimate a binary treatment effect, propensity score methodologies seem well suited for this case.

4.2 Results.

We had 10 continuous and 9 categorical covariates. Tables 8 and 9 below summarize the distributions of those variables, including how they differ among treatment group (Table 8) and outcome (Table 9). The columns labeled "p" indicate the p-value of a test for whether there is a statistically significant difference in the distributions between the two groups. The columns labeled "test" indicate the type of statistical test that was performed. For continuous variables, a one-way ANOVA (equivalent to a t-test, since there are only two groups) was performed if the distribution was approximately normal, otherwise the nonparametric Kruskal-Wallis Rank Sum Test was performed. For categorical variables, a chi square test was performed unless the cell counts were too low, in which case a Fisher's Exact Test was performed.

In Table 8 below, we can see that the majority of the covariates are fairly well balanced at the outset relative to the treatment group (relatively high p-values, and low standardized mean differences), however, 9 of the 19 do have SMDs above 0.20.

Table 8. Baseline distribution of covariates by TREATMENT group.

		TREATMENT				
		ac	her	p	test	SMD
	n	67	28			
CONTINUOUS VARIABLES	Kar.Score (median [IQR])	100.00 [90.00, 100.00]	90.00 [90.00, 100.00]	0.218	KW	0.212
	Age (median [IQR])	49.00 [39.50, 58.00]	51.50 [45.75, 57.00]	0.443	KW	0.158
	Ht.in (mean (sd))	64.32 (2.28)	64.26 (4.11)	0.928	ANOVA	0.018
	Wt.lb (mean (sd))	162.19 (33.09)	175.03 (43.77)	0.121	ANOVA	0.331
	BMI (median [IQR])	26.90 [23.70, 30.00]	29.50 [23.15, 34.10]	0.200	KW	0.359
	HR (mean (sd))	80.53 (12.99)	73.36 (10.48)	0.011	ANOVA	0.608
	Sys.BP (mean (sd))	123.67 (15.25)	123.79 (14.10)	0.972	ANOVA	0.008
	Dias.BP (mean (sd))	76.39 (9.32)	75.29 (11.50)	0.624	ANOVA	0.106
	Leisure.Indx (mean (sd))	1.04 (0.31)	1.09 (0.36)	0.435	ANOVA	0.171
	Sport.Indx (median [IQR])	2.00 [0.67, 3.33]	2.01 [0.33, 3.08]	0.226	KW	0.240
CATEGORICAL VARIABLES	Race = White (%)	59 (88.1)	26 (92.9)	0.743	Chi-Sq	0.164
	Base.BC.Stage (%)			0.005	Chi-Sq	0.833
	Not staged	5 (7.5)	0 (0.0)			
	Stage I	8 (11.9)	12 (42.9)			
	Stage II	30 (44.8)	10 (35.7)			
	Stage III	24 (35.8)	6 (21.4)			
	Hist.Hypertension = yes (%)	19 (28.4)	10 (35.7)	0.642	Chi-Sq	0.158
	Hist.Hyperlipidemia = yes (%)	16 (23.9)	9 (32.1)	0.563	Chi-Sq	0.185
	Hist.Arryhthmia.CSD = yes (%)	7 (10.4)	1 (3.6)	0.429	Exact	0.272
	Hist.Diab.Mellitus = yes (%)	5 (7.5)	2 (7.1)	1.000	Exact	0.012
	Fam.Hist.Dilatd.Card.Myop = yes (%)	20 (29.9)	3 (10.7)	0.065	Exact	0.490
	Current.Beta.Block = yes (%)	10 (14.9)	3 (10.7)	0.749	Exact	0.126
	Current.ACE.ARB = yes (%)	9 (13.4)	6 (21.4)	0.363	Exact	0.212

NOTES: SMD=Standardized Mean Difference. KW=Kruskal-Wallis Rank Sum Test. ANOVA=One-way Analysis of Variance.
Chi-Sq=Chi-square test of independence. Exact=Fisher's Exact Test.
Continuous variables: Normally Distributed -- mean and standard deviation within treatment group are shown.
Non-normally distributed -- median and Inter-Quartile Range are shown.
Categorical variables: Frequency counts and percents within each treatment group are shown.

In Table 9 below, we are not concerned with balance of covariates between levels of the outcome variable. However, we are interested in an initial assessment of association between the covariates and the outcome, because we will want to include any covariates that do have such an association in our outcome regression model. However, Karnofsky Score is the covariate with the lowest p-value for this test of association, and its p-value is 0.164, not close to the traditional threshold of 0.05. While this may be a result of a true lack of association with the outcome for all of these covariates, it may also be an artifact of the fairly low sample size involved, a total of 95 observations, with only

18 in the positive outcome group. Accordingly, in building our outcome regression model, we were fairly liberal in testing multiple variables for inclusion in the model.

Table 9. Baseline distribution of covariates by OUTCOME group.

