THE USE OF PROPENSITY SCORE METHODS TO ADDRESS CONFOUNDING BY PROVIDER

Bradley Gordon Hammill

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2015

Approved by:

Amy H. Herring

John S. Preisser

Gary G. Koch

Matthew L. Maciejewski

Sean M. O'Brien

ABSTRACT

Bradley Gordon Hammill: The Use of Propensity Score Methods to
Address Confounding by Provider
(Under the direction of Amy H. Herring)

For research questions regarding the real-world effectiveness and safety of medical therapies and devices, researchers must often rely on observational data. Unlike controlled clinical trials, the assignment of treatment to patients in routine medical practice is not randomized. One class of methods used extensively by researchers to address this selection problem is propensity score methods. The role of the healthcare provider has not typically been accounted for when propensity score methods are employed, despite the fact that provider, by imparting an effect on both patient-level treatment assignment and patient-level outcomes, is a potential confounding factor.

When a healthcare provider has measurable impacts on both a patient's treatment assignment and their downstream outcomes, simulation results demonstrated that not accounting for these provider effects could lead to biased estimates of treatment effect when using propensity score methods. This was true specifically when a provider's direct effect on treatment was correlated with their effect on outcome; a situation that occurs when providers having better patient outcomes use therapies at higher (or lower) rates than other providers. Propensity score methods that incorporated provider were able to control this error.

Even when provider effects on treatment and outcome were uncorrelated, it was still important to account for provider in the propensity score treatment model. Failure to do so resulted in confidence intervals around the estimated treatment effect that were either substantially too wide or too narrow, depending on the estimation methods used.

A criticism of typical 1:1 propensity score matching, whether stratified by provider or not, is that the data from many patients are not utilized in the outcomes analysis. Full matching addresses this issue by optimally assigning all treated patients and all comparison patients into variably-sized matched sets. The result is closer matches between study groups than those obtained by other matching methods. Full matching is not currently utilized frequently because it is difficult to implement. A macro to perform full matching by leveraging SAS optimization procedures is presented.

To my wife, Azot, and son, Sebastian,
for their constant love and support.

To everyone for their patience.

TABLE OF CONTENTS

LIST OF TABLES ix

LIST OF FIGURES xi

LIST OF ABBREVIATIONS xii

CHAPTER 1: INTRODUCTION AND LITERATURE REVIEW 1

Background 1

Propensity Score Methods 2

Theory 2

Estimation and Application of Propensity Scores 4

Assessing Covariate Balance 10

The Role of the Provider in Clinical Research 11

Provider Effects on Treatment and Outcome 11

Provider as a Confounding Factor 13

Accounting for Provider in Propensity Score Methods 15

Prior Research on Clustering and Propensity Score Methods 17

Objectives of the Current Research 21

CHAPTER 2: DATA CHARACTERISTICS AND THE PERFORMANCE OF PROPENSITY SCORE METHODS IN THE PRESENCE OF CONFOUNDING BY PROVIDER 24

Introduction 24

Simulation Study 25

Data Generation Process 26

Simulation Scenarios 29

Propensity Score Methods 30

Metrics 31

Simulation Study Results 32

Clinical Example 34

Clinical Example Results 36

Discussion 38

CHAPTER 3: STANDARD ERROR OF TREATMENT EFFECT ESTIMATES FROM PROPENSITY SCORE METHODS IN THE PRESENCE OF CONFOUNDING BY PROVIDER 51

Introduction 51

Simulation Study 52

Data Generation Process 52

Simulation Scenarios 55

Propensity Score Methods 55

Estimation Methods 57

Metrics 63

Simulation Study Results 64

Clinical Example 67

Clinical Example Results 71

Discussion 74

CHAPTER 4: A SAS MACRO FOR OPTIMAL MATCHING AND FULL MATCHING ON PROPENSITY SCORES 89

Introduction 89

Matching Methods 91

SAS Macro Details 100

Clinical Example..102

Clinical Example Results 104

Discussion 109

CHAPTER 5: CONCLUSION 125

APPENDIX 1: MISCELLANEOUS SAS CODE FOR ESTIMATING
PROPENSITY SCORES, CALCULATING WEIGHTS, PERFORMING
MATCHING, AND ESTIMATING APPROPRIATE TREATMENT
EFFECT ESTIMATES 130

APPENDIX 2: SAS CODE FOR THE FULL MATCHING MACRO, FMATCH.SAS......138

REFERENCES 152

LIST OF TABLES

Table 2.1 - Simulation results for inverse probability of treatment weighting methods. Mean estimated treatment difference, relative bias, variance, and mean squared error of treatment effect estimators for each scenario by cluster size distribution. True treatment difference = 2.0. 46

Table 2.2 - Simulation results for propensity score matching methods (not conditional on cluster). Mean estimated treatment difference, relative bias, variance, and mean squared error of treatment effect estimators for each scenario by cluster size distribution. True treatment difference = 2.0. 47

Table 2.3 - Simulation results for propensity score matching methods (within cluster). Mean estimated treatment difference, relative bias, variance, and mean squared error of treatment effect estimators for each scenario by cluster size distribution. True treatment difference = 2.0. 48

Table 2.4 - Characteristics of ADHERE-HF patients receiving low-dose and high-dose intravenous diuretics 49

Table 2.5 - Estimated relative risk and 95% confidence interval for association between receipt of high-dose diuretic and in-hospital mortality by propensity score method and treatment model specification..50

Table 3.1 - List of Estimation Methods Utilized 79

Table 3.2 - Simulation results for propensity score matching methods on data generated without cluster effects, by match type and estimation method. True treatment difference = 2.0. 80

Table 3.3 - Simulation results for inverse probability of treatment weighting methods on data generated without cluster effects, by match type and estimation method. True treatment difference = 2.0..81

Table 3.4 - Simulation results for propensity score matching methods. Mean estimated treatment difference and mean squared error, by match type and estimation method. True treatment difference = 2.0. 82

Table 3.5 - Simulation results for propensity score matching methods. Mean width of 95% confidence interval and coverage probability, by match type and estimation method. 83

Table 3.6 - Simulation results for inverse probability of treatment weighting methods. Mean estimated treatment difference and mean squared error, by estimation method. True treatment difference = 2.0...84

Table 3.7 - Simulation results for inverse probability of treatment weighting methods. Mean width of 95% confidence interval and coverage probability, by estimation method. 85

Table 3.8 - Characteristics of ADHERE-HF patients receiving milrinone and dobutamine....86

Table 3.9 - Estimated relative risk and 95% confidence interval from inverse probability of treatment weighting methods for association between milrinone (vs. dobutamine) and in-hospital mortality. 87

Table 3.10 - Estimated relative risk and 95% confidence interval from propensity score matching methods for association between milrinone (vs. dobutamine) and in-hospital mortality. 88

Table 4.1. - Network characteristics associated with optimal matching and full matching, performed with and without stratification....119

Table 4.2 - Ratio bounds for full matching associated with different combinations of thinning caps, thickening caps and treatment group proportions 120

Table 4.3 - Patient characteristics included as independent variables in the treatment model 121

Table 4.4 - Metrics describing match results, by matching method 122

Table 4.5 - Minimum and maximum treated-to-comparison ratios in matched sets resulting from full matching 123

Table 4.6 - Standardized differences for selected covariates between matched treatment and control groups, by matching method used 124

LIST OF FIGURES

Figure 1.1 - Potential relationships between treatment (A), outcome (Y), measured patient factors (X), unmeasured patient factors (U), and healthcare provider (P).....23

Figure 4.1 - Network representation of unstratified matching. Panel A shows the nodes and available links. Nodes include a flow source (α), a flow sink (β), treated patients (1–5) and comparison patients (a–e). Gray lines indicate potential links between nodes. Panel B shows an example solution for optimal matching. Panel C shows an example solution for full matching......112

Figure 4.2 - Network representation of stratified matching. Panel A shows the nodes and available links. Nodes include a flow source (α), a flow sink (β), treated patients (1–5) and comparison patients (a–e). Gray lines indicate potential links between nodes. Panel B shows an example solution for optimal matching. Panel C shows an example solution for full matching.....113

Figure 4.3 - Network representation of matching with calipers. Panel A shows the nodes and available links. Nodes include a flow source (α), a flow sink (β), treated patients (1–5) and comparison patients (a–e). Gray lines indicate potential links between nodes. All potential links shown between patients in the treatment (T) and comparison (C) groups are those with distance less than the given caliper width. Panel B shows an example solution for optimal matching with calipers, which required the use of an excess node (ε)......114

Figure 4.4 - Proportion of differences on the matching variable greater than 0.2SD within matched sets at each site. Panel A shows these proportions by number of treated patients at the site. Panel B shows these proportions by number of total patients at the site, with sites additionally categorized by number of treated patients.....115

Figure 4.5 - Sample macro call for the full matching macro, FMATCH.SAS.....116

Figure 4.6 - Sample records from the link-level output dataset generated by the full matching macro, FMATCH.SAS.....117

Figure 4.7. Sample records from the record-level output dataset generated by the full matching macro, FMATCH.SAS.....118

LIST OF ABBREVIATIONS

ADHERE	Acute Decompensated Heart Failure Registry
CI	Confidence interval
ETT	Effect of treatment on the treated
GEE	Generalized estimation equations
GLM	General linear model
HF	Heart failure
ICC	Intraclass correlation
IPTW	Inverse probability of treatment weighting
PS	Propensity score
SD	Standard deviation

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Background

For research questions regarding the real-world effectiveness and safety of medical therapies and devices, researchers must often rely on observational data. Unlike controlled clinical trials, the assignment of treatment to patients in real world settings is not randomized. This differential selection of patients to treatment leaves analyses susceptible to confounding, which can result in biased effect estimates unless properly addressed.

One class of methods used extensively by researchers to address this treatment selection problem is propensity score methods (Austin, 2008; Sturmer et al., 2006). The goal of propensity score methods, in short, is to balance confounding factors between the treated and comparison groups (Rosenbaum & Rubin, 1983). Assuming all confounding factors are measured, this balance leads to consistent estimates for the effect of the treatment on the response.

The role of the healthcare provider has not typically been accounted for in clinical research when propensity score methods are employed. This despite the fact that provider, by imparting an effect on both patient-level treatment assignment and patient-level outcome, is a potential confounding factor. In analyses that utilize propensity score methods for data scenarios where providers act as confounder factors, important questions remain about how best to incorporate provider into the analysis and about the costs of ignoring provider in the analysis.

In this chapter, we review the theoretical basis of propensity scores and describe how they are typically used in analyses. We discuss the issues surrounding providers in healthcare research and possible extensions of propensity score methods to address the problem of confounding by provider. And we will examine prior research that has addressed similar ideas about clustered data within propensity score analyses.

Propensity Score Methods

Theory

The theoretical justification for propensity score methods is based on the Rubin causal model (Rubin, 1974). In this model, it is supposed that every experimental unit has multiple potential outcomes, one for each experimental condition. Suppose A represents a point exposure, with $A = 1$ when the subject is exposed to the experimental treatment of interest ("treatment") and $A = 0$ when the patient is not exposed to the experimental treatment ("comparison"). Whether the comparison group is simply unexposed to the experimental treatment or exposed to an alternative treatment is not a critical distinction, theoretically. For each subject, there is an outcome associated with each condition. We let Y_1 denote the outcome the subject would have experienced if they received treatment and Y_0 denote the outcome the subject would have experience if they did not receive treatment. The outcome, Y, we actually observe for each subject corresponds to one of these potential outcomes. For subjects in the treatment group, $Y = Y_1$and for subjects in the comparison group, $Y = Y_0$. Even though we can only observe one of these outcomes per subject, we rely on the existence of the other unobserved outcome, the counterfactual, to draw conclusions about causation.

