8-2015

Evaluating the Effects of Standardized Patient Care Pathways on Clinical Outcomes

Anna V. Romanova
University of Tennessee - Knoxville, aromanov@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Anna V. Romanova entitled "Evaluating the Effects of Standardized Patient Care Pathways on Clinical Outcomes." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Business Administration.

Russell Zaretki, Major Professor

We have read this dissertation and recommend its acceptance:

Charles Noon, Randy Bradley, Bogdan Bichescu

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Evaluating the Effects of Standardized Patient Care Pathways on Clinical Outcomes

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Anna V. Romanova
August 2015

DEDICATION

To my Mom. You gave me the best education you could, and made so many sacrifices for this dream to come true. Thank you for everything. I would not be here without you.

Для Мамы. Ты дала мне самое хорошее образование, которое было возможно. Ты стольким пожертвовала на пути к этой мечте. Спасибо за все. Благодаря тебе я там, где мечта претворилась в реальность.

ACKNOWLEDGEMENTS

I would like to express my special appreciation and thanks to my advisor Dr. Russell Zaretzki, who has been a tremendous mentor for me. I would like to thank Dr. Zareztki for encouraging my research and for allowing me to grow as a research scientist. His advice on both research as well as on my career have been extremely valuable. I would also like to thank my committee members, Dr. Charles Noon, Dr. Bogdan Bichescu, and Dr. Randy Bradley for serving as my committee members and for their brilliant comments and suggestions. I would especially like to thank Julia VanZyl, MD, and all the medical professionals that I have met at the University of Tennessee Medical Center for their support with data collection and keen insights. Their dedication to improving patient care is truly inspiring. A special thanks to my family and friends, who supported me in writing and incented me to strive towards my goal.

ABSTRACT

The main focus of this study is to create a standardized approach to evaluating the impact of the patient care pathways across all major disease categories and key outcome measures in a hospital setting when randomized clinical trials are not feasible. Toward this goal I identify statistical methods, control factors, and adjustments that can correct for potential confounding in observational studies. I investigate the efficiency of existing bias correction methods under varying conditions of imbalanced samples through a Monte Carlo simulation. The simulation results are then utilized in a case study for one of the largest primary diagnosis areas, chronic obstructive pulmonary disease (COPD) at the University of Tennessee Medical Center.

The analysis of the COPD pathway effects on the readmission rates showed a significant positive impact, with reduction in the probability of readmissions between 12% and 16%. The reduction in the length of stay was reported across all the models with historical controls, but the effect was not statistically significant.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The Affordable Care Act (ACA) of 2010 attempts to drive the health care system away from its current fee-for-service model towards a system that attempts to reimburse providers based on the quality of the outcomes that they achieve. In response to these changing demands, healthcare providers such as the University of Tennessee Medical Center (UTMC) are experimenting with new approaches to delivery of healthcare.

One such approach, standardized patient care pathways, more concisely "pathways", attempts to improve care delivery by standardizing treatment protocols for a wide range of hospital treated ailments. In addition to the development of pathways, organizations such as the Agency for Healthcare Research and Quality are promoting the careful measurement and evaluation of new programs.

The main goal of the current work is to create a standardized approach to assessing the impact of the UTMC pathways across all major disease categories and key outcome measures. Toward this goal I identify models, control factors, and adjustments that can correct for potential confounding in observational studies. I also address the issues of handling missing data. In addition, I implement the methodology and provide an actual analysis for one of the largest primary diagnoses areas, chronic obstructive pulmonary disease (COPD).

Since pathway assignment is not random and may be influenced by the patient's acuteness level and a physician's choice, commonly used statistical methods may result in biased estimates of pathways effects due to the presence of selection bias, or confounding between pathway assignment and clinical outcomes. To correct for potential selection bias, it is necessary to apply a different set of models that employ procedures to achieve data balancing before assessing treatment effects. These models originate from both biostatistics and econometric

studies of observational data and include the OLS regression with a treatment dummy variable, propensity score matching (PSM), the Abadie and Imbens non-parametric matching estimator (AI), the doubly robust matching estimator (DBR), and the Heckman treatment effect model (HE).

This study attempts to investigate the accuracy of the estimates for the average treatment effect (ATE) obtained through different corrective methods while taking into account the problem of imbalanced samples where the number of control units is much larger that the number of treated cases. The impact of sample imbalance is evaluated using a Monte Carlo simulation under three different data generation scenarios that emphasize the underlying model assumptions. The simulation results are then utilized in the analysis of the patient level data from the UTMC to estimate the ATE of pathways on the length of stay (LOS) and readmission rates.

The simulation study shows that the regression model and the PSM are expected to produce accurate estimates of the ATE, if all the confounders are included in the model. The choice of potential confounding variables was largely determined by the nature of the problem, review of existing medical and statistical studies, and data availability. The confounding factors that have been measured and included in the analysis are comorbidities, patient acuteness level on admission, severity of illness, and hospital congestion. It is also important to control for patient's age and gender, vital signs on admission, conditions and complications not present on admission, and payer information when estimating pathways effects on clinical outcomes.

While earnest efforts have been made to carefully measure and construct potential confounding variables, there still exists a possibility of the model misspecification. The best candidate to correct for an omitted variable bias, according to the simulation results, is the DBR estimator. Implementing the doubly robust method for estimating the ATE requires paying close

attention to the sample imbalance. Keeping the proportion of controls below 75% will ensure its accuracy and provide the necessary robustness against model misspecification.

This investigation is based on a rigorous analysis of the data and attempts to provide accurate and reliable estimates of pathway effects by carefully constructing treatment and control groups, taking into account confounding factors, and controlling for other variables that affect clinical outcomes. It can be extended to more general applications of hospital performance improvement, and provide benchmarks for future program evaluation.

The rest of this dissertation is organized as follows. In Chapter 1 I introduce the definition criteria for pathways and review the procedures for evaluating pathway effectiveness employed in randomized controlled trials and controlled before and after studies. I discuss the hierarchy of research designs in the medical field and present arguments in favor of the use of observational studies for evaluating treatment effects in the absence of randomization. I provide examples of observational studies in business and medical literature, and identify methods and techniques commonly used for modeling treatment effects and outcome variables in business and in clinical settings.

Chapter 2 focuses on statistical methodology for estimating average treatment effects in observational studies. I address the issues of overt and hidden bias, compare statistical methods for bias correction with econometric models, identify model assumptions, and describe the common pitfalls of currently existing methods.

Chapter 3 explores the effectiveness of several estimators of treatment effects through a Monte Carlo simulation under different settings for sample size and proportions of treated and controls. I employ three data generation scenarios that mirror the type of the selection bias and allow evaluating the performance of the estimators in the presence of overt and hidden bias.

Chapter 4 describes the estimation of the COPD pathway effects on the length of stay and 30-day readmission using regression adjustment, propensity score modeling, and the doubly robust estimator. It starts with a detailed description of the study design and the data, addresses the choices of confounders and control variables, and details the steps of the estimation process and balance assessment. Chapter 4 concludes with a discussion of the estimation results and their implications.

The concluding section of the dissertation contains a review of the key contributions and limitations of the study, identifies potential areas for improving the study design and outlines methodological issues for future research.

CHAPTER 1

LITERATURE REVIEW

The concept of clinical pathways (CPWs) first appeared in 1985 in the New England Medical Center (Boston, MA) following the introduction of the diagnosis related groups (DRG) system in 1983. In 2003, more than 80% of hospitals in the US used clinical pathways for at least some of their interventions (Saint et al., 2003), and their numbers are growing rapidly. The driving forces behind the new approach to patient care delivery are the shift in decision making in hospitals from opinion-based to evidence-based and the current policy changes that aim at improving the quality of care and reducing the costs.

CPWs provide more than just general clinical guidelines for the treatment of specific health conditions. They translate the general recommendations of clinical guidelines into the local systems and detail the steps and time frames to address these recommendations.

Pathways development and implementation requires a significant amount of resources, yet the effectiveness of clinical pathways remains highly debatable. Individual studies have shown the results that are varied and contradictory due to the lack of a uniformly accepted definition of a clinical pathway, clinical variability, and methodological quality. The purpose of this literature review is to summarize the methods used to analyze the effect of CPWs on clinical and financial outcomes in existing studies, and to identify the common issues for the observational study design.

1.1 Clinical Pathways: An overview of Existing Studies

In 2010, the Cochrane Collaboration published an extensive review of studies evaluating the effects of clinical pathways on professional practice, clinical and financial outcomes in

healthcare institutions (Rotter et al., 2010). Out of 260 studies assessed, only 27 were included in the review, meeting both the definition and methodological criteria as defined by Practice and Organization of Care (EPOC).

The Cochrane report uses the five criteria suggested by Kinsman et al. (2010) to determine whether an intervention constitutes a clinical pathway. These criteria are the following:

1. The intervention is a structured multidisciplinary plan of care.
2. The intervention is used to channel the translation of guidelines or evidence into local structures.
3. The intervention detailed the steps in a course of treatment or care in a plan, pathway, algorithm, guideline, protocol or other 'inventory of actions'.
4. The intervention had timeframes or criteria-based progression (that is, steps were taken if designated criteria were met).
5. The intervention aimed to standardize care for a specific clinical problem, procedure or episode of healthcare in a specific population.

The majority of the studies included in the review (19 out of 27) were randomized controlled trials (RCTs), two studies were controlled clinical trials (CCTs) with quasi-random allocation, another two used interrupted time series analysis (ITSs), and four were controlled before and after studies (CBAs). Most of the studies that met the definition criteria but were excluded from the review were simple before and after studies characterized by a high risk of bias due to the lack of control.

RCTs are controlled experiments in which participants are randomly allocated between the treatment and control groups. CCTs are studies where the allocation process is quasi-random

(e.g. based on alternation, date of birth, patient ID). CBAs involve nonrandom treatment assignment and include a baseline period for outcome assessment. To meet the minimum requirements for inclusion of CBAs in EPOC reviews, the studies must be based on the same pre and post intervention periods for treatment and control groups and have a minimum of two comparable intervention and two control sites. ITSs estimate the change in trend for a dependent variable by breaking its time series into pre and post intervention periods and comparing the means of a dependent variable in two periods. At least three data points before and three after the intervention are necessary to meet the EPOC methodological requirements[1].

The outcome measures reported in the studies of CPWs can be divided into the following groups:

1) patient outcomes (e.g. length of stay (LOS), mortality rate, readmissions, hospital acquired complications, adverse events, ICU admissions, and discharge destination),
2) professional practice outcomes (e.g. staff satisfaction, adherence to evidence based practice, quantity and quality of documentation, and pathway specific quality measures such as time to mobilization post surgery),
3) financial outcomes (e.g. hospital costs, hospital charges, and resource utilization measures).

The most commonly reported outcome is the LOS measured in hours or days, with the majority of studies showing a significant positive effect[2]. 12 out of 15 studies in the Cochrane

[1] The ITSs are still vulnerable to several important validity threats, one of them being *history*, or the possibility of confounding between the intervention and other events around the time of the intervention. To correct for plausible bias in the estimates of the observed effect, it is recommended to use control series that were not subject to the intervention.

[2] Length of stay reflects hospital practices with respect to hospitalization, and as such may not always reflect a positive outcome. In some instances, an increased LOS may indicate better care (e.g. when mortality decreases).

report that examined the effect of CPWs on LOS showed significant reductions in LOS (Delaney, 2003; Dowsey, 1999; Gomez, 1996; Smith, 2004, and others). A reverse effect, or an increased LOS, was associated with a CPW for stroke rehabilitation but did not reach statistical significance as reported by Falconer (1993) and Sulch (2000). Studies carried out in the US, where hospital LOS is historically lower, reported smaller decreases in LOS (weighted mean difference ,WMD of -0.8 days) compared to Australian (WMD of -1.6 days) and Japanese studies (WMD of -3.1 days). The report also points out that invasive conditions showed slightly stronger effect of CPWs on LOS (WMD of -1.4 days versus -1.1 for noninvasive conditions), which is consistent with health economic theories according to which invasive treatments have lower treatment variance and are more easily standardized than noninvasive procedures (Shluechtermann, 2005).

Several other key findings of the Cochrane review are worth noting. One of them is a significant reduction in hospital acquired complications associated with CPWs. For patients recovering from surgery and managed on a CPW, the pooled result of an absolute risk reduction for the studies included in the review was 5.6%, which corresponds to a prevention of one complication for every 17 patients. Another important conclusion of the Cochrane review is CPWs contribution to improved documentation, which was achieved without negatively impacting LOS and hospital costs. However, the effects of CPWs on readmission and mortality, according to the Cochrane review, were not statistically significant.

Rotter et al. (2010) point out that more evidence is needed to provide insights about the key elements of CPWs and the mechanisms through which CPWs affect economic and patient

---

LOS can be considered as a quality indicator only when other patient outcomes are taken into account to avoid misleading conclusions.

outcomes. They recommend that future systematic reviews group studies by pathway condition in order to reduce clinical and statistical heterogeneity and to provide reliable conclusions.

1.2. Relative Importance of RCTs and Observational Studies

While randomized trials are considered to be the gold standard for identifying the effects of an intervention, only four of the RCTs included in the Cochrane review were assessed to have a low risk of bias (Bauer 2006; Cole 2002; Kollef 1997; Marelich 2000). Other studies had a moderate risk of bias with exception of one low risk CBA (Smith, 2004). Sources of bias that may exist in RCTs include concealment of allocation, blinded assessment of outcomes, incomplete outcome data, selective reporting, and contamination of the control professionals.

Allocation concealment is a procedure used to ensure random treatment assignment in an RCT setting. Standard methods of allocation concealment include sequentially numbered, opaque, sealed envelopes (SNOSE), sequentially numbered containers, pharmacy controlled randomization, and central randomization (Piaggio et al., 2006). In practice, allocation concealment mechanisms may not always be effective. Clinical investigators in RCTs often find it hard to maintain impartiality in taking care of individual patients and interfere into the random treatment assignment process. Treatment allocation may become evident to investigators or patients due to treatment related side-effects thereby introducing bias or influencing any subjective parameters collected by investigators or requested from subjects. Even though it is recommended that allocation concealment methods be described in detail and included in an RCT protocol, most RCTs have unclear allocation concealment in their protocols and in their publications (Pildal et al., 2005).

Blinding is a set of "procedures that prevent study participants, caregivers, or outcome assessors from knowing which intervention was received" (Wood et al., 2008). Unlike allocation concealment, blinding can be inappropriate or impossible to perform in an RCT. If an RCT requires patient's active participation in a treatment such as physical therapy, participants cannot be blinded to the intervention. Another example is an RCT that involves the use of CPWs, where caregivers are the active participants and cannot be blinded. RCTs without blinding tend to be biased toward beneficial effects if the RCTs' outcomes were subjective as opposed to objective (Wood et al., 2008). Noseworthy et al. (1994) showed that in an RCT of treatments for multiple sclerosis, unblinded neurologists felt that the treatments were beneficial, while blinded neurologists did not.

RCTs require significant amount of time and resources. The conduct of an RCT takes several years until being published, which restricts the medical community from new knowledge and may be of less relevance at the time of publication. In 2006, Johnston et al. investigated the public return on investment in medical research by evaluating the effects of 28 RCTs totaling $335 million in cost on medical care and health. Their analysis showed that only four (14%) of the RCTs resulted in cost savings to society, and only six trials (21%) resulted in measurable improvements in health.

In 2011, Kessler et al. proposed to speed translation of healthcare research into practice by going as far as suggesting a moratorium on RCTs for the next decade to allow for a the shift in current research paradigms. They advocate the need for "pragmatic, transparent, contextual, and multilevel designs that include replication, rapid learning systems and networks, mixed methods, and simulation and economic analyses to produce actionable, generalizable findings that can be implemented in real-world settings". Peek et al. (2014) emphasize that medical

research often fails to find its way into practice or policy in a timely manner, and propose “the 5 R’s” as a new emerging standard for research in the medical field. The 5 R’s stand for the research that is relevant to stakeholders, rapid and recursive in application, redefines rigor, reports on resources required and is replicable. The approach proposed by Peek et al., largely motivated by the recent policy changes, is an attempt to address the needs of the Triple Aim by improving care and health outcomes and reducing cost.

Concato et al. (2000) cast doubt on the idea that RCTs' results are “evidence of the highest grade" and point out that in 99 reports evaluated “the average results of the observational studies were remarkably similar to those of the randomized, controlled trials.” They find substantial variation in the results of RTCs, and argue that observational studies are less prone to heterogeneity in results due to broader representation of the population and fewer opportunities for differences in management of subjects. Specific inclusion and exclusion criteria for coexisting illnesses and severity of disease in RCTs may result in creating a distinct group of patients, whose treatment protocol may not be representative of clinical practice. Benson et al. (2000) reach similar conclusions based on their analysis of 136 reports about 19 diverse treatments. Kessler et al. (2011) also point out that intensive interventions delivered by world-class experts in leading medical centers administered to a very specific patient population cannot be expected to work equally well in other public health settings.

Vandenbroucke (2008) suggests another line of reasoning that questions RCTs' contribution to scientific knowledge beyond other types of studies. He argues that if study designs are ranked by their potential for new discoveries, then anecdotal evidence would be at the top of the list, followed by observational studies, followed by RCTs. Glasziou et al. (2007) investigate treatments with dramatic and rapid effects and come to the conclusion that RCTs may

be unnecessary for these types of treatments. Einhorn (2002) draws an example of such treatment from a 1974 nonrandomized study where combination chemotherapy including cisplatin for metastatic testicular cancer increased the cure rate from 5% to 60%.

Observational studies can be designed with enough rigor to approximate randomization conditions by adopting the principles of experimental design (Concato et al., 2000). The "restricted cohort" design (Horwitz et al., 1990) identifies a "zero time" for determining patient's eligibility and base-line features, uses inclusion and exclusion criteria similar to those in clinical trials, adjusts for differences in base-line susceptibility to the outcome, and uses statistical methods such as intention-to-treat analysis similar to those in RCTSs. Use of appropriate statistical methodology for causal inference and careful study design are the key aspects in strengthening an observational study and reducing potential bias in the absence of randomization.

1.3. Treatment Effect Studies in Nonmedical Fields.

Data from other scientific disciplines doesn't support the hierarchy of research designs that currently exists in the medical field. In a comprehensive review of 302 meta-analyses Lipsey et al. (1993) compared the results of RCTs and observational studies of various psychological, educational, and behavioral treatments. Using a unit-free measure of the intervention effect to allow for comparisons across different topics and outcome variables, the authors were able to find evidence against the contention that observational designs consistently overestimate treatment effects as compared with RCTs. The numerous studies of treatment effects across various disciplines suggest that existing bias correction methods can provide reliable results in

the absence of randomization, but also raise a lot of questions about the choice of methodology and study design.

The models for bias correction come from two strands of literature that differ conceptually in the underlying assumptions about the selection mechanism. The statistical tradition assumes that the treatment assignment is exogenous and random conditional on specified covariates. This assumption is referred to, interchangeably, as unconfoundedness (Rosenbaum and Rubin, 1983), selection on observables (Barnow, Cain, and Goldberger, 1980), and conditional independence (Lechner, 1999), and implies that that treatment assignment is independent of the potential outcome if all covariates are observed and held constant. In contrast, econometricians often model treatment selection as a nonrandom choice and then use the conditional probability of receiving treatment to control for selection bias in the outcome analysis and therefore do not require the selection on observables assumption.

Tucker (2010), in her meta-analysis of selection bias studies in accounting and finance, discusses the use of the propensity score matching (PSM) and the Heckman treatment effect model (HE). The PSM addresses selection bias on observables, while the HE is only appropriate in situations when the selection bias is due to unobservables. In business, examples of observable differences are firm size and growth. Unobservable differences arise when researchers do not have access to information that is available to managers and market participants (e.g. information revealed by a financial audit but not accessible by some market participants). Tucker points out the importance of understanding the generating process of the non-experimental data in deciding between the two settings and choosing the correct methodology. When unobservables are not the primary concern, it is still necessary to check the covariates' balance and the sensitivity of findings to the effects of unobservables. Only one study out of 17 included in the meta-analysis

reports on common support, and none of them include the sensitivity analysis. The author also draws attention to several studies where the HE method is misused due to its lacking robustness to model specifications. It requires strong assumptions for the outcome regression (must be linear), the selection equation (must be modeled as probit) and the error terms (must follow bivariate normal distribution), and she finds that many of these assumptions are violated by the researchers.

Clatworthy, Makepeace, and Peel (2009) examine the limitations of the HE model using a sample of 36,636 UK private companies to estimate the large auditor (Big Four) premium, and arrive to the conclusion that the HE estimates are highly sensitive to changes in sample and model specification, particularly to the omission of a key identifying variable. The authors also refer to three different studies of UK private companies that have produced three different sets of results analyzing the impact of unobservable variables on premiums using the HE method, but reported similar findings using standard models. The results obtained from the PSM and portfolio matching by Clatworthy, Makepeace, and Peel were consistent with the majority of previous studies.

Dehejia and Wahba (2002) show that the PSM method can yield accurate estimates of the treatment effect when the treated group differs substantially from the potential pool of controls. They use Lalonde's (1986) dataset and compare the obtained estimates of the treatment effect to the benchmark results from the experiment. The authors address the three important issues in implementing matching based on propensity scores: whether or not to match with replacement, how many matches to use for each treated unit, and which matching method to use. They demonstrate that when there is a sufficient overlap in the distribution of the propensity scores between treated and controls, most of the matching methods will produce similar results.

Matching with replacement minimizes the distance between the matched pairs, and is beneficial in terms of bias reduction. Matching with replacement is a also better alternative when there are very few relevant comparison units. Matching without replacement increases bias and can produce results that are sensitive to the order in which the matches are done, but can improve the precision of the estimates. 1-to-1 matching also produces the smallest distance between the matched pairs, while matching one-to-many improves the precision of the estimates, but at the cost of increased bias.

Shadish , Clark and Steiner (2008) conducted a randomized experiment comparing random and nonrandom treatment assignments. To avoid confounding assignment method with other study features, they randomly assigned participants to be in a randomized experiment or nonrandomized experiment. Participants of the randomized experiment we randomly assigned to mathematics or vocabulary training, and the participants of a nonrandomized experiment chose which training they wanted. All participants were treated identically and attended the same training sessions. After training, all participants were assessed on both mathematics and vocabulary outcomes. The study showed that regression adjustment in the nonrandomized experiment reduced the estimated bias by 84-94%, and PSM method reduced bias by 58-96% depending on the outcome measure and adjustment method. The authors mention that the methods may have worked well in part because of a very rich set of covariates, well measured and related to both selection process and the outcome measures. PSM adjustments performed poorly when the scores were constructed from predictors of convenience (sex, age, marital status, and ethnicity) rather than from a broader set of covariates. They emphasize that a lack of covariate richness may reduce the accuracy of adjustments.

Another important finding by Shadish et al. is the sensitivity of the PSM adjustments to variations in how the propensity scores are constructed, particularly to which balance criteria are used. They point out that significance testing (Rosenbaum and Rubin, 1984) may confuse successful balance with low power, and recommend using the size of the imbalance proposed by Rubin in 2001 as a more desirable criterion. They also report that estimation results were sensitive to how missing data in the predictors were managed.