		OUTCOME				
		0	1	p	test	SMD
	n	77	18			
CONTINUOUS VARIABLES	Kar.Score (median [IQR])	100.00 [90.00, 100.00]	90.00 [90.00, 100.00]	0.164	KW	0.168
	Age (median [IQR])	50.00 [42.00, 59.00]	47.50 [39.25, 55.25]	0.330	KW	0.186
	Ht.in (mean (sd))	64.32 (2.91)	64.22 (3.02)	0.897	ANOVA	0.034
	Wt.lb (mean (sd))	164.92 (37.32)	170.47 (35.07)	0.568	ANOVA	0.153
	BMI (median [IQR])	27.00 [23.00, 31.60]	28.60 [25.13, 31.30]	0.445	KW	0.168
	HR (mean (sd))	78.12 (12.50)	79.67 (13.72)	0.645	ANOVA	0.118
	Sys.BP (mean (sd))	124.24 (15.38)	121.39 (12.40)	0.465	ANOVA	0.204
	Dias.BP (mean (sd))	75.59 (10.22)	78.11 (8.74)	0.336	ANOVA	0.265
	Leisure.Indx (mean (sd))	1.05 (0.33)	1.06 (0.29)	0.964	ANOVA	0.012
	Sport.Indx (median [IQR])	2.00 [0.67, 3.33]	2.50 [0.75, 3.33]	0.633	KW	0.123
CATEGORICAL VARIABLES	Race = White (%)	68 (88.3)	17 (94.4)	0.736	Chi-Sq	0.22
	Base.BC.Stage (%)			0.273	Chi-Sq	0.574
	Not staged	5 (6.5)	0 (0.0)			
	Stage I	17 (22.1)	3 (16.7)			
	Stage II	29 (37.7)	11 (61.1)			
	Stage III	26 (33.8)	4 (22.2)			
	Hist.Hypertension = yes (%)	24 (31.2)	5 (27.8)	1.000	Chi-Sq	0.074
	Hist.Hyperlipidemia = yes (%)	21 (27.3)	4 (22.2)	0.888	Chi-Sq	0.117
	Hist.Arryhthmia.CSD = yes (%)	7 (9.1)	1 (5.6)	1.000	Exact	0.136
	Hist.Diab.Mellitus = yes (%)	5 (6.5)	2 (11.1)	0.614	Exact	0.164
	Fam.Hist.Dilatd.Card.Myop = yes (%)	18 (23.4)	5 (27.8)	0.762	Exact	0.101
	Current.Beta.Block = yes (%)	12 (15.6)	1 (5.6)	0.451	Exact	0.331
	Current.ACE.ARB = yes (%)	11 (14.3)	4 (22.2)	0.474	Exact	0.207

NOTES: SMD=Standardized Mean Difference. KW=Kruskal-Wallis Rank Sum Test. ANOVA=One-way Analysis of Variance.
Chi-Sq=Chi-square test of independence. Exact=Fisher's Exact Test.
Continuous variables: Normally Distributed -- mean and standard deviation within treatment group are shown.
Non-normally distributed -- median and Inter-Quartile Range are shown.
Categorical variables: Frequency counts and percents within each treatment group are shown.

The crude estimate of treatment effect in terms of risk difference is -0.016, indicating almost no meaningful difference in risk of cardiotoxicity between the two treatment groups.

As shown in Table 8 above, 9 out of 19 covariates were initially unbalanced (defined as having an SMD greater than 0.20) across the two treatment groups. Our initial propensity score model entered all 19 covariates as main effects only into a logistic regression model with the treatment variable (chemotherapy type) as the outcome.

Checking the propensity score-adjusted balance by means of the SMD revealed that only four covariates remained unbalanced, and three of those had SMDs very close to the 0.20 threshold. The SMD for the fourth unbalanced covariate, Heart Rate, decreased substantially from 0.608 initially to 0.359 after adjustment by this specification of the propensity score. This demonstrates how the adjustment by the propensity score generally improves the balance of the covariates that are included in the propensity score model.

Nevertheless, in our simulation study above, we reaffirmed the theme in the literature that including X3 (and by extension X4) type variables in the propensity score model increases the variance of our ATE estimators. Therefore, we would prefer to have a propensity score model specification that includes only X1 and X2 type variables. We went through the list of 19 covariates and classified each as X1 (related to both treatment and outcome), X2 (related to outcome only), X3 (related to treatment only), or X4 (related to neither). It should be noted that we were fairly generous in our determination of whether a covariate was related to treatment or outcome because, as shown in Tables 8 and 9, there are very few strong associations from any of the covariates. We then revised the propensity score model to include only X1 and X2 type variables. Table 10 below summarizes the SMD for each covariate before any adjustment and after adjustment under the two different propensity score models. Shaded cells in the table indicate SMDs greater than 0.20.

Table 10. Standardized Mean Diffs before and after adjustment on propensity score.

		Covariate Type	Propensity Score Model 0	1	2
CONTINUOUS VARIABLES					
	Kar.Score	**X1**	**0.212**	**0.044**	**0.036**
	Age	**X2**	**0.158**	**0.203**	**0.099**
	Ht.in	X4	0.018	0.034	0.013
	Wt.lb	X3	0.331	0.054	0.366
	BMI	X3	0.359	0.042	0.356
	HR	X3	0.608	0.359	0.539
	Sys.BP	**X2**	**0.008**	**0.102**	**0.005**
	Dias.BP	**X2**	**0.106**	**0.215**	**0.019**
	Leisure.Indx	X4	0.171	0.091	0.376
	Sport.Indx	X4	0.24	0.052	0.233
CATEGORICAL VARIABLES					
	Race	X4	0.164	0.164	0.158
	Base.BC.Stage	**X1**	**0.833**		
	Not staged				
	Stage I			**0.151**	**0.053**
	Stage II			**0.145**	**0.020**
	Stage III			**0.185**	**0.046**
	Hist.Hypertension	X4	0.158	0.109	0.056
	Hist.Hyperlipidemia	X4	0.185	0.147	0.078
	Hist.Arryhthmia.CSD	X4	0.272	0.207	0.366
	Hist.Diab.Mellitus	X4	0.012	0.077	0.156
	Fam.Hist.Dilatd.Card.Myop	X3	0.49	0.125	0.433
	Current.Beta.Block	**X2**	**0.126**	**0.199**	**0.174**
	Current.ACE.ARB	X4	0.212	0.008	0.123

NOTES: Shaded cells indicate SMDs greater than 0.20.
Bold font rows indicate covariates categorized as either type X1 or type X2.
PS Model 0 is unadjusted; Model 1 includes all 19 covariates as main effects only; Model 2 includes only X1 and X2 type covariates.