If it were possible for all subjects in the population to belong to both the treatment and comparison groups, the population average treatment effect, Δ, would be easily estimated as the average of each subject's individual treatment effect, as $E(Y_1 - Y_0) = E(Y_1) - E(Y_0) = \mu_1 - \mu_0 = \Delta$, where μ_0 and μ_1 are population outcomes for the comparison and treatment conditions. Instead, what is estimable is dependent on the treatment actually received, or $E(Y_1|A = 1) - E(Y_0|A = 0)$ which typically does not equal Δ due to confounding. Meaning, if subjects have known values for covariates $\boldsymbol{X}$ that affect outcomes, and treatment is assigned with respect to those covariates, then $E\{E(Y_1|A = 1, \boldsymbol{X})\} \neq E(Y_1)$. Randomized trials are able to properly estimate the average treatment effect by assigning treatment independent of a subject's covariates, and therefore, potential outcomes. Formally, in these cases, $(Y_1, Y_0) \perp A|\boldsymbol{X}$, where $\perp$ indicates independence.

Rosenbaum and Rubin (1983) introduced the propensity score as a balancing score which would lead to this same conditional independence when formal randomization was not present. The propensity score $e(\boldsymbol{X}) = P(A = 1|\boldsymbol{X})$ is the probability of receiving treatment for a particular set of covariate values. Use of the propensity score requires sufficient conditions. First, assuming $\boldsymbol{X}$ includes all confounders, then $(Y_1, Y_0) \perp A|e(\boldsymbol{X})$, meaning all potential responses are conditionally independent of the treatment given the measured variables. This is the condition of no unmeasured confounding. Second, there must be a non-deterministic probability of receiving each treatment at all values of the measured variables, or $0 < P\big(A = a|e(\boldsymbol{X})\big) < 1$. When these both hold, then $E(Y_1|A = 1, e(\boldsymbol{X})) = E(Y_1|e(\boldsymbol{X}))$ and $E(Y_0|A = 0, e(\boldsymbol{X})) = E(Y_0|e(\boldsymbol{X}))$ which allows unbiased estimation of the treatment effect. Different propensity score methods achieve this in slightly different ways, as will be discussed below.

Estimation and Application of Propensity Scores

The propensity score is a subject-specific probability of treatment. It is usually estimated with a logistic regression model having treatment as the dependent variable and other measured factors as the independent variables. Prior research has demonstrated that the most important factors to include in this treatment model are those that confound the relationship between treatment and outcome. If important confounders are not included, the eventual estimate of the treatment effect will be biased (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006). Inclusion of other factors, those related only to the treatment or only to the outcome, may be helpful, but can lead to fewer matches being made, if propensity score matching is used, and may result in a treatment effect estimate with reduced efficiency (Austin et al., 2007; Brookhart et al., 2006; Bhattacharya & Vogt, 2007).

A number of methods have been proposed for utilizing propensity scores in ways that induce the covariate balance between treatment groups that is so critical. Three were proposed by Rosenbaum and Rubin (1983) in their manuscript that initially described the basis and use of propensity scores. These methods are stratification (or subclassification), model-based adjustment, and matching. The newest application of propensity scores is inverse probability of treatment weighting, proposed by Robins, Hernán, and Brumback (2000) and Hirano and Imbens (2001). Propensity score matching and inverse probability of treatment weighting are the two methods most commonly applied in clinical research today and will be discussed in more detail than the other two methods.

Stratification, or subclassification, on the propensity score is done by first creating equally sized strata of subjects based on quantiles of the estimated propensity score distribution, then estimating the treatment effect within each stratum (Rosenbaum & Rubin, 1984). These individual stratum-specific estimates may be combined using Mantel-Haenszel methods to arrive at an overall treatment effect estimate. Quintiles are often used to define the group, because Cochran (1968) showed that five groups is often adequate to reduce 90% of the bias for many distributions. For stratification to yield accurate treatment effect estimates, there needs to be balance on the covariates between treated and comparison groups within each stratum. This should happen because propensity scores should be similar between the treated and comparison groups within each of these strata. Stratification has the advantage of utilizing all subjects in the data. It is also very easy to implement and balance between study groups within the strata is easy to assess. In practice, however, it has been demonstrated that in the extreme strata, containing the highest and lowest propensity score estimates, there is often residual imbalance between study groups that results in poor performance of the resulting effect estimates (Austin et al., 2007).

Model-based adjustment involves simply replacing the covariates in a traditional regression model with the estimated propensity score or some function of the estimated propensity score, such as the linear predictor from the treatment model. While this method is simple to implement and uses data from all available study subjects, it does not allow for evaluation of covariate balance between study groups. It is also the method with the weakest theoretical basis. Rosenbaum and Rubin (1983), themselves, urge caution using this method and list several scenarios common within observational data analysis where it may perform poorly. For example, if the covariance matrices, for the observed covariates, are not equal between the study groups then this method can lead to bias in estimates of treatment effect.

Propensity score matching seeks to create covariate balance between the treated and comparison groups more directly than stratification, by matching individual patients from each group to each other. The goal of any matching scheme is to identify appropriate comparison patients for all treated patients such that the only difference between the group of treated patients and comparison patients, after matching, is the treatment itself. This independence between treatment assignment and potential outcomes is required for the causal model presented earlier to be valid, as:

$$E\{E(Y_1|A=1, e(\boldsymbol{X})) - E(Y_0|A=0, e(\boldsymbol{X}))\} = E\{E(Y_1|\, e(\boldsymbol{X})) - E(Y_0|\, e(\boldsymbol{X}))\}$$

$$= E(Y_1 - Y_0)$$

If it were possible to find exact matches for all treated patients on all covariates, that would be ideal. However, given the number of covariates typically used in clinical research, that goal is often unattainable. Matching on the propensity score, or some function of the propensity score, instead of on individual covariates allows for multivariate matching through the use of a scalar balancing score.

The goal listed above contain a few, sometimes competing ideas. First, the distance between covariates for individual sets of matched patients should be as minimal as possible. Second, the total distance between covariates for the matched groups should be as minimal as possible. And third, all treated patients should be matched. Regarding the first two, there are many ways to specify how propensity score matching is accomplished (Rosenbaum & Rubin, 1985a), some that prioritize patient-level distance and some that prioritize group-level distance. These will be discussed below. But the third idea is also important. Incomplete matching occurs when there are treated patients without suitable matches from the comparison group. Rosenbaum and Rubin (1985) demonstrated that incomplete matching opens the door for bias in the resulting effect estimates, especially if the response curves for each study group are not parallel across all levels of the propensity score.

The most commonly used matching method is greedy matching. To perform a greedy match, patients in the treatment group are matched, one by one, to the closest matching patient in the comparison group. Once a match is made, both patients are removed from the pool of eligible patients used for matching and there is no reconsideration of the complete matches. An alternative to greedy matching is optimal matching. Instead of prioritizing close matches at the patient-level, optimal matching methods seek to minimize the total distance between the treated and comparison groups among the matched patients. This process is more intensive computationally in that it is iterative and does not often have a closed-form solution. Research has found some advantage of using optimal matches to achieve balance, although greedy matching has been found to balance study groups adequately (Gu & Rosenbaum, 1993; Rosenbaum, 1989).

Other issues related to the mechanics of matching include the use of calipers (D'Agostino, 1998) and the number of treated and comparison patients in the matched sets. Using calipers when matching means that two patients whose propensity scores are farther apart than the set caliper size cannot be matched. There is some evidence to indicate that this may be beneficial for balance when any of the covariates to be matched are continuous (Austin, 2011), although this usually means that some treated patients will not have any eligible comparison patients for matching. And while most matched sets include one treated and one comparison patient, matching multiple comparison patients to a single treated patient in a fixed *m*:1 ratio has shown mixed results. Austin (2010) found that matching multiple patients from the comparison group to each treated patient, if sample size allows, may lead to better efficiency of the treatment effect estimate. Hansen (2004), on the other hand, has shown that the use of multiple comparison patients per treated patients can actually introduce substantial imbalance on covariates between the matched sets. As a more flexible option, Gu and Rosenbaum (1993) showed that "full" matching—using all records in the data and allowing for unequal sized matching sets—may be an even better strategy.

Once a matched sample has been created, the outcomes analysis proceeds using usual methods. Analysis with unequally sized matched sets, resulting from methods like "full" matching, require conditional statistical methods, accounting for the matched sets. In fact, some insist that conditional statistical methods should be used in all cases (Austin, 2008), while others disagree that it is essential for 1:1 or 1:*m* matching (Hill, 2008). Advantages of propensity score matching include ease of implementation and analysis. As usually performed, disadvantages include reduced sample size that is representative of the treated population, not the overall population. Estimates generated from this method are not the average treatment effect of the population, but the average treatment effect among the treated, which can be a different quantity.

Propensity score weighting is more appropriately called inverse probability of treatment weighting (IPTW). Each patient is weighted by the inverse of the estimated probability that they would have, based on their covariate, received the treatment they were assigned. For treated patients, this weight is just the inverse of the estimated propensity score. For comparison patients, this weight is the inverse of 1 minus the propensity score, as:

$$w_i = \frac{1}{I(A_i = 1)\widehat{e_i}(\boldsymbol{X}) + I(A_i = 0)(1 - \widehat{e_i}(\boldsymbol{X}))}$$

This weighting creates pseudo-populations in which the covariates are no longer associated with the outcome. There is a direct relationship between these weights and post-stratification weights based on Horvitz-Thompson estimators in the survey sampling literature (Horvitz & Thompson, 1952). The weight above is based on a point exposure. IPTW methods can be extended for exposures that vary over time. If appropriate time-varying covariates are also available, the marginal structural model (Robins et al., 2000) is one such extension that entails re-estimation of the propensity score and reweighting at multiple time points during a follow-up period.

Once weights are assigned, they can be used within usual analytic methods. Because the weights are based on estimated propensity scores that have their own variance, however, standard errors should be estimated appropriately, using bootstrapping or derived formulas (for example, Lunceford & Davidian, 2004). An advantage of IPTW is that it utilizes all study subjects. Unlike matching, this means the estimand associated with IPTW is the average treatment effect, so the results are generalizable back to the source population that generated the treated and comparison patients. Of course, because all patients are utilized, problems can arise if there are values of the propensity score for which no treated (or comparison) patients can be found, as this violates one of the basic assumptions of propensity score analysis. A disadvantage of IPTW is that results are subject to extreme weights, which occur when patients who are very likely to be treated are not treated and vice versa. Stabilized versions of the weight may alleviate some of these problems, by rescaling all weights around 1.0 to prevent very large weights from affecting the calculations and results.

Assessing Covariate Balance

For valid inference, all propensity score methods aim to make the treatment conditionally independent of the potential outcomes given measured covariates. One way to check the success of the propensity score method chosen is to measure the covariate balance between the treatment and comparison groups after matching on the propensity score (or on some function of the propensity score) or after weighting by the inverse probability of treatment.

The measure most frequently recommended for balance is the standardized difference, (Rosenbaum & Rubin, 1985a). The standardized difference is a metric for determining the distance between two samples for a given covariate free of the effects of sample size. For continuous variables, the standardized difference, *d*, is defined as:

$$d = \frac{(\bar{x}_T - \bar{x}_C)}{\sqrt{\frac{s_T^2 + s_C^2}{2}}}$$

Note that this difference depends only on the sample means and a pooled estimate of standard deviation. The analogous measure for dichotomous variables (Austin 2009) is often given as:

$$d = \frac{(p_T - p_C)}{\sqrt{\frac{p_T(1 - p_T) + p_C(1 - p_C)}{2}}}$$

Variables are usually assessed one at a time and said to be balanced if the standardized difference is less than 0.10 between groups. It is often helpful to show the standardized difference in both the original sample and the matched or weighted sample, to demonstrate the reduction in imbalance that resulted from the specific propensity score methods being utilized. Standard hypothesis testing is not recommended for the assessment of balance since the sample size in the matched cohort or weighted cohort could vary from the original sample, resulting in potential differences in statistical significance due to sample size alone.