In some situations, when only few covariates are available, simple matching methods might be useful in detecting treatment effects. Barber and Lyon (1996) provide an example of a matching study that is focused on detecting firms with abnormal operating performance following major corporate events or decisions. They create a model of expected operating performance by matching firms on industry (using two- or four-digit SIC code), size within industry, and pre-event performance. Shafer and Moeller (2012) investigate the impact of adopting Six Sigma on corporate performance by comparing Six Sigma firms to overall industry performance benchmarks and to the performance of a portfolio of control firms. The matching is done based on industry and past performance. Zhao (2004) compares PSM with covariate matching estimators and indicates that PSM methods usually have a small bias but a large standard error compared to covariate matching methods. His simulation results indicate that matching without replacement produces a larger bias and a smaller standard error than matching with replacement, and show that PSM methods work best with large sample sizes and when the correlations between covariates and the treatment indicator variable are high.

1.4 Studies Evaluating Treatment Effects in Medical Literature

Austin and Mamdani (2006) provide a detailed description of several PSM methods for estimating the effectiveness of a medical treatment, in particular the use of statins post-AMI. Their study is largely motivated by the growing interest in using observational data to evaluate the impact of medical treatments or interventions on clinical outcomes with no consensus as to which propensity score method is preferable. They carry out a detailed propensity score analysis and discuss the most commonly used PSM methods including propensity score matching, stratification, covariate adjustment, and weighting using the propensity score. To assess the differences in characteristics between the treated and control patients, they follow Normand et al. (2001) and use the standardized differences in the means of covariates as a tool for balance assessment. They emphasize the importance of checking the balance before and after matching, as well as the need for a structured approach to the construction of the propensity score model described by Rosenbaum and Rubin. In their example, matching on propensity scores achieves greater balance than stratifying on the quintiles of the propensity scores, but at the cost of a reduced sample size, since many treated patients did not have an appropriate match among the controls. They use a greedy-matching algorithm with a caliper width of 0.2 standard deviations of the logit of the estimated propensity score, but do not provide the reasoning behind this choice of the matching algorithm and other specifics.

Residual imbalance in the propensity scores between groups within stratum or within a matched sample could indicate residual imbalance in measured covariates. Austin and Mamdani suggest using QQ plots to assess this imbalance and argue that QQ plots might be more sensitive to detecting residual imbalance between treated and controls than box plots traditionally used for this purpose.

These authors also refer to the classic tradeoff between variance and bias when choosing between matching and stratifying analysis. Stratification uses the entire sample but may results in greater bias dues to residual confounding within stratum. Matching uses a smaller subset of patients, thus diminishing the precision of the estimated ATE. Austin and Mamdani' empirical findings are consistent with this idea. They show that the estimated effects were attenuated in the matched analysis relative to the stratified analysis due to the differences between the matched sample and the overall sample.

Once a matched sample is obtained, there are several approaches to estimating the ATE. The two considered in Austin and Mamdani's study are matching on the propensity score and covariate adjustment using the propensity score. In case of matching on propensity score, the authors fit a logistic regression model with an intercept and a treatment dummy variable to estimate the effect of statin therapy on mortality. Covariate adjustment translates into fitting multivariate logistic regression models that include a combination of the estimated propensity scores, the treatment variable, and other relevant patient characteristics as covariates. A limitation of the covariate adjustment is the requirement for the regression model to be specified correctly and, according to Austin and Mamdani, researchers rarely examine the fit of their model compared to more complex models in practice.

The authors draw attention to the differences between the absolute and relative treatment effects and the adjusted model-based estimates, and advocate the use of the former. The logic behind the absolute and relative treatment effect estimates is nested in the theory of counterfactuals and the ability of the propensity score methods to replicate the design of an RCT given that the propensity score model is specified correctly. RCTs allow for direct calculation of

the ATE. The direct calculation of both the absolute and relative ATE can be described mathematically as

$$ATE_{Abs} = \frac{1}{n_0}\sum_{i=1}^{n_0} Y_{0i} - \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i} \quad (1.4.1)$$

$$ATE_{Rt} = \frac{(\frac{1}{n_0})\sum_{i=1}^{n_0} Y_{0i}}{(\frac{1}{n_1})\sum_{i=1}^{n_1} Y_{1i}} \quad (1.4.2)$$

where Y_{0i} is the observed outcome for the ith control case, Y_{1i} is the observed outcome for the ith treated case in the matched sample, and n_0 and $n_1$are the number of treated and control cases in the matched sample, respectively.

Austin and Mamdani arrive at the conclusion that in their data set matching on propensity scores was optimal as it was based on a matched sample with no imbalance in measured covariates and resulted in an estimate closest to the one obtained from a meta-analysis of statin RCTs. However, they caution against using the results of meta-analysis as a gold standard for measuring the performance of the competing propensity score methods and advise paying close attention to potential inherent differences in observational study designs.

In another recent study, Khwaja et al. (2011) estimate the effects of catheterization on patients with AMI on their 1-year mortality outcome and adopt a novel approach in the absence of the true treatment effect. They develop a dynamic structural model of hospital and treatment choice and the consequent mortality outcomes, and then use the estimated model to simulate data with known treatment effects.

The data for the dynamic structural model come from the Cooperative Cardiovascular Project (CCP). It consists of randomly selected Medicare patient records for patients admitted to non-federal acute-care hospitals in the USA with a principal diagnosis of AMI over an 8-month period and combines them with detailed clinical chart records for each patient. The authors add

hospital variables pertaining to availability of facilities for cardiac catheterization, angioplasty, and open-heart surgery to the data set and calculate annual hospital heart surgery volume. Excluding certain categories of patients that have not met the requirements of the study design leaves a sample of 114,818 patients. The covariates available for the analysis of the effect of catheterization on 1-year mortality are demographic characteristics, hospital characteristics and detailed patient characteristics that include Charlson comorbidity index, Killip class (a classification method for assessing the likelihood of congestive heart failure), and blockage status. The authors argue that including detailed patient characteristics captures the severity of illness, which is the primary factor for determining treatment for AMI patients. Failure to condition on severity of illness leads to biased estimates of the treatment effect on mortality outcome. The benefits of including clinical data in the analysis is one of the key findings demonstrated in Khwaja et al., and has important implications for observational studies. The study shows that the bias correction estimators perform well when measures of heterogeneity such as clinical variables are included in the regression analysis, but when the data are poor in such measurements, the bias becomes evident. The flexible logit estimator is closest to the true ATE, when clinical data are included in the model, followed by the fully interacted OLS, nearest-neighbor propensity score matching and OLS matching estimator. The HE estimators are not stable showing a lot of variation, with the fully interacted HE's performance ranked the lowest among all the matching estimators.

Another study by Austin et al. (2005) also focuses on the use of detailed clinical data for estimating treatment effects in observational studies. They examine several PSM models constructed using only administrative data and compared their performance with the models built using only clinical data. According to Rosenbaum and Rubin's theory, the propensity score is

only designed to balance measured covariates, and the authors' empirical findings are consistent with that theory. They show that propensity scores developed using administrative data do not necessarily balance unmeasured clinical confounders.

In addition to investigating the impact of different data sources, Austin et al. raise an important question about the use of previously validated risk-adjusted outcome models for evaluating the treatment-outcome relationship. In their study, the effectiveness of treatment was magnified when only variables from a previously validated model were used with clinical data compared to when all the variables of the PSM model were included in the analysis. However, fitting these two models to the administrative data produced similar estimates of the ATE, which can be explained by a relative scarcity of information contained in administrative data. While they point out the differences in the obtained estimates of the ATE, the discussion about the accuracy of the estimates remains open.

Identifying and measuring confounding factors is a necessary condition for obtaining unbiased estimates of the ATE, and while researchers are making every attempt to account for confounding factors, they are often limited by the data availability. Stephen and Berger (2003) evaluated the effects of an accelerated clinical pathway after elective colon resection on the LOS, readmissions, and costs per patient. Even though clinical charts were available for all the patients included in the study, only the data on age, gender, indication for operation, type of operation, and postoperative complications were collected. The authors used regression analysis with adjustments for age, sex, diagnosis and type of operation to estimate CPW effect of the LOS. The estimation methods for readmissions and costs were not clearly specified. The authors note that one surgeon operated on all the patients included in the study over the period of two and a

half years, and that an increase in surgical experience may be confounded with the number of postoperative complications.

Statistical methods for modeling LOS and other outcome variables are often ill-specified, which creates difficulties in assessing the validity of the results and compromises a study's replicability. Lin et al. (2011) investigate the effectiveness of CPW in coronary artery bypass surgery by estimating the differences in the LOS, postoperative complications, and costs. They collect data on patient demographic characteristics, patient surgical risk indicator (EuroSCORE), and surgery specific characteristics such as operation procedure and the number of surgical graft bypass. The authors note that mean comparisons, multiple regression analyses, and logistic regressions were performed for evaluation purposes, but do not report the estimation details. The validity and applicability of the study results remain unclear.

Kelly et al. (2013) provide a better example of modeling LOS in their study of patients who had radical prostatectomy. They use logistic regression to identify factors that predict prolonged LOS (LOS>9), and specify three variable groups to consider for inclusion in the model. The population-based, cancer registration data used in the study contain a variety of clinical characteristics for the patients as well as hospital volume and surgeon volume. Clinicians frequently work in both private and public sectors, and the availability of national cancer registry allowed for an accurate measure of the entire volume of cases for each surgeon, which is one of the major strengths of the study by Kelly et al. The authors report marital status, number of comorbidities, disease stage, hospital volume and surgeon volume to be significant predictors of a prolonged LOS in prostate cancer patients undergoing radical prostatectomy in public hospitals.

Smith et al. (2004) investigated the effects of a COPD CPW on readmission and mortality rates in a CBA study that was included in the Cochrane review (Rotter et al., 2010). Four public teaching hospitals in South Australia were included in the study, two of which were assigned as control hospitals and two as intervention hospitals. Eligible subjects had a principal diagnosis of COPD based on daily hospital admission records. 92% of these patients had COPD reported as their principal or secondary diagnosis at discharge. The researchers identified a preintervention phase (May to November 1998) with 721 COPD admissions, and a postintervention phase (late November 1998 to June 1999) with 509 admissions. The study design allowed for a comparison of subjects in control and treated groups, adjusted for pre-intervention differences between the two groups of hospitals. The authors use regression adjustment method to identify the effects of PW on the outcome variables with age, gender, type of admission (emergency/elective) and the number of comorbidities (defined as the number of ICD-10 secondary discharge diagnoses) as potential confounders. Smith et al. report a significant increase in elective readmission rates for the intervention group, and in emergency readmissions for the control group, which may be indicative of a transformative effect of the CPW on health care delivery, changing it from reactive to a more proactive management of the COPD. The decrease in mortality rates for the intervention group was not significant at $p \leq 0.05$. The study indicates no changes in mortality rates for the control group, as well as no changes in the LOS for both groups. The authors also mention that female gender and number of comorbidities were associated with an increase in the LOS, without providing estimation details.

The authors point out several factors that could have potentially contributed to diluting CPW effects. One of them is the difficulty in diagnosing the COPD correctly on admission, which results in mismatched placement of COPD CPW to suitable COPD patients. Another

factor is the limited knowledge of the precise CPW implementation process. Only 76% of CPWs had some evidence of use, according to entry and tick boxes on the guideline sheets. Yet staff practice may have been influenced by COPD CPWs through care of other COPD patients, which could have demonstrated in the lack of difference between intention-to-treat and per-protocol analysis.

Doherty and Jones (2006) offer some insights into CPW implementation process, and demonstrate significant improvements in compliance with evidence-based care in a CBA study of asthma patients in small rural district hospitals in Australia. Using a cluster design, they matched 8 hospitals pair wise based on RRMA rating and hospital size, and allocated one hospital in each pair to the experimental group. 98 patients were allocated to intervention hospitals, and 89 to control hospitals. There were no baseline differences in asthma severity between the groups.

Assessment of severity is crucial in management of asthma patients because asthma guidelines have different treatment strategies for different degrees of severity. The study shows that documentation of severity improved by 54% with CPW usage, and the effect was statistically significant. Spirometry use (the preferred method for diagnosing and monitoring the progress of asthma) increased from 25% to 62% in the CPW hospitals with no change in the control group. The authors report statistically significant improvement in other targeted outcomes of CPW usage such as use of systemic steroids, inappropriate use of antibiotics and use of STAMP. However the applicability of these results in other settings may be limited due to the specifics of the study.

Marrie et al. (2000) performed an efficacy analysis of pneumonia CPW in a multicenter CCT that involved 20 hospitals, and demonstrated that there were no differences in patients'

quality of life and adverse clinical outcomes between groups as well as a significant reduction in resource utilization for patients who were admitted at CPW institutions. The median LOS in their study was reported to be lower at CPW institutions (5.0 days vs 6.7 days at control sites, $p = 0.01$). CPW patients also received 1.7 fewer days of intravenous antibiotic therapy ($p = .01$) and were more likely to be treated with a single class of antibiotic (67% vs 27%, $p < 0.001$). The exclusion restrictions in this experiment were quite severe. Patients with an immune deficiency, patients who experienced shock, and who required intubation or direct admission to ICU, patients with alcohol addiction and chronic renal failure, among others, were ineligible for the study. Another criticism and potential source of bias is the unit analysis error as the randomization unit (hospital) was different from the unit of analysis (patient).

Two RTCs included in the Cochrane review (Gomez et al., 1996 and Roberts et al., 1997) investigated AMI pathway effects on LOS and hospital costs. The focus of both studies was the accelerated diagnosis of patients with chest pain in the ED. Patients with low risk for AMI were identified as the most relevant group for cost/benefit analysis since their admission to a coronary care unit is not likely to be cost effective, while discharge may be unsafe, and only low risk patients were included in the trials. The control and treated groups were selected based on well defined inclusion and exclusion criteria, and no significant differences were present in the baseline characteristics of the groups in either study. Roberts et al. had slightly larger sample sizes: 82 CPW patients and 83 controls, while in the study by Gomez et al. both groups were of size 50. The results of the two RCTs are consistent, showing significant reduction in the LOS and costs for patients on accelerated diagnostic ED CPW, but neither one is completely free of bias. Participants and investigators were not blinded, and since the intent was to increase

efficiency, attending physicians for patients in the control group may have been biased toward ordering briefer, more economic evaluations, thus reducing true differences between groups.

1.5. Summary of Empirical Findings

While the existing bias correction methods in observational studies show promising results when contrasted to randomized experiments (Dehejia and Wahba, 2002; Shadish et al.,2007), some of their shortcomings, such as sensitivity to sample size and model specifications, become evident and require special consideration. The sensitivity of the Heckman treatment effect model in this regard appears to be more pronounced than that of OLS regression. The propensity score models are expected to replicate the design of a randomized experiment, but their estimation results may be influenced by how the propensity scores are constructed as well as by the method chosen for balance assessment. In addition, the classic tradeoff between bias and precision is unavoidable when matching either with or without propensity scores.

The appropriate study design is a key aspect in successful estimation of treatment effects when randomization is not feasible, and includes choosing the control and treated groups, using relevant inclusion and exclusion criteria, and identifying potential confounders and control variables. Several studies demonstrate that the use of clinical data for estimating treatment effects provides a richer set of covariates and may increase the accuracy of the estimates. A rich set of covariates in clinical settings often includes patients' demographic characteristics, vital signs and laboratory values, comorbidities, severity of illness indicators, and relevant hospital characteristics. Including all important covariates in the model is crucial when using the methods based on the selection on observables assumption.

CHAPTER 2
METHODOLOGY

An observational study attempts to draw inferences about the effects caused by a treatment or intervention when subjects are not randomly assigned to treatment as they would be in a randomized trial but instead are assigned treatment by other means. In these studies, subjects decide which, if any, of the treatment levels to receive, and as a result of this self-selection, the potential for an unmeasured confounding variable to impact the outcome cannot be ruled out. If such a confounding variable exists, there is a considerable possibility of badly biased estimates of treatment effects and invalid conclusions.

Two types of bias should be considered when estimating a model based on an observational study. The first type is overt bias. Rosenbaum (2002) defines overt bias as "one that can be seen in the data at hand", which means that it is related to observable or measured variables in a study. Overt bias can result from either omission of observable variables in the model or from the specification of an improper functional form for the relationship between observable variables and the outcome variable. In contrast, hidden bias is associated with the omission of unobservable variables (i.e. correlated omitted variables).

There are two long-standing traditions in econometrics and statistics that offer solutions for bias correction methods depending on the bias type, but no consensus exists as to which method is preferable and how different settings in observational studies affect the modeling choices for treatment effects. This study, in part, is an attempt to delineate the translation of the differences between the two existing schools of thought into applied research, and to address potential issues that require further clarification.

2.1 The Advantages of Proper Randomization

Randomized experiments, whereby the assignment to treatment and control groups is random, represent an objective and robust approach to estimating treatment effects. This random external treatment assignment mechanism ensures that treated and control groups are comparable, with respect to both measured and latent characteristics. As a result, randomization ensures that differences between treated and control groups (before and after treatment) are due to chance, tend to be small, and are, therefore, not confounded with the treatment assignment indicating that estimates will not be biased.

An ideal randomization procedure would maximize statistical power and minimize confounding and selection bias. However, no single randomization procedure meets those goals in every circumstance as many reviews of RCTs demonstrate (e.g., Rotter et al., 2010). Unfortunately, randomized experiments often raise complex, sometimes insurmountable, challenges when applied to both business and clinical scenarios, as such settings do not conform to the embedded assumptions of randomized assignment, and in many instances are not a feasible option due to huge costs both in terms of time and money as well as potential impacts on ongoing processes.

2.2 The Counterfactual Model Framework

The existing bias correction methods for observational studies where treatment assignment is not random are best described by employing the counterfactual model framework and Pearl's directed acyclic graphs (DAGs) (Pearl, 2000). The key assumption of the counterfactual model is that each individual in the population has a potential outcome under each treatment state, which can be observed by an individual's exposure to each state. However, at

any point in time only one treatment state and potential outcome can be observed. For example, in a study of the effect of seat belts on fatalities in automobile accidents, drivers who were wearing a seat belt at the time of an accident have a theoretical what-if probability of a fatal accident under the state "no seat belt", and drivers who did not have their seat belt on have a theoretical probability of a fatal accident under "seat belt" state. These theoretical potential outcomes are not actually observed and, hence, referred to as counterfactuals.

Let Y_i^1 and Y_i^0 denote potential outcomes for observation i under treatment and control, respectively. By necessity, a researcher must analyze an observed outcome variable Y_i. We can then define the observable outcome variable Y_i as

$$Y_i = W_i Y_{i1} + (1 - W_i) Y_{i0}, \tag{2.2.1}$$

where $W_i \in \{0,1\}$ is a treatment assignment dummy variable. Rearranging terms and expressing potential outcomes as deviations from their means, the equation for Y_i takes the following form:

$$\begin{aligned} Y_i &= E(Y_0) + W_i[E(Y_1) - E(Y_0)] + u_{i0} + W_i(u_{i1} - u_{i0}) \\ &= E(Y_0) + W_i E(\delta) + u_i, \end{aligned} \tag{2.2.2}$$

where $u_i = u_{i0} + W_i(u_{i1} - u_{i0})$, $u_{i0} = Y_{i0} - E(Y_0)$ and $u_{i1} = Y_{i1} - E(Y_1)$. For a consistent estimate of the true average treatment effect E(δ), W_i and u_i must be uncorrelated.

Heckman and Robb (1985) propose a supplemental equation, known as the assignment or selection equation, that determines W_i. The treatment selection is modeled by specifying a latent continuous variable W_i*:

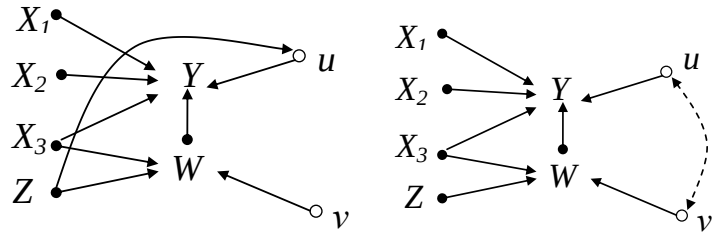
$$W_i^* = Z_i \alpha' + v_i, \tag{2.2.3}$$

where Z_i represents all observed variables that affect treatment assignment, α_i is a vector of coefficients, and v_i is an error term that captures both systematic unobserved factors that affect the assignment process and completely random determinants of treatment selection. The

treatment dummy variable W_i is determined by the following rule: $W_i = 1$ if $W_i^* > c$ and $W_i = 0$ if $W_i^* < c$ (c is an arbitrary cutoff value).

When all the systematic determinants of treatment selection have been observed as Z, treatment assignment is ignorable or, in other words, selection is on observables. In this case Z_i and u_i are correlated, but v_i and u_i are not. Under this condition OLS regression estimates of (2.2.2) will be unbiased if all the variables in Z are sufficiently included in the model. In panel (a) Figure 2.1 there is a back-door path $W \leftarrow Z \rightarrow u \rightarrow Y$, which means that the effect of W on Y is confounded by Z, (u being random noise).

When the observed variables in Z are only a subset of the factors that affect treatment selection, the unobserved components enter into the treatment selection latent variable W_i^* through the error term, v_i. Now Z_i and u_i are not correlated, but v_i and u_i are, and the condition is known as selection on unobservables, or a nonignorable treatment assignment. Under this setting, the OLS estimates from (2.2.2) will be biased and inconsistent. The bidirected edge between u and v in panel (b) Figure 2.1 indicates that u and v are mutually dependent on one or more unobserved common causes, and since u has a direct effect on Y, and v on W, but neither u, nor v can be measured, mutual dependence on a common unmeasured cause exists between W and Y.



(a) Selection on Observables (b) Selection on Unobservables

Figure 2.1 Selection mechanisms in statistics and econometrics

Selection on observables with an omitted variable shown in Figure 2.2 may look similar to selection on unobservables when mapped on a DAG, but the two mechanisms differ in underlying assumption about the error terms. Selection on observables assumes that the error terms will be uncorrelated once the omitted variable is included in the model, or, in other words, conditioning on *Z* would close the back-door path between *W* and *Y*. In selection on unobservables, the mutual cause of *u* and *v* cannot be measured, and including *Z* in the model does not change its error structure.

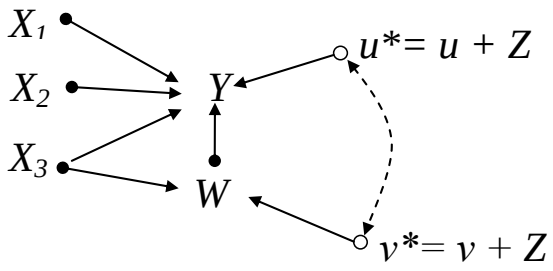


Figure 2.2 Selection on Observables with an Omitted Variable

2.3 Conditioning to balance versus conditioning to adjust

Treatment assignment models in statistics fall under two categories: matching techniques and regression implementations of conditioning. Matching techniques attempt to balance the determinants of the outcome variables, while regression models adjust for other causes of the outcome. Morgan and Winship (2007) demonstrate that both techniques can be considered variants of each other and explain the different ways in which they are used in applied research.