Model 2 constituted our final propensity score model, because all of the X1 and X2 type covariates are balanced, as assessed by the SMD. Having settled on a propensity score model, we proceeded to estimate the ATE. Given the results of our simulation study, we focused our efforts on the Doubly Robust propensity score application method. This entailed fitting a logistic regression model in which our binary outcome, cardiotoxicity, was regressed on the treatment variable, chemotherapy type, as well as all

X1 and X2 type covariates, that is, all covariates related to outcome, whether or not also related to treatment. Of course, this is the same set of covariates as those used in building our propensity score model.

We used a stepwise selection procedure applied to the full set of X1 and X2 covariates (the six covariates identified in bold font in Table 10) to find the logistic regression model with the best fit for estimating the outcome variable, while requiring that the treatment variable be included in the model. It turned out that *no* covariates were selected into the model! Regardless of what subset of X1 and X2 variables was tried, no other covariates were deemed significant. We therefore expanded our scope and tested *all* 19 covariates for inclusion in the outcome regression model along with the treatment variable. Again, no covariates were selected into the model with the treatment variable.

The one exception to the above depended on the form of specifying the covariate indicating the subject's breast cancer stage. If this was entered in the model as a 4-level factor variable, then it was not selected in the stepwise procedure. If, however, it was entered as a series of three binary dummy variables, then "Stage2" was selected into the model. In this regression model, Stage2 had a p-value of 0.077, and the treatment variable had a p-value of 0.981, with a point estimate of -0.014 (i.e., an estimated odds ratio of 0.986, so essentially no treatment effect). Without the adjustment for "Stage2," the model contained the treatment variable only. In this case, the treatment variable had a p-value of 0.861, and a point estimate of -0.102 (odds ratio = 0.903).

Under the Doubly Robust propensity score application method, the formula for the estimated risk difference is:

$$\widehat{ATE}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\frac{Z_i Y_i - (Z_i - \hat{e}_i)m_1(\boldsymbol{X_i}, \widehat{\boldsymbol{\alpha}})}{\hat{e}_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1 - Z_i)Y_i - (Z_i - \hat{e}_i)m_0(\boldsymbol{X_i}, \widehat{\boldsymbol{\alpha}})}{(1 - \hat{e}_i)},$$

Using this formula in the case where the outcome regression model includes *only* the treatment variable, we obtain an estimate of the treatment effect of $\widehat{ATE} = -0.0103$.

This compares to the crude estimate of -0.0155. There really is not much difference between the crude and the adjusted estimates in this case, but of course, there are not any additional covariates strongly related to the outcome either that might have been confounding the treatment effect.

The bootstrap standard error of our estimate is 0.098, and the 95% normal confidence interval of the ATE is (-0.198, 0.185). The bootstrap ATE estimates do look normally distributed in a histogram, so this 95% CI seems appropriate (the percentile method for the confidence interval results in a very similar range). In conclusion, we do not reject the null hypothesis that the risk difference is zero.

CHAPTER V.

DISCUSSION

5.1 Simulation.

Observations. Our simulation exercise provided several insights into propensity score methodologies. In some cases, features of these methodologies that have been highlighted in the literature were reproduced and confirmed; in other cases, we derived insights that affected our own choices regarding how to analyze our case study data. The first observation from the simulation is that the sample size of the dataset can have a substantial impact on the variance of the estimators. This is, of course, exactly what we know to expect from basic statistical theory. Nevertheless, understanding the scale of the difference and how it varies across the different propensity score application methods was beneficial, particularly as our case study data had a low sample size as well (discussed more below). In our simulation scenarios, the coefficient of variation (standard deviation divided by the mean) of our effect estimates ranged from 1.4 to 2.8 for the low sample size datasets, and it ranged from 0.13 to 0.25 for our large sample size datasets.

Second, our simulation illustrated the process of checking the balance of covariates after adjusting on the propensity score. We focused on inspecting the

standardized mean differences, but other methods could be used as well. We showed that adjusting for the propensity score can result in balancing the covariates across treatment groups.

Our simulation also provided an illustration of the doubly robust property of the DR method. Even when the propensity score model was misspecified, our ATE estimates were unbiased; this was because the outcome regression model was correctly specified. In the DR.MIS method, which had a misspecified outcome regression model, moderate bias resulted if the propensity score model did not include all confounders. In this simulation setup, negligible bias resulted under the DR.MIS method as long as X1 variables were included in the propensity score model. Nevertheless, it is likely that in more complex relationships, more substantial bias would result when both outcome model and propensity score model are misspecified and the departures of those misspecifications from the true underlying models becomes more severe.

Next, the simulation analysis demonstrated that, under these conditions at least, the DR method is preferable to IPTW and Stratification. Even if we assume a misspecified outcome model is settled upon, the MSE of the resulting effect estimates are lower than the MSEs under the alternative methods. There is a bias-variance tradeoff; stratification improves on variance but is deficient on bias, while IPTW has the opposite effect. Combining these different goals into the MSE measure leads us to prefer the DR method.

Our simulation also illustrated the increase in variance of estimators if X3 type variables are included in the propensity score model. However, in most of our scenarios that increase was very modest. This supports the notion that the "cost" of including such

variables is low, and so to simplify analysis, or in cases where it is difficult to determine what the true relationships are, it may be reasonable to include these variables. Of course, our simulation had only one specific X3 type variable, with a straightforward relationship to the treatment variable. In real life, these relationships could be more complex, and the strength of the relationships can range from very low to very high.