It may be more appropriate to assess the distance between the matched (or weighted) treatment and comparison samples across all measured variable simultaneously, instead of a single variable at a time. A multivariate distance metric like the Mahalanobis distance,

$$d = \sqrt{(\mathbf{x}_{\boldsymbol{T}} - \mathbf{x}_{\boldsymbol{C}})^{T}\mathbf{S}^{\mathbf{-1}}(\mathbf{x}_{\boldsymbol{T}} - \mathbf{x}_{\boldsymbol{C}})}$$

based on the mean vectors $\mathbf{x}_T$ and $\mathbf{x}_C$ for the treated and comparison groups, respectively, and the pooled covariance matrix **S** (Mahalanobis, 1936), would result in a scalar that could be used for this purpose. Currently, however, there is no guidance on the use of such a measure within propensity score-based analyses.

The Role of the Provider in Clinical Research

Provider Effects on Treatment and Outcome

In clinical practice, a provider is the individual or collection of individuals that provides healthcare to patients. This could be an individual physician, a physician practice, a clinic, or even a hospital. Practically, in research evaluating specific healthcare treatments, providers are often whatever identifiable unit available in the data is most proximate to the assignment of that treatment to the patient. Substantial evidence exists to demonstrate that providers can have profound effects on both treatment assignment and outcomes.

The Dartmouth Health Atlas is a primary source of information about how the likelihood of treatment varies by location, beyond that which would be expected by patient characteristics alone. They note that some care is preference-sensitive—varying because of physician preferences for different alternative treatments—and some care is supplysensitive—varying because of differences in physician availability or technological capacity (Wennberg, 2002). Recent research based on Dartmouth Health Atlas data has demonstrated that there is, often substantial, regional variability in rates of joint replacement procedures (Fisher et al., 2010), interventional carotid procedures (Goodney et al., 2010), and even prescription drug utilization (Munson et al., 2013).

Diffusion of technological and pharmaceutical innovations is another factor in the differences between treatment rates by provider. Research has suggested that certain provider factors are associated with adoption of newer treatments. For example, providers associated with an oncology research network were more likely to implement novel diagnostic procedures (Carpenter et al., 2011); larger hospitals and teaching hospitals were more likely than other hospitals to adopt robotic surgical technology (Barbash et al., 2014); and prescription of novel schizophrenia drugs varied by location in ways that correlated with the distribution of ethnic minorities (Horvitz-Lennon, Alegría, & Normand, 2012). All of these mechanisms of diffusion affect the probability that a patient will receive the treatment and may also be relevant for patient outcomes.

There is also abundant research demonstrating that providers have effects on outcomes. Some of the earliest organized programs to publicly report provider quality was undertaken by cardiac surgeons. Patients in New York (New York State Department of Health, 2012), Massachusetts (Massachusetts Department of Public Health, 2013), and Pennsylvania (Pennsylvania Health Care Cost Containment Council, 2013), among others, have access to reports detailing physician-specific risk-adjusted mortality rates associated with common cardiovascular surgical procedures. Similarly, the Hospital Compare program was initiated by the Centers for Medicare and Medicaid Services (web) to report on and draw attention to hospital-level differences in outcomes experienced by patients hospitalized for specific conditions. Published reports based on these data demonstrate the hospital-level variability in short-term mortality and readmission outcomes, after controlling for patient factors, among patients admitted for pneumonia, heart failure and acute myocardial infarction (Bernheim et al., 2010; Krumholz et al., 2009; Lindenauer et al., 2009). And the Dartmouth Health Atlas, in addition to documenting different in treatment rates, reports differences in outcomes by geography (Goodman, Fisher, & Chang, 2011).

Health services researchers have delved into reasons why certain providers may have different outcome profiles. They have found differences associated with whether or not the hospital was located in an urban or rural setting (Casey, Burlew, & Moscovice, 2010; Goldman & Dudley, 2008); whether or not the hospital was a teaching facility (Shahian et al., 2012); whether or not the hospital was considered a safety net hospital treating primarily uninsured or underinsured patients (Ross et al., 2007); and to what extent the hospital invested in major medical equipment and information technology (Coye & Kell, 2006). Even within these broad categories of hospitals, however, there was still broad variation.

Provider as a Confounding Factor

Figure 1.1 shows the potential role of providers in an analysis of a treatment on an outcome. If, controlling for measured patient factors, providers do not exert an effect on either treatment assignment or outcome (panel A), then provider is not a confounding factor. If, on the other hand, provider exerts either a direct or indirect effect on both treatment and outcome (panel B), provider is a confounding factor. Direct effects on outcomes could be provider-specific factors such as the skill or experience of a surgeon or the care processes in place at a hospital. Indirect effects on outcomes could arise if providers don’t have a direct influence on the outcome, but the patient population they serve is distinct in ways that may advantage or disadvantage their outcomes. Factors that may differentiate patient populations could include socioeconomic status, disease severity, or cultural attitudes toward healthcare, in general. As examples, providers who serve distinct patient populations could include those at safety net hospitals that treat the uninsured and those at exemplary hospitals that attract the most difficult cases for a specific condition. A key point is that whatever is generating the provider effect is otherwise unmeasured. To account for provider is to control, implicitly, for all these other factors that are common to their patient population, but different from other providers’ patient populations.

There are numerous examples within the epidemiology literature where the idea that providers have an effect on outcome is explicitly discounted. Arguing that this provider-outcome link is ignorable (Walker, 2013) allows them to leverage the variability in provider treatment rates as instrumental variables (Brookhart & Schneeweiss, 2007). Such preference-based instruments have been used to estimate the safety and effectiveness of specific pharmaceutical therapies (Rassen et al., 2010; Schneeweiss et al., 2006), radiation therapy among prostate cancer patients (Sheets et al., 2012), and drug-eluting stents among patients undergoing coronary interventions (Venkitachalam et al., 2011).

Of course, these epidemiologic studies may not be wrong to ignore the provider-outcome link. The specific treatments, outcomes, and healthcare settings involved may affect whether or not the analysis question suffers from confounding by provider. For example, outcomes of surgical treatments are subject to direct physician effects in ways that outcomes of pharmaceutical treatments are not. Similarly, brief office visits may not be associated with the intensity of care that is associated with hospital stays. And finally, provider effects may be of primary interest in the study of short-term outcomes, as compared to long-term outcomes. Indirect provider effects—those associated with the patient population served—should be less differentiated by specific treatment or setting, however.

As noted by Bhattacharya and Vogt (2007), because there is no test to determine if a variable is an instrument, researchers must use their own understanding of the problem to guide them. And because providers have been documented as affecting both treatment and outcomes, it is reasonable to at least consider them as potential confounding factors in any analysis examining treatment effectiveness. Clinical trials routinely stratify randomization by provider to account for potential confounding at the provider level (Friedman, Furberg, & DeMets, 2010); and the importance of accounting for clustering by provider within observational clinical studies is recognized (Localio, Berlin, Ten Have, & Kimmel, 2001). Specific to the propensity score methods of interest here, excluding provider when it is truly a confounding factor should result in biased estimates of treatment effect just as it would if we excluded any confounding factor (Brookhart et al., 2006; Austin et al., 2007).

Accounting for Provider in Propensity Score Methods

Assuming for a particular clinical question that provider acts as a confounding factor between an exposure and an outcome, then it needs to be accounted for in the methods. As with any confounding factor in a propensity score analysis, the goal is to achieve balance in the distribution of that factor between treatment groups. Currently, there is no consensus about how best to incorporate provider into an analysis that utilizes propensity score methods. Note that because propensity score matching and inverse probability of treatment weighting are the most theoretically sound and most commonly utilized propensity score-based methods, these will be the only methods discussed in this section.

For inverse probability of treatment weighting, the decision to incorporate provider effects into the treatment model should be all that is needed. This will be typically be done using a set of provider-specific indicator variables as covariates, specified as either fixed effects or random effects in the logistic regression predicting treatment. These covariates may allow for provider-specific intercepts and/or provider-specific covariate effects. Typical model fit metrics can be used to guide the specification of the model. A full review of the choice between using fixed or random effects will not be undertaken here, but factors such as the number of providers, the average number of records per provider, and the overall treatment prevalence should guide the decision. Weights based on probability of treatment estimates from such a model should balance providers between study groups along with balancing all other covariates. There should be no assumption that provider will balance between study groups if not included in the treatment model.

For propensity score matching, there are two ways to incorporate provider into the analysis. First, in the treatment model, as described above. Second, as strata within which the matching takes place. This leads to a few possible strategies: (1) Matching within provider based on a treatment model that ignores provider; (2) matching within provider based on a treatment model that incorporates provider; or possibly (3) matching across providers based on a treatment model that incorporates provider. The first strategy could be problematic. If provider is not included in the treatment model, the parameter estimates associated with the other covariates will be biased (Neuhaus, Hauck, & Kalbfleisch, 1992) and will not reflect the actual, within-site treatment assignment mechanism. This may be a problem for matching when there are substantial between-site differences in the distribution of covariates at the patient level. The second strategy would account for this potential problem. The third strategy does not guarantee balance on providers by study group. Since matching necessarily only results in a fraction of the original study sample being utilized, it may be that treated patients at providers with high treatment rates may match to comparison patients at providers with low treatment rates.

Prior Research on Clustering and Propensity Score Methods

Much of the previous research on incorporating clustering into propensity score analyses has taken place within the social science literature to evaluate educational interventions applied to students within schools.   Hong and Raudenbush (2006) were some of the first researchers to attempt to incorporate school membership into a propensity score-based analysis. Their question of interest was whether or not grade retention (i.e. holding a student back a grade), compared to social promotion, led to increased academic learning. Because some schools are more likely than others to retain students making slow progress, they incorporated school effects into the grade retention (treatment) model. The resulting propensity scores were used to create strata of students within which outcomes were compared. There were other aspects of their analysis—later formalized (Hong, 2010) and called marginal mean weighting through stratification—that went beyond the standard propensity score approach.

Similar to Hong and Raudenbush, Thoemmes & West (2011) studied the effect of early grade retention on future test scores using propensity score stratification. They performed substantial simulation work that examined stratifying within schools and across schools using treatment probabilities from both pooled models and hierarchical models.

When the intraclass correlation of the covariates was low, there were very few differences in the results by stratification method or treatment model. When intraclass correlation was high, the within-school stratification methods outperformed the across-school methods, as long as a treatment model that allowed for school-specific effects was used.

Arpino and Mealli (2008) and Kelcey (2011) approach the problem of cluster-level confounding as a missing variable problem. Although this may implicitly be the scenario described earlier, both of these researchers explicitly generated data for their simulation work using contextual, or cluster-level, factors that were then treated as unmeasured. Using propensity score matching, their results were similar to those from Thoemmes and West, above, in that methods were least biased if the treatment model incorporated cluster in some way. Additionally, they found that within-cluster matching based on a pooled treatment model did not perform well.

Kim and Seltzer (2007) was interested in evaluating a program that provided enrichment during high school to certain students, with the goal of promoting enrollment in post-secondary education. As with many researchers in the education literature, one of their primary focuses was estimating school-specific treatment effects. [The only example found in the clinical literature that used propensity score methods and considered estimating provider-specific treatment effects was Griswold, Localio, and Mulrow (2010), summarized below.] While they do not present any simulation results, they discuss and demonstrate matching on propensity scores within site; noting that treatment models which allow for random slopes and intercepts will balance student-level factors within schools better than treatment models that do not.