In a randomized experiment, treatment status is expected to be independent of all observed and unobserved variables that determine the outcome. In this case, the data are balanced with respect to *X* as shown in (2.3.1),

$$\Pr[X|W = 1] = \Pr[X|W = 0] \quad (2.3.1)$$

which requires that the probability distribution of W be the same within the treatment and control groups.

In Figure 2.3, a back-door path $Z \leftrightarrow X \rightarrow Y$ is present from W to Y, where Z represents a complete set of all observable variables that are direct causes of treatment assignment, and X represents a complete set of all observable variables other than W that are direct causes of Y. The bidirected edge between Z and X means that they are mutually caused by some set of common unobserved factors.

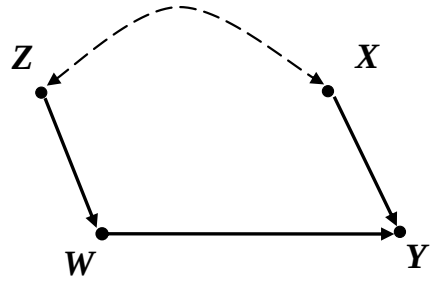


Figure 2.3. Conditioning on Z versus X

All back-door paths signified by the bidirected edge in Figure 2.3 can be blocked by conditioning on either Z or X because none of them is a collider[3], meaning that causal effects of other factors do not collide with each other at Z or X. Conditioning on Z is considered a balancing technique whereas conditioning on X is considered an adjustment-for-other-causes conditioning strategy. Conditioning on Z ensures that the variables in Z and W are no longer associated within subgroups defined by the conditioning. The treatment and control groups will

[3] Morgan and Winship note that the same conclusions will hold if Z and X include several variables within them such that some members of X cause W directly and some members of Z cause Y directly. However, there must be at least one variable in one set that causes W but not Y and at least on variable in the other that causes Y but not W.

be balanced with respect to *Z*. Alternatively, conditioning on *X* makes the resulting subgroup differences in *Y* across *W* within *X* attributable to *W* alone.

Though the distinction between balancing and adjustment for other causes may seem somewhat artificial, the distinction is important. There is an unobserved set of systematic causes that generates the relationship between *Z* and *X*, and conditioning on either *Z* or *X* is necessary to identify the treatment effect. In many applied research situations these two sets of variables may be quite different.

2.4 Regression Adjustment for Estimating Causal Effects

The OLS regression is considered an adjustment-for-other-causes conditioning strategy. We can identify the regression model as $E(Y|W,X)$ as it depends on observed data. We now rewrite the regression model as

$$E(Y|W,X) = \beta_0 + \beta_W W + X'\beta_x, \tag{2.4.1}$$

where, again, *W* is a binary variable indicating treatment condition ($W = 1$ if treated, and $W = 0$ otherwise). In general, $E(Y|W = 1, X)$ is the regression among treated, and $E(Y|W = 0, X)$ among controls.

Averaging over all possible values of X (both treatments) is equivalent to

$$E\{E(Y|W = 1, X)\} = E\{E(Y_1|W = 1, X)\} = E\{E(Y_1|X)\} = E(Y_1), \tag{2.4.2}$$

and, in the same way,

$$E\{E(Y|W = 0, X)\} = E(Y_0). \tag{2.4.3}$$

Then the average treatment effect (ATE) can be estimated as

$$E(Y_1) - E(Y_0) = E\{E(Y|W = 1, X)\} - E\{E(Y|W = 0, X)\} =$$
$$= E\{E(Y|W = 1, X) - E(Y|W = 0, X)\} \tag{2.4.4}$$

For a continuous outcome, the ATE can be estimated directly from fitting the OLS regression model. Suppose the true regression is

$$E(Y|W,X) = \beta_0 + \beta_W W + X'\beta_x. \quad (2.4.5)$$

Then

$$E(Y|W=1,X) - E(Y|W=0,X) = \beta_0 + \beta_W(1) + X'\beta_x - \beta_0 - \beta_W(0) - X'\beta_x = \beta_W. \quad (2.4.6)$$

Thus, if there are no unmeasured confounders, i.e. if the model is correctly specified, $\hat{\beta}_W$ is the unbiased estimate of the ATE.

For a binary outcome, if the true regression is

$$E(Y \Big| W,X) = \frac{\exp(\beta_0+\beta_W W+X'\beta_x)}{1+\exp(\beta_0+\beta_W W+X'\beta_x)}, \quad (2.4.6)$$

$$E(Y|W=1,X) - E(Y|W=0,X) = \frac{\exp(\beta_0+\beta_W+X'\beta_x)}{1+\exp(\beta_0+\beta_W+X'\beta_x)} - \frac{\exp(\beta_0+X'\beta_x)}{1+\exp(\beta_0+X'\beta_x)} \quad (2.4.7)$$

Logistic regression yields $\hat{\beta}_0, \hat{\beta}_W, \hat{\beta}_X$. The ATE is then estimated by averaging the differences in the predicted values of *Y* obtained using the estimated model under each state across all observed X_i:

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N} \frac{\exp(\hat{\beta}_0+\hat{\beta}_W+X'\hat{\beta}_x)}{1+\exp(\hat{\beta}_0+\hat{\beta}_W+X'\hat{\beta}_x)} - \frac{\exp(\hat{\beta}_0+X'\hat{\beta}_x)}{1+\exp(\hat{\beta}_0+X'\hat{\beta}_x)}. \quad (2.4.8)$$

2.5 Abadie and Imbens Bias-Corrected Matching Estimator (AI)

Matching is a data balancing technique that aims to achieve independence between the outcome and the treatment conditional on a set of covariates that are considered to be determinants of the treatment assignment. To overcome the dimensionality problem, matching estimators use a vector norm to calculate distances on the observed covariates between a treated case and each of its potential matches among the control units. The vector norm is used to select the outcome of a control case with the shortest distance on covariates as a counterfactual for the

treated case. Similarly, the matching estimators can choose the outcome of a treated case to serve as a counterfactual for the control case.

The vector norm is calculated based on either the inverse of the sample variance matrix or the inverse of the sample variance-covariance matrix. The latter approach is known as computing Mahalanobis distances, and the method gets the name of Mahalanobis metric matching (Rubin, 1973).

Mahalanobis metric matching randomly orders study participants and then calculates the Mahalanobis distances between the first treated unit and all controls (Guo, 2010):

$$d(i,j) = (U - V)^T C^{-1}(U - V), \quad (2.5.1)$$

where *U* and *V* are values of the matching variables for treated unit i and control unit *j*, and *C* is the sample covariance matrix of the matching variables from the full set of controls. The algorithm selects the control unit *j* that has the minimum distance *d(i, j)* as a match for the treated unit i. This process is repeated until matches are found for all treated cases.

When many covariates are included in the model, it may be difficult to find close matches because Mahalanobis metric matching is not based on a one-dimensional score. Another limitation of this method that arises from the curse of dimensionality problem is that the average Mahalanobis distance between observations always increases when the number of covariates becomes larger.

A simple matching estimator of the ATE uses the matched data set obtained though Mahalanobis metric matching to impute the missing potential outcomes by computing the average outcome for units with similar values on observed covariates. Matching is done with replacement, and the final inference of matching estimators may depend on the number of matches chosen for each unit. Abadie et al. (2004) recommend using 4 control matches for each

treated case as a rule of thumb to avoid incorporating bad quality matches and not to rely on too little information as would be the case with 1-to-1 matching.

Abadie and Imbens (2002) demonstrate that when the matching is not exact, the simple estimator is biased in finite samples. With *k* continuous covariates, the simple matching estimator will have a bias term that corresponds to the differences in covariates between treatment and control groups. They propose to use regression adjustment to correct for the bias that remains after matching and develop a bias-corrected matching estimator.

This estimator consists of two steps. First all units in both treated and control groups are matched with replacement, so that the results are not order dependent. After matching all units some of the remaining bias is removed through regression on a subset of the covariates, with the subvector denoted by Z_i.

In step 1, they create matches for all units. For each *i*, the set of indices for the closest match, *J(i)* can be defined as

$$J(i) = \{1, \dots, N | W_i \neq W_j, d(X_i, X_j) = min_{m:W_m \neq W_i} d(X_i, X_m)\}, \qquad (2.5.2)$$

where *m* is the number of matched per unit with replacement, the distance be based on the Mahalanobis metric $d(x,z) = (x-z)'\widehat{\Omega}_X^{-1}(x-z)$, and $\widehat{\Omega}_X = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})'$, with $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$.

Given the sets *J(i)*, Abadie and Imbens define $\hat{Y}_i(0)$, $\hat{Y}_i(1), \hat{X}_i(0),$ and $\hat{X}_i(1)$ as follows:

$$\hat{Y}_i(0) = \begin{cases} Y_i, & if\ W_i = 0, \\ \frac{1}{\#J(i)}\sum_{j \in J(i)} Y_j, & if\ W_i = 1, \end{cases} \qquad (2.5.3)$$

$$\hat{Y}_i(1) = \begin{cases} \frac{1}{\#J(i)}\sum_{j \in J(i)} Y_j, & if\ W_i = 0, \\ Y_i, & if\ W_i = 1, \end{cases} \qquad (2.5.4)$$

$$\hat{X}_i(0) = \begin{cases} X_i, & if\ W_i = 0, \\ \frac{1}{\#J(i)} \sum_{j \in J(i)} X_j, & if\ W_i = 1, \end{cases} \tag{2.5.5}$$

$$\hat{X}_i(1) = \begin{cases} \frac{1}{\#J(i)} \sum_{j \in J(i)} X_j, & if\ W_i = 0, \\ X_i, & if\ W_i = 1. \end{cases} \tag{2.5.6}$$

This leads to a matched sample, with N pairs, where each pair is characterized by a quintuple: $(\hat{Y}_i(0), \hat{Y}_i(1), \hat{X}_i(0), \hat{X}_i(1), W_i)$.

The simple matching estimator, $\hat{\tau}_{sm} = \frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_i(1) - \hat{Y}_i(0))$, has a bias of $O_p\left(N^{-\frac{1}{k}}\right)$, where *K* is the dimension of the covariates. To improve its properties the authors suggest using linear regression to adjust for biases associated with differences in covariate values.

In step 2, they run two OLS regressions on *N* units to obtain the least squares estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{Y}_i(0) = \alpha_c + \beta_c' \hat{X}_i(0) + \varepsilon_{ci}, \tag{2.5.7}$$

and

$$\hat{Y}_i(1) = \alpha_t + \beta_t' \hat{X}_i(1) + \varepsilon_{ti}. \tag{2.5.8}$$

Next they adjust the imputed potential outcomes as

$$\hat{Y}_i^{adj}(0) = \begin{cases} Y_i, & if\ W_i = 0, \\ \frac{1}{\#J(i)} \sum_{j \in J(i)} Y_j + \beta_c'(\hat{X}_i(1) - \hat{X}_i(0)), & if\ W_i = 1, \end{cases} \tag{2.5.9}$$

$$\hat{Y}_i^{adj}(1) = \begin{cases} \frac{1}{\#J(i)} \sum_{j \in J(i)} Y_j + \beta_t'(\hat{X}_i(0) - \hat{X}_i(1)), & if\ W_i = 0, \\ Y_i, & if\ W_i = 1. \end{cases} \tag{2.5.10}$$

The bias-adjusted estimator becomes

$$\hat{\tau}_{match}(Y, X, W) = \frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_i^{adj}(1) - \hat{Y}_i^{adj}(0)). \tag{2.5.11}$$

Abadie and Imbens point out that the proposed matching estimator does not eliminate all biases associated with differences in the covariates in large samples sufficiently fast to achieve root-N convergence. However, they claim that in practice, the linear regression adjustment eliminates a large part of the bias that remains after the simple matching.

2.6 The Propensity Score Model (PSM)

The main advantage of the propensity score methods over conventional matching is dimensionality reduction. The vector *X* may include many covariates, which makes finding a good match from a control group for a given treated unit challenging when using conventional matching methods. The propensity score solves the problem of dimensionality by reducing the dimension of *X* to a single one-dimensional score.

Rosenbaum and Rubin (1983) define a propensity score as a conditional probability of treatment assignment given a set of observed covariates:

$$e(X) = P(W = 1|X) = E\{I(W = 1)|X\} = E(W|X). \quad (2.6.1)$$

It is customary to estimate the probability of treatment, *e(X)*, by fitting a logistic regression model:

$$P(W = 1|X) = e(X, \alpha) = \frac{\exp(\alpha_0 + X'\alpha_x)}{1 + \exp(\alpha_0 + X'\alpha_x)}. \quad (2.6.2)$$

The predicted values from this logistic regression are known as propensity scores.

Under the ignorable treatment assignment assumption,

$$(Y_0, Y_1) \perp\!\!\!\perp W|e(X). \quad (2.6.3)$$

Given that the strong ignorable treatment assignment assumption holds and *e(X)* is a true balancing score, the expected difference in observed outcomes of the two treatment conditions at

e(X) is the ATE at *e(X)*. In terms of counterfactual framework, the following expression is an unbiased estimate of the ATE, conditional on the propensity score, *e(X)*:

$$E[E(Y_1|e(X), W = 1) - E(Y_0|e(X), W = 0)\,] = E[Y_1 - Y_0|e(X)] \quad (2.6.4)$$

Guo and Fraser (2010) describe propensity score modeling as a three-step sequenced analysis. Step 1 involves specification of a logistic regression model, step 2 is resampling to create a matched data set based on estimated propensity scores, and step 3 is postmatching analysis.

One of the key points in propensity score methods is finding the best model for estimating propensity scores when the true propensity scores are not known. When the propensity score model is misspecified, the ignorable treatment assignment assumption does not hold and the estimate of the ATE that is no longer unbiased. The best logistic regression should take into consideration covariate balance and meet the general requirements for the goodness of fit.

There is a strong emphasis in the literature on PSM methods regarding the importance of including carefully chosen and appropriate conditioning variables in their correct functional form in the model for estimating propensity scores. Smith and Todd (2005) argue that including more conditioning variables may exacerbate the common support region problem. Rosenbaum and Rubin (1984, 1985) recommend expanding the propensity score model to include high-order polynomial terms and interactions and use stepwise regression to select variables based on a Wald statistic and its associated p-value. Rosenbaum (2002) suggests that the logistic regression model should include all covariates for which group differences meet a low threshold for significance ($|t| > 1.5$). Eichler and Lechner (2002) propose to use a variant of Rosenbaum and Rubin's measure, which is based on standardized differences in means between treated and

controls for each variable in *X*, squares of each variable in *X*, and first-order interaction terms for each pair of variables in *X*. The standardized difference is defined as

$$d = \frac{100(\bar{x}_T - \bar{x}_C)}{\sqrt{(s^2{}_T - s^2{}_C)/2}} \text{ for continuous variables} \quad (2.6.5)$$

and by

$$d = \frac{100(p_T - p_C)}{\sqrt{[\,p_T(1-p_T) + p_C(1-p_C]/2}} \text{ for categorical variables.} \quad (2.6.6)$$

According to Normand et al. (2001), covariates with a standardized difference of greater than 10% are indicative of a meaningful imbalance between treatment groups.

Checking for covariate imbalance is necessary in a full data set when selecting the variables for the propensity score model, and in a matched data set, created using the estimated propensity scores. McCaffrey et al. (2004) propose an algorithm that minimizes the sample average standardized difference (ASAM) in the covariates by modifying the generalized boosting modeling criterion (GBM). Rosenbaum and Rubin (1984) define the search for the best propensity score model as a reiterative process of fitting a logistic regression model, matching, checking for data imbalance in a matched data set, and refitting a logistic regression model if the imbalance still exists.

Once the best logistic regression model is determined, the estimated propensity scores can be used to match treated units to controls, or, alternatively, as sampling weights in further analysis without matching. Matching treated and controls on propensity scores balances the data, but typically reduces the sample size because the common support region is now defined by the propensity scores and might not cover the whole range of observations in the original data set.

Available matching algorithms include nearest neighbor, kernel, and interval matching procedures. Nearest neighbor matching randomly orders treated units and then selects a matched

control unit j for a treated unit i based on the smallest absolute difference of propensity scores among all possible pairs:

$$C(P_i) = \min_j \left\| P_i - P_j \right\|, j \in I_0, \tag{2.6.7}$$

where $C(P_i)$ is the neighborhood of matched control units, P_i and P_j are the propensity scores for treated and control units, and I_0 is the set of control units. Nearest neighbor matching may results in poor matches for some treatment case. Another variant of this algorithm, nearest neighbor matching with a caliper, is designed to remedy the possibility of bad matches by restricting matches to some maximum distance, or a specified caliper width. Rosenbaum and Rubin (1985) recommend setting the caliper width at $0.25\sigma_p$, where σ_p is the standard deviation of the estimated propensity scores. The algorithm can be run with or without replacement. With replacement, each control case can be matched to more than one treated unit. Without replacement, a control unit is taken out of the pool of possible matches once it is matched. One of the weaknesses of matching without replacement is that the estimate of the ATE will vary depending on the initial ordering of the treatment cases.

Kernel matching, developed by Heckman et al. (1998), constructs the counterfactuals of each treatment unit using all control units weighted based on the their distance from the treated case. The weights, w_{ij} are calculated with a kernel function, $G(.)$ that transforms the distance between the target treatment case and all control cases in the data set. Using propensity scores to measure the distance, kernel matching estimators define the weight as

$$w_{ij} = \frac{G\left[\frac{\hat{p}(s_j) - \hat{p}(s_i)}{a_n}\right]}{\sum_j G\left[\frac{\hat{p}(s_j) - \hat{p}(s_i)}{a_n}\right]}, \tag{2.6.8}$$

where a_n is a bandwidth parameter that scales the difference in the estimated propensity scores based on the sample size and $\hat{p}(.)$ is the estimated propensity score as a function of its argument. The denominator is a scaling factor equal to the sum of all transformed distances across control cases, and is needed so that the sum of w_{ij} is equal to 1 when all control units are matched to each target treatment unit. The main criticism of kernel matching is its high probability of producing bad quality matches since all control cases are used in creating matches for each treated case.

Interval matching divides the treated and control cases into segments based on the estimated propensity scores, and then calculates the treatment effect within these intervals (Rosenbaum and Rubin, 1984). This method strongly resembles nearest neighbor caliper matching when each interval includes exactly one treated case. When there are several treated units in each interval, it does not allow for covariate adjustment in postmatching analysis and limits the researcher to mean comparisons of outcomes between groups.

The choice of a particular matching algorithm is not clear, and is likely to be application dependent. Morgan and Winship (2007) demonstrate that different matching algorithms and software routines yield different estimates of the ATE, and recommend further investigation of these issues in future research.

Assuming that propensity score matching yields balanced data and the ignorable treatment assignment assumption holds, researchers can undertake covariate adjustments for the matched sample to estimate the ATE. Any regression-type model may be used at this stage to estimate the ATE by using a dichotomous explanatory variable indicating treatment conditions. Morgan (2001) conducts a regression analysis to estimate the effect of Catholic schools on learning following caliper matching. Smith (1997) uses a hierarchical regression model to

estimate the effects on mortality of an organizational innovation within hospital after nearest neighbor matching.

Propensity scores can be also used without matching. This technique is known as adjustment by inverse weighting. Rather than using the difference of simple averages, $\bar{Y}_1 - \bar{Y}_0$, the ATE is estimated by the difference of inverse propensity score weighted averages (Rosenbaum, 1987):

$$\widehat{IPW} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i Y_i}{e(X_i,\hat{\alpha}_X)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)Y_i}{1-e(X_i,\hat{\alpha}_X)}. \quad (2.6.9)$$

Inverse probability weights are calculated as the inverse of the conditional probability of receiving the actual treatment: 1/PS for the treated and 1/(1-PS) for the controls. As such, inverse probability weighting creates a pseudopopulation in which the distributions of confounders among the treated and controls are the same. It eliminates an association between the confounders and treatment, so that the weighted averages reflect the averages in the true population (Funk et al., 2010).

By the law of large numbers, $\frac{1}{N}\sum_{i=1}^{N}\frac{W_i Y_i}{e(X_i,\hat{\alpha}_X)}$ should estimate the mean of a term in the sum with $\hat{\alpha}$ replaced by the quantity it estimates:

$$E\left(\frac{WY}{e(X)}\right) = E\left(\frac{WY_1}{e(X)}\right) = E\left[E\left\{\frac{WY_1}{e(X)}|Y_1, X\right\}\right] \quad (2.6.10)$$

since $WY = W\{Y_1 W + Y_0(1-W)\} = W^2 Y_1 + W(1-W)Y_0$ and $W^2 = W, W(1-W) = 0$ due to W being a binary variable. By ignorable treatment assignment assumption, $(Y_0, Y_1) \perp\!\!\!\perp W|e(X)$,

$$E\left[E\left\{\frac{WY_1}{e(X)}|Y_1, X\right\}\right] = E\left\{\frac{Y_1}{e(X)}E(W|Y_1, X)\right\} = E(\frac{Y_1}{e(X)}E(W|X)). \quad (2.6.11)$$

Then

$$E\left(\frac{WY}{e(X)}\right) = E(\frac{Y_1}{e(X)}E(W|X)) = E(\frac{Y_1}{e(X)}e(X)) = E(Y_1). \quad (2.6.12)$$

Similarly, $E\left(\frac{(1-W)Y}{1-e(X)}\right) = E(Y_0)$.

The estimate of ATE obtained by inverse weighting is not robust to the model misspecifications, and will only estimate the true ATE correctly if the postulated propensity score model is identical to the true propensity score.

2.7 The Doubly Robust Estimator (DBR)

The doubly robust estimator (DBR), first proposed by Robins et al. (1994) and reviewed by Davidian et al. (2010), is a relatively new method of estimating the ATE that combines the outcome regression model with propensity scores to estimate the ATE, and is designed to correct the bias that occurs when regression and/or propensity score models are misspecified. It can be viewed as augmenting the inverse weighted estimator. Following the notation in Funk et al. (2011), the formula for a doubly robust estimator can be expressed as:

$$\widehat{DBR} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i Y_i}{e(X_i,\hat{\alpha}_X)} m_1(X_i,\hat{\beta}_1) - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)Y_i}{1-e(X_i,\hat{\alpha}_X)} m_0(X_i,\hat{\beta}_0), \qquad (2.7.1)$$

where $m_1(.)$ and $m_0(.)$ are the predicted values from regression models on the baseline covariates for the treatment and control groups, respectively. The augmentation component is formed by taking the product of two bias terms—one from the propensity score model and one from the outcome regression model. If either bias term equals zero (as is the case when one of the models is correct), then it ‘‘zeros out’’ the other, nonzero bias term from the incorrect model. In other words, the DBR estimator will be unbiased if at least one of the models (regression or PSM) is specified correctly.