Our analysis of propensity score standard errors using bootstrap estimates generally leads to confidence intervals with empirical coverage rates from one-half to three percentage points less than the nominal 95% coverage rate. The stratification method consistently had the lowest empirical coverage rate. This makes sense given that it generally had the highest bias, yet was among the lowest on variance. This finding is also consistent with the results of a simulation study by Austin, et al. (2010).

Limitations. In any simulation analysis, there are always more interesting questions that arise than can be covered within a reasonably defined scope. In our case, and with respect to connecting our simulation with our case study, we have two categories of limitations that future work could examine. First, the number and distributions of our covariates were both limited. Not that we must mimic our case study dataset to derive relevant insights, but some questions may be more illuminated with a more complex simulation setup. For example, it was stated above that the DR method is preferable to IPTW and stratification because, even assuming the misspecified DR model, the overall MSE was better. However, if the outcome actually has a more complex relationship to the treatment and, say, to multiple other covariates (more than just the 4 in our setup), perhaps one of the other two methods would be preferred to the DR method. After all, IPTW and stratification do not make further use of the covariates

after the estimation of the propensity scores. If an outcome regression model is too far off the mark, it could be the case that the bias introduced would lead to one of the other two methods being preferable.

Second, our setup included exactly one X3 type variable. As explained above, some authors have highlighted that including X3 type variables in the propensity score model increases the variance of the estimators. Other authors have suggested that the cost of including these variables is low, so an analyst should simply include all measured covariates from the outset. Our setup did not provide a means to truly evaluate this issue, for example by doing sensitivity testing assuming various numbers and distributions of X3 variables, and different scenarios related to their inclusion or exclusion in the propensity score models. It would interesting to determine if more instructive guidance could be determined on this subject.

5.2 Case study.

Observations. The most poignant conclusion from our propensity score analysis of the case study data is that there is no evidence of a non-zero treatment effect, even after adjustment and balancing on covariates. The difference between the crude treatment effect estimate and the propensity score-adjusted estimate is minimal, and both of these estimates are not statistically significantly different from zero. Few variables had strong univariate associations with the treatment group, and *none* of the variables had a strong univariate association with the outcome.

The propensity score balancing approach worked as expected for the most part. Some covariates, such as Heart Rate, remained unbalanced even after trying variations of

it in the propensity score model, such as a quadratic function, or an interaction with other variables. In this case, it is useful to return to the idea of only including X1 and X2 type variables in the propensity score model. Heart Rate was categorized as an X3 type variable: it is related to treatment ($p = 0.011$), but not to outcome ($p = 0.645$). Therefore, we really need not include it in the propensity score model, and we do not need to insist on balance for that variable.

Limitations. An important dynamic running throughout this case study analysis, however, is the very low sample size of the dataset. There were only 95 observations, and only 18 observations having the outcome of cardiotoxicity. Meanwhile, there were quite a large number of covariates – 19 included in the specific analysis here. Nearly half of these are categorical variables, meaning that there is serious sparseness in the data. While this analysis does not reveal any associations, it could well be that collecting a larger dataset or monitoring the subjects for a longer period of time may help reveal actual effects that do exist. Particularly as cardiotoxicity may be a condition that takes a fairly long amount of time to develop, or at least to manifest itself, the extended monitoring of subjects may be insightful.

Another limitation of the case study analysis has to do with missing data within the set of covariates. In the case study, the original dataset consisted of over 1200 variables. Many of these were redundant in some respects, and others were effectively summarized in other variables, such as "Sport Index" or "Leisure Index," which both served as summaries of multiple responses from the CHAMPS questionnaire set of variables. Nevertheless, much missing data existed in the dataset. Of the over 1200 variables, only about 20 had sufficient completion rates to be included in our analysis.

Within this subset, about half still had missing values, with the frequency of missing values ranging from 1 to 8 out of the 95 observations. In this analysis, all missing values were simplistically assigned to be the mean or median or mode of their respective distributions, as seemed most appropriate given the particular variable. A study focused solely on the most accurate estimation of the treatment effect would require more attention be given to this missing data issue. Specifically, multiple imputation methods would be conducted for estimating the actual values of the missing data, based on other existing covariates, and sensitivity analyses would be conducted for estimating how much the missing data assumptions and models affect the final estimates of treatment effects. For our purposes, we are primarily interested in the question of propensity score methodologies, and in particular how best to model the propensity score and how best to apply estimated propensity scores; detailed missing data analysis is beyond our scope. Furthermore, with the estimated treatment effect being so close to zero, the low number of observations affected by the missing data, and the minimal apparent associations in the dataset as a whole, it is exceedingly unlikely that conclusions regarding the treatment effect would change.

A final limitation of the case study data analysis is that there was in fact fairly low overlap in propensity scores between the treatment groups. According to some authors (for example, Little and Rubin, 2000), this implies that there may be little comparability between the groups, and so we cannot actually draw causal inferences from the analysis. While this may be, it is apparent that there is not even associational inference to made with this dataset, let alone causal inference!

5.3 General comments on propensity score methodology.

Propensity score methodologies are not a panacea. Many challenges exist in correctly applying and drawing insights from them. The specification of the propensity score model is not a trivial task. The choice of propensity score application method is important and will lead to a set of follow-on decisions. If the doubly robust method is chosen, then the specification of the outcome regression model is an important task. If stratification is chosen, then ensuring overlap of propensity scores between the treatment groups and deciding where stratum boundaries should be will both require attention. If IPTW method is chosen, although it is perhaps the simplest method to execute, then one must be aware of the potential for high weights (which may lead to the need for trimming) and, in general, realize that the variance of the estimators may be higher than under other methods.