There has been some work on or discussion about incorporating clustering into propensity score analyses within the clinical literature as well. One of the earliest articles that discussed the role of provider as a potential confounding factor within clinical research was Joffe, et al. (2004). From their perspective, there was widespread understanding that provider may affect outcome, but there was less understanding that they may also affect treatment assignment. They propose that marginal structural models or inverse probability of treatment weighting could be done using weights based on provider-specific probabilities of treatment and demonstrate this within an analysis of the complication rates associated with different type of coronary percutaneous interventions. They found substantial differences in results when provider-specific weights were used compared to weights from a pooled treatment model was used.

In a short editorial, Griswold, et al. (2010) noted that a safety study regarding proton pump inhibitors (Ray et al., 2010) used propensity score methods without accounting for provider. They reanalyzed the data incorporating provider in the treatment model in multiple ways—as fully stratified models and as hierarchical models with provider-specific random effects. They used propensity score covariate adjustment and found no differences in the overall conclusions. They concluded by suggesting that researchers using propensity score methods run sensitivity analyses that incorporate provider, to see if the results are robust.

A working group report from the Mini-Sentinel program (Cook, et al. 2012) examined whether or not effect estimates from analyses that used inverse probability of treatment weighting where the treatment model was estimated correctly (i.e. incorporating provider) using all data were comparable to an IPTW analysis where the treatment model was fully stratified by provider. This work answers a slightly different question than other research that compared methods that did or did not ignore provider. Their research question was less about whether or not providers (or data partners, in their project) were sources of confounding, and more about whether or not combining treatment estimates generated by each provider using IPTW would result in similar estimates from a pooled approach. This question is important for their program since patient-level data cannot leave the data partner, requiring all analyses to be stratified. They found that results between methods were similar when estimating the risk difference.

Li, Zaslavsky, and Landrum (2013) examined, in more theoretical detail, methods to incorporate clustering into analyses that use inverse probability of treatment weighting. In simulation studies, they found that modeling the propensity score correctly, by including cluster-specific parameters, resulted in less bias than methods that ignore cluster. They also showed, as may be expected, that estimates were more efficient when there were large clusters, as opposed to small clusters. Their data generation process was complicated in many ways. Similar to work described above, they generated data including a cluster-level covariate that was then ignored in the estimation of the treatment model. In the outcome model, they also specified cluster-specific treatment effect heterogeneity in addition to cluster-specific intercepts. Finally, the clinical example presented was unusual. They estimated the "average controlled difference" in receipt of breast cancer screening between black patients and white patients. Race is the type of non-manipulable exposure that fails to satisfy the requirements of the potential outcomes framework (Hernán, 2005).

Drawing conclusions from this prior work is challenging. Results from simulations seem to indicate that incorporating cluster into the treatment model is important for achieving balancing and minimizing bias in the treatment effect estimate when there is strong cluster-level confounding and/or large differences in covariate distributions between clusters. Results from actual clinical questions analyzed with and without provider-specific propensity score methods sometimes demonstrated a difference in the results and sometimes did not. It is likely that the characteristics of the data used for these questions were responsible for these findings, but relatively little is known about the data conditions that should lead researchers to anticipate problems.

Objectives of the Current Research

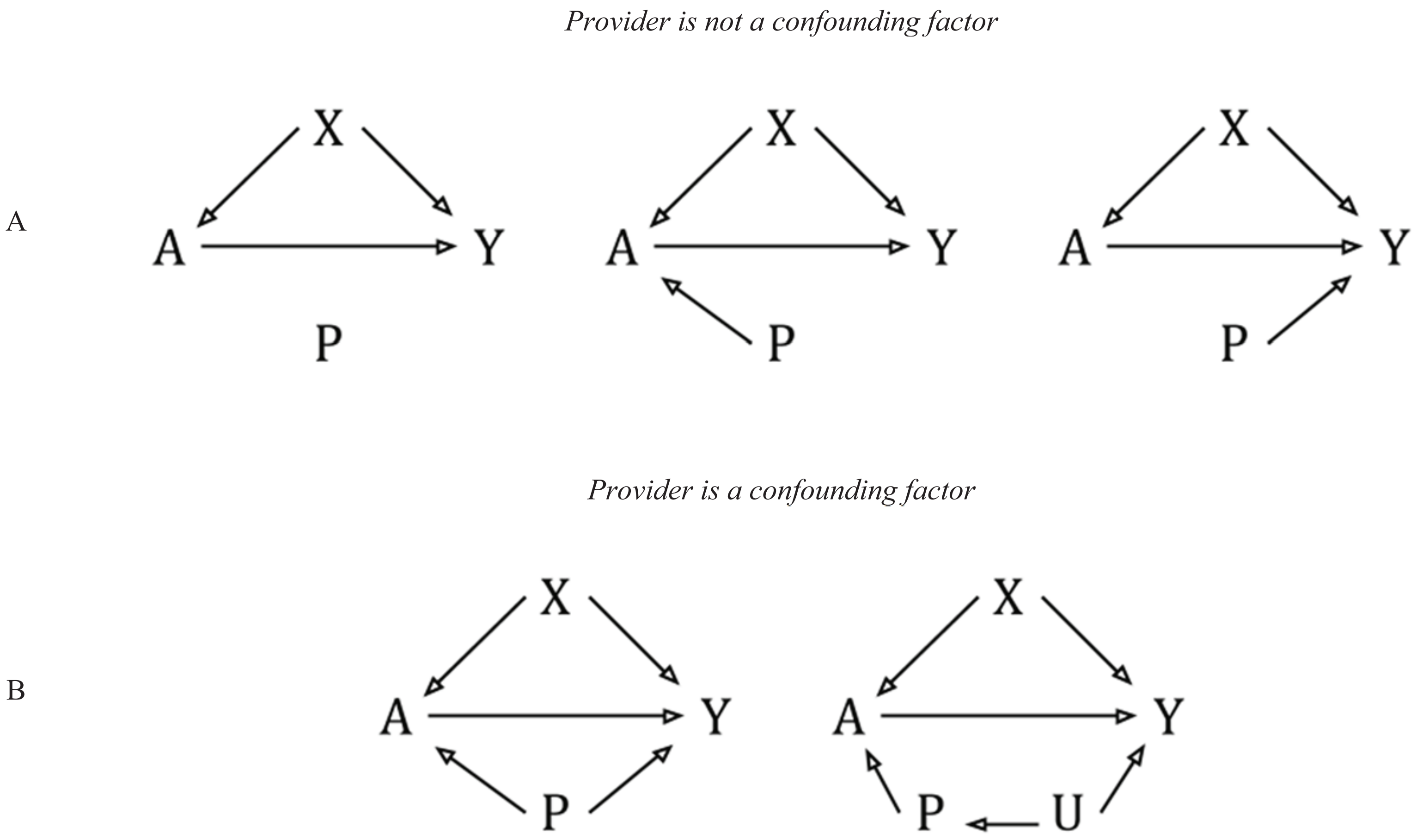
There are a number of unanswered questions regarding the use of propensity score methods in situations where confounding by provider exists. The chapters that follow will address a few key, practical issues that should guide researchers. Chapter 2 will examine the impact that different data scenarios have on the results of propensity score analyses when confounding by provider exists. Specifically, it is assumed providers exhibit a distribution on treatment rate, outcome rate, population size, and distribution of patient characteristics. When using propensity score matching or inverse probability of treatment weighting methods, it is not known whether the mere existence of these differences leads to bias or if these factors need to be correlated for there to be an important effect on the results.

Chapter 3 will address the question of whether or not there are risks to incorporating provider into propensity score methods when the provider effects on treatment and outcomes are not correlated. In addition, we will explore whether or not the standard errors associated with estimates based on different propensity score methods properly account for the clustering of patients within providers. Most researchers, when using multivariable regression models, will incorporate provider in some manner if clustering by provider exists in the data, so perhaps that should be a default for propensity score analyses as well.

Chapter 4 will present a SAS macro for performing optimal matching and full matching on the propensity score, as alternatives to less optimal greedy matching methods. Full matching, in particular, does not result in reduced sample size and should lead to more efficient treatment estimates compared to greedy matching. The macro includes a provision to match within strata, like providers. While full matching was proposed as a propensity score matching method decades ago, its use is limited due to the complexity of implementation. This goal of the SAS macro and of this paper is to make full matching methods available to researchers.

Each of these chapters will provide an applied clinical research example. In addition, Chapters 2 and 3 will include results from Monte Carlo simulations.

Figure 1.1. Potential relationships between treatment (A), outcome (Y), measured patient factors (X), unmeasured patient factors (U), and healthcare provider (P)


Provider is not a confounding factor

Provider is a confounding factor

CHAPTER 2

DATA CHARACTERISTICS AND THE PERFORMANCE OF PROPENSITY SCORE METHODS IN THE PRESENCE OF CONFOUNDING BY PROVIDER

Introduction

In comparative effectiveness research, confounding may arise at the level of the healthcare provider because both treatment assignment and outcome can vary by provider. Numerous studies have documented the variability of treatment rates by geography (e.g. Wennberg, 2002; Fisher et al., 2010; Munson et al., 2013) or provider (e.g. Carpenter et al., 2011; Barbash et al., 2014). Treatment rates may differ for a variety of reasons, including differential rates of adoption for new therapies or simple provider preferences for a specific therapy. In addition, there is evidence that outcomes differ by provider. Provider profiling reports produced by Hospital Compare (Centers for Medicare and Medicaid Services, web) and regional programs (e.g. New York State Department of Health, 2012; Massachusetts Department of Public Health, 2013; Pennsylvania Health Care Cost Containment Council, 2013) among others, demonstrate that substantial provider-level variation exists in different patient populations after controlling for patient risk. These differential outcomes may result from provider differences in the quality of patient care or from unmeasured differences in provider case-mix.

Previous research on the use of propensity score methods in the presence of confounding by some clustering level, such as provider, has demonstrated the risks of ignoring cluster in the analysis (Thoemmes & West, 2011; Arpino & Mealli, 2008; Kelcey, 2011; Li, Zaslavsky, & Landrum, 2013). These risks include inefficient or biased treatment effect estimates. Previous applications of propensity score methods, however, have been less conclusive about the benefit of conditioning on provider (Joffe, 2004; Griswold, Localio, & Mulrow, 2010). Compared to results from pooled analyses that ignore provider, sometimes the results have differed and sometimes they have been similar. It is very likely that the characteristics of the data, for both simulation work and real-world clinical examples, dictate when incorporating provider into propensity score methods is most essential.

Relevant provider-level data characteristics include treatment rate, outcome rate, patient population size, and average patient characteristics. For there to be confounding by provider, it is necessary that providers exhibit a distribution on treatment rate and outcome rate beyond that expected by the characteristics of the patients they care for. Additionally, it is reasonable to expect that providers differ by the size of their patient population and by the distribution of patient characteristics. I conducted a simulation study to explore whether or not the mere existence of these differences leads to bias and inefficiency in the treatment effect estimate or if these factors need to be correlated for there to be important effects on the results. We then explored the characteristics of provider data in a clinical example.

Simulation Study

We used Monte Carlo methods to simulate situations where patients were clustered within healthcare providers, and where those providers exhibited effects on both the treatment assignment and the resulting outcome of patients, independent of the observed patient-level covariates. We were specifically interested in examining different cluster-level specifications within the data generation process.