Emsley et al. (2008) outline the following steps for implementing the DBR estimator:

1. Fit a logistic (or probit) regression model for treatment conditional on the baseline variables (time-dependent variables can be included if required for longitudinal analysis). The predicted values from this regression give the estimated propensity scores, PS.
2. Fit a regression model for the outcome on the baseline variables for the treatment group only (W_i = 1), and obtain predicted values for the whole sample. This gives the value for $m_1(X_i)$.
3. Fit the same regression model for the outcome on the baseline variables for the control group only (W_i = 0), and obtain the predicted values for the whole sample. This gives the value for $m_0(X_i)$.
4. Substitute the predicted values of propensity scores,$m_1(X_i)$, and $m_0(X_i)$ into the expression for the double-robust estimator as defined by (2.7.1).

While the DBR estimator has been described in the statistical literature, it is not yet well known among the broader research community. Prior simulations have confirmed that the doubly robust estimator is unbiased when a confounder is omitted from one but not both of the component models (Bang and Robins, 2005; Lunceford and Davidian, 2004). As with any new method, caution is warranted. According to Lunceford and Davidian (2004), the DBR estimator is generally less efficient than the maximum likelihood estimator with a correctly specified model. Thus, when choosing the DBR estimator over regression or PSM methods, it is important to consider a trade-off between potentially reducing bias at the expense of precision.

2.8 Heckman Treatment Effect Model (HE)

Bias correction methods that address selection on unobservables were first developed by Lee (1978) and Heckman (1979). Heckman proposes a two-stage approach for evaluating the

effects of programs with binary treatment choices where the outcomes depend on a linear combination of observable and unobservable covariates. The Heckman treatment effect model addresses the bias due to selection on unobservables by estimating a bias correction term in the first stage through the choice model and adding it in the second-stage outcome regression.

The estimate of the ATE can be calculated by comparing the average difference in outcomes of treated and controls as defined in (2.8.1) and (2.8.2):

$$E(Y_{1i}|W_i = 1) = \beta_0^1 + X_i^{'}\beta + E(u_{1i}|W_i = 1) = \beta_0^1 + X_i^{'}\beta + E(u_{1i}|v_i > -Z_i\alpha) \quad (2.8.1)$$

and

$$E(Y_{0i}|W_i = 0) = \beta_0^0 + X_i^{'}\beta + E(u_{0i}|W_i = 0) = \beta_0^0 + X_i^{'}\beta + E(u_{0i}|v_i \leq -Z_i\alpha). \quad (2.8.2)$$

Assuming that (u_1, v) and (u_0, v) are binormally distributed with zero means and variances $\sigma_{u_1 v}$ and $\sigma_{u_0 v}$ (σ_v is normalized at 1) and following the properties of truncated binormal distributions as described in Green (2003), the following holds true:

$$E(u_{1i}|v_i > -Z_i\alpha) = \sigma_{u_1 v}\frac{\varphi(-Z_i\alpha)}{1-\Phi(-Z_i\alpha)} = \sigma_{u_1 v}\frac{\varphi(Z_i\alpha)}{\Phi(Z_i\alpha)}, \quad (2.8.3)$$

$$E(u_{0i}|v_i \leq -Z_i\alpha) = \sigma_{u_0 v}\frac{-\varphi(-Z_i\alpha)}{\Phi(-Z_i\alpha)} = \sigma_{u_0 v}\frac{-\varphi(Z_i\alpha)}{1-\Phi(Z_i\alpha)} \quad (2.8.4)$$

Plugging these expressions into equations for $E(Y_{1i}|W_i = 1)$ and $E(Y_{0i}|W_i = 0)$ and differencing them results in (2.8.5):

$$E(Y_{1i}|W_i = 1) - E(Y_{0i}|W_i = 0) = (\beta_0^1 - \beta_0^0) + (\sigma_{u_1 v}\frac{\varphi(Z_i\alpha)}{\Phi(Z_i\alpha)} - \sigma_{u_0 v}\frac{-\varphi(Z_i\alpha)}{1-\Phi(Z_i\alpha)}), \quad (2.8.5)$$

where $\beta_0^1 - \beta_0^0$ is the true ATE, and $\sigma_{u_1 v}\frac{\varphi(Z_i\alpha)}{\Phi(Z_i\alpha)} - \sigma_{u_0 v}\frac{-\varphi(Z_i\alpha)}{1-\Phi(Z_i\alpha)}$ is the selection bias due to unobservables.

The estimation is performed in two stages:

Stage 1: Obtain $\hat{\alpha}$ by estimating the selection equation $W_i^* = Z_i\alpha^{'} + v_i$.

Stage 2: Estimate $Y_{1i} = \beta_0^1 + X_i'\beta + \frac{\varphi(Z_i\widehat{\alpha})}{\Phi(Z_i\widehat{\alpha})} + u_{1i}$ and $Y_{0i} = \beta_0^0 + X_i'\beta + \frac{-\varphi(Z_i\widehat{\alpha})}{1-\Phi(Z_i\widehat{\alpha})} + u_{0i}$. The difference between $\hat{\beta}_0^1$ and $\hat{\beta}_0^0$ will be the estimate of the ATE in this model.

What makes it possible to estimate the hidden (unobserved) bias is the fact that only the knowledge of the mean effect of the unobservable factors in the treatment selection on the outcome given the observed data is necessary. This effect can be calculated from truncated bivariate distributions of the unobservables as long as the distributions are specified. It is estimated in the first stage and added to the second stage for error correction.

The Heckman treatment effect model is highly parameterized and requires strong distributional assumptions for both the outcome regression and the treatment selection model. The error correction variable had the correct functional form only when the outcome regression is linear, the treatment selection is modeled as probit, and the error terms follow a bivariate normal distribution. If any of these requirements are not met, the error correction variable will not correct the selection bias.

2.9 Summary of the current methodological issues in observational studies

Selection bias presents a great challenge in observational studies. While several bias correction methods are well described in the literature, no unified approach exists in identifying the preferred method under specific settings of applied research. Statistical methods address bias correction when selection is on observables, and econometric models are designed to correct the hidden bias when selection is on unobservables. The main difficulty arises from the fact the key assumptions about the selection process that warrant the unbiased estimates of the ATE cannot be empirically tested. Researchers can only attempt to build a convincing case that all important covariates have been assessed or employ actual randomization of treatment assignment, which

ensures that all observed and unobserved covariates are on average balanced prior to treatment administration.

A thorough understanding of the sources of selection bias is critical for estimating treatment effects in observational studies. In simulation settings, when all observable covariates are accounted for, the Heckman treatment effect model yields a highly biased estimator of the ATE, while regression adjustment, PSM, and AI estimator meet the expectations of correcting the overt bias quite adequately (Guo, 2010). Guo also demonstrates that all four methods yield biased estimates under selection on observables with an omitted variable, with the Heckman treatment effect model bias being the largest of the four.

The treatment effect model requires the assumption about a nonzero correlation of error terms in selection and outcome equations and strongly depends on correct model specification. This requirement is more pronounced that that of OLS regression. With no definite procedure to test conditions under which the assumptions of the Heckman model are violated, the estimation results should be interpreted with caution.

Balancing methods and regression adjustment require the ignorable treatment assignment assumption to produce unbiased estimates of the ATE, and are designed to remedy the overt bias, but fail to provide accurate estimates when hidden bias is present. The PSM methods and regression adjustment rely heavily on correct model specification. Matching methods without propensity scores do not involve estimation of unknown functional forms and are easy to implement, but due to dimensionality problem, their applicability is limited to situations when the number covariates is small.

Matching with and without propensity scores has several weaknesses. The decision to use matching with or without replacement as well as choosing the number of matches for each

treated unit is a tradeoff between precision and bias. Matching with replacement is a better alternative when very few relevant control units are available for comparison. It minimizes the distance between the matched pairs, and is beneficial in terms of bias reduction. Matching without replacement increases bias and can produce results that are sensitive to the order in which the matches are done, but improves the precision of the estimates. 1-to-1 matching produces the smallest distance between the matched pairs and reduces bias. At the same time, the precision of estimates with 1-to-1 matching suffers because large amount of information available from the data is discarded in the process.

Matching algorithms commonly used in propensity score matching are nearest neighbor with caliper, kernel and interval matching. Morgan and Winship (2007) demonstrate that the choice of the matching algorithm affects the estimation results when everything else is held constant. The performance of these matching algorithms remains debatable, with little evidence as to which algorithm is more efficient in particular settings.

The existing models for bias correction may be sensitive to the sample size and to the ratio of treated and controls in the sample. Kennedy (2003), for example, casts doubts about the accuracy of the estimation results from the Heckman treatment effect model when the sample size is small. To gain more insights into this issue, I investigate the performance of several bias-correction methods under different settings for sample size and sample imbalance in Chapter 3.

CHAPTER 3

SIMULATION STUDIES: RESULTS AND IMPLICATIONS

The nature of the data available for observational studies involving causal inference and sample selection bias may often lead to a problem of imbalanced samples where the number of control units is much larger than the number of treated cases. To investigate the impact of sample imbalance on the accuracy of estimates of ATE obtained through different corrective methods, I conducted a Monte Carlo simulation study under a variety of settings for the sample size and proportions of treated and control cases in a sample.

I compared five models: the OLS regression, propensity score matching with a postmatching regression analysis, the doubly robust matching estimator, the Abadie and Imbens matching estimator, and the Heckman treatment effect model using maximum likelihood estimation. The models selected for the Monte Carlo study are designed to estimate the ATE, as opposed to, for example, kernel based matching that estimates only the average treatment effect for the treated.

As discussed in Chapter 2, the first four models require the assumption that the treatment assignment is exogenous and random conditional on specified covariates, or, in other words, that that selection is on observables. It implies that that treatment assignment is independent of the potential outcome if all covariates are observed and held constant. In contrast, in the Heckman treatment effect model treatment selection is viewed as a nonrandom choice. It employs the conditional probability of receiving treatment to control for selection bias in the outcome analysis, and therefore the Heckman treatment effect model does not require the selection on observables assumption.

To emphasize the importance of underlying model assumptions, I adopted the data generation scenarios that mirror both types of selection bias as suggested by Guo and Fraser (2010). The data generation process adopted allows comparing the performance of the estimators under different types of selection bias as well as under different settings for the sample size and proportion of control cases in a sample.

3.1 Research Questions for the Simulation Study

I use two data generation settings to mimic the two types of selection bias: selection on observables and selection on unobservables. The two types of selection bias and their implications for the ATE estimation are discussed in detail in Chapter 2 of this dissertation. The sample sizes are set at $n_1 = 1000$, $n_2 = 500$, and $n_3 = 200$. One setting for proportion of controls is $p_C = 0.5$, which represent a balanced sample. The other two settings, $p_C = 0.75$ and $p_C = 0.9$, reflect the varying degrees of imbalance. The number of repetitions for the Monte Carlo simulation is set at 10,000.

The goal of this Monte Carlo simulation is to compare the performance of the estimators of the ATE across the five models under different settings for the bias selection, sample sizes, and degrees of sample imbalance in terms of the ratio of treated and controls. It aims to address the following 4 research questions:

1) Within each setting for the selection bias, given a balanced sample, which model performs the best, and how are the five models ranked on bias and mean square error criteria?
2) What is the effect of sample size on the accuracy of the estimates within each setting for the selection bias?

3) Within each setting for the selection bias, how sensitive are the models to sample imbalance in terms of bias and mean square error criteria?
4) Do the increased sample size and perfect sample balance help improve the accuracy of the estimates when the model assumptions are violated?

It is worth noting that this Monte Carlo study simulates very limited settings of data generation, and its conclusions cannot be generalized to other settings. The main purpose of the study is to demonstrate that the performance of the models under a common setting of data generation will vary, and the variation in performance will be magnified in smaller samples with higher degrees of imbalance between treated and control groups.

3.2 Data Generation

The data generation process adopted here is based on the counterfactual model framework presented in Section 2.2 of this dissertation, and is designed to account for selection on observables, selection on unobservables, and selection on observables with an omitted variable.

Setting I: Selection on observables

To approximate selection on observables, Z_i, the covariate that affects the treatment assignment, should be correlated with u_i, the error term of the outcome equation, while u_i and v_i, the error term of the selection equation, are uncorrelated. Following Guo (2010), I use three covariates (x_1, x_2, x_3) that affect the outcome *y*, allow *z* to determine the treatment assignment *w* only, and x_3 to affect both the outcome and the treatment assignment. The outcome Y is generated as

$$Y = 100 + .5x_1 + .2x_2 - .05x_3 + .5W + u. \quad (3.2.1)$$

The true selection equation is

$$W^* = .5Z + .1x_3 + \nu \tag{3.2.2}$$

The covariates x_1, x_2, x_3, and Z and the error term u are random variables that are normally distributed with a mean vector (3 2 10 5 0), standard deviation vector (.5 .6 9.5 2 1) and the symmetric correlation matrix as defined in (3.2.3):

$$r_{(x1,x2,x3,Z,u)} = \begin{bmatrix} 1 & & & & \\ .2 & 1 & & & \\ .3 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & .4 & 1 \end{bmatrix}. \tag{3.2.3}$$

The error term of the selection equation, ν, is a random variable from a standard normal distribution, and $W = 1$, if $W^* > Median(W^*)$, and $W = 0$ otherwise.

This specification creates a correlation of .4 between Z and u, and a zero correlation between u and ν. This correlation structure meets the requirements for simulating selection on observables, as shown in Figure 2.2.1a in Section 2.2 of this dissertation. Under this specification, the true ATE in the population is known and equal to .5, as shown in (3.2.1).

Setting II: Selection on unobservables

Selection on unobservable requires Z_i and u_i to be uncorrelated while nonzero correlation exists between u_i and ν_i. Selection on unobservables is shown in Figure 2.2.1b in Section 2.2 For the second setting to mimic selection on unobservables, the outcome and treatment assignment are generated by the same processes as described by (3.2.1) and (3.2.2), with a few modifications in the error structure of the selection equation and in the correlation matrix.

The error term ν of the selection equation now follows

$$\nu = \delta + .15\varepsilon, \tag{3.2.4}$$

where δ is a standard normal random variable, and ε is a zero mean normally distributed random variable, which is correlated with u. The correlation matrix $r(x_1, x_2, x_3, Z, u, \varepsilon)$ is now defined as

$$r_{(x1,x2,x3,Z,u,\varepsilon)} = \begin{bmatrix} 1 & & & & & \\ .2 & 1 & & & & \\ .3 & 0 & 1 & & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & .7 & 1 \end{bmatrix} \tag{3.2.5}$$

Setting III: Selection on observables with an omitted variable

For the third setting, selection on observables with an omitted variable, I omit Z , the covariate affecting selection equation, from all models, which creates overt selection bias while the data generation process remains the same as in setting 1.

3.3 Model Specifications

For selection on observables and selection on unobservables settings all five models have the same specification. OLS regression includes all four covariates and is modeled as shown by (3.3.1):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 Z + \hat{\tau} W, \tag{3.3.1}$$

where $\hat{\tau}$ is the estimate of the ATE in each repetition.

The propensity scores are estimated using the logistic regression model in (3.3.2):

$$\hat{e}(x) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\hat{\beta}_3 x_3+\hat{\beta}_4 Z)}}. \tag{3.3.2}$$

The estimated propensity scores are then used to match each treated case to a control case using nearest neighbor with caliper. The caliper width is set at a quarter of the standard deviation of the estimated propensity scores, and matching is performed without replacement. For postmatching analysis, I fit the OLS regression model described in (3.3.1) using the matched sample.

The DBR estimator uses the estimated propensity scores as defined by (3.3.2) and the predicted values from the two OLS regressions as specified in (3.3.1). The first regression model is fitted for the treatment group only (W_i = 1), and then the predicted values are obtained for the whole sample. This gives the values for $m_1(X_i)$, the regression augmentation term. The second OLS regression model is fitted for the control cases in a similar way, and its predicted values form $m_0(X_i)$. The ATE for the DBR estimator is calculated using (2.7.1) in each iteration of the Monte Carlo simulation.

The outcome regression equation for the Heckman treatment effect model is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\tau} W. \quad (3.3.3)$$

The selection equation is

$$W^* = \gamma Z + \nu,$$

$$W = 1 \; if \; W^* > Med(W^*), and \; W = 0 \; otherwise \quad (3.3.4)$$

The conditional probabilities are defined as $Prob(W = 1|Z) = \Phi(Z\gamma)$ and $Prob(W = 0|Z) = 1 - \Phi(Z\gamma)$, and the model is estimated by the maximum likelihood estimation. The selection equation in the HE model includes only Z, which is different from the PSM model where the logistic regression employs x_1, x_2, x_3, and Z. When all the covariates are included in the HE model, it does not converge. However, the current specification captures the main features of selection on observables and is the best possible model in these settings.

The covariates for the Abadie and Imbens matching estimator include x_1, x_2, x_3, and Z. The vector norm is calculated based on the inverse of the sample variance matrix using R-package “Match”.

3.4 Criteria for Performance Assessment

Following Guo (2010), I use two criteria to assess the performance of the bias correction methods. One criterion is the estimated bias. I average the estimated values of the ATE obtained for each of the 10,000 samples, and because the true ATE is known, the difference between the average estimated ATE and the true ATE provides an estimation bias for a given model.

The second criterion is the estimated mean square error (MSE), which is estimated as follows:

$$MSE = \sum_{i=1}^{10,000}\left(\tau_{MODEL_i} - .5\right)^2/10{,}000\ . \qquad (3.4.1)$$

MSE provides a measure of the variation of the sampling distribution for the estimated treatment effects. A small MSE value as defined in (3.4.1) indicates low variation.

3.5 Simulation Results I: Selection on Observables

The simulation results obtained under selection on observables are summarized in Table 3.1. Overall, The OLS regression produced the best results, followed closely by the propensity score model and the AI estimator. On average, the bias for the ATE remained well below ±1% even in smaller samples with only 10% of treated cases when the OLS and PSM models were used for estimation. The PSM model performed slightly better than the OLS regression in highly imbalanced samples based on the estimated bias, but showed higher variability.

The superior performance of the OLS and the PSM models is due to the fact that the model assumptions were satisfied by the data generation process, and x_3 and Z, the main variables that determine treatment assignment, were controlled for in the analysis. In practice, x_3 and Z might not be the only factors affecting treatment assignment, may not be available or collected, and the error term of the outcome equation u may be correlated with another variable

Table 3.1. Simulation Results under Selection on Observables

	Mean ATE			MSE			Bias		
	n=200	**n=500**	**n=1000**	**n=200**	**n=500**	**n=1000**	**n=200**	**n=500**	**n=1000**
OLS									
P_c= .5	0.4991	0.5002	0.4988	0.0297	0.0116	0.0056	-0.18%	0.05%	-0.25%
P_c= .75	0.4988	0.5027	0.4995	0.0351	0.0142	0.0070	-0.25%	0.53%	-0.09%
P_c = .9	0.4973	0.5010	0.5001	0.0604	0.0241	0.0118	-0.54%	0.20%	0.02%
PSM									
P_c= .5	0.4982	0.5005	0.4985	0.0335	0.0128	0.0061	-0.36%	0.10%	-0.29%
P_c= .75	0.4989	0.5023	0.5000	0.0466	0.0172	0.0083	-0.22%	0.45%	-0.01%
P_c = .9	0.5005	0.5007	0.4997	0.1327	0.0373	0.0166	0.09%	0.14%	-0.06%
DBR									
P_c= .5	0.5229	0.5190	0.5144	0.0647	0.0278	0.0144	4.58%	3.80%	2.89%
P_c= .75	0.5605	0.5506	0.5391	0.1274	0.0601	0.0319	12.10%	10.11%	7.82%
P_c = .9	0.6288	0.6244	0.6036	0.4674	0.2267	0.1355	25.75%	24.88%	20.71%
AI									
P_c= .5	0.4982	0.5001	0.4984	0.0523	0.0214	0.0108	-0.36%	0.02%	-0.33%
P_c= .75	0.4958	0.5024	0.5014	0.1253	0.0555	0.0288	-0.84%	0.48%	0.28%
P_c = .9	0.4849	0.5007	0.5031	0.6025	0.2520	0.1400	-3.01%	0.15%	0.63%
HE									
P_c= .5	1.9120	1.8793	1.8748	2.8637	2.2331	2.0523	282.40%	275.86%	274.96%
P_c= .75	1.9643	1.8866	1.8752	3.5012	2.3646	2.1073	292.86%	277.33%	275.03%
P_c = .9	2.2189	1.9594	1.9272	6.9040	3.2034	2.5303	343.77%	291.88%	285.44%

instead of *Z*. These conditions are restrictive, and should be carefully considered in observational studies.

The Heckman treatment effect model failed in all settings with true selection on observables overestimating the true effect by 275%-344%, on average, with very high MSE's ranging between 2.05 and 6.90. These results were consistent with other empirical findings that emphasize the HE model sensitivity to model assumptions, and with the underlying theory. The error terms of the outcome and selection equations in the HE model are required to have a nonzero correlation. This requirement was violated by the data generation process under selection on observables, which resulted in highly biased and completely unreliable estimates of the ATE. The findings of the simulation study emphasize the fact that the assumption of nonzero correlation of the two error terms in the HE model is crucial for obtaining unbiased estimates of the ATE using this method.

Balanced samples with equal proportions of treated and control cases produced the most stable estimates of the ATE across all five models. Overall, smaller sample size and increasing imbalance between the numbers of treated and control cases resulted in larger MSE for all the estimates of the ATE. The AI estimator worked well with a balanced ratio of treated and controls and was ranked third among the five models. It became increasingly unstable with a high proportion of controls ($p_c = 0.9$) even in larger samples compared to the OLS regression and the PSM model. However, the results of the AI estimation were still reliable when proportion of controls was at 0.75, with an estimated bias of less than ±1% .

The results of the DBR estimator were somewhat unexpected. The DBR estimator is designed to be unbiased if at least one of the models (regression or PSM) is specified correctly. The augmentation component of the DBR estimator of the ATE is constructed in such a way that

if either regression or PSM bias term is equal to zero, then it removes the other, nonzero bias term from the incorrect model. When selection on observables was simulated, both the regression and the PSM model had a correct specification and performed well, while, surprisingly, the DBR estimator produced noticeably higher bias. In particular, with proportions of controls above 0.5, the method overestimated the true effect by 8% to 26%. It also appeared to be much more sensitive to the sample imbalance compared to the top three models and deteriorated at a significantly faster rate than the AI estimator. Kang and Schaffer (2007) discuss this phenomenon and point out that when the regression model is specified correctly, adding additional augmented terms results in overfitting the model and does not improve the OLS estimates.