Nevertheless, the benefits of the propensity score approach are substantial. Reducing a high-dimensional dataset into a single scalar score can assist in figuring out how similar or disparate the dataset observations are. Using this score for adjustment helps, in a straightforward way, to "level the playing field" with respect to the treatment groups. Considering the difficulties and limitations of experimental data, the ability to derive *causal* inferences from observational data is extremely valuable. There is no doubt, therefore, that propensity score methodologies represent an important tool in scientific research.

REFERENCES

Austin, P. C. (2010). "The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies." Stat Med **29**(20): 2137-2148.

Austin, P. C. (2011). "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." Multivariate Behav Res **46**(3): 399-424.

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

Brookhart, M. A., S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn and T. Sturmer (2006). "Variable selection for propensity score models." Am J Epidemiol **163**(12): 1149-1156.

Cochran, W. G. (1968). "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." Biometrics **24**(2): 295-313.

Cole, S. R. and M. A. Hernan (2008). "Constructing inverse probability weights for marginal structural models." Am J Epidemiol **168**(6): 656-664.

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

Emsley, R., M. Lunt, A. Pickles and G. Dunn (2008). "Implementing double-robust estimators of causal effects." Stata Journal **8**(3): 334-353.

Funk, M. J., D. Westreich, C. Wiesen, T. Sturmer, M. A. Brookhart and M. Davidian (2011). "Doubly robust estimation of causal effects." Am J Epidemiol **173**(7): 761-767.

Geisberg, C. and D. B. Sawyer (2010). "Mechanisms of Anthracycline Cardiotoxicity and Strategies to Decrease Cardiac Damage." Current hypertension reports **12**(6): 404-410.

Goodman, N. (1947). "The Problem of Counterfactual Conditionals." The Journal of Philosophy **44**(5): 113-128.

Harder, V. S., E. A. Stuart and J. C. Anthony (2010). "Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research." Psychol Methods **15**(3): 234-249.

Hirano, K. and G. W. Imbens (2001). "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." Health Services and Outcomes Research Methodology **2**(3): 259-278.

Hitchcock, C. (2012). Probabilistic Causation. The Stanford Encyclopedia of Philosophy. E. N. Zalta. URL = http://plato.stanford.edu/archives/win2012/entries/causation-probabilistic/.

Holland, P. W. (1986). "Statistics and Causal Inference." Journal of the American Statistical Association **81**(396): 945-960.

Hume, D. (1748). An Enquiry Concerning Human Understanding.

Ichikawa, J. J. and M. Steup (2014). The Analysis of Knowledge. The Stanford Encyclopedia of Philosophy. E. N. Zalta. URL = http://plato.stanford.edu/archives/spr2014/entries/knowledge-analysis/.

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

Imbens, G. (2004). "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." Review of Economics and Statistics.

Lewis, D. (1973). "Causation." The Journal of Philosophy **70**(17): 556-567.

Lewis, D. (2000). "Causation as Influence." The Journal of Philosophy **97**(4): 182-197.

Little, R. J. and D. B. Rubin (2000). "Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches." Annual Review of Public Health **21**: 121-145.

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

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

Menzies, P. (2014). Counterfactual Theories of Causation. The Stanford Encyclopedia of Philosophy. E. N. Zalta. URL = http://plato.stanford.edu/archives/spr2014/entries/causation-counterfactual/.

Millimet, D. L. and R. Tchernis (2009). "On the Specification of Propensity Scores, With Applications to the Analysis of Trade Policies." Journal of Business & Economic Statistics **27**(3): 397-415.

Rosenbaum, P. R. (1987). "Model-Based Direct Adjustment." Journal of the American Statistical Association **82**(398): 387-394.

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

Rosenbaum, P. R. and D. B. Rubin (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." Journal of the American Statistical Association **79**(387): 516-524.

Rubin, D. B. (2004). "On principles for modeling propensity scores in medical research." Pharmacoepidemiol Drug Saf **13**(12): 855-857.

Wikipedia. (7 November 2015 13:12 UTC). "Counterfactual conditional." Retrieved 20 March 2016 16:03 UTC, 2016, from https://en.wikipedia.org/w/index.php?title=Counterfactual_conditional&oldid=689481105.

Williamson, E., R. Morley, A. Lucas and J. Carpenter (2012). "Propensity scores: from naive enthusiasm to intuitive understanding." Stat Methods Med Res **21**(3): 273-293.

APPENDIX

6.1 Supplementary Tables

Table A1. Summary of outcome vs treatment, representative simulated dataset.

	Y=0	Y=1	Ttl	P(Y=1|Z=z)
Z=0	272	240	512	0.469
Z=1	104	384	488	0.787
Ttl	376	624	1000	
			Crude Risk Diff.	**0.318**

Table A2. Summary of outcome vs treatment, case study dataset.

	Y=0	Y=1	Ttl	P(Y=1|Z=z)
Z=0=ac	54	13	67	0.194
Z=1=her	23	5	28	0.179
	77	18	95	
			Crude Risk Diff.	**-0.015**

Table A3. MSE, Bias, and Variance for all four propensity score application methods, for n=100 vs n=1000[7].