Data Generation Process

For each provider *j* in the simulated data, we first generated provider-level information that was subsequently used to generate patient-level data. Specifically, we generated five random variables—$n_j, a_{1j}, b_{4j}, u_j, v_j$—distributed as:

$$\ln(n_j) \sim N(\delta, 0.5)$$

$$a_{1j} \sim N(-1, 0.5)$$

$$b_{4j} \sim N(0, 1)$$

$$u_j \sim N(0, 1)$$

$$v_j \sim N(0, 1)$$

These variables were generated with the following correlations:

$$\begin{pmatrix} 1 & \rho_{na} & \rho_{nb} & \rho_{nu} & \rho_{nv} \\ \rho_{na} & 1 & \rho_{ab} & \rho_{au} & \rho_{av} \\ \rho_{nb} & \rho_{ab} & 1 & \rho_{bu} & \rho_{bv} \\ \rho_{nu} & \rho_{au} & \rho_{bu} & 1 & \rho_{uv} \\ \rho_{nv} & \rho_{av} & \rho_{bv} & \rho_{uv} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \psi_{nb} & \psi_{nu} & \psi_{nv} \\ 0 & 1 & 0 & 0 & 0 \\ \psi_{nb} & 0 & 1 & 0 & 0 \\ \psi_{nu} & 0 & 0 & 1 & \psi_{uv} \\ \psi_{nv} & 0 & 0 & \psi_{uv} & 1 \end{pmatrix}$$

The correlations that remain unspecified are noted below in the specific data scenarios.

Each of these provider-level variables has a specific purpose. The number of patients per provider was set by n_j, which was rounded to the nearest integer. The distribution of provider sizes is log-normally distributed and the mean of the log-distribution (δ) was set to either 5.0 to generate "large" providers or 3.5 to generate "small" providers. The large providers range in size (based on ±2 standard deviations) from about 55 to 400 patients. The small providers range in size from about 12 to 90 patients. The proportion of patients within each provider having characteristic X_1, described below, is determined by a_{1j}. Provider-level proportions for this variable ranged from about 12% to 50%. The provider-level mean value of X_4, described below, is determined by b_{4j}. And the values of u_j and v_j are deviations about the overall intercept in the treatment and outcome models, respectively. The terms a_{1j}, b_{4j}, u_j, and v_j all induce intraclass correlation for the associated covariates, for the treatment, or for the outcome.

Next, for each patient i within provider j, we generated four random variables—$X_{1ij}, X_{2ij}, X_{3ij}, X_{4ij}$—distributed as:

$$X_{1ij} \sim Bern\left(\text{logit}^{-1}(a_{1j})\right)$$

$$X_{2ij} \sim N(b_2, 1)$$

$$X_{3ij} \sim N(b_3, 1)$$

$$X_{4ij} \sim N\left(b_{4j}, 1\right)$$

The terms a_{1j} and b_{4j} are provider-specific terms described above and resulted in intraclass correlations of about 0.10 and 0.50 for X_1 and X_4 respectively. The terms b_2 and b_3 are set based on the value of X_{1ij}. When $X_{1ij} = 1$, then $b_2 = 0.5$ and $b_3 = -0.5$. When $X_{1ij} = 0$, then $b_2 = -0.5$ and $b_3 = 0.5$. The three normal random variables were generated with the following correlations:

$$\begin{pmatrix} 1 & \rho_{X_2,X_3} & \rho_{X_2,X_4} \\ \rho_{X_2,X_3} & 1 & \rho_{X_3,X_4} \\ \rho_{X_2,X_4} & \rho_{X_3,X_4} & 1 \end{pmatrix} = \begin{pmatrix} 1 & -.4 & -.1 \\ -.4 & 1 & .1 \\ -.1 & .1 & 1 \end{pmatrix}$$

The idea was to generate a set of covariates for each patient that exhibited interesting correlation and were not all independent of each other.

We then randomly assigned each patient a treatment, A_{ij}, as a Bernoulli random variable having a mean parameter equal to probability $p_{ij,A}$, which was determined by the following function:

$$\text{logit}\left(p_{ij,A}\right) = \alpha_0 + u_j + \alpha_1 X_{1ij} + \alpha_2 X_{2ij} + \alpha_3 X_{3ij} + \alpha_4 X_{4ij}$$

The parameters $[\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4]$ were fixed within all simulations to the values [–1.5, ln(1.5), ln(1.5), ln(0.8), ln(0.67)]. This yielded a treatment rate of just over 20%. We also randomly generated an outcome, Y_{ij}, for each patient as a normal random variable having a standard deviation equal to 1 and a mean parameter equal to $\mu_{ij,Y}$, which was determined by the following function:

$$\mu_{ij,Y} = \beta_0 + v_j + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_{TRT} A_{ij}$$

The terms u_j and v_j are provider-specific terms described above, leading to intraclass correlations of about 0.25 and 0.50 for the treatment and outcome, respectively. While some of the intraclass correlations indicated may be higher than typically seen in actual data, they serve to make patterns in the results more recognizable. The parameters $[\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_{TRT}]$ were fixed within all simulations to the values [0, –1, –1, 1, 1, 2]. This resulted in an observed outcome equal to about 0.8 in the unexposed group. Note that covariates associated with higher probability of treatment were associated with lower outcomes values, and vice versa. If we assume that a higher outcome value is optimal, then the treatment is working to improve outcomes and is assigned most frequently to the patients that would otherwise have poor outcomes.

Simulation Scenarios

We generated 1000 data sets with “large” providers and 1000 data sets with “small” providers for six different scenarios involving correlations between the provider-level quantities generated—provider size, mean of X_4, deviation from average treatment rate, and deviation from average outcome. Scenario #1 specified zero correlation between each of these quantities. Scenario #2 specified a positive correlation between provider size and deviation from average treatment rate ($\psi_{nu} = 0.5$); meaning large providers were more likely than average to assign treatment. Scenario #3 specified a positive correlation between provider size and deviation from average outcome ($\psi_{nv} = 0.5$); meaning large providers were more likely than average to be associated with better outcomes. Scenario #4 specified a positive correlation between provider size and mean of X_4 ($\psi_{nb} = 0.5$); meaning larger providers were more likely to have higher averages for that covariate. Scenario #5 specified a positive correlation between a provider’s deviation from average treatment rate and it’s deviation from average outcomes ($\psi_{uv} = 0.5$); meaning providers that were more likely to assign treatment were also more likely to have better outcomes. And scenario #6 specified each of these correlations simultaneously.

Within each generated data set, we applied different of propensity score matching and inverse probability of treatment weighting methods. After matching or weighting, we estimated the balance between treatment and comparison groups with respect to both covariates and providers and we estimated the treatment difference.

Propensity Score Methods

We first estimated three different treatment regression models. All models were specified as generalized linear models with logit links and binary error distributions having treatment A as the dependent variable and patient-level covariates X_1 to X_4 as predictors. They differed in how provider was incorporated. The first, or pooled, model ignored provider. The second model incorporated provider through the specification of provider-specific fixed effects. The third model incorporated provider through the specification of random effects as provider-specific deviations around the intercept. These random effects were assumed to be normal with mean 0 and this third model was estimated using generalized linear mixed model methods.

Using the predicted probabilities of treatment from each of the models described above, we applied two general propensity score-based methods—propensity score matching (Rosenbaum & Rubin, 1985) and inverse probability of treatment weighting (Robins, Hernán, & Brumback (2000); Hirano & Imbens, 2001). Inverse probability of treatment weighting (IPTW) creates pseudo-populations in which the other measured covariates are not associated with treatment, allowing outcomes between the weighted study groups to be compared directly. Patient-level weights were calculated as the inverse of the estimated probability of receiving the treatment that the patient actually received. For a patient that received treatment, this weight was the inverse of the predicted probability generated from the treatment models. For a patient in the comparison group, this weight was the inverse of 1 minus the model-based predicted probability. Propensity score matching creates sets of treated and comparison patients on the basis of their estimated propensity scores. Outcomes can be compared between the matched sets. We made 1:1 matches between treated and comparison patients both within provider and ignoring provider. Matches were made on the linear predictor from each treatment model using a greedy matching algorithm within calipers having a width of 0.2 SD of the linear predictor (Austin, 2011).

To estimate the adjusted effect of treatment on outcome when using IPTW methods, we calculated the difference between the weighted mean outcomes for each study group. To estimate the adjusted effect of treatment on outcome when using propensity score matching, we calculated the difference between the mean outcomes for each study group among the matched patients. For comparison to the treatment effect estimated by these propensity score methods, we estimated the observed treatment effect by taking the difference between the mean outcomes for each study group in the full data set prior to weighting or matching.

Metrics

Within each simulation scenario and for each combination of propensity score method and treatment model, we describe the validity and efficiency of the treatment effect estimates. Metrics to assess the treatment effect estimates include the mean, bias, variance, and mean squared error of each estimator, defined as:

Mean $= S^{-1} \sum_{s=1}^{S} \widehat{\Delta}_s = \overline{\Delta}$

Bias $= \overline{\Delta} - \Delta_{TRT}$

Variance $= (S-1)^{-1} \sum_{s=1}^{S} [\widehat{\Delta}_s - \overline{\Delta}]^2$

Mean squared error $= S^{-1} \sum_{s=1}^{S} [\widehat{\Delta}_s - \Delta_{TRT}]^2$

where S is the number of simulated data sets; Δ_{TRT} is the true treatment effect; and $\widehat{\Delta}_s$ is the estimated treatment effect for data set s.

All simulations were conducted in SAS version 9.3 (SAS Institute Inc, Cary, North Carolina). Sample SAS code for estimating propensity scores, calculating weights, performing matching, and estimating treatment effects and standard errors is shown in Appendix 1.

Simulation Study Results

Table 2.1 reports the simulation results for inverse probability of treatment weighting. The most striking result occurred for the scenarios where the provider-level average treatment rate was correlated with the outcome rate, even after controlling for patient characteristics (#5 and #6). In these situations, using weights based on a treatment model that does not incorporate provider effects yielded substantially biased estimates of treatment effect. Including provider-specific effects into the treatment model, as either fixed or random effects, largely ameliorated this problem. These findings were true for both small providers and large providers.

For the other simulation scenarios (#1 through #4), when the provider sizes were small, the results were less clear. In each of these four scenarios, weights based on the pooled treatment model led to treatment effect estimates with the lowest mean squared error. Use of weights based on a treatment model that included provider-specific random effects tended to produce estimates that were less biased than those based on other weights. But these estimates also tended to have higher variance, resulting in little, if any, reduction in mean squared error compared to estimates based on other weights. Weights based on a treatment model with fixed provider-specific effects were consistently biased.

When provider sizes were large, some of these inconsistencies disappeared. Weights based on a treatment model that included provider-specific random effects were least biased, and these estimates exhibited similar mean squared error to those from weights based on a treatment model with fixed provider effects. Results from using a pooled treatment model always had highest mean squared error.

Table 2.2 reports the simulation results for propensity score matching that was not conditional on provider. Similar to results based on weighting, results from unconditional matches based on estimates from the pooled treatment model were severely biased when the provider-level outcome rate was correlated with the provider-level treatment rate. Unlike above, this problem did not always disappear when matching was done based on estimates from provider-specific treatment models.

For small providers, unconditional matches made based on results from treatment models that incorporated providers as fixed effects resulted in treatment effect estimates that exhibited substantial bias across all scenarios. Using random effects instead did not completely solve the problem for small providers either. In fact, within small providers, the least biased estimates for Scenarios #1 through #4 were those from matching based on estimates from the pooled treatment model. The variability of these estimates, however, was always higher.

Matching without regard to provider, when provider sizes were large, exhibited more predictable behavior. Estimates from matching based on treatment models with provider-specific random effects had the lowest mean squared error. And estimates from matching based on pooled treatment models had the highest mean squared error—sometimes five or times as high as others within the same simulation scenario. This inflated MSE was not due to higher bias, but rather to substantially more variable estimates.