3.6 Simulation Results II: Selection on Unobservables

In selection on unobservables, the HE model worked relatively well in terms of the estimation bias, on average, but the MSE's for the HE estimates were the highest among the five models (Table 3.2). Given that in this particular setting all the model assumptions for the HE model were satisfied, the results are not encouraging. In balanced samples, the bias for the HE estimates was between -2% and .5%, but the MSE with small samples was 0.98, as opposed to the MSE of 0.06 for the OLS regression model. The HE model was not robust to sample imbalances and exhibited a significant decline in performance when proportion of controls increased to 0.75. At $p_c = 0.9$, the HE overestimated the true ATE by 8% to 14%, but the model became unstable yielding MSE's as high as 4.32 for $n = 200$. Overall, the performance of the HE model was the worst with small sample sizes for both balanced and unbalanced samples, which is consistent with other empirical findings suggesting that the HE model is not recommended for small samples (Kennedy, 2003).

Table 3.2. Simulation Results under Selection on Unobservables

	Mean ATE			MSE			Bias		
	n=200	**n=500**	**n=1000**	**n=200**	**n=500**	**n=1000**	**n=200**	**n=500**	**n=1000**
OLS									
P_c= .5	0.66	0.67	0.67	0.0607	0.0412	0.0350	32.74%	33.17%	33.62%
P_c= .75	0.66	0.66	0.66	0.0669	0.0410	0.0331	31.79%	31.51%	31.65%
P_c = .9	0.65	0.65	0.65	0.0954	0.0515	0.0379	30.88%	30.59%	30.53%
PSM									
P_c= .5	0.68	0.68	0.68	0.0795	0.0498	0.0406	35.04%	35.19%	35.44%
P_c= .75	0.68	0.68	0.68	0.1091	0.0576	0.0451	35.74%	35.34%	35.71%
P_c = .9		0.69	0.68		0.0983	0.0631		37.41%	36.94%
DBR									
P_c= .5	0.73	0.72	0.72	0.1245	0.0812	0.0658	45.28%	44.50%	44.25%
P_c= .75	0.80	0.78	0.77	0.2343	0.1403	0.1081	59.20%	55.74%	54.30%
P_c = .9	0.96	0.94	0.90	0.7491	0.4438	0.2971	92.78%	88.23%	79.28%
AI									
P_c= .5	0.69	0.69	0.70	0.0968	0.0627	0.0519	38.19%	38.76%	39.43%
P_c= .75	0.71	0.72	0.72	0.1933	0.1080	0.0814	42.64%	43.28%	44.01%
P_c = .9	0.78	0.78	0.77	0.7805	0.3699	0.2358	56.10%	55.97%	54.67%
HE									
P_c= .5	0.50	0.49	0.50	0.9757	0.3808	0.1872	0.37%	-1.94%	0.49%
P_c= .75	0.52	0.53	0.52	1.4367	0.5344	0.2589	4.51%	6.15%	4.54%
P_c = .9	0.54	0.57	0.56	4.3242	1.2846	0.6060	7.67%	13.72%	11.43%

The performance of the OLS regression worsened under selection on unobservable. On average, the ATE estimates were highly biased (overestimation of 30.5% to 33.6%), with MSE's much higher than under selection on observables, though still lower than the MSE's produced by other models in this setting. The results from the PSM model were similar to the ones from the OLS regression, but the bias and MSE were slightly higher for the PSM model. As with selection on observables, both the OLS regression and the PSM model were robust against sample imbalance and their performance did not decline significantly in smaller samples.

The AI and DBR estimators broke down as the data became more unbalanced in terms of the number of treated and control cases. The estimated bias increased from 38% to 56%, accompanied by a higher variability (the MSE went up from 0.1 to 0.8) for the AI estimator as the samples become more imbalanced. The deterioration in the quality of the DBR estimates was even more pronounced with a twofold increase in the estimated bias. The DBR model overestimated the true ATE by 79% to 93% when p_c was set at 0.9. The results of the simulation suggest that using fewer controls may help achieve superior performance for these two estimators.

3.7 Simulation Results III: Selection on Observables with an Omitted Variable

Setting 3, selection on observables with an omitted variable, is likely to be a more realistic scenario for many applications. It describes the case of overt bias that occurs when researchers are not able to include all the relevant confounders in the model. This violates model assumptions for the OLS regression, the PSM model, as well as for the AI estimator, and creates biased estimates of the ATE. The results of the simulation obtained in setting 3 were consistent

with the theory, and brought the bias-correction properties of the DBR estimator to the spotlight (Table 3.3).

The DBR estimator outperformed the rest of the models considerably in terms of both the estimated bias and the MSE. On average, it overestimated the true effect only by 4% or 5% for well balanced samples, and by 8% to 13% for larger samples with 25% of treated cases, but was not robust against highly unbalanced data even for large samples.

The OLS regression, PSM model, and AI estimator failed to produce reliable results in this setting. All three models consistently overestimated the true effect by 93%-96%, and by over 100% when samples were highly unbalanced. The MSE's produced by these three models were also much higher than the ones from the DBR estimator.

Because the OLS and PSM models produced very similar bias estimates, I reran the simulation using several settings of coefficients for the outcome equation to generate the data. In each case the same pattern of common bias estimates occurred with the same direction of bias. Furthermore, the results of the DBR estimator were not significantly affected by those changes.

The HE model showed a lot of variability in the estimation of the ATE did not converge under misspecified selection on observables using R package "sampleSelection".

3.8 Implications of the Simulation Results for the Estimation of Treatment Effects

According to the results of this simulation study, no single model works well in all scenarios. The quality of the results strongly depends on the fit between the assumptions embedded in a model and the process of data generation. While the results obtained in the first two settings were consistent with what one would expect given the data generation process, particularly with the OLS and the PSM models, there are no guarantees that the data at hand fits perfectly under either selection on observables or selection on unobservables scenario.

Table 3.3. Simulation Results under Selection on Observables with an Omitted Variable

	Mean ATE			MSE			Bias		
	n=200	**n=500**	**n=1000**	**n=200**	**n=500**	**n=1000**	**n=200**	**n=500**	**n=1000**
OLS									
P_c= .5	0.97	0.97	0.97	0.2418	0.2278	0.2230	93.30%	93.41%	93.45%
P_c= .75	0.97	0.98	0.98	0.2559	0.2393	0.2334	94.78%	95.26%	95.34%
P_c = .9	1.00	1.01	1.01	0.3131	0.2856	0.2748	100.65%	102.31%	102.54%
PSM									
P_c= .5	0.97	0.97	0.97	0.2509	0.2312	0.2239	93.27%	93.49%	93.36%
P_c= .75	0.98	0.98	0.98	0.2807	0.2484	0.2383	95.90%	95.94%	95.83%
P_c = .9	1.03	1.03	1.03	0.4092	0.3266	0.2982	105.73%	106.23%	105.24%
DBR									
P_c= .5	0.53	0.52	0.52	0.0644	0.0274	0.0148	5.10%	3.82%	3.57%
P_c= .75	0.56	0.55	0.54	0.1294	0.0581	0.0324	12.77%	9.16%	7.90%
P_c = .9	0.64	0.62	0.61	0.4773	0.2306	0.1356	27.37%	24.58%	21.62%
AI									
P_c= .5	0.97	0.98	0.98	0.2526	0.2372	0.2324	94.74%	95.06%	95.24%
P_c= .75	1.01	1.01	1.01	0.3130	0.2841	0.2745	102.14%	102.83%	102.88%
P_c = .9	1.10	1.11	1.12	0.5558	0.4442	0.4160	120.24%	122.40%	123.51%
HE									
P_c= .5	0.09	0.78	0.74	0.7937	0.9930	0.5351	-81.93%	56.79%	47.20%
P_c= .75									
P_c = .9									

In majority of applications the information regarding the tenability of model assumptions is not available as it is often not known if a study omits important covariates. Therefore, empirical findings must be conditioned on a discussion of model assumptions. The assumptions that ensure the unbiasedness of the estimates of the treatment effects should be disclosed and the conditions that may compromise the estimation should be discussed.

The HE model failed to produce accurate estimates of the ATE even under selection to unobservables due to high variability, and was extremely unreliable under selection on observables. Overall, the HE model appears to be more sensitive to the embedded model assumption, the sample size and the degree of sample imbalance in terms of the ratio between the treated and control cases than the rest of the models included in the study. Its poor performance under the ideal data generation settings with small samples raises serious concerns about the validity of estimation results.

The OLS regression and the PSM model appear to work well, but only in very restrictive settings that require previous knowledge about the main sources of selection bias as well as the availability of all relevant covariates and their correct specification in the model. It is worth noting, that the OLS regression and the PSM model outperformed the matching estimator under all three settings, with the OLS regression model always coming first among the three.

In the light of the tenability of model assumptions, the DBR estimator deserves special attention. It performed relatively well in the presence of overt bias, when all the other models failed to provide reasonably unbiased estimates of the ATE. Given that the presence of an omitted variable is a strong possibility in estimating pathway effects, the results of the DBR estimation should be considered together with those obtained through regression adjustment and

PSM. The DBR estimator may offer protection against the overt bias if the proportion of controls does not exceed 0.75 and the sample size is sufficiently large.

The degree of sample imbalance and the sample size appear to increase the variability of the estimates across all models, which has been particularly evident with the AI and the DBR estimators, while the OLS and the PSM models have shown more robust results. All models showed an increase in the estimated bias due to sample imbalance, and for some the loss of accuracy was very pronounced. One of the key findings of this simulation study is the improved performance of the models in the samples with similar proportions of treated and control cases. This finding should be taken into consideration at the stage of a study design when identifying control and treated groups, and has several implications for the model selection when samples are imbalanced.

While larger sample size did not always translate into increased accuracy of the estimates in terms of the estimated bias, the effect of the sample size on their variability was evident across all models in all settings. Reduced MSE's were reported for all the estimates of the ATE when sample size increased. Even though researchers often do not have control over the sample size, it should be given serious consideration at the model selection stage and included into the discussion of the estimation results.

CHAPTER 4

A CASE STUDY: BIAS CORRECTION MODELS FOR ESTIMATING PATHWAY EFFECTS

This chapter focuses on the analysis of pathway effects on the LOS and 30-day readmission rate based on a sample of COPD patients admitted to the University of Tennessee Medical Center (UTMC) between January 1, 2012 and June 30, 2014. It opens with a discussion of the motivation behind the analysis and then provides a detailed description of the study design, including the choice of the two control groups, data description and limitations, and identification of potential confounding factors and important covariates for adjustment.

The estimation of the effects of the COPD pathway is performed using three methods of bias correction: regression adjustment, propensity score matching with postmatching regression adjustment, and the doubly robust estimator. The choice of the estimation methods is governed by the tenability of model assumptions and by the results of the simulation study presented in Chapter 3. The estimation results obtained using the original sample and the data with imputed missing values are compared across the three models, followed by a discussion about the implications of the study.

4.1 Motivation for the Study

The Affordable Care Act, signed into law on March 23, 2010, aims to ensure wider access to healthcare, and contains many provisions to improve healthcare outcomes and reduce costs through increased competition, regulation, and incentives to expand the use of information technologies and streamline the delivery of healthcare. One such provision is implementation of electronic medical records (EMR) and the corresponding computerized physician order entry

(CPOE) system, which standardize much of the billing and health records, allowing for secure transferability and access at a lower cost.

The use of informational technologies also facilitates the creation of new tools for clinicians that seek to improve quality of care and deliver better outcomes. The standardized patient care pathway is one example of such tools. It is a multidisciplinary evidence-based care plan that defines and optimizes the essential steps in the care of a specific group of patients with a predictable clinical course. The goals of clinical pathways are defined by applying process management thinking to patient care and include limiting undesirable variation in patient care, maximizing clinical efficiency, and creating a standardized approach built around best practices and optimal resource allocation.

In September of 2013 the UT Medical Center launched a new model of patient care delivery based on the concept of standardized patient care pathways. The pathways are built electronically to be used together with the CPOE system, and are being implemented as a hospital-wide policy. The key research goal of this study is to develop a rigorous evaluation methodology and use it analyze the effects of a COPD pathway on the length of stay (LOS) and 30-day readmission rates, in order to determine if the use of a pathway improves healthcare outcomes for COPD patients.

The absence of randomization in pathway assignment in this study opens up basic analyses methods to potential biases as discussed in Chapter 2 and 3. To address this I apply bias correction methods for the estimation of pathway effects that are rooted in the theory of the counterfactuals. The key idea in the theory of counterfactuals is that each subject has a potential outcome under each treatment state, which in this study translates into two sets of clinical outcomes for each patient: one under care received with a pathway approach and the other

resulting from a traditional care delivery. Because it is impossible to observe both sets of outcomes for any one patient, causal effects cannot be calculated at the individual level. Instead, the counterfactual model estimates the average treatment effects at the population level using the outcomes from the treated and control groups as estimates for the counterfactual, or unobserved, outcomes. Patients who did not have pathways assigned to them will form a control group, and patients who received care under a pathway model will be considered the treated group.

4.2 Data Description

As mentioned above, the data set for the analysis includes COPD patients admitted to the UTMC during the period from January 1, 2012 through June 25, 2014 (with the latest patient discharge date being June 30, 2014). Patients admitted during the period of September 1, 2013 - December 31, 2013, when pathways were used without a compliance tracker, were excluded from the analysis.

Since misclassification of the COPD diagnosis is likely on admission (Smith et al., 2004), only patients with principal diagnosis of COPD at discharge were included in this study. Among the total of 970 COPD patients, there were 737 patients from the pre pathway period (January 2012 – August 2013), 172 patients that were assigned a COPD pathway (January 2014 – June 2014), and 61 patients that did not have a COPD pathway assigned when pathways were already in use (January 2014 – June 2014).

The data collected for this study comes from both the clinical and the administrative databases of the hospital, and includes patient demographic characteristics (age, gender, race, insurance type), clinical characteristics (vital signs on admission, classical COPD risk factors, procedures and secondary diagnoses) as well as hospital characteristics (number of beds available and patient days). Table 4.1 contains the summary statistics for all of the continuous

Table 4.1. Summary statistics for the COPD patients, January 1, 2012 – June 30, 2014

	LOS, days	*LOS, hours*	*Age*	*Temp*	*HR*	*BPs*	*BPd*	*Oxygen*	*BMI*	*eGFR*	*Braden*	*HCa*	*HCd*	*CCI*
n	970	970	970	970	970	970	970	970	970	970	970	970	970	970
n, missing	0	0	0	1	1	0	0	1	37	54	2	12	0	0
min	1	9	31	95	48	13	38	46	12	4	7	0.58	0.57	0
max	50	1184	98	104	166	238	180	100	75	437	23	0.94	0.94	9
range	49	1175	67	9	118	225	142	54	63	433	16	0.36	0.37	9
median	4	101	67	98	93	140	74	95	26	66	20	0.76	0.75	1
mean	5.02	129.75	66.59	98.08	94.83	143.15	76.97	93.33	27.64	69.17	19.68	0.75	0.75	1.34
std.dev	4.40	104.30	11.28	0.95	19.38	28.39	15.83	6.12	8.62	32.74	2.74	0.0538	0.0541	1.39

variables available for the analysis based on the entire sample of the COPD patients initially considered for the study.

The LOS, reported in days, was recalculated to be measured in hours based on the exact admission and discharge time in the clinical risk reports for each patient. The LOS measured in hours was used as the outcome variable in the analysis. The average LOS for the entire sample of the COPD patients during the study period was 5.02 days, or 129.75 hours, 1 day was reported as the shortest LOS, and the LOS of 50 days was the longest in the sample. Since a LOS of more than 9 days is considered unusual for most DRGs, the outlier diagnostics are necessary when assessing the estimation results.

The measures of hospital congestion on admission (HCa) and at discharge (HCd) were calculated as a daily ratio of patient days and observation days to beds available, and as such, higher values for HCa and HCd denote higher daily hospital congestion. On average, the hospital was operating at a 75% capacity level over the study period, with some days reaching as high as 94% of total capacity.

The distributions of age, temperature, heart rate, and blood pressure appear to be normal, while Oxygen, BMI, glomerular filtration rate, eGFR (a measure of renal function,) and Braden score exhibit sample distributions that are highly skewed, and might result in leverage points which have a detrimental impact on the estimation. Charlson comorbidity index, CCI, is expected to have a right skewed distribution with a low mean (Hall, 2005), and the observed distribution of the CCI for the COPD patients in the sample is consistent with the theory.

Table 4.2 summarizes the categorical variables available for the analysis for the entire sample and for the control and treated groups. It shows that more than a third of patients in the

Table 4.2 Summary Statistics for categorical data

	All COPD patients		**Historic Controls**		**Contemp. controls**		**Pathway patients**		**Historic Controls**	**Contemp. controls**
	Count	*Freq.*	*Count*	*Freq.*	*Count*	*Freq.*	*Count*	*Freq.*	*Std Diff, %*	*Std Diff, %*
Gender										
Female	481	0.52	374	0.53	27	0.47	80	0.50	-3.89	4.88
Male	446	0.48	335	0.47	31	0.53	80	0.50	3.89	-4.88
Race										
Black	55	0.06	42	0.06	3	0.05	10	0.06	0.96	3.28
Other	5	0.01	2	0.00	0	0.00	3	0.02	10.94	13.82
White	867	0.94	665	0.94	55	0.95	147	0.92	-5.27	-8.40
Insurance										
Self pay	18	0.02	16	0.02	2	0.03	0	0.00	-15.19	-18.90
Private	296	0.32	227	0.32	16	0.28	53	0.33	1.67	8.53
Medicare	407	0.44	318	0.45	30	0.52	59	0.37	-11.51	-21.38
Medicaid	123	0.13	87	0.12	6	0.10	30	0.19	12.71	16.98
Readmission										
No	585	0.63	421	0.59	46	0.79	118	0.74	21.79	-9.30
Yes	342	0.37	288	0.41	12	0.21	42	0.26	-21.79	9.30
Tobacco										
never a smoker	129	0.14	104	0.15	6	0.10	19	0.12	-5.83	3.44
former smoker	333	0.36	261	0.37	25	0.43	47	0.29	-11.21	-20.40
current smoker	385	0.42	304	0.43	14	0.24	67	0.42	-1.43	27.16
Total	927	1.00	709	1.00	58	1.00	160	1.00		

sample were readmitted to the hospital within 30 days. The 30-day readmission rate is one of the measures used for the inpatient quality reporting program. The Hospital Readmissions Reduction Program, mandated by the Affordable Care Act, requires the Centers for Medicare and Medicaid (CMS) to reduce payments to hospitals with excess readmissions. The first penalty affecting payment was for discharges beginning October 1, 2012, and these penalties increase yearly up to a maximum of 3% reached in the fiscal year of 2015.

In the light of the recent policy changes, and taking into account the fact that almost 60% of the COPD patients admitted to the UTMC during the study period had Medicare or Medicaid insurance, understanding the effects of the COPD pathway on readmission rates becomes increasingly important.

Tobacco use is a behavioral risk factor for COPD, and must be controlled for in modeling the LOS and readmission rate. The Tobacco variable is constructed from the clinical data reports and has three categories that denote patient use of tobacco: never a smoker, a former smoker, and a current smoker. 78% of the COPD patients in the data set were either former or current smokers, while 80 patients (about 9%) have missing values for tobacco use.

There are two control groups identified in the study. The first control group, historical controls, consists of 709 patients from the pre pathway period, who were admitted to the hospital between January 1, 2012 and August 31, 2013. The second group, 58 patients who were not assigned a pathway during the period of January 1, 2014 – June 30, 2014, are referred to as the contemporaneous control group. Patients that had a COPD pathway assigned during the first half of the year 2014 become the treated group.

Potential weaknesses exist in the use of either control group to estimate the untreated potential outcome. Contemporaneous controls might overestimate the untreated outcome due to

from a spillover effect from existing pathways, and thus lead to underestimation of the pathway effects. In addition, if pathways increase the efficiency of hospital operations, the contemporaneous controls might have better outcomes due to the overall improvement in the system. If the same clinicians and nurses treating patients with and without pathways may implement certain pathway guidelines they consider efficient even when a pathway has not been initiated.

Historical controls, clearly free of existing pathways influences on outcome measures, may reflect the effects of other factors (e.g. changes in hospital discharge policy, implementation of new quality improvement tools in different areas of the hospital, effects of new legislative regulations in healthcare). These factors may either contribute to driving the magnitude of the estimated pathway effect up, or, on the contrary, diminish it.

The study considers both designs (with contemporaneous and historical controls) since measuring pathways spillover effects in a hospital setting and quantifying the effects of all possible changes over time is rather challenging, if not impossible.

Reduction in the LOS for expired patients is not indicative of a positive effect of the pathway, and as such should be singled out from the pathway effect on the LOS. The analysis of the LOS should be accompanied by examining the effects of pathway on the in-hospital mortality, as the LOS tends to decrease, on average, when mortality rates are higher. Currently, it is not feasible to investigate the effects of the COPD pathway on in-hospital mortality due to the insufficient sample sizes. In total, there were 43 expired patients total, 28 expired in the historical control group, 12 in the treated group, and only 3 among the contemporaneous controls.

After excluding expired patients from the dataset, the sample sizes for all three groups are slightly smaller: 709 historical controls, 160 treated cases, and 58 contemporaneous controls.

Table 4.3 describes summary statistics for the categorical variables in the control and treated groups after exclusion of expired patients.

The covariate imbalance prior to matching is evident in both designs. The categorical variables with high standardized differences for both the historical and contemporaneous controls are Insurance and Tobacco. The continuous variables with significant mean differences between groups in the design with the historical controls include Age, Braden score, HCa and HCd, and CCI. Their standardized mean differences, shown in the last two columns of Table 4.3, are greater than 10% , which also suggests that the imbalance is considerable (Normand et al., 2001).

The imbalance in the design with the contemporaneous controls is mainly due to the Braden score and CCI. The mean differences are highly significant for the Braden score and the CCI. The corresponding standardized mean differences for these covariates are around 40%, and need to be addressed by regression adjustment and propensity score matching.

The existing covariate imbalance between the groups provides strong evidence that pathway assignment may be confounded with factors that are prognostic of the LOS and readmission rate for the COPD patients. The propensity score matching aims at balancing the data by matching the treated and control cases based on the predicted values for conditional probabilities of a patient being assigned a pathway, and the importance of assessing the covariate imbalance in a matched sample should be emphasized. According to the theory, the propensity score model that is specified correctly should balance the distribution of all the covariates between the treated and control groups in the matched sample, regardless of their inclusion in the logistic regression model used for estimating the propensity score. The model fit and data balancing are the two guiding principles in the model selection for propensity score matching.

Table 4.3 Summary statistics for the COPD patients by control and treated groups.