			Propensity Score Model							
			Int. only	Confounders Included				No Confounders		
Propensity Score Application Method			NULL	X1	X1X2	X1X3	X1X2X3	X2	X3	X2X3
IPTW	MSE	n100	0.2829	0.1237	0.1192	0.2810	0.3069	0.2746	0.3564	0.3593
		n1000	0.0216	0.0011	0.0009	0.0026	0.0025	0.0214	0.0278	0.0276
	BIAS	n100	0.1409	-0.0004	0.0001	-0.0017	0.0004	0.1416	0.1573	0.1604
		n1000	0.1441	0.0011	0.0009	0.0008	0.0005	0.1438	0.1637	0.1633
	VAR	n100	0.2630	0.1237	0.1192	0.2810	0.3069	0.2546	0.3316	0.3336
		n1000	0.0008	0.0011	0.0009	0.0026	0.0025	0.0007	0.0010	0.0009
DR	MSE	n100	0.0806	0.0870	0.0882	0.1399	0.1715	0.0807	0.0948	0.0965
		n1000	0.0007	0.0008	0.0008	0.0012	0.0013	0.0007	0.0009	0.0009
	BIAS	n100	-0.0003	-0.0010	-0.0015	-0.0032	-0.0031	-0.0003	-0.0015	-0.0011
		n1000	-0.0003	-0.0001	0.0000	-0.0009	-0.0009	-0.0003	-0.0006	-0.0006
	VAR	n100	0.0806	0.0870	0.0882	0.1399	0.1715	0.0807	0.0948	0.0965
		n1000	0.0007	0.0008	0.0008	0.0012	0.0013	0.0007	0.0009	0.0009
DR.MIS	MSE	n100	0.1022	0.1026	0.0903	0.1673	0.1830	0.0930	0.1221	0.1138
		n1000	0.0021	0.0010	0.0008	0.0015	0.0013	0.0020	0.0026	0.0025
	BIAS	n100	0.0338	-0.0005	-0.0017	-0.0034	-0.0040	0.0341	0.0364	0.0379
		n1000	0.0352	0.0003	0.0001	-0.0002	-0.0004	0.0350	0.0398	0.0395
	VAR	n100	0.1010	0.1026	0.0903	0.1673	0.1830	0.0918	0.1208	0.1124
		n1000	0.0009	0.0010	0.0008	0.0015	0.0013	0.0008	0.0010	0.0009
STRAT	MSE	n100	0.2868	0.1131	0.0991	0.1420	0.1221	0.2830	0.3452	0.3487
		n1000	0.0216	0.0012	0.0010	0.0016	0.0014	0.0214	0.0271	0.0269
	BIAS	n100	0.1407	0.0115	0.0127	0.0120	0.0127	0.1425	0.1555	0.1599
		n1000	0.1441	0.0141	0.0141	0.0163	0.0159	0.1439	0.1619	0.1617
	VAR	n100	0.2670	0.1130	0.0989	0.1418	0.1219	0.2627	0.3211	0.3231
		n1000	0.0008	0.0010	0.0008	0.0013	0.0011	0.0007	0.0009	0.0008

[7] The "exception" mentioned in the text on page 29 in Section III refers to the fact that the reduction in MSE for the DR and DR.MIS methods is higher than the factor of 13, as mentioned for the IPTW method for models excluding X1. The reduction is actually by a factor of a little over 100, for DR, and around 47, for DR.MIS. But this is consistent with the rest of the explanation in that part of the text. Because DR method has virtually no bias, MSE is entirely driven by variance; any reduction in variance, therefore, translates to a virtually identical reduction in MSE. DR.MIS, meanwhile, has some bias, but much less than the IPTW or STRAT methods. Therefore, the reduction in variance is ameliorated to some extent when viewing the reduction in MSE, but it is not ameliorated to the degree that it is for the IPTW and STRAT methods. In those methods, the reduction in MSE was by a factor of around 13, despite a reduction in variance by a factor of over 300. For DR.MIS, the reduction in variance by a factor of a little over 100 resulted in a reduction of MSE by a factor of about 47.

Table A4. Description of case study variables.

CONTINUOUS VARIABLES	Kar.Score	Karnofsky Score*
	Age	Age at baseline
	Ht.in	Height in inches
	Wt.lb	Weight in pounds
	BMI	Body Mass Index
	HR	Heart rate
	Sys.BP	Systolic blood pressure
	Dias.BP	Diastolic blood pressure
	Leisure.Indx	Leisure index, a composite score summarizing several answers from the CHAMPS physical activity questionnaire
	Sport.Indx	Sport index, a composite score summarizing several answers from the CHAMPS physical activity questionnaire
CATEGORICAL VARIABLES	Race	Subject's race, either White or Non-white
	Base.BC.Stage	Baseline breast cancer stage
	Not staged	
	Stage I	
	Stage II	
	Stage III	
	Hist.Hypertension	History of hypertension
	Hist.Hyperlipidemia	History of hyperlipidemia
	Hist.Arryhthmia.CSD	History of arrhythmia or conduction system disease
	Hist.Diab.Mellitus	History of diabetes mellitus
	Fam.Hist.Dilatd.Card.Myop	Family history of dilated cardiomyopathy
	Current.Beta.Block	Currently on beta blocker
	Current.ACE.ARB	Currently on ACE inhibitor

* A measure assessing functional impairment, ranging from 0% (dead) to 100% (no complaints/no evidence of disease)

6.2 Background and Context: Logical Inference from Scientific Research.

Philosophy of knowledge and theories of causation. Why do we engage in scientific research? Quite simply, to gain knowledge about our world. Rarely, however, is knowledge sought as nothing more than a snapshot of current conditions; rather, we want to understand how conditions change over time and, in particular what *causes* those changes, and *how* causes and effects relate to each other. We want *to know how things work.* It is probably only a slight oversimplification to say that scientific research is, fundamentally, intended to identify and describe "cause-and-effect" relationships in our world.