Table 2.3 reports the results for propensity score matching performed within provider. Across all of the simulation scenarios and all of the treatment model specifications, results were remarkably similar. In general, both bias and variance of the treatment estimates were minimal—especially when compared to results obtained from unconditional matching (Table 2.2). While the mean squared error for estimates from matching within provider based on results from the pooled treatment model was always higher than those from matching based on results from either provider-specific treatment model, the differences were not substantial. The problem noted above for the other two methods—when provider-level treatment and outcome rates were correlated and a pooled treatment model was used—was not found here.

Clinical Example

To demonstrate the potential impact of these different statistical methods, we present an analysis of the association between receipt of high-dose intravenous loop diuretics and in-hospital mortality among a population of patients admitted to the hospital for acute decompensated heart failure. Intravenous diuretics are a recommended therapy to address volume overload in patients with decompensated heart failure. Loop diuretics act in the kidney to block sodium and water reabsorption, which leads to effective symptomatic relief and decreased blood pressure. The benefits associated with diuretic use need to be balanced against the potential harms. It is suggested that diuretics, potentially because of the drop in blood pressure they induce, may lead to increased risk of renal dysfunction, which can lead to increased morbidity or mortality. For this reason, dosing of diuretics is important and should be limited to the lowest effective dose. Receipt of high doses of diuretics are often observed in a hospitalized heart failure population because response to diuretics decreases as heart failure severity progresses.

Higher mortality associated with high-dose diuretics has previously been reported by Peacock, et al. but that analysis did not consider the role of provider (Peacock, 2009). For this study, providers were hospitals. It is of interest to see if the strength or direction of the association changes when provider is taken into account. Many aspects of this research question make it plausible that provider is a confounding factor. Across different hospitals, it is likely that case-mix differs in systematic ways, leading patients who are more like each other to be clustered. The exposure is directly under the control of the hospital and, at the time of patient enrollment, there were no guidelines regarding the ideal dosing of diuretics in decompensated heart failure patients. Finally, it is possible that hospitals differ in the care they provide to heart failure patients, which may result in differences in outcomes by hospital.

For this analysis, we used data from heart failure hospitalizations that occurred between January 2001 and December 2003 and were entered into the Acute Decompensated Heart Failure National Registry (ADHERE) (Adams, et al., 2005). Similar to the previous study, we identified patients aged 65+ years old, not on vasoactive therapy, who received diuretics within one day of presenting to the hospital. The treatment of interest was receipt of high-dose diuretics. The dosing window of interest was the 24 hours following initiation of intravenous diuretics. A high dose was defined as ≥160 mg. The comparison group comprised those who received <160 mg of diuretic during that 24 hour dosing period. The outcome of interest was in-hospital mortality.

All of the propensity score-based methods and treatment model specifications used in the simulation study were used for this study as well. The estimation of treatment effect from these data differed from the simulation study since the outcome here was binary and the treatment effect of interest was the relative risk. To estimate the relative risk directly, we used a generalized linear mixed model with a log link and the binary error distribution. When applying IPTW methods, we estimated this model on the weighted patient data and requested robust standard errors with clusters defined by patient. When applying propensity score matching, we estimated this model using only patients in the resulting matched sets. Generalized estimating equation methods with an exchangeable working correlation matrix were used to account for the potential correlation of patients outcomes within the matched sets. We present the relative risk estimates and 95% confidence intervals from each method. All analyses were conducted in SAS v9.3.

Clinical Example Results

There were 43,434 patients within 236 providers in the study population. The number of patients per provider ranged from 9 to over 1,000, with the median (Q1, Q3) equal to 144 (63, 259). Among study patients, 9,469 (21.8%) were treated with a high dose of diuretics. High-dose diuretic treatment rates across providers ranged from 1.6% to 55.6% with median (Q1, Q3) rates equal to 19.7% (12.3%, 29.8%). Controlling for patient demographics, the intraclass correlation of treatment was 0.14. This intraclass correlation coefficient was estimated as $\rho = \frac{\sigma_u^2}{\sigma_u^2 + \pi^2/3}$, where σ_u^2 was the estimated variance of the provider random effects about the overall intercept in a hierarchical logistic model (Rodrıguez & Elo, 2003).

The observed in-hospital mortality rate in the high-dose group was 3.3%, about 30% higher than the rate among the low-dose group. Outcome rates for comparison patients across providers ranged from 0% to over 10%. The estimated intraclass correlation of the outcome, estimated among the comparison patients, was 0.03. The correlation between the estimated provider effects on treatment and outcomes was about –0.10.

Table 2.4 describes the sample characteristics and outcome rates for the two study groups. There were some differences between the groups with respect to age and gender. The high-dose group was, on average, two years younger than the low-dose group and had a higher proportion of males. Many of the medical history variables were similar between the groups. The largest differences were seen for chronic renal insufficiency and diabetes mellitus, both of which were more prevalent in the high-dose group. The high-dose group also had substantially higher rates of edema at initial evaluation and higher average blood urea nitrogen (BUN) values than the low-dose group.

Table 2.4 also shows the intracluster correlation of the patient characteristics. For the medical history and laboratory variables, these correlations ranged from near 0 to just over 0.06, except for hemoglobin, which had an intraclass correlation equal to 0.155. And while age and gender had low intraclass correlations, those associated with race were quite high, around 0.40 for each category shown. Clustering by provider was more pronounced for the initial evaluation characteristics. Indications of rales and congestion had intraclass correlations over 0.10 and the value associated with fatigue was 0.20.

The relative risks estimated by each different propensity score method and treatment model specification are shown in **Table 2.5**. Estimates from application of inverse probability of treatment weighting methods differed by treatment model used. Estimates of the risk associated with treatment were higher when weights were based on results from models that incorporated provider. The results from both matching methods did not show this same pattern. However, relative risk estimates from matches made without regard to provider were substantially lower than estimates from matches made within provider. The variability of these estimates did not appear to differ by propensity score method or treatment model specification.

Discussion

When a healthcare provider has measurable impacts on both a patient's treatment assignment and their downstream outcomes, we found that not accounting for these provider effects could lead to biased estimates of relative risk when using propensity score methods. This was true for our simulation study specifically when a provider's direct effect on treatment was correlated with their effect on outcome; a situation that occurs when providers with better patient outcomes use therapies at higher (or lower) rates than other providers. Propensity score methods that incorporated provider in some manner were able to control this error.

We examined the performance of propensity score matching and inverse probability of treatment weighting for estimating treatment effects across a number of data scenarios. For both large and small provider sizes, we examined the impact that correlation had on different combinations of provider-level characteristics. We were specifically interested in provider treatment rate, outcome rate, patient population size, and average patient characteristics. Only scenarios that included a correlation between provider treatment rate and outcome rate led to bias in the resulting effect estimates when provider was not incorporated into the propensity score analysis. When provider was ignored in other scenarios, variance of the effect estimates tended to be inflated when compared to effect estimates from analyses that incorporated provider.

These findings are consistent with prior research that found that ignoring provider, or some other sort of clustering, in propensity score methods led to biased estimates of treatment effect (Arpino & Mealli, 2008; Kelcey, 2011; Li, Zaslavsky, & Landrum, 2013). In the simulation work reported by these authors, the problem of cluster-level confounding was approached as a missing variable problem. Data was generated that included contextual, or cluster-level, factors which had consistent effects on both treatment and outcome. In the propensity score methods applied to these data, these cluster-level factors were treated as unmeasured, but cluster-level indicators were used instead. Naturally, the effect of these cluster-level indicators on treatment and their effect on outcomes were correlated as a result of the data generation process. That is the situation we found that led to the most bias if not properly handled.

Our findings are also consistent with results from Griswold, Localio, & Mulrow (2010). In a re-analysis of data from a safety study of proton pump inhibitors, they found that including provider effects into the propensity score treatment model did not lead to a different estimate of effects compared to results based on a treatment model without provider effects. This result may be expected, as we have demonstrated, when the provider effects on treatment are not correlated with provider effect on outcome.

Clusters, generally, can be incorporated into propensity score analyses in different ways. First, in the treatment model, through the inclusion of either fixed or random cluster-specific effects. If using inverse probability of treatment weighting, creating weights from the results of these conditional treatment models suffices. In our simulations, estimates based on these weights performed well when clusters were large, even when cluster effects on treatment and outcome were correlated. For smaller clusters, IPTW may not always be the best choice.

Second, cluster can be controlled directly through the use of conditional matching. In our simulations, estimates based on matching within cluster performed well regardless of the treatment model specification and regardless of the cluster size. The estimates having the least error, in general, were those from matching within cluster based on treatment models that incorporated cluster effects. When the treatment rate is low and there are sufficient number of comparison patients available for matching within cluster, it may make sense to use this method as a primary strategy.

There are times when within-cluster matching may not be the best strategy though. First, if there are many providers with high treatment rates, it may be impossible to match all treated patients with comparison patients, even using 1:1 matching. More flexible matching, where the number of treatment and comparison patients within each matched set is not rigidly fixed, should be explored in these situations. Second, within-provider matching requires that each provider, or at least the substantial majority of providers, contribute a relatively large number of patients to the data. A study having a large number of providers, each with a small number of patients, is not well-suited to this this approach since many providers may not be represented at all in the final sample. Third, exclusive providers, those with 0% or 100% treatment rates, will fall out of the analysis. Theoretically, the loss of these providers is sound, because their patients will have a true probability of treatment equal to 0 or 1, which violates a basic assumption underlying propensity score analysis.

It is difficult to recommend the use of a pooled treatment model when strong cluster effects on both treatment and outcome are found to exist, even though propensity score methods that used treatment probabilities from these models did not always fare poorly. Our simulations were primarily designed to identify problems due to correlations among specific cluster-level quantities, and thus we did not vary the strength of the associations between covariates and treatment by cluster. It is possible that substantial differences in the treatment mechanism across clusters would lead to problems when using a pooled treatment model, since it may not lead to the balance expected when propensity score methods conditional on cluster are used.

In a clinical example examining diuretic dosing and in-hospital mortality among hospitalized heart failure patients, these different propensity score methods and treatment model specifications yielded quite different estimates of risk. While we did not have a gold standard against which to compare, it should be noted that the results from IPTW methods indicated higher risk when cluster was included in the treatment model. Similarly, all estimates associated with within-provider matching were higher than estimates that matched across providers.

The presumption that providers can act as a confounding factor has intuitive appeal. They certainly act to affect the treatment rates of patients and can, through different direct or indirect pathways, be associated with differential outcome rates for their patients. A provider's effect on patient-level treatment assignment, as modeled in this study, can be thought of as a simple preference for one treatment over another. More complicated provider effects could result if different providers weighed certain patient characteristics differentially when determining a treatment strategy. Provider effects on outcomes could be direct or indirect. Indirect effects would include unmeasured factors associated with the provider's patient case-mix. Urban safety net hospitals, for example, serve a very different patient population than other hospitals, and it's possible that those patients would have had less favorable outcomes regardless of the provider they saw. Direct provider effects would include factors that are, in some way, under the control of the provider. This could be the provision of high-quality care by a hospital.

Indirect provider effects on outcomes, due to systematic differences in case-mix, are possible in all analyses. These effects may be relatively weak, but may be expected to persist regardless of the length of follow-up in the study. Direct provider effects on outcome, on the other hand, may only arise in specific settings or for specific treatments and may have a time-limited effect. For example, there may be no plausible direct provider effect associated with outcomes of a medication that was prescribed in an office setting. This differs from the likely provider effect associated with outcomes of a complex surgical procedure or medical device implant that requires hospitalization; and it is likely that these effects are most pronounced during the hospitalization and during the period immediately following discharge.

Ignoring the role of providers in the face of this confounding can therefore lead to effect estimates for a treatment that are contaminated with provider effects. When we account for provider properly, we estimate a treatment effect that controls for any otherwise unmeasured provider-level factors common to both the treated and comparison groups. The use of these methods does not, however, relieve us of the assumption that there is no unmeasured confounding. If providers assign treatment according to criteria that are related to outcome and not fully measured, resulting estimates of treatment effects can be biased.