	LOS, days	*LOS, hours*	*Age*	*Temp*	*HR*	*BPs*	*BPd*	*Oxygen*	*BMI*	*eGFR*	*Braden*	*HCa*	*HCd*	*CCI*
n_{CH}	709	709	709	709	709	709	709	709	683	677	708	700	709	709
n_{CH}, missing	0	0	0	0	0	0	0	0	26	32	1	9	0	0
min	1	9	31	96	48	13	38	61	13	4	8	0.58	0.57	0
max	41	985	93	103	154	238	145	100	64	437	23	0.86	0.86	9
median	4	98	67	98	93	140	75	95	26	66	20	0.75	0.74	1
mean	4.75	123.46	66.62	98.08	94.65	143.98	77.11	93.43	27.58	69.55	19.54	0.74	0.74	1.35
std.dev	3.85	92.03	11.36	0.91	19.47	28.90	15.39	5.92	8.50	34.15	2.74	0.0520	0.0507	1.37
std diff, %	-7.52	-4.73	-18.1	8.9	2.0	-8.8	3.9	-11.5	3.0	-1.5	43.9	61.2	95.0	-20.3
n_T	160	160	160	159	159	160	160	159	152	144	160	160	160	160
n_T, missing	0	0	0	1	1	0	0	1	8	16	0	0	0	0
min	1	20	43	96	51	80	48	46	13	24	13	0.68	0.68	0
max	27	666	92	104	145	207	137	100	59	220	23	0.94	0.94	9
median	4	101	64	98	93	139	76	94	27	68	22	0.77	0.79	1
mean	4.48	119.35	64.61	98.17	95.03	141.56	77.72	92.70	27.84	69.10	20.67	0.77	0.78	1.08
std.dev	3.41	81.73	10.88	1.10	17.71	26.41	16.07	6.70	8.69	26.81	2.38	0.0501	0.0489	1.31
n_{CH}	58	58	58	58	58	58	58	58	58	55	57	58	58	58
n_{CH}, missing	0	0	0	0	0	0	0	0	0	3	1	0	0	0
min	1	32	40	96	59	89	47	73	12	17	14	0.69	0.69	0
max	16	363	89	101	140	208	115	100	75	155	23	0.87	0.93	5
median	4	97	69	98	90	139	73	94	28	72	20	0.79	0.79	2
mean	5.29	132.19	67.26	98.04	92.72	140.31	73.60	93.29	29.79	69.28	19.74	0.78	0.78	1.66
std.dev	3.93	87.33	10.22	1.06	19.41	28.28	15.38	5.28	10.76	29.97	2.47	0.0522	0.0483	1.54
std diff, %	-22.1	-15.2	-25.1	12.3	12.4	4.6	26.2	-9.8	-20.0	-0.6	38.4	-10.6	3.7	-40.2

When pathway assignment is not random, simple comparison of the means across treated and control groups will provide biased results due to confounding of the pathway effects with other factors. The unadjusted mean differences for the LOS are shown in Table 4.4. The purpose of this study is to address the problem of selection bias through identifying the important confounders of the pathway effects and employing several statistical methods for bias correction, while ensuring that the identified confounding factors are sufficiently included in the models.

Table 4.4 The Unadjusted Average Differences in the LOS and Readmission Rates.

	ATE	*SE(ATE)*
LOS, Historical controls	-4.11	0.0547
LOS, Contemporaneous controls	12.84	0.0912
Readmission, HistoricalControls	-0.20	0.2554
Readmission, Contemporaneous Controls	0.03	0.4828

4.3 Identifying Potential Confounders and Controls

Factors such as patient demographic and clinical characteristics, severity of illness, comorbidities, insurance, and hospital congestion can be viewed as potentially affecting both the clinical outcomes and a physician's decision to initiate a pathway, and require serious consideration.

A vast majority of medical studies emphasize the importance of adjusting for comorbidities when evaluating treatment effects and modeling clinical outcomes. Comorbidities are diseases or disorders that coexist with a disease. Comorbid conditions may delay diagnosis, influence treatment choices, affect treatment progress, and confound the analysis. When selection bias exists, patients need to be stratified by risk for statistical analysis, and when bias is

related to comorbidities, a valid measurement of comorbid illnesses is essential for estimating the treatment effects (Hall, 2005).

A comorbidity index reduces all the coexisting conditions and the severity of those conditions to a single numeric score, thus facilitating comparisons across patients. While several general comorbidity indices are available for measuring the impact of comorbidities, the Charlson Comorbidity Index (CCI) is the only index designed using statistical methodology. Another advantage of the CCI is that it creates a continuous variable for scoring. The CCI in this study was computed based on the ICD-9 codes as described in Table 4.5 excluding the codes for the chronic pulmonary disease, the principal diagnosis for the patients in the data set.

Table 4.5 Charlson Comorbidity Index

Reported ICD-9 CM Codes	Condition	CCI
410 – 410.9	Myocardial Infarction	1
428 – 428.9	Congestive Heart Failure	1
433.9, 441 – 441.9, 785.4, V43.4	Peripheral Vascular Disease	1
430 – 438	Cerebrovascular Disease	1
290 – 290.9	Dementia	1
490 – 496, 500 – 505, 506.4	Chronic Pulmonary Disease	1
710.0, 710.1, 710.4, 714.0 – 714.2, 714.81, 725	Rheumatologic Disease	1
531 – 534.9	Peptic Ulcer Disease	1
571.2, 571.5, 571.6, 571.4 – 571.49	Mild Liver Disease	1
250 – 250.3, 250.7	Diabetes	1
250.4 – 250.6	Diabetes with Chronic Complications	2
344.1, 342 – 342.9	Hemiplegia or Paraplegia	2
582 – 582.9, 583 – 583.7, 585, 586, 588 – 588.9	Renal Disease	2
572.2 – 572.8	Moderate or Severe Liver Disease	3
042 – 044.9	AIDS	6

The type of insurance should be considered as a potential confounder, as it is likely to affect both the outcome variables (Fisher et al., 2001) and the pathway assignment. Reimbursement for Medicare and Medicaid patients that make up nearly 60% of the COPD patients in the data set requires hospitals to provide detailed reporting on procedures, providers, and billing schedules. If pathways are viewed as instrumental to improving documentation quality, physicians might be more inclined to initiate a pathway for Medicare and Medicaid patients than for privately insured or uninsured patients. The social and behavioral characteristics of Medicaid patients make them likely candidates for readmissions and for a prolonged hospital stay. Medicare patients are either at least 65 years old, or under 65 and disabled, and their LOS and readmission rates might differ significantly from the rest of the patient population. The proportions of readmissions, for example, differ significantly among the COPD patients in the data set based on their insurance type (*p*-value of 0.0014 for the chi square test).

Many studies suggest using a severity of illness indicator as an important covariate for the estimation of treatment effects (e.g., Khwaja et al., 2011, Kelly et al., 2013, and Marrie et al., 2000) Mechanical ventilation is a commonly used indicator for the severity of illness in COPD patients (Brattebo et al., 2002), but poor reporting on procedures in 2014 UTMC data prevented the use of this variable in the analysis. 109 patients out of 233 in the period from January 1, 2014 to June 30, 2014 did not have any procedures reported, while for 47 of them the recorded LOS was between 5 and 18 days. 63 of these patients were assigned a COPD pathway, and 46 were not. The severity of illness indicator would be missing for more than 75% of the patients in the contemporaneous control group given its size of 61 patients, and for more than a third of the pathway patients. Since this variable is not missing at random, applying traditional data

imputation algorithms is not recommended, and therefore the variable indicating the use of mechanical ventilation is not included in the model.

The Braden score is a tool that aims to help health care professionals assess a patient's risk of developing a pressure ulcer (Kozier, 2008). Braden score is calculated based on examining the ability of a patient to cognitively react to pressure-related discomfort, the degree of moisture the skin is exposed to as well as the degree of friction and sheer, the levels of physical activity and mobility, and a patient's nutritional status. The end-stage COPD patients are characterized by significantly lower levels of physical activity and mobility, as well as by a poorer nutritional status, and are more likely to have low scores on the Braden scale corresponding to higher risk. Therefore, Braden scores can be used as a proxy for the severity of illness indicator in COPD patients.

The standardized differences for the group means for the Braden score were around 40% for both historical and contemporaneous controls, and the *t* tests were highly significant as well ($p<0.0001$). The differences in the distribution of Braden scores for the pathway patients and the two control groups shown in Figure 4.1 suggest that Braden score may affect both the outcome variables and the pathway assignment, or, in other words, be a potential confounder. Pathway patients, on average, have higher Braden scores than patients in both the historical and contemporaneous control groups, which may be indicative of a higher severity of the disease among the control patients. Therefore, failure to adjust for the Braden score in the model could create a bias in the estimates of the pathway effects.

While it is reasonable to assume that the pathway assignment could be driven by an individual physician's preference, an individual physician's effect is not accounted for in this study. Including it in the model would be challenging from a modeling perspective due to a large

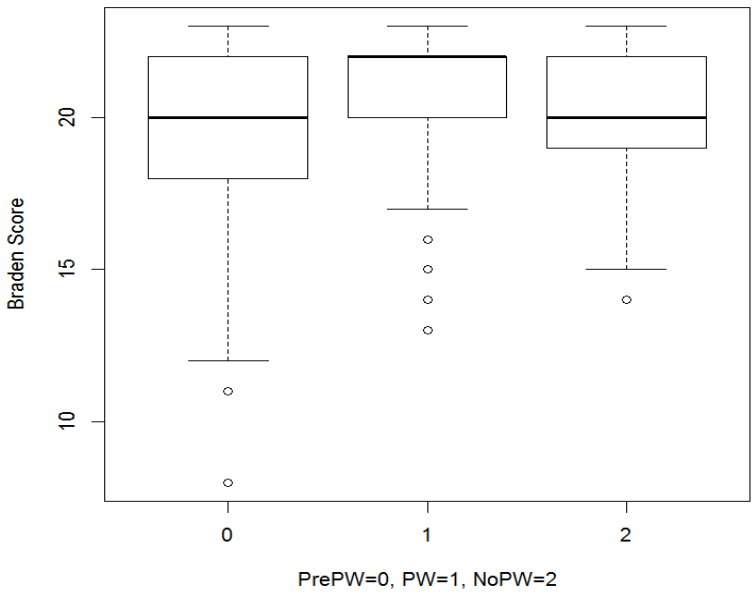


Figure 4.1 Box Plots of Braden Score for Treated and Control Groups.

number of categories, and potentially misleading given the current process of the data collection.

The attending physician that is reported for each patient is their attending physician at discharge, and it is likely that the pathway was assigned by a different attending physician present on admission. The reporting on attending physicians is not always accurate, as many records assign the role of an attending physician to an ER doctor in place of a specialist. Another challenge in capturing an individual physician's effect is a high turnover rate of medical professionals at the UTMC, which results in a small area of overlap when comparing the treated group to the historical controls.

4.4 Model Specifications

Based on the extensive review of the current methodology for observational studies and taking into consideration the results of the simulation study discussed in Chapter 3 of this dissertation, I identified three bias correction methods for estimating the effects of the CPOD

pathway on the LOS and readmission rates. The three methods employed in the current analysis of pathway effects are regression adjustment, propensity score matching with postmatching regression adjustment, and the doubly robust estimator.

Regression adjustment and propensity score matching performed very well under selection on observables, including small and imbalanced samples. Given the small sample size (n=218) in the study design that uses contemporaneous controls, and assuming that all the important covariates are included in the model, regression and PSM are the best candidates for estimating pathway effects on the LOS and readmission rate.

While every effort has been made to account for potential confounders of the pathway effect and the outcome variables, the possibility of omitted variable bias cannot be ruled out. If such a variable exists, the results of regression and propensity score estimation will be biased due to violation of the model assumptions. Using the doubly robust estimator may remedy the bias in the estimates of the pathway effect given that at least one model is specified correctly. According to the results of the simulation study, the DBR estimator works relatively well when proportion of controls does not exceed 75% even in small samples. The proportion of historical controls is around 80%, and given a large sample size of 889, the DBR estimator might provide considerable protection against the overt bias. In the second study design, the contemporaneous controls make up about 40% of the sample, and the results of the DBR estimator should be contrasted to those from the regression adjustment and the PSM.

To estimate the effects of the COPD pathway on the LOS through regression adjustment, a regression model for count data is required. While both Poisson and negative binomial regression models are designed to analyze count data, the two regression models differ in regards to their assumptions of the conditional mean and variance of the dependent variable. Poisson

models assume that the conditional mean and variance of the distribution are equal. Negative binomial regression models do not assume an equal mean and variance and particularly correct for overdispersion in the data, which is when the variance is greater than the conditional mean (Simonoff, 2003; Faraway, 2006). The negative binomial regression produces a better fit for the LOS, and therefore, is a better modeling choice.

The probability distribution for a negative binomial variable that allows for different means μ_i for each is y_i can be expressed as follows:

$$f(y_i; \mu_i, \upsilon) = \frac{\Gamma(y_i+\upsilon)}{y_i!\Gamma(\upsilon)}\left(\frac{\upsilon}{\upsilon+\mu_i}\right)^{\upsilon}\left(\frac{\mu_i}{\upsilon+\mu_i}\right)^{y_i}. \quad (4.4.1)$$

The means are based on the logarithmic link, $\boldsymbol{\mu} = \exp(\boldsymbol{X\beta})$. The negative binomial parameters β and α, where $\alpha = \frac{1}{\upsilon}$ can be estimated using maximum likelihood. The asymptotic variance of $\widehat{\boldsymbol{\beta}}$ can be estimated using

$$\hat{V}(\widehat{\boldsymbol{\beta}}) = \left(X' diag\left[\frac{\widehat{\mu}_i}{1+\widehat{\alpha}\widehat{\mu}_i}\right]X\right)^{-1}. \quad (4.4.2)$$

The covariates for the negative binomial regression model for the LOS include age, temperature, blood pressure, oxygen, BMI, eGFR, Braden, CCI, tobacco use, insurance type, and hospital congestion on admission and at discharge. The probability of a readmission within 30-days is estimated using a logistic regression with a similar starting set of covariates as the one identified for modeling the LOS.

The propensity score models are used to predict the probability that a patient would be assigned a COPD pathway on admission to the hospital. The list of factors that could potentially affect pathway assignment includes patient characteristics such as age, vital signs on admission, COPD risk factors such as tobacco use, eGFR, Braden score, and CCI, and external factors, mainly hospital congestion on admission and insurance type.

To develop a propensity score model that balances the measured covariates, I used a structured, iterative approach, described Rosenbaum and Rubin (1984). In the first stage, for each group of controls, I estimate one-variable logistic regression models for main effects, interaction terms, and quadratic terms for continuous variables. Once significant predictors from one-variable models are identified, I fit a logistic regression model that includes only those significant predictors using stepwise variable selection to identify the variables for the propensity score model. In the next step, the control cases are matched to the treated cases, and the matched data set is assessed for the covariate imbalance. If the covariate imbalance is not achieved, insignificant higher order terms are dropped from the model until the new matched data set is balanced.

The DBR estimator uses the results of the regression adjustment model and propensity predicted values from the propensity score model in the augmentation term as described in (2.7.1). The standard error for the DBR estimator of the ATE is calculated using the following expression:

$$SE(DBR) = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N} \begin{matrix} \left(\frac{W_i Y_i}{e(X_i,\hat{\alpha}_X)} - \frac{W_i - e(X_i,\hat{\alpha}_X)}{e(X_i,\hat{\alpha}_X)} m_1(X_i,\hat{\beta}_1) - \right. \\ \left. \left[\frac{(1-W_i)Y_i}{1-e(X_i,\hat{\alpha}_X)} + \frac{W_i - e(X_i,\hat{\alpha}_X)}{\mathbf{1}-e(X_i,\hat{\alpha}_X)} m_0\left(X_i,\hat{\beta}_0\right)\right] - \widehat{DBR}\right)}} \quad (4.4.3)$$

4.5 Estimation Results

The estimation results for the study design with historical controls are shown in Table 4.6. Using both the unmatched sample and the sample matched on the estimated propensity scores, neither negative binomial (NB) regression for the LOS found that the pathway coefficient was significant. The estimated ATE for both samples showed reduction in the LOS (by 2.3 hours

Table 4.6. Estimation Results for the LOS with Historical Controls

	NB model		**NB with PS matching**			
	Estimate	SE	*PS*	*SE(PS)*	Estimate	SE
(Intercept)	4.2630 .	2.5320	-35.63**	12.0258	0.2172	3.6489
Age	0.0003	0.0029	-0.0044	0.0118	-0.0008	0.0045
GenderMale	0.0047	0.0534			-0.0264	0.0784
Insurance1	0.3814 *	0.1565			0.4459*	0.2151
Insurance2	0.3190 *	0.1557			0.3929.	0.2145
Insurance3	0.2722 .	0.1624			0.2223	0.2279
HCa	0.0722	0.4553	12.13***	2.5422	-0.4439	0.7710
HCd	1.2670 **	0.4656			2.0810**	0.6940
Temp	-0.0078	0.0247	0.23*	0.1166	0.0232	0.0346
HR	0.0025 .	0.0013			0.0031	0.0019
BPS	-0.0003	0.0011	-0.01*	0.0054	-0.0005	0.0018
BPD	0.0004	0.0021	0.02*	0.0094	0.0009	0.0030
Oxygen	0.0053	0.0037	-0.03	0.0172	0.0096.	0.0050
BMI	0.0007	0.0030			0.0083.	0.0047
eGFR	0.0000	0.0010	-0.01**	0.0051	0.0014	0.0019
Braden	-0.0419 ***	0.0089	0.28***	0.0558	-0.0302*	0.0149
Tobacco1	0.0755	0.0684			-0.0364	0.1012
Tobacco2	-0.0055	0.0715			-0.2374*	0.1061
charlson	0.0226	0.0180	-0.12	0.0932	0.0429	0.0266
PW1	-0.0187	0.0682			-0.0735	0.0766
ATE	**-2.3**				**-9.1**	

in unmatched sample, and by 9.1 hours in the matched sample), but, again, these results were not statistically significant. The percent change in the LOS based on the confidence interval for the PW coefficient under regression adjustment was between -12% and 14% for the PW group compared to historical controls.

Insurance type was significant in the NB model in both samples, with private insurance, Medicare and Medicaid insurance contributing to a longer LOS in comparison to the uninsured category. The sign of the coefficient for the hospital congestion at discharge was positive suggesting that it increases the LOS (p value < 0.001). The current tobacco use category was significant in the model predicting the LOS in the matched sample. The negative sign for the current tobacco use should not be interpreted as a positive prognostic effect of smoking for the COPD patients and can be explained by the smoke-free campus policy at the UTMC.

Braden score is a variable that requires special consideration. While usually not included in the studies of COPD patients, it was a significant predictor for the LOS in the NB models with historical controls in both samples as well as in the logistic regression model. Omitting the Braden score from the model negative impacted the fit of the model in both samples based on the AICc values. In the unmatched sample, the $AICc_{Braden}$ value of 7,481 and the $AICc_{No\ Braden}$ value of 7,501 were observed indicating that the variable is a significant predictor. Omitting Braden score from the model had a high impact on the magnitude of the ATE, which suggests that it is an important confounder for the LOS and the COPD pathway treatment. Moreover, the Braden score was a better predictor for the LOS of COPD patients than the CCI, based on the AICc values and stepwise variable selection.

Figure 4.2 shows Pearson residuals from the NB model for the unmatched sample with historical controls. The outlier diagnostics identified several observations in the unmatched

sample that had unusually high values for the LOS in both the treated and control groups. Removing the outliers helped improve the model fit, but did not change the significance of the pathway coefficient. Since pathways aim to standardize care for patients with a specific clinical problem, the effects of pathways are expected to be more pronounced among patients that fall under the general guidelines of a treatment protocol as opposed to patients with particularly complex and unique cases. Excluding patients whose hospital stay was over 9 days from the analysis of the COPD pathway effect on the LOS is equivalent to some exclusion restriction used in randomized controlled trials and is instrumental to improving the accuracy of the estimates of the ATE.

After removing the outliers in the LOS, Insurance type was no longer significant, but Braden score and hospital congestion at discharge were still indentified as important predictors by the NB model. The ATE became lower in magnitude (-1.62 hours), and had a much smaller standard error (SE) of 0.0063. The pathway coefficient also remained insignificant after removing the outliers.

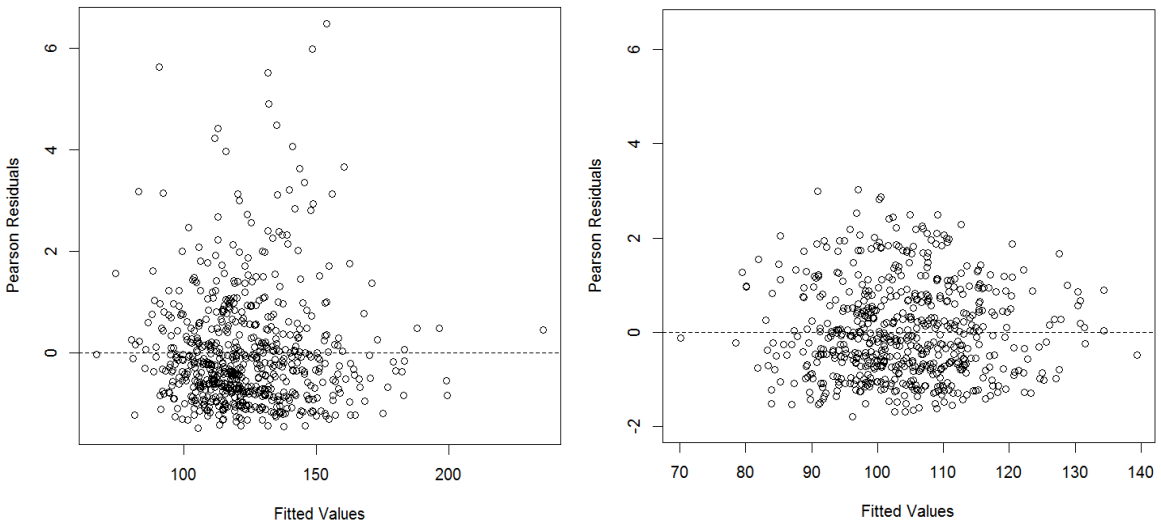


Figure 4.2 Pearson Residuals vs Fitted Values Before and After Removing the Outliers

The impact of the influential points shown in Figure 4.3 was assessed for each model in the study in terms of the model's fit and the estimates of the ATE. The identified points were likely to describe the end-stage COPD patients, and therefore, could potentially contribute to the "washing out" of the pathway effect, given that pathways are designed to treat specific conditions as opposed to managing the end-stage phase of a disease. The magnitude of the estimated pathway effects appeared to be affected by these points, but their significance level was not.

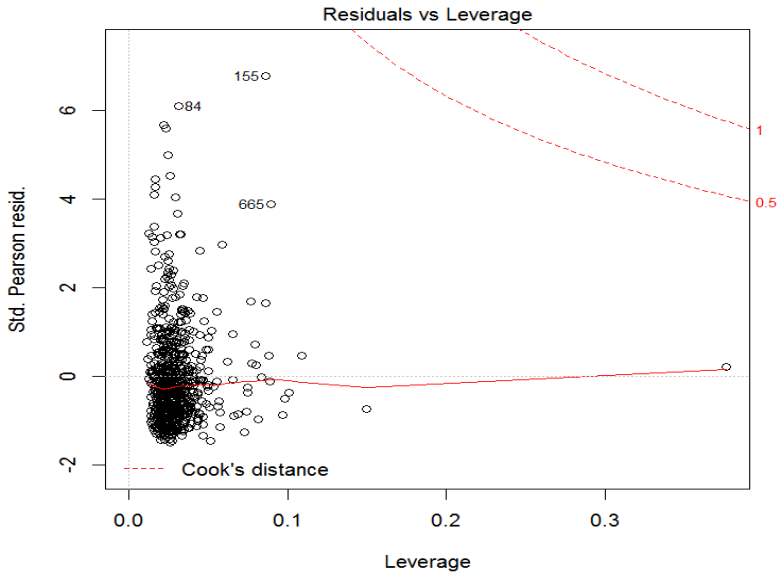


Figure 4.3 Leverage Points from the NB model in the unmatched sample

The propensity score model in Table 4.6 was identified as the one providing the highest degree of the covariate balance improvement. Using the predicted values of propensity scores from the logistic regression described in Table 4.6 and *1-to-4 matching with replacement* improved the overall covariate balance by 65% to 97%. The standardized differences below 10% were observed for all the variables in the matched sample with exception of the Braden score and the hospital congestion on admission and at discharge. Adjusting for those variables in the NB

regression model in the postmatching analysis was necessary to single out the effect of the COPD pathway.