Knowledge as justified true belief. There is a long and rich history in both philosophy and science (and the philosophy of science!) of exploring exactly what is meant by knowledge, and causation, and inference. One traditional theory of knowledge is the "JTB" theory of knowledge: we have real knowledge about something when we have a Justified, True Belief about that thing. It would not make sense to say we *knew* something if we did not *believe* it. If we believed something, but our belief was false, it certainly does not seem like we should say that we know it. Finally, if we have a belief, and the belief is true, but we are not *justified* in holding our belief (we are right about the belief only due to "luck," say), then again, it does not seem like we should claim to have knowledge. An example for this latter condition might be the following situation: you come home from work in the evening and see that the grass is wet, therefore you develop a belief that it rained earlier in the day. Perhaps it did in fact rain in the morning, so your belief happens to be true; but the grass dried after the rain, and it is now wet again actually because your neighbor watered his lawn in the afternoon. In this case, if the *only* justification for your belief about it raining earlier is the currently wet grass, then we would say that you do not have knowledge; your belief is true only from luck, and you are not justified in holding that belief[8].

Even though the JTB theory of knowledge does have its limitations and deficiencies, for our purposes, it is a sensible and pragmatic theory and provides a solid

[8] Certainly, there could be additional evidence leading to your belief that it rained. Perhaps the sky is very overcast. Perhaps not just the grass, but the sidewalk, and street, and cars are all wet too. Perhaps you actually *saw* it raining. It is the full body of evidence available, along with the set of alternative explanations for that full body of evidence, that would determine whether a particular belief is justified sufficiently to count as "knowledge."

basis for proceeding to ideas related to causal inference[9]. "Causal inference" is the process by which we use evidence to draw conclusions related to cause-and-effect relationships. This subject, too, has a rich history, spanning fields of philosophy and every branch of science, including in particular, the social sciences, biological sciences, and economics. Statistics, as a branch of science fundamentally concerned with deriving accurate and justified conclusions from empirical data, runs throughout *all* other sciences and is organically bound with philosophical ideas about logical inference. We will explore one particular line of causal theories, because it constitutes the foundation of logical reasoning underlying much – perhaps even all – scientific research.

Briefly, however, we distinguish two primary types of logical reasoning. *Deductive* reasoning takes premises and valid rules as implying, with certainty, particular conclusions. For example: In the morning, the sun is always in the East. It is now morning; therefore, the sun is in the East. If the premises are all true, then the conclusion must be true. This approach has been roughly characterized as reasoning "from the general to the specific." *Inductive* reasoning takes premises and evidence and *infers* conclusions. This is reasoning "from the specific to the general." Importantly, even under sound inductive reasoning, possibilities exist for alternative explanations than those inferred; standing theories are strong insofar as they are coherent with existing empirical data and support predictions about future empirical data, but they may be refined, qualified, and even at times, rendered inaccurate as our true knowledge of the world gets updated and expands.

[9] A good place to start for exploring more about the benefits and drawbacks of the JTB theory of knowledge, and about subsequent theories of knowledge, is the Stanford Encyclopedia of Philosophy Ichikawa, J. J. and M. Steup (2014). The Analysis of Knowledge. The Stanford Encyclopedia of Philosophy. E. N. Zalta. URL = http://plato.stanford.edu/archives/spr2014/entries/knowledge-analysis/..

Irrefutable certainty then, in a strictly logical sense, is *not* the outcome of the inferential process, or of inductive reasoning. Consequently, irrefutable certainty is not the result of the scientific process, either, since that process relies in large part on inductive reasoning. (Indeed, this is why it is sometimes claimed that scientific theories are never fully *proven*.) Nevertheless, we certainly can and do have real knowledge about our world, if we can live with the traditional idea of knowledge as justified, true belief. But the degree to which our beliefs constitute true knowledge is highly tied to *how true* those beliefs are, and *how justified* those beliefs are.

Causal theories and counterfactuals. Realizing then that many of our beliefs stemming from scientific research are beliefs about causes and effects, in order to think about how we can assess the truth status and the justification status for our cause-and-effect beliefs, we now explore certain theories of causation. The particular line of causal theories (i.e., what it means to say something is a cause, and how to establish cause-and-effect relationships) we will focus on begins with David Hume in his *Enquiry Concerning Human Understanding* in 1748. In Section VII of that treatise, Hume writes:

> "[W]e may define a cause to be *an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second. Or, in other words, where, if the first object had not been, the second never had existed.*" (Italics in original.)

This is important because it is, apparently, the first time that causation is mentioned in a way that would later become known as "counterfactuals" (Menzies 2014). Counterfactuals are statements that invert statements about current or past conditions. For example, let us suppose that it rained this morning, so we may reasonably say something like, "It rained this morning, so the grass got wet." A counterfactual would invert the antecedent condition, and this counterfactual is then typically followed by

some different (often opposite) effect. Thus: "If it had not rained this morning, the grass would not have gotten wet." The difference between the true condition and its counterfactual, together with the difference between the actual outcome and the hypothesized outcome under the counterfactual condition, lead us to make inferences about causes and effects. In the example, our inference is that the rain *caused* the grass to get wet[10].

As will be seen below, the idea of, and reliance on, counterfactual reasoning is absolutely central to our inferential process regarding causes and effects within the set of propensity score methodologies (indeed, it underlies most scientific research, as well as common inferences we make in our everyday lives). Interestingly, as pointed out by Menzies, these two sentences by Hume actually presuppose rather different (though not incompatible) theories of causation. Implicit in the first sentence is a "regularity" theory of causation, what Hume refers to as "constant conjunction." With a regularity theory of causation, particular causes are *always* followed by their effects (Hitchcock 2012). Indeed, in subsequent exposition of his definition of causation, Hume emphasizes the requirement of "constant conjunction."[11] On the other hand, Hume's second sentence in the quote above deals in hypotheticals, by reversing or contradicting an existing condition. Although Menzies claims that this sentence by Hume is the "first explicit definition of causation in terms of counterfactuals," Hume does not proceed to explore or

[10] Note that we are not concerned here currently with evaluating the truth status of the claim, "If it had not rained this morning, the grass would not have gotten wet." Such evaluation is important for establishing the justification for our inference, but at this point, we are simply defining terms and demonstrating the process of counterfactual reasoning.