One argument against conditioning analyses on provider is that treatment decisions are highly protocol-driven and consistently made by applying the same criteria to all patients. If this happens, assuming all confounders are measured, then the treated and comparison groups within a provider should be clinically distinct groups and have propensity score distributions that do not overlap, rendering within-provider propensity score methods unusable. Ignoring provider in this scenario is not the solution, since any systematic differences in outcomes by provider would remain unaccounted for. In some way, this suggests that we rely on inconsistency within provider regarding treatment assignment in order to properly estimate that treatment's effect on outcome. A provider's consistency in treatment assignment may differ by type of therapies, whether due to supply of the therapy—which is relatively unlimited for pharmaceuticals and can be limited for major devices or surgical procedures—due to the level of patient involvement in the treatment decision, or due to a provider's equipoise regarding treatment options. Inconsistency may also arise when the provider in an analysis is actually a practice group, hospital, or other conglomeration of physicians, since physicians within such groups may not always act in concert.

Related to the above scenario, consider a situation where providers are consistently making treatment decisions based on information that is not measured and that may also affect outcomes. As an example, analyses that use administrative health data (e.g. Medicare claims) often do not have information about patient frailty or disease severity. Even though some providers may decide to treat sicker patients more frequently than other providers would, there is no way around the fact that this variation is not benign, but systematic. Whether or not provider is taken into account during the analysis, this is a basic example of unmeasured confounding, which violates a fundamental assumption of propensity score analyses (Rosenbaum & Rubin, 1983) and will almost certainly yield incorrect treatment effect estimates.

There are a number of issues that a researcher will want to consider before using conditional propensity score methods. The most important of which is the plausibility of confounding by provider for the research question being asked. Certain treatments, outcomes, and healthcare settings are more likely than others to require attention. For example, outcomes of surgical treatments are patient to direct physician effects in ways that outcomes of pharmaceutical treatments are not. Similarly, brief office visits may not be associated with the intensity of care and effect on outcomes that are more likely to be associated with hospital stays. And provider effects may be a more important and relevant factor in the study of short-term outcomes, as compared to long-term outcomes. Indirect provider effects on outcomes—those associated with the patient population served by the provider—should be less differentiated by specific treatment or setting, however. It is possible to estimate the observed provider effects, as simple intraclass correlation coefficients, on treatment and outcomes using hierarchical regression models. Non-zero intraclass correlation coefficients for both treatment and outcome may suggest confounding and favor the use cluster-specific methods for propensity score analyses. Our clinical example, for example, had non-zero intraclass correlation coefficients for both treatment and outcome, controlling for other patient covariates. The hospital setting and short-term nature of the outcome made the possibility of confounding by provider plausible. The success of cluster-specific propensity score methods may also depend on cluster sizes. Study data that has information from a very large number of very small clusters may be inappropriate for these methods. In the simulation work, our “small clusters” still included over 25 patients, on average, which was sufficient to give consistent estimates of treatment effects when propensity score matching within clusters was used. With larger clusters, both conditional matching and inverse probability of treatment weighting methods performed very well.

In conclusion, when the possibility of confounding by provider exists, we recommend estimating propensity scores using a provider-specific treatment model. Appropriate estimates of treatment effect can then be found using either within-provider matching or inverse probability of treatment weighting.

Table 2.1 - Simulation results for inverse probability of treatment weighting methods. Mean estimated treatment difference, relative bias, variance, and mean squared error of treatment effect estimators for each scenario by provider size distribution. True treatment difference = 2.0.

Simulation Scenario / Treatment Model Specification	Treatment Different Estimates for Small Providers				Treatment Different Estimates for Large Providers			
	Mean	Relative Bias	Variance (x1000)	MSE (x1000)	Mean	Relative Bias	Variance (x1000)	MSE (x1000)
Scenario #1: No correlation between provider-level quantities								
Pooled	1.988	–0.60%	9.68	9.83	1.975	–1.25%	6.55	7.17
Provider fixed effects	1.888	–5.62%	6.78	19.42	1.959	–2.03%	2.44	4.09
Provider random effects	1.987	–0.63%	14.05	14.21	1.992	–0.41%	3.27	3.34
Scenario #2: Correlation between provider size and deviation from average treatment rate								
Pooled	1.972	–1.40%	7.59	8.37	1.975	–1.23%	5.64	6.24
Provider fixed effects	1.880	–6.01%	5.65	20.12	1.965	–1.73%	2.10	3.30
Provider random effects	1.977	–1.13%	9.87	10.38	2.001	0.05%	3.03	3.03
Scenario #3: Correlation between provider size and deviation from average outcome								
Pooled	1.973	–1.37%	10.15	10.90	1.973	–1.34%	6.80	7.52
Provider fixed effects	1.893	–5.36%	8.35	19.85	1.973	–1.33%	2.03	2.73
Provider random effects	2.007	0.36%	16.67	16.72	2.006	0.32%	3.08	3.12
Scenario #4: Correlation between provider size and mean of X_4								
Pooled	1.972	–1.41%	11.28	12.08	1.970	–1.49%	6.61	7.50
Provider fixed effects	1.894	–5.32%	12.11	23.45	1.970	–1.49%	3.98	4.86
Provider random effects	2.009	0.45%	18.84	18.92	2.005	0.24%	5.39	5.41
Scenario #5: Correlation between a deviation from average treatment rate and deviation from average outcomes								
Pooled	2.385	19.26%	9.89	158.28	2.387	19.36%	7.95	157.93
Provider fixed effects	1.998	–0.10%	6.86	6.86	1.998	–0.09%	2.19	2.20
Provider random effects	2.011	0.55%	10.78	10.90	1.999	–0.06%	3.13	3.13
Scenario #6: All mentioned correlations simultaneously								
Pooled	2.403	20.15%	14.12	176.50	2.404	20.20%	11.81	175.09
Provider fixed effects	2.011	0.53%	6.57	6.68	2.008	0.38%	1.92	1.98
Provider random effects	2.036	1.78%	11.17	12.44	2.002	0.10%	2.69	2.69

Abbreviation: MSE = Mean squared error

Table 2.2 - Simulation results for propensity score matching methods (not conditional on provider). Mean estimated treatment difference, relative bias, variance, and mean squared error of treatment effect estimators for each scenario by provider size distribution. True treatment difference = 2.0.

Simulation Scenario / Treatment Model Specification	Treatment Different Estimates for Small Providers				Treatment Different Estimates for Large Providers			
	Mean	Relative Bias	Variance (x1000)	MSE (x1000)	Mean	Relative Bias	Variance (x1000)	MSE (x1000)
Scenario #1: No correlation between provider-level quantities								
Pooled	2.002	0.09%	5.62	5.62	1.992	–0.40%	5.64	5.70
Provider fixed effects	2.206	10.29%	3.19	45.53	2.039	1.95%	0.63	2.15
Provider random effects	1.944	–2.82%	4.01	7.20	1.988	–0.61%	0.67	0.82
Scenario #2: Correlation between provider size and deviation from average treatment rate								
Pooled	1.990	–0.52%	6.33	6.44	1.990	–0.52%	5.14	5.25
Provider fixed effects	2.169	8.45%	2.84	31.41	2.032	1.61%	0.61	1.65
Provider random effects	1.933	–3.33%	3.44	7.89	1.983	–0.85%	0.48	0.77
Scenario #3: Correlation between provider size and deviation from average outcome								
Pooled	1.992	–0.42%	6.76	6.83	1.991	–0.45%	6.03	6.11
Provider fixed effects	2.193	9.66%	3.99	41.30	2.033	1.63%	0.71	1.78
Provider random effects	1.936	–3.22%	2.26	6.40	1.989	–0.54%	0.67	0.79
Scenario #4: Correlation between provider size and mean of X_4								
Pooled	1.994	–0.31%	7.64	7.68	1.991	–0.45%	5.75	5.83
Provider fixed effects	2.200	10.00%	2.68	42.67	2.043	2.17%	0.70	2.59
Provider random effects	1.939	–3.07%	3.49	7.25	1.983	–0.83%	0.84	1.11
Scenario #5: Correlation between a deviation from average treatment rate and deviation from average outcomes								
Pooled	2.394	19.70%	6.57	161.84	2.398	19.92%	5.92	164.61
Provider fixed effects	2.237	11.86%	3.56	59.79	2.052	2.61%	0.74	3.47
Provider random effects	1.948	–2.60%	2.70	5.40	1.992	–0.41%	0.77	0.84
Scenario #6: All mentioned correlations simultaneously								
Pooled	2.421	21.05%	12.34	189.53	2.421	21.03%	9.72	186.63
Provider fixed effects	2.208	10.42%	3.28	46.74	2.047	2.33%	0.62	2.79
Provider random effects	1.929	–3.56%	3.06	8.14	1.985	–0.74%	0.67	0.88

Abbreviation: MSE = Mean squared error

Table 2.3 - Simulation results for propensity score matching methods (within provider). Mean estimated treatment difference, relative bias, variance, and mean squared error of treatment effect estimators for each scenario by provider size distribution. True treatment difference = 2.0.

Simulation Scenario / Treatment Model Specification	Treatment Different Estimates for Small Providers				Treatment Different Estimates for Large Providers			
	Mean	Relative Bias	Variance (x1000)	MSE (x1000)	Mean	Relative Bias	Variance (x1000)	MSE (x1000)
Scenario #1: No correlation between provider-level quantities								
Pooled	1.985	–0.76%	2.10	2.33	1.987	–0.65%	0.44	0.61
Provider fixed effects	1.992	–0.39%	1.99	2.05	1.990	–0.48%	0.33	0.42
Provider random effects	1.990	–0.50%	1.94	2.04	1.991	–0.47%	0.37	0.46
Scenario #2: Correlation between provider size and deviation from average treatment rate								
Pooled	1.975	–1.23%	2.28	2.88	1.979	–1.04%	0.37	0.80
Provider fixed effects	1.985	–0.73%	1.98	2.20	1.986	–0.70%	0.26	0.46
Provider random effects	1.985	–0.76%	1.86	2.10	1.985	–0.74%	0.23	0.45
Scenario #3: Correlation between provider size and deviation from average outcome								
Pooled	1.976	–1.19%	1.18	1.75	1.982	–0.92%	0.36	0.70
Provider fixed effects	1.985	–0.75%	1.46	1.69	1.987	–0.65%	0.37	0.53
Provider random effects	1.986	–0.71%	1.29	1.49	1.987	–0.67%	0.33	0.51
Scenario #4: Correlation between provider size and mean of X_4								
Pooled	1.975	–1.27%	1.34	1.99	1.982	–0.92%	0.45	0.79
Provider fixed effects	1.980	–1.01%	1.32	1.73	1.990	–0.49%	0.37	0.47
Provider random effects	1.983	–0.87%	1.39	1.70	1.991	–0.47%	0.38	0.47
Scenario #5: Correlation between a deviation from average treatment rate and deviation from average outcomes								
Pooled	1.977	–1.13%	1.91	2.43	1.983	–0.86%	0.40	0.70
Provider fixed effects	1.986	–0.70%	1.52	1.72	1.989	–0.53%	0.28	0.39
Provider random effects	1.981	–0.93%	1.56	1.90	1.989	–0.55%	0.27	0.39
Scenario #6: All mentioned correlations simultaneously								
Pooled	1.981	–0.94%	1.55	1.90	1.981	–0.94%	0.35	0.71
Provider fixed effects	1.985	–0.73%	1.49	1.71	1.987	–0.67%	0.38	0.56
Provider random effects	1.984	–0.79%	1.36	1.61	1.986	–0.68%	0.34	0.53