Employing other matching mechanisms, such as matching without replacement and 1-to-1 matching, did not help improve data balancing in the matched samples. The pathway coefficient in the NB models for the matched samples obtained through those matching mechanisms did not reach statistical significance. The estimate of the ATE, while still not statistically significant, exhibited a lot of variation depending on the matching mechanism applied. In *1-to-1 matching with replacement*, the ATE of the COPD pathway was -13.4 hours, with the standard error (SE) of 0.2891. When *1-to-1matching without replacement* was used, the data balance worsened considerably for Oxygen, and the postmatching analysis produced the ATE estimate of -4.2 hours, with the SE of 0.0841. *Matching 1 treated case to 4 control cases without replacement* resulted in a matched sample with different group distributions for Oxygen as well, while yielding the ATE of -8.3 hours, with the SE of 0.0978.

The variation in the magnitude of the ATE estimates can be attributed to the lack of statistical significance. The inability of the models to detect the effect of the COPD pathway on the LOS does not necessarily imply that the pathway was not efficient in reducing the LOS. Given that all the models predicted the same direction of the effect (reduction in the LOS), using a larger sample size for both the control and the treated groups might be beneficial in identifying the effect of the pathway in the future.

Matching without replacement improved the precision of the estimates as the observed standard errors were smaller in both models when matching without replacement was used. The bias-correction properties of the matching mechanisms cannot be verified empirically based on the estimation results when the true ATE is unknown, and require further investigation.

The propensity score model with contemporaneous controls was estimated using the same iterative approach as discussed in Section 4.3. The estimated coefficients for the propensity scores are shown in Table 4.7. Given the small number of control units available for matching, only matching with replacement was considered. Using 1-to-4 matching with replacement resulted in a matched sample with 25 control cases matched to 60 treated cases. Even though the propensity score model increased covariate balance by 88% to 93%, the standardized differences larger than 10% in absolute value still remained for variables Insurance, BPD, Braden score, HCd, and CCI. The distribution of propensity scores in the treated and control groups were slightly different, as more treated cases had high propensity scores greater than 0.8 (Figure 4.4).

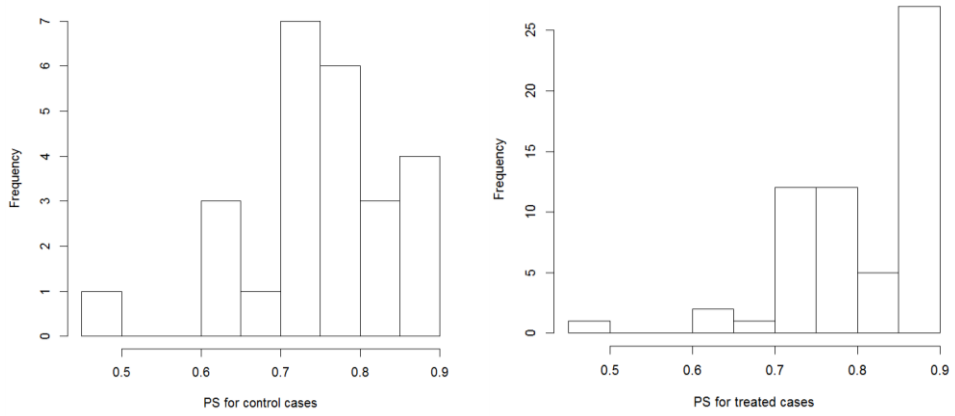


Figure 4.4. Distribution of Propensity Scores for the Contemporaneous Control and Treated Cases in the Matched Sample

The effect of the COPD pathway was negative showing the reduction in the LOS, but was not statistically significant in either sample. The CCI was a stronger predictor for the LOS than the Braden score in the unmatched sample with contemporaneous controls, but lost its significance in the matched sample. Insurance type, Medicare category in particular, and hospital congestion at discharge were significant in the matched sample. The sign of the HCd coefficient

Table 4.7. Estimation Results for the LOS with Contemporaneous Controls

	NB model		**NB with PS matching**			
	Estimate	SE	*PS*	*SE(PS)*	Estimate	SE
(Intercept)	-0.9178	4.5336	-3.150013	2.50947	-3.1687	4.8598
Age	0.0061	0.0071	-0.003996	0.020683	0.0029	0.0077
GenderMale	0.0172	0.1108			0.0500	0.1258
Insurance1	-0.2589	0.5855				0.1178
Insurance2	-0.5448	0.5758			-0.2270.	0.1939
Insurance3	-0.4945	0.5810			-0.1919	1.3291
HCa	0.9318	1.1455			-0.5087	1.2618
HCd	1.2254	1.0878			3.1308*	0.0459
Temp	0.0453	0.0427			0.0799 .	0.0039
HR	0.0040	0.0031			0.0032	0.0030
BPS	-0.0038	0.0026			-0.0038	0.0057
BPD	0.0110 *	0.0046			0.0159**	0.0072
Oxygen	-0.0062	0.0073			-0.0128 .	0.0082
BMI	0.0103 .	0.0062			0.0089	0.0029
eGFR	-0.0017	0.0026			-0.0066*	0.0394
Braden	-0.0362	0.0227	0.2374**	0.086331	-0.0758.	0.1687
Tobacco1	0.1749	0.1464			0.0665	0.1734
Tobacco2	-0.0608	0.1602			-0.1781	0.0464
CCI	0.0683 *	0.0343	-0.2899*	0.130286	0.0319	0.1203
PW1	-0.0819	0.1187			-0.0810	0.3487
ATE	**-10.51**				**-9.75**	

remained the same as in the models with the historical controls, suggesting that the LOS, on average, was likely to be longer when the hospital was operating at high capacity levels. Other factors found to be prognostic for the LOS in the sample of the COPD patients were vital signs on admission and the eGFR index. Higher eGFR values correspond to a better renal function, and the negative sign for the eGFR coefficient in the model for the LOS was consistent with clinical expectations.

The estimates of the ATE did not reach statistical significance in the samples with the contemporaneous controls. The magnitude of the effect was higher compared to the samples with the historical controls. The observed unadjusted mean difference in the LOS for these comparison groups was reported at 12.8 hours, while the ATE obtained using regression adjustment, though not statistically significant, showed the reduction in the LOS of about 10 hours in both the unmatched and matched samples for the pathway group. The confidence interval for the PW coefficient under regression adjustment showed that the percent change in the LOS in the PW group compared to the contemporaneous controls was between -24% and 22%.

The DBR estimator did not perform as well as expected in the analysis of the COPD pathway effects on the LOS exhibiting extreme volatility of the ATE estimates, and, thus, yielding results that were not reliable. The estimates of the ATE obtained with the doubly robust method are presented in Table 4.8. The simulation study presented in Chapter 3 suggests that the accuracy and precision of the DBR estimator is highly sensitive to the small sample sizes and high degrees of imbalance between proportions of treated and control cases. Row-wise deletion of the missing observations in the NB and the PSM models resulted in reduced sample sizes. The sample with the contemporaneous controls had a total of 140 patients, with a low proportion of

control cases (p_C = .27). While the sample size in the design with the historical controls was larger (n_{HC} = 668), the imbalance between the treated and control case was even higher (p_C = .85).

Table 4.8 The Doubly Robust Estimator of the ATE of the COPD Pathway on the LOS

	ATE	SE(ATE)
DR HC missing data	-3.71	20.21
DR CC missing data	-8.21	19.12
DR HC imputed data	-4.00	10.88
DR CC imputed data	-3.07	13.04

While the existing sample imbalances and the small sample size could explain the poor performance of the DBR estimator, the possibility of misspecifying both the regression and the PSM model cannot be entirely ruled out. The bias correction property of the DBR estimator requires that at least one model, either the regression model, or the propensity score model, is specified correctly. In the presence of hidden or overt bias, when an important confounder is either unobservable or not accounted for in the model, both the regression and the PSM models would be misspecified, and the bias correction properties of the DBR estimator would not hold in this scenario.

The logistic regression model for the 30-day readmission rate was estimated only for the sample with historical controls. The contemporaneous controls group did not have sufficient data on readmissions. Table 4.9 shows the number of readmissions for the pathway and the contemporaneous control groups.

The historical control group had a much higher readmission rate compared to the PW group (41% vs 22%), while readmission rates in the PW and contemporaneous control groups

Table 4.9 Number of Readmissions by Treated and Control Groups

	HC	CC	PW
Readmitted = 0	333	31	80
Readmitted = 1	233	7	22
Percent readmitted	41.17%	18.42%	21.57%

were very close (22% vs 18%). Similar readmission rates between the contemporaneous control group and the PW group can be attributed to the PW spillover effect. The COPD PW at the UTMC contains a prompt for an inhaler that many of the COPD patients receive at discharge, and since the same clinicians treated both pathway and non pathway patients, patients in both groups were likely to be discharged with an inhaler.

The pathway coefficient was highly significant in the model for the hospital readmission (p-value < 0.001). The estimates of the ATE from the logistic regression model in both the unmatched and matched samples showed a 14% to 16% reduction in the probability of readmission (34% to 38% improvement in the PW group over historical control group), and had low standard errors (Table 4.10). The DBR estimator for the ATE of the COPD pathway on readmission using the historical controls showed a 20% reduction, with the SE of 0.0618.

The propensity score model identified earlier for the sample with historical controls and described in Table 4.6, was used to create a matched sample for estimating the pathway effect on readmission. The matched sample contained 209 control cases and 95pathway patients.

The estimation results between in the two samples were similar. The estimated coefficient for temperature, heart rate, blood pressure, and eGFR had the same signs and were close in magnitude and significance level in both models. Age and insurance type were strong predictors of the probability of a readmission in the unmatched sample, but did not reach statistical significance in the matched sample.

Table 4.10 Logistic Regression Models for 30-day Readmissions with Historical Controls

	Logistic Regression		**Logistic regression with PSM**	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	13.4117	9.9002	13.7476	876.0895
Age	0.0262 *	0.0109	0.0242	0.0192
GenderMale	-0.0504	0.2015	-0.0024	0.3440
Insurance1	2.8557 *	1.3742	16.0924	875.9258
Insurance2	3.3300 *	1.3762	16.9036	875.9258
Insurance3	3.4460 *	1.3832	16.4677	875.9259
HCa	1.0998	1.6815	5.7200.	3.1968
HCd	-0.6634	1.7270	0.4634	2.8592
Temp	-0.2047 *	0.0960	-0.3563 *	0.1619
HR	0.0161 ***	0.0048	0.0255 **	0.0080
BPS	0.0017	0.0041	0.0016	0.0076
BPD	-0.0176 *	0.0079	-0.0287 *	0.0137
Oxygen	-0.0005	0.0136	-0.0264	0.0222
BMI	0.0177	0.0113	0.0119	0.0209
eGFR	0.0109 **	0.0039	0.0157 *	0.0077
Braden	-0.0309	0.0329	-0.0581	0.0647
Tobacco1	0.2977	0.2497	0.2681	0.4111
Tobacco2	-0.2455	0.2652	-0.6613	0.4312
charlson	0.0691	0.0668	-0.0143	0.1131
PW1	-0.7925 **	0.2788	-0.8392 *	0.3390
ATE	**-0.1583**		**-0.1503**	

4.6 Estimation Results with Imputed Missing Values

In most models, including NB and logistic regression, R automatically excludes all cases in which any of the inputs are missing. This can limit the amount of information available in the analysis, especially if the model includes many inputs with significant numbers of missing values.

The presence of the missing values in the data collected for the COPD patients had a significant impact on the sample size. A total of 221 observations were lost due to the missing values in the estimation process, including 146 patients from the historical control group, 58 pathway patients, and 20 patients from the contemporaneous control group. The variables with the highest number of missing values were Insurance (88 missing values), Tobacco (86), eGFR (54), and BMI (37). Temperature, HR, and Oxygen had 1 missing value each, and two values were missing for the Braden score.

The pattern of the missing values shown in Figure 4.6 and the missing data analysis suggest that the missingness in the data is random, or, in other words, that the probability that a variable is missing depends only on available information. The assumption that missingness is random allows imputing missing values through multiple imputation methods available for analysis of incomplete multivariate data.

Rubin (1987) suggests a Monte Carlo technique for multiple imputations in which the missing values are replaced by $m>1$ simulated versions, where m is typically small (e.g. 3-10). In Rubin's method for repeated imputation inference, each of the simulated complete datasets is analyzed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing-data uncertainty.

The R package “Amelia” implements a new expectation-maximization (EM) with bootstrapping algorithm (Honaker et al., 2011). The algorithm first bootstraps a sample dataset with the same dimensions as the original data, estimates the sufficient statistics by EM, and then imputes the missing values for the sample. It repeats this process *m* times to produce the *m* complete datasets where the observed values are the same and the unobserved values are drawn from their posterior distributions. Honaker et al. assert that their package works faster and with larger numbers of variables than various Markov chain Monte Carlo approaches, and gives essentially the same answers.

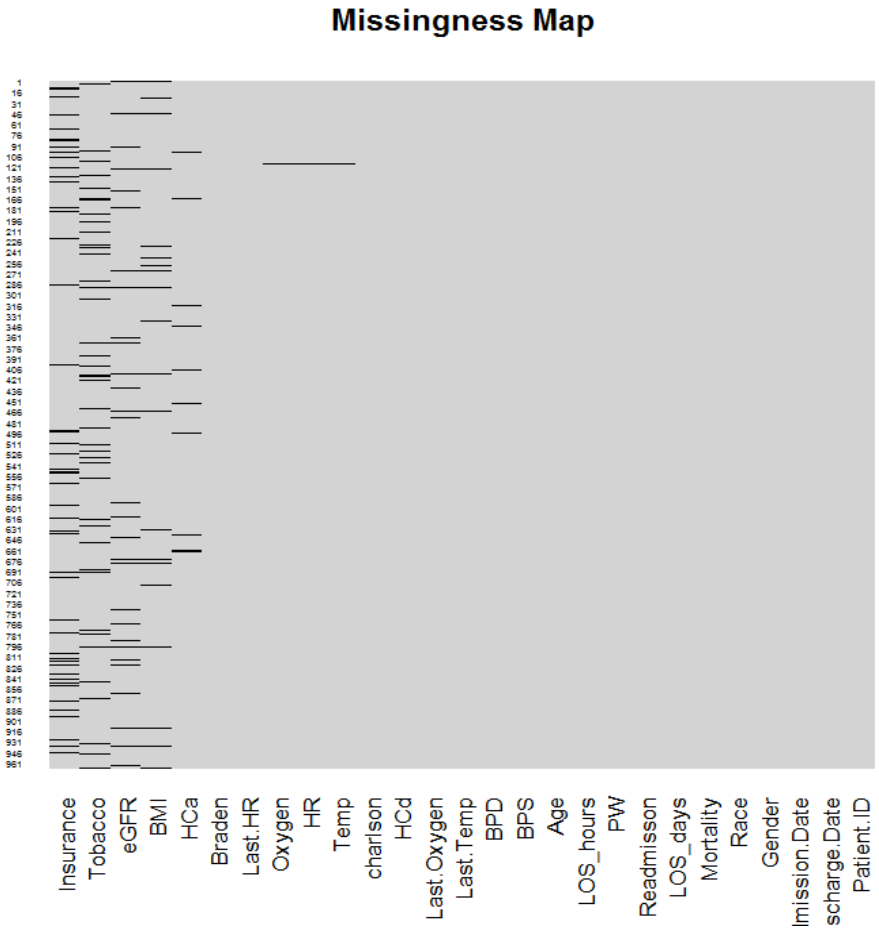


Figure 4.5 Missingness Map

I used “Amelia” package in R to impute the missing values for the variables Insurance, Tobacco, eGFR, BMI, Temp, Oxygen, HR, and Braden, and compared listwise deleted regression results to results pooled from the same regression run on 25 imputed data sets. The estimation results for the LOS with historical controls obtained using imputed missing values are presented in Table 4.11.

The model results with imputed missing values using historical controls were essentially similar to the results obtained when the missing values were deleted. The ATE estimate was still not statistically significant as well as the estimate of the pathway coefficient in both samples. With imputed missing values the regression adjustment in the unmatched and matched samples produced estimates of the ATE that were closer in magnitude (-3.4 and -3.1 hours, respectively), as opposed to the results with the deleted missing values (-2.3 hours and -9.1 hours), but the standard errors of the ATE were much higher. Hospital congestion at discharge, Insurance, Braden score and CCI were identified as significant predictors for the LOS, in a similar manner, but the Tobacco variable was no longer significant.

When contemporaneous controls were used for estimating pathway effects with imputed missing values, the direction of the effect changed. The ATE was 1.6 and 4.8 hours in the unmatched and in the matched sample, respectively, but the SE were unreasonable high (12.0008 for the matched sample).

The performance of the DBR estimator (Table 4.8) improved significantly in larger samples with imputed missing values. The standard errors reduced by almost one half in the sample with historical controls, and by about one third with contemporaneous controls.

Table 4.11 Estimation Results for the LOS with Historical Controls with imputed missing values

	NB Regression		**NB regresson with PSM**	
	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	4.7118 *	2.2609	5.1896	3.1634
Age	-0.0003	0.0025	-0.0012	0.0038
GenderMale	-0.0160	0.0451	-0.0286	0.0663
Insurance1	0.2891 .	0.1581	0.4211 .	0.2421
Insurance2	0.2352	0.1536	0.3519	0.2348
Insurance3	0.2668 .	0.1603	0.3091	0.2511
HCa	0.1658	0.4349	-0.4022	0.7213
HCd	0.8972 *	0.4012	0.9863 .	0.5800
Temp	-0.0086	0.0217	-0.0117	0.0291
HR	0.0026 *	0.0011	0.0027	0.0017
BPS	-0.0011	0.0010	-0.0012	0.0014
BPD	0.0013	0.0018	0.0014	0.0027
Oxygen	0.0055	0.0034	0.0076	0.0050
BMI	0.0000	0.0027	-0.0011	0.0042
eGFR	0.0000	0.0008	-0.0006	0.0015
Braden	-0.0434 ***	0.0077	-0.0398 **	0.0143
Tobacco1	0.0739	0.0617	0.0547	0.0885
Tobacco2	-0.0611	0.0642	-0.0888	0.0917
CCI	0.0372 *	0.0158	0.0303	0.0237
PW1	-0.0281	0.0574	-0.0261	0.0672
ATE	**-3.4124**		**-3.1487**	

The model for readmission with imputed data (Table 4.12) yielded very similar results. The pathway coefficient was negative and highly significant, and the estimate of the ATE was only slightly lower in magnitude (-.12) with a low standard error.

4.7 Discussion of the Results and Limitations of the Study of Pathway Effects

While all the models employed in the current analysis agreed on the direction of the COPD pathway effect showing a reduction in the LOS with exception of the regression adjustment with the imputed data, the ATE for the LOS did not achieve statistical significance in any model. The estimated reduction in the LOS with historical controls was between 2.3 and 8.2 hours, and, while not statistically significant, was consistently reported across the three methods in both the unmatched and matched samples. The estimates of the ATE obtained with contemporaneous controls showed higher variability both in terms of the magnitude and the direction of the effect (fluctuations from -15.6 hours to 1.8 hours), and had higher standard errors compared to the estimates obtained using historical controls.

The effects of the COPD pathway were statistically significant in all the models for 30-day readmission, suggesting a positive effect of the pathway on the probability of a readmission within 30 days. The estimated reduction in the probability of readmission was between 12% and 16%. These estimation results are not unexpected, given that the COPD pathway at the UTMC contains a detailed section on discharge instructions, including a prompt for an inhaler that many COPD patients should receive at discharge, and designates a patient follow-up specialist whose role is to advise patients on follow up care, remind them to refill their prescription, and provide additional assistance.

The bias correction methods used for the analysis of the pathway effects in this study

Table 4.12. Logistic Regression Models for 30-day Readmissions with Historical Controls with Imputed Missing Values

	Logistic Regression			
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.5814	8.5143	0.7730	0.4395
Age *	0.0213	0.0095	2.2514	0.0244
GenderMale	-0.1129	0.1709	-0.6609	0.5087
Insurance1	1.4521	1.2395	1.1716	0.2414
Insurance2	1.9472	1.2398	1.5706	0.1163
Insurance3	2.0260	1.2585	1.6099	0.1074
HCa	1.2415	1.4622	0.8491	0.3958
HCd	0.0728	1.4825	0.0491	0.9608
Temp	-0.1190	0.0814	-1.4606	0.1441
HR ***	0.0143	0.0042	3.3955	0.0007
BPS	0.0011	0.0036	0.3067	0.7591
BPD	-0.0095	0.0068	-1.4003	0.1614
Oxygen	-0.0056	0.0123	-0.4527	0.6508
BMI	0.0123	0.0101	1.2252	0.2205
eGFR **	0.0099	0.0034	2.9304	0.0034
Braden	-0.0482	0.0283	-1.6992	0.0893
Tobacco1	0.3636	0.2408	1.5096	0.1311
Tobacco2	-0.2374	0.2473	-0.9603	0.3369
CCI **	0.1556	0.0577	2.6979	0.0070
PW1 **	-0.5789	0.2225	-2.6011	0.0093
ATE	**-0.12**			

assume that the selection bias is due only to some observable factors, and including these factors in the model should provide unbiased estimates of the ATE. However, proving empirically the tenability of this assumption is not possible, and given the data limitations discussed in Section 4.2, the possibility of overt bias that is not accounted for in the model still remains.

The analysis presented in this dissertation was the first attempt to evaluate a new program in the beginning phase of its implementation, and a lot of adjustments were necessary throughout the study period. One of them was adding the missing compliance tracker as a discrete element to the pathway structure, which was discovered and addressed four months after the pathways had been launched. As a result, the first four months of pathway usage could not be included in the analysis, and the final data set was significantly smaller than expected. Increasing the sample size should improve the precision of the estimates, and would be particularly beneficial for the DBR estimator, that showed significant improvement with imputed data in larger samples.

Other areas of improvement include constructing a new metric for the hospital congestion that incorporate resource utilization in terms of nurse hours as well as hospital capacity levels, refining the existing measures for patient clinical characteristics, such as oxygen levels and comorbidities, and addressing the poor quality reporting that was identified in the analysis.