[11] Two other criteria for causation besides constant conjunction, according to Hume, are spatial/temporal contiguity, and temporal succession Hume, D. (1748). An Enquiry Concerning Human Understanding.. These latter two criteria remain important even under a counterfactual theory of causation, which will be the foundation for the logical inference employed in this paper.

develop any counterfactual theory of causation. Perhaps he viewed this sentence merely as amplifying the previous sentence, and did not in fact realize that later thinking would distinguish a truly separate condition, the truth status of which must be evaluated and analyzed for its implications related to causality. At any rate, Holland (1986) seems to agree that Hume is the first philosopher whose analysis of causation is strongly relevant to our "potential outcomes framework" and "Rubin's Causal Model," both concepts which will be defined and discussed in more depth shortly.

It was some time before other philosophers gave serious attention to counterfactuals, although John Stuart Mill did attempt to analyze the truth relationships among such statements (Menzies 2014). The first use of the actual term "counterfactual" was apparently by Nelson Goodman in 1947 (Wikipedia). However, Goodman's focus in that paper was primarily on determining what other statements needed to hold, and what relationships among the statements needed to exist, in order to ascertain the truth status of a counterfactual claim (Goodman 1947); thus, his subject was much more in the line of Mill's. David Lewis, by contrast, formulated the "best known and most thoroughly elaborated counterfactual theory of causation" (Menzies 2014). His (1973) paper presented his original theory, but he eventually made substantial revisions (2000) to address certain deficiencies that were revealed over the several decades following his initial paper. His theory encompasses both deterministic causation, in which causes are followed by their effects with certainty, and "chancy causation," in which the absence of the causal event *c* would have decreased the chance of the subsequent event *e* occurring.

6.3 Glossary and Acronyms

ATE – Average Treatment Effect – the estimated or actual treatment effect assuming the treatment is administered to the entire population as compared to if the treatment is withheld from the entire population

ATT – Average Treatment effect among Treated – the estimated or actual treatment effect for the subpopulation of units that actually receives (or were to receive) the treatment. This differs from ATE if it is the case that those individuals in the population who actually receive (or would receive) the treatment have a smaller or larger treatment effect than those individuals who do not (or would not) receive the treatment.

DR – Doubly Robust – a propensity score application method combining propensity score weights with an outcome regression model

IPTW – Inverse Probability of Treatment Weighting – a propensity score application method that weights treatment observations by the inverse of the propensity score, and weights control observations by the inverse of (1 – propensity score)

JTB – Justified True Belief – a traditional theory of knowledge within epistemology

PS – Propensity Scores

RCT – Randomized Controlled Trial

SITA – Strongly Ignorable Treatment Assignment – the assumption within propensity score methodologies that, conditional on covariates, a unit's response is independent of treatment assignment

SMD – Standardized Mean Difference – A measure used to assess the balance of a covariate between the treatment and control group

SUTVA – Stable Unit Treatment Value – the assumption within propensity score methodologies that a unit's treatment effect is independent of all other units' treatment assignments

X1 – Confounders, the category of covariates related to both treatment and outcome

X2 – The category of covariates related to outcome only, also called prognosticators

X3 – The category of covariates related to treatment only

X4 – The category of covariates related to neither treatment nor outcome

CURRICULUM VITA

NAME:	John Anthony Craycroft
ADDRESS:	University of Louisville School of Public Health and Information Sciences 485 E. Gray St Louisville, KY 40202
DOB:	Louisville, KY – December 26, 1975
INTERNET:	john.craycroft@louisville.edu john_craycroft@yahoo.com
EDUCATION & TRAINING:	B.S., Statistics The George Washington University 1994-1998
	B.A., Philosophy The George Washington University 1994-1998
	MBA Emory University 2002-2004
PROFESSIONAL MEMBERSHIPS:	American Statistical Association (ASA), member since 2013
PRESENTATIONS:	"Bringing Value: Market Share Analysis That Goes Deeper" Poster presentation, American Statistical Association Conference on Statistical Practice, San Diego, CA, February 18, 2016
	"Introduction to SAS Topics" Guest lecturer presentation to Biostatistical Methods I classes, September, 2014, and October, 2015
	"Effect of Sample Design on Precision of Estimates" To FedEx Strategic Market Analysis division, March 12, 2008

VOLUNTEER EXPERIENCE:	Inaugural Vice President University of Louisville Biostatistics Club 2015-16 academic year
	Little League, T-ball, and Soccer coach Spring 2014, 2015, 2016; and Fall 2015
	Corporate chairman FedEx Operation Feed 2005-2007, in support of the Memphis Food Bank (raised average of $165,000 per year, plus thousands of pounds of food donations each summer for the Food Bank)
	Teaching Assistant Decision Information Analysis, Goizueta Business School 2003-04 academic year
	Honors Council Goizueta Business School 2003-04 academic year
	Reader for the Kentucky Recording for the Blind and Dyslexic
	Junior Achievement teacher 1999
AWARDS & HONORS:	Revenue Sciences Achievement Award FedEx Services April, 2015
	Rising Star Teamwork Award FedEx Global Marketing Spring, 2014
	Multiple “Bravo Zulu” awards, general FedEx award for outstanding performance, service, collaboration, etc. 2004 – 2015
	Kullback Prize in Statistics The George Washington University Spring, 1998
HONOR SOCIETIES	Phi Beta Kappa Beta Gamma Sigma