CONCLUSION

The goal of this dissertation was to create a standardized approach to assessing the impact of the UTMC pathways across all major disease categories and key outcome measures. To accomplish this, I identified models, control factors, and adjustments to correct for potential confounding in pathway assignment and the outcome measures, and provided a case study for one of the largest primary diagnoses areas, chronic obstructive pulmonary disease (COPD). I also addressed the issues of handling missing data and investigated the effects of sample size and sample imbalance on the performance of the bias correction methods. I review these contributions below in brief detail before discussing limitations of the study and a variety of goals for future work in this area.

The widely accepted study designs for evaluating the effects of a treatment or an intervention in medical literature, such as RCT, CBA, and ITS (Rotter et al., 2010), require a set of conditions that are not always available for researchers. The current study of the pathway effects on clinical outcomes was characterized by the absence of randomization for pathway assignment, unavailability of multiple intervention and control sites, and a relatively short study period, and as such, called for a different approach that was identified through an intensive methodological review of the bias correction techniques for observational studies.

I used the following three methods to evaluate the effects of the COPD pathway on the clinical outcomes under the settings described above: regression adjustment, propensity score matching with postmatching regression adjustment, and the doubly robust estimator. The choice of these methods was based on the tenability of model assumptions, the robustness of the models to small sample sizes and sample imbalance, as well as their sensitivity to the model specifications.

In the simulation study presented in Chapter 3, the OLS regression and the PSM model appeared to work well when the selection on observables assumption was satisfied, conditional on the availability of all relevant covariates and their correct specification in the model, and exhibited less sensitivity to model specifications and sample size and imbalance than other methods that were included in the study. The doubly robust estimator was expected to perform well in the presence of overt bias given a sufficiently large sample size (n>500) and a low degree of sample imbalance (the proportion of controls ≤ 0.75).

The effect of the COPD on the LOS was not statistically significant in all three models with both historical and contemporaneous controls. The direction of the effect, though no statistically significant, was consistently reported to be negative by the models employed in the study using historical controls. The estimated ATEs with historical controls showed a reduction in the LOS of 2.3 to 8.2 hours. Higher variability of the estimates observed with the contemporaneous control group can be explained by a smaller sample size available for contemporaneous study design.

The pathway coefficient was statistically significant in the models for the 30-day readmission, and the estimated ATEs showed a reduction in the probability of readmissions between 12% and 16%. The results obtained with imputed missing values were consistent with these findings, showing a reduction of 12% in the probability of a readmission due to pathway usage.

In an attempt to account for all the important observable covariates in the models, the data collected for the study included a rich set of patient clinical and demographic characteristics, vital signs and laboratory values, comorbidities, severity of illness indicators, and

relevant hospital characteristics. The following results pertaining to the choice of confounders and control factors require special consideration:

1) The significance of the Braden score both in the regression model for the LOS and in the propensity score model suggests that it likely to be an important confounder and should be considered as an important covariate in predictive modeling of clinical outcomes and in evaluating treatment effects for COPD patients and other DRGs.
2) Hospital congestion on admission was shown to have a significant effect on pathway assignment, and hospital congestion at discharge was an important predictor for the LOS. These results were consistent with other empirical findings suggesting the importance of including hospital characteristics in the model for estimating the effects of a treatment or an intervention in observational studies (Kelly et al., 2013).
3) As suggested by previous empirical research (Fisher et al., 2001), insurance type was a significant predictor for both the LOS and readmissions, and as such, should be included in future studies of pathway effects.

The models employed in the study are expected to produce unbiased estimates of the ATE provided that the assumption of selection on observables is satisfied. The tenability of this assumption cannot be empirically tested, and thus, the possibility of hidden or overt bias cannot be entirely ruled out. The poor performance of the DBR estimator in the models for the LOS could be attributed to the tenability of model assumptions, and suggests that the results of the regression adjustment and propensity score matching should be interpreted with caution.

Another factor that contributed to the high volatility of the DBR estimator of the pathway effect on the LOS was a small sample size of the contemporaneous study design, and a highly imbalanced sample with historical controls, where the proportion of control cases after row-wise

deletion of the missing values reached 85%. These factors are a likely cause for the loss of accuracy and efficiency in the DBR estimator, according to the results of the simulation study presented in this dissertation.

Several issues that raise questions about the accuracy of the PSM estimates still remain, even if the propensity score model were specified correctly. The PSM model results appeared to be sensitive to the choice of the matching algorithm, caliper width, number of matches specified for each treated case, and to whether matching was done with or without replacement. The details of the matching process require further investigation and are identified as potential areas for future research.

The study design presented in the current work included two control groups, a historical control group and a contemporaneous control group, as an attempt to single out the effects of pathways from other factors that could not be measured. And while potential weaknesses exist in the use of either control group for estimating the untreated potential outcome, both comparison groups should be considered in future analyses as an added protection against pathways spillover effects and the effects of certain changes occurring over time.

Other areas of improvement include increasing the sample size, including nurse hours in a hospital congestion metric, refining the existing measures of patient clinical characteristics, such as oxygen levels and comorbidities, and addressing the issue of missing values in the reports on procedures performed.

Future research for this work will be focused on several key steps in the methodological development. Several issues related to matching, such as the choice of a matching algorithm, should be investigated further. The simulation study presented in this dissertation can be extended to address the degree of the overlap between treated and control cases and its impact of

the performance of the bias correction methods. Another goal is model extension for multilevel treatments to accommodate the analysis of pathway effects when more than one pathway is used, as it is expected to be the case for patients with several comorbidities in the future.

RFERENCES

Abadie, A., Drukker, D., Herr, J. L., & Imbens, G. W. (2004). Implementing matching estimators for average treatment effects in Stata. *Stata journal, 4*, 290-311.

Abadie, A., & Imbens, G. (2002). Simple and bias-corrected matching estimators for average treatment effects: National Bureau of Economic Research Cambridge, Mass., USA.

Armstrong, C. S., Jagolinzer, A. D., & Larcker, D. F. (2010). Chief executive officer equity incentives and accounting irregularities. *Journal of Accounting Research, 48*(2), 225-271.

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in medicine, 25*(12), 2084-2106.

Austin, P. C., Mamdani, M. M., Stukel, T. A., Anderson, G. M., & Tu, J. V. (2005). The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Statistics in medicine, 24*(10), 1563-1578.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61*(4), 962-973.

Barber, B. M., & Lyon, J. D. (1996). Detecting abnormal operating performance: The empirical power and specification of test statistics. *Journal of financial Economics, 41*(3), 359-399.

Barnow, B., Cain, G., & Goldberger, A. (1980). Issues m the analysis of selection bias. *Institute tor Research on Poverty Discussion Papers, 600*.

Bauer, M. S., McBride, L., Williford, W. O., Glick, H., Kinosian, B., Altshuler, L., . . . Sajatovic, M. (2006). Collaborative care for bipolar disorder: part I. Intervention and implementation in a randomized effectiveness trial. *Psychiatric Services, 57*(7), 927-936.

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine, 342*(25), 1878-1886.

Brattebø, G., Hofoss, D., Flaatten, H., Muri, A. K., Gjerde, S., & Plsek, P. E. (2002). Quality improvement report: Effect of a scoring system and protocol for sedation on duration of patients' need for ventilator support in a surgical intensive care unit. *BMJ: British Medical Journal, 324*(7350), 1386.

Chaney, P. K., Jeter, D. C., & Shivakumar, L. (2004). Self-selection of auditors and audit pricing in private firms. *The Accounting Review, 79*(1), 51-72.

Chu, Y.-T., Ng, Y.-Y., & Wu, S.-C. (2010). Comparison of different comorbidity measures for use with administrative data in predicting short-and long-term mortality. *BMC health services research, 10*(1), 140.

Clatworthy, M. A., Makepeace, G. H., & Peel, M. J. (2009). Selection bias and the Big Four premium: new evidence using Heckman and matching models. *Accounting and business research, 39*(2), 139-166.

Cole, M. G., McCusker, J., Bellavance, F., Primeau, F. J., Bailey, R. F., Bonnycastle, M. J., & Laplante, J. (2002). Systematic detection and multidisciplinary care of delirium in older medical inpatients: a randomized trial. *Canadian Medical Association Journal, 167*(7), 753-759.

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine, 342*(25), 1887-1892.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics, 84*(1), 151-161.

Delaney, C. P., Zutshi, M., Senagore, A. J., Remzi, F. H., Hammel, J., & Fazio, V. W. (2003). Prospective, randomized, controlled trial between a pathway of controlled rehabilitation with early ambulation and diet and traditional postoperative care after laparotomy and intestinal resection. *Diseases of the colon & rectum, 46*(7), 851-859.

Doherty, S., & Jones, P. (2006). Use of an'evidence-based implementation'strategy to implement evidence-based care of asthma into rural district hospital emergency departments. *Rural Remote Health, 6*(1), 529.

Dowsey, M. M., Kilgour, M. L., Santamaria, N. M., & Choong, P. (1999). Clinical pathways in hip and knee arthroplasty: a prospective randomised controlled study. *The Medical Journal of Australia, 170*(2), 59-62.

Eichler, M., & Lechner, M. (2002). An evaluation of public employment programmes in the East German state of Sachsen-Anhalt. *Labour Economics, 9*(2), 143-186.

Einhorn, L. H. (2002). Curing metastatic testicular cancer. *Proceedings of the National Academy of Sciences, 99*(7), 4592-4595.

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

Falconer, J., Roth, E., Sutin, J., Strasser, D., & Chang, R. (1993). The critical path method in stroke rehabilitation: lessons from an experiment in cost containment and outcome improvement. *QRB. Quality review bulletin, 19*(1), 8-16.

Faraway, J. J. (2005). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*: CRC press.

Fisher, W. H., Barreira, P. J., Lincoln, A. K., Simon, L. J., White, A. W., & Sudders, M. (2001). Insurance status and length of stay for involuntarily hospitalized patients. *The journal of behavioral health services & research, 28*(3), 334-346.

Funk, M. J., Westreich, D., Weisen, C., & Davidian, M. (2010). Doubly robust estimation of treatment effects. *Analysis of Observational Health Care Data Using SAS*, 85-103.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American journal of epidemiology, 173*(7), 761-767.

Gili, M., Sala, J., López, J., Carrión, A., Béjar, L., Moreno, J., . . . Sánchez, G. (2011). Impact of comorbidities on in-hospital mortality from acute myocardial infarction, 2003-2009. *Revista Española de Cardiología (English Edition), 64*(12), 1130-1137.

Glasziou, P., Chalmers, I., Rawlins, M., & McCulloch, P. (2007). When are randomised trials unnecessary? Picking signal from noise. *BMJ: British Medical Journal, 334*(7589), 349.

Gomez, M. A., Anderson, J. L., Karagounis, L. A., Muhlestein, J. B., & Mooders, F. B. (1996). An emergency department-based protocol for rapidly ruling out myocardial ischemia reduces hospital time and expense: results of a randomized study (ROMIO). *Journal of the American College of Cardiology, 28*(1), 25-33.

Goodney, P. P., Stukel, T. A., Lucas, F. L., Finlayson, E. V., & Birkmeyer, J. D. (2003). Hospital volume, length of stay, and readmission rates in high-risk surgery. *Annals of surgery, 238*(2), 161.

Grunau, G. L., Sheps, S., Goldner, E. M., & Ratner, P. A. (2006). Specific comorbidity risk adjustment was a better predictor of 5-year acute myocardial infarction mortality than general methods. *Journal of clinical epidemiology, 59*(3), 274-280.

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

Hall, S. F. (2006). A user's guide to selecting a comorbidity index for clinical research. *Journal of clinical epidemiology, 59*(8), 849-855.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153-161.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies, 65*(2), 261-294.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics, 30*(1), 239-267.

Honaker, J., Joseph, A., King, G., Scheve, K., & Singh, N. (2001). Amelia: A Program for Missing Data (Windows version) Cambridge, MA: Harvard University.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software, 45*(7), 1-47.

Horwitz, R. I., Viscoli, C. M., Clemens, J. D., & Sadock, R. T. (1990). Developing improved observational methods for evaluating therapeutic effectiveness. *The American journal of medicine, 89*(5), 630-638.

Johnston, S. C., Rootenberg, J. D., Katrak, S., Smith, W. S., & Elkins, J. S. (2006). Effect of a US National Institutes of Health programme of clinical trials on public health and costs. *The Lancet, 367*(9519), 1319-1327.

Kelly, M., Sharp, L., Dwane, F., Kelleher, T., & Comber, H. (2012). Factors predicting hospital length-of-stay and readmission after colorectal resection: a population-based study of elective and emergency admissions. *BMC health services research, 12*(1), 77.

Kelly, M., Sharp, L., Dwane, F., Kelleher, T., Drummond, F. J., & Comber, H. (2013). Factors predicting hospital length-of-stay after radical prostatectomy: a population-based study. *BMC health services research, 13*(1), 244.

Kennedy, P. (2003). *A guide to econometrics*: MIT press.

Kessler, R., & Glasgow, R. E. (2011). A proposal to speed translation of healthcare research into practice: dramatic change is needed. *American journal of preventive medicine, 40*(6), 637-644.

Khwaja, A., Picone, G., Salm, M., & Trogdon, J. G. (2011). A comparison of treatment effects estimators using a structural model of AMI treatment choices and severity of illness information from hospital charts. *Journal of Applied Econometrics, 26*(5), 825-853.

Kollef, M. H., Shapiro, S. D., Silver, P., John, R. E. S., Prentice, D., Sauer, S., . . . Baker-Clinkscale, D. (1997). A randomized, controlled trial of protocol-directed versus physician-directed weaning from mechanical ventilation. *Critical care medicine, 25*(4), 567-574.

Kozier, W. (2008). Erb's. *Fundamental of Nursing: Eight edition New York, Person International Edition*.

Lee, D. S., Donovan, L., Austin, P. C., Gong, Y., Liu, P. P., Rouleau, J. L., & Tu, J. V. (2005). Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. *Medical care, 43*(2), 182-188.

Lee, L.-F. (1978). Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International economic review*, 415-433.

Lin, Y.-K., Chen, C.-P., Tsai, W.-C., Chiao, Y.-C., & Lin, B. Y.-J. (2011). Cost-effectiveness of clinical pathway in coronary artery bypass surgery. *Journal of medical systems, 35*(2), 203-213.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *American psychologist, 48*(12), 1181.

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

Marelich, G. P., Murin, S., Battistella, F., Inciardi, J., Vierra, T., & Roby, M. (2000). Clinical investigations in critical care-Protocol Weaning of Mechanical Ventilation in Medical and Surgical Patients by Respiratory Care Practitioners and Nurses: Effect on Weaning Time and. *Chest, 118*(2), 459-467.

Marrie, T. J., Lau, C. Y., Wheeler, S. L., Wong, C. J., Vandervoort, M. K., Feagan, B. G., & Investigators, C. S. (2000). A controlled trial of a critical pathway for treatment of community-acquired pneumonia. *Jama, 283*(6), 749-755.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods, 9*(4), 403.

Morgan, S. L. (2001). Counterfactuals, causal effect heterogeneity, and the Catholic school effect on learning. *Sociology of education*, 341-374.

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference*: Cambridge University Press.

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

Noseworthy, J. H., Ebers, G. C., Vandervoort, M. K., Farquhar, R., Yetisir, E., & Roberts, R. (1994). The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology, 44*(1), 16-16.

Pearl, J. (2000). Models, reasoning and inference: Cambridge University Press Cambridge, UK:.

Peek, C., Glasgow, R. E., Stange, K. C., Klesges, L. M., Purcell, E. P., & Kessler, R. S. (2014). The 5 R's: an emerging bold standard for conducting relevant research in a changing world. *The Annals of Family Medicine, 12*(5), 447-455.

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., Evans, S. J., & Group, C. (2006). Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *Jama, 295*(10), 1152-1160.

Pildal, J., Chan, A.-W., Hróbjartsson, A., Forfang, E., Altman, D. G., & Gøtzsche, P. C. (2005). Comparison of descriptions of allocation concealment in trial protocols and the published reports: cohort study. *Bmj, 330*(7499), 1049.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-, 19*(1), 1-16.

Roberts, R. R., Zalenski, R. J., Mensah, E. K., Rydman, R. J., Ciavarella, G., Gussow, L., . . . McDermott, M. F. (1997). Costs of an Emergency Department—Based Accelerated Diagnostic Protocol vs Hospitalization in Patients With Chest Pain: A Randomized Controlled Trial. *Jama, 278*(20), 1670-1676.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods, 23*(8), 2379-2412.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*(398), 387-394.

Rosenbaum, P. R. (2002). *Observational studies*: Springer.

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

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics*, 103-116.

Rotter, T., Kinsman, L., James, E., Machotta, A., Gothe, H., Willis, J., . . . Kugler, J. (2010). Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs. *Cochrane Database Syst Rev, 3*(3).

Rotter, T., Kugler, J., Koch, R., Gothe, H., Twork, S., van Oostrum, J. M., & Steyerberg, E. W. (2008). A systematic review and meta-analysis of the effects of clinical pathways on length of stay, hospital costs and patient outcomes. *BMC health services research, 8*(1), 265.

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

Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 543-546.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3-4), 169-188.

Saint, S., Hofer, T. P., Rose, J. S., Kaufman, S. R., & McMahon Jr, L. F. (2003). Use of critical pathways to improve efficiency: a cautionary tale. *The American journal of managed care, 9*(11), 758-765.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods, 13*(4), 279.

Schlüchtermann, J., Sibbel, R., Prill, M.-A., & Oberender, P. (2005). Clinical Pathways als Prozesssteuerungsinstrument im Krankenhaus. *Clinical pathways: Facetten eines neuen Versorgungsmodells*, 43-57.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*(484), 1334-1344.

Shafer, S. M., & Moeller, S. B. (2012). The effects of Six Sigma on corporate performance: An empirical investigation. *Journal of Operations Management, 30*(7), 521-532.

Simonoff, J. S. (2013). *Analyzing categorical data*: Springer Science & Business Media.

Smith, B., Cheok, F., Heard, A., Esterman, A., Southcott, A., Antic, R., . . . Ruffin, R. (2004). Impact on readmission rates and mortality of a chronic obstructive pulmonary disease inpatient management guideline. *Chronic respiratory disease, 1*(1), 17-28.

Smith, H. L. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological methodology, 27*(1), 325-353.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of econometrics, 125*(1), 305-353.

Steiner, P. M. (2010). S. Guo & MW Fraser (2010). Propensity Score Analysis: Statistical Methods and Applications. *Psychometrika, 75*(4), 775-777.

Stephen, A. E., & Berger, D. L. (2003). Shortened length of stay and hospital cost reduction with implementation of an accelerated clinical care pathway after elective colon resection. *Surgery, 133*(3), 277-282.

Steyerberg, E. W. (2008). *Clinical prediction models: a practical approach to development, validation, and updating*: Springer Science & Business Media.

Sulch, D., Perez, I., Melbourn, A., & Kalra, L. (2000). Randomized controlled trial of integrated (managed) care pathway for stroke rehabilitation. *Stroke, 31*(8), 1929-1934.

Tucker, J. W. (2010). Selection bias and econometric remedies in accounting and finance research. *Journal of Accounting Literature, Winter*.

Vandenbroucke, J. P. (2008). Observational research, randomised trials, and two views of medical science. *PLoS medicine, 5*(3), e67.

Wood, L., Egger, M., Gluud, L. L., Schulz, K. F., Jüni, P., Altman, D. G., . . . Sterne, J. A. (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *Bmj, 336*(7644), 601-605.

Zhao, Z. (2004). Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and statistics, 86*(1), 91-107.

APPENDIX

Summary of Analysis Elements

1. Choosing the Bias Correction Methods

When researchers have strong reasons to believe that all confounders and controls are accounted for and included in the model, regression adjustment and propensity score matching are the best candidates for estimating the average treatment effect. Both methods are robust to small sample sizes and sample imbalance in terms of the number of treated and control cases, and outperform other estimators of the ATE when the ignorable treatment assignment assumption is satisfied.

The Heckman treatment effect model is designed to correct for hidden bias, but requires the assumption about a nonzero correlation of error terms in selection and outcome equations and strongly depends on correct model specification. Its sensitivity to model assumptions is more pronounced than that of OLS regression, and therefore, with no definite procedure to test conditions under which the assumptions of the Heckman model are violated, its estimation results should be interpreted with caution.

While the PSM methods and regression adjustment rely heavily on correct model specification, matching methods without propensity scores do not involve estimation of unknown functional forms and are easy to implement, but due to dimensionality problem, their applicability is limited to situations when the number covariates is small. They also appear to be more sensitive to sample imbalances and small sample sizes.

The doubly robust estimator should be used when an important confounder is likely to be omitted from the model to correct for overt bias. Given that the presence of an omitted variable is a strong possibility in many applications, the results of the DBR estimation should be considered together with those obtained through regression adjustment and PSM. The DBR estimator may

offer protection against the overt bias when the proportion of controls does not exceed 0.75 and the sample size is sufficiently large.

2. Common Issues with Matching

Matching with and without propensity scores has several weaknesses. The decision to use matching with or without replacement as well as choosing the number of matches for each treated unit is a tradeoff between precision and bias. Matching with replacement is a better alternative when very few relevant control units are available for comparison. It minimizes the distance between the matched pairs, and is beneficial in terms of bias reduction. Matching without replacement increases bias and can produce results that are sensitive to the order in which the matches are done, but improves the precision of the estimates. 1-to-1 matching produces the smallest distance between the matched pairs and reduces bias. At the same time, the precision of estimates with 1-to-1 matching suffers because large amount of information available from the data is discarded in the process.

Matching algorithms commonly used in propensity score matching are nearest neighbor with caliper, kernel and interval matching. Morgan and Winship (2007) demonstrate that the choice of the matching algorithm affects the estimation results when everything else is held constant. The performance of these matching algorithms remains debatable, with little evidence as to which algorithm is more efficient in particular settings.

3. Study Design

To approximate randomization conditions, observational studies should be designed with enough rigor by adopting the principles of experimental design. Identifying the control and treated groups, zero time for determining patient's eligibility and base-line features, using inclusion and exclusion criteria similar to those in clinical trials, adjusting for differences in

base-line susceptibility to the outcome are important elements of a successful study design in clinical settings.

VITA

Anna V. Romanova is a Ph.D. candidate and graduate teaching associate in Business Analytics and Statistics at the University of Tennessee. She received B.S. degrees in Economics and Linguistics from Kurgan State University in Russia in 1999. Anna then worked as an economist in the defense industry in Russia. In 2002 Anna returned to school to pursue her M.S. in economics and later on in statistics at the University of Tennessee. In 2009, Anna began work on her Ph.D. in Statistics. Anna has taught several semesters of Microeconomics, Macroeconomics, Monetary Theory, and Introduction to Statistics while working on her graduate degrees at UT, and worked on several research projects with the Oak Ridge National Laboratory and the UT Medical Center. After obtaining a Ph.D., Anna will move to Charlotte, NC where she will serve as an Assistant Professor of Quantitative Methods at Winthrop University.