Abbreviation: MSE = Mean squared error

Table 2.4 - Characteristics of ADHERE-HF patients receiving low-dose and high-dose intravenous diuretics

Variable		Low-Dose (N = 33,965)	High-Dose (N = 9,469)	Standardized Difference, %	Intracluster Correlation
Demographics	Age (years), Mean (SD)	80.0 (7.9)	78.0 (7.6)	25.3	0.036
	Gender, Male	13,782 (40.6%)	4,432 (46.8%)	12.6	0.051
	Race			11.9	
	White	26,889 (79.2%)	7,190 (75.9%)		0.398
	Black	4,396 (12.9%)	1,621 (17.1%)		0.460
	Other/unknown	2,680 (7.9%)	658 (6.9%)		0.409
Medical History	Anemia	17,763 (52.3%)	5,639 (59.6%)	14.7	0.054
	Atrial fibrillation	12,157 (35.8%)	3,692 (39.0%)	6.6	0.025
	Coronary artery disease	19,172 (56.4%)	5,887 (62.2%)	11.7	0.029
	Chronic renal insufficiency	7,952 (23.4%)	3,366 (35.5%)	26.9	0.041
	Chronic obstructive pulmonary disease/Asthma	10,296 (30.3%)	3,290 (34.7%)	9.5	0.026
	Diabetes mellitus	12,912 (38.0%)	4,677 (49.4%)	23.1	0.011
	Hyperlipidemia	11,009 (32.4%)	3,408 (36.0%)	7.5	0.058
	Hypertension	24,749 (72.9%)	7,058 (74.5%)	3.8	0.041
	Prior myocardial infarction	9,861 (29.0%)	3,015 (31.8%)	6.1	0.034
	Peripheral vascular disease	5,897 (17.4%)	1,994 (21.1%)	9.4	0.063
	Prior stroke/Transient ischemic attack	6,297 (18.5%)	1,859 (19.6%)	2.8	0.030
	Current smoker	2,502 (7.4%)	740 (7.8%)	1.7	0.046
Initial Evaluation	Fatigue	10,559 (31.1%)	2,790 (29.5%)	3.5	0.204
	Rales	24,281 (71.5%)	7,037 (74.3%)	6.4	0.127
	Edema	22,311 (65.7%)	7,277 (76.9%)	24.9	0.041
	Congestion	24,131 (71.0%)	6,870 (72.6%)	3.3	0.103
	Ejection fraction, <40%	11,036 (32.5%)	3,427 (36.2%)	8.5	0.044
Laboratory results	Systolic blood pressure (mmHg), Mean (SD)	146.4 (29.9)	144.8 (30.3)	5.3	0.019
	BUN (mg/dL), Mean (SD)	28.8 (16.8)	34.3 (20.6)	29.6	0.013
	Serum sodium (mmol/L), Mean (SD)	138.3 (4.7)	138.4 (4.6)	1.9	0.054
	Hemoglobin (g/dL), Mean (SD)	12.4 (2.5)	12.0 (2.3)	15.9	0.155

Values for Low-Dose and High-Dose groups presented as N (%) unless otherwise specified
Abbreviation: SD = Standard deviation

Table 2.5 - Estimated relative risk and 95% confidence interval for association between receipt of high-dose diuretic and in-hospital mortality by propensity score method and treatment model specification

Propensity Score Method / Treatment Model Specification	Relative Risk (95% CI)
Inverse probability of treatment weighting	
Pooled	1.32 (1.22, 1.43)
Provider fixed effects	1.53 (1.42, 1.66)
Provider random effects	1.56 (1.45, 1.69)
Propensity score matching (unconditional)	
Pooled	1.15 (0.98, 1.36)
Provider fixed effects	1.12 (0.94, 1.32)
Provider random effects	1.19 (1.00, 1.41)
Propensity score matching within provider	
Pooled	1.30 (1.08, 1.55)
Provider fixed effects	1.26 (1.06, 1.52)
Provider random effects	1.33 (1.11, 1.60)

Abbreviation: CI = Confidence interval

CHAPTER 3

STANDARD ERROR OF TREATMENT EFFECT ESTIMATES FROM PROPENSITY SCORE METHODS IN THE PRESENCE OF CONFOUNDING BY PROVIDER

Introduction

Data used for clinical research are often collected among patients clustered within providers. These providers may affect the assignment of the treatment and may be associated with differential outcomes, above and beyond those expected by patient characteristics. We demonstrated that propensity score methods that do not account for provider resulted in biased treatment effect estimates when provider effects on treatment are correlated with provider effects on outcome. This could happen, for example, if certain hospitals that are more likely to offer certain treatments also tend to have lower adverse events rates than other hospitals.

Such correlation may not be the most common scenario encountered in practice, however. It may be more likely that these provider effects are present, but uncorrelated. In these scenarios, we showed that treatment effect estimates resulting from either propensity score matching or inverse probability of treatment weighting are generally unbiased. We did not check to see how appropriate the related standard error estimates were for these treatment effects, though. Typically, when clustered data are utilized for analysis, statistical methods must account for the correlation of subjects within the cluster (Gelman & Hill, 2006). It is not clear when using propensity score methods (Rosenbaum & Rubin, 1983) if matching within provider or including provider-specific effects in the estimation of the treatment model is adequate to yield standard error estimates that have the proper coverage of nominal 95% confidence intervals.

We sought to understand how well various propensity score methods estimated the standard error of treatment effects in the presence of confounding by provider. Using simulation studies and a clinical example, we aim to identify the analysis strategies that yield both unbiased estimation of the treatment effect and appropriate inference.

Simulation Study

We used Monte Carlo methods to simulate situations where patients were clustered within healthcare providers, and where those providers exhibited effects independent of the observed patient-level covariates on both patient-level treatment assignment and outcomes. We were specifically interested in examining different estimators based on data from situations where the provider effects on treatment and outcome were uncorrelated. The only data generation parameter we varied between scenarios was the strength of the intraclass correlation (ICC) of the outcome.

Data Generation Process

For each provider *j* in the simulated data, we first generated provider-level information that was subsequently used to generate patient-level data. Specifically, we generated five independent random variables—$n_j, a_{1j}, b_{4j}, u_j, v_j$—distributed as:

$$\ln(n_j) \sim N(5, 0.5)$$

$$a_{1j} \sim N(-1, 0.5)$$

$$b_{4j} \sim N(0,1)$$

$$u_j \sim N(0,1)$$

$$v_j \sim N(0,\sigma_v^2)$$

The number of patients per provider is set by n_j, which was rounded to the nearest integer. The distribution of provider sizes is log-normally distributed and the mean of the log-distribution was set to 5.0 to generate providers that ranged in size from about 55 to 400 patients. The proportion of patients within each provider having characteristic X_1, described below, is determined by a_{1j}. Provider-level proportions for this variable averaged about 35% and ranged from about 12% to 50%. The provider-level mean value of X_4, described below, is determined by b_{4j}. And the values of u_j and v_j are deviations around the overall intercept in the treatment and outcome models, respectively. The terms u_j and v_j induce intraclass correlation. The value of σ_v^2 was allowed to vary.

For each patient i within provider j, we generated four random variables—$X_{1ij}, X_{2ij}, X_{3ij}, X_{4ij}$—distributed as:

$$X_{1ij} \sim Bern\left(\text{logit}^{-1}(a_{1j})\right)$$

$$X_{2ij} \sim N(b_2, 1)$$

$$X_{3ij} \sim N(b_3, 1)$$

$$X_{4ij} \sim N\left(b_{4j}, 1\right)$$

The provider-specific quantities a_{1j} and b_{4j} resulted in intraclass correlations of about 0.10 for X_1 and 0.50 for X_4 respectively. The terms b_2 and b_3 are set for each patient based on that patient's value for X_{1ij}.

$$b_2 = \begin{cases} 0.5, & X_{1ij} = 1 \\ -0.5, & X_{1ij} = 0 \end{cases}$$

$$b_3 = \begin{cases} -0.5, & X_{1ij} = 1 \\ 0.5, & X_{1ij} = 0 \end{cases}$$

The three patient-level normal random variables, X_2, X_3, and X_4 were generated with the following correlations:

$$\begin{pmatrix} 1 & \rho_{X_2,X_3} & \rho_{X_2,X_4} \\ \rho_{X_2,X_3} & 1 & \rho_{X_3,X_4} \\ \rho_{X_2,X_4} & \rho_{X_3,X_4} & 1 \end{pmatrix} = \begin{pmatrix} 1 & -.4 & -.1 \\ -.4 & 1 & .1 \\ -.1 & .1 & 1 \end{pmatrix}$$

We then randomly assigned each patient to a treatment, A_{ij}, as a Bernoulli random variable having a mean parameter equal to probability $p_{ij,A}$, determined by the following function:

$$\text{logit}(p_{ij,A}) = \alpha_0 + u_j + \alpha_1 X_{1ij} + \alpha_2 X_{2ij} + \alpha_3 X_{3ij} + \alpha_4 X_{4ij}$$

The parameters $[\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4]$ were fixed within all simulations to the values [–1.0, ln(1.5), ln(1.5), ln(0.8), ln(0.67)]. This yielded a treatment rate of just over 30%. Due to the provider-specific term, u_j, the intraclass correlation of the treatment was about 0.25.

Finally, we randomly generated an outcome, Y_{ij}, for each patient as a normal random variable having a standard deviation equal to 1 and a mean parameter equal to $\mu_{ij,Y}$, which was determined by the following function:

$$\mu_{ij,Y} = \beta_0 + v_j + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4ij} + \beta_{TRT} A_{ij}$$

The provider-specific term, v_j, also resulted in intraclass correlation for the outcome. The magnitude of this correlation was allowed to vary by scenarios described below.

The parameters $[\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_{TRT}]$ were fixed within all simulations to the values [0, –1, –1, 1, 1, 2]. This resulted in an observed outcome equal to about 0.8 in the unexposed group. These parameters were set such that covariates associated with a higher probability of treatment were associated with a lower outcomes value, and vice versa. By defining higher outcome values as optimal, the treatment is seen to improve the outcome and is assigned more frequently to patients, based on their covariate values, that would otherwise had poor outcomes.

Simulation Scenarios

We generated 1000 data sets with 50 providers for three different values of σ_v^2, the parameter that controls the provider-level intraclass correlation of the outcome. At $\sigma_v^2 = 0.00$, there was no intraclass correlation for the outcome ($r = 0.0$). At $\sigma_v^2 = 0.05$, the intraclass correlation for the outcome was weak ($r \approx 0.05$). And at $\sigma_v^2 = 0.33$, the intraclass correlation for the outcome was relatively strong ($r \approx 0.25$). In clinical research, it would likely be unusual to observe an intraclass correlation stronger than this for most outcomes. We applied multiple propensity score methods to each data set and calculated both the treatment effect point estimate and its standard error.

The data generation specifications described above include many parameters that are provider-specific. As a comparison to these data, we thought it would be helpful to additionally generate 1000 data sets that did not rely on any provider-specific quantities. To do this, we generated data sets with 4000 records where $X_1, X_2, X_3, X_4, A,$ and Y were created as described above, but with the quantities $a_{1j}, b_{4j}, u_j,$ and v_j all equal to zero. We applied the same pooled propensity score and estimation methods described below to these data sets.

Propensity Score Methods

We first estimated two treatment regression models that differed in how provider was incorporated. Both models were specified as generalized linear models with logit links and binary error distributions having treatment *A* as the dependent variable and patient-level covariates X_1–X_4 as predictors. The first model ignored provider. This is referred to as the pooled model. The second model was fit using generalized linear mixed model methods and incorporated provider through the specification of provider-specific intercepts, estimated using random effects. These effects were assumed to be normal with mean 0. This model is referred to as the provider-specific model.

We applied multiple propensity score-based methods to the predicted probabilities from each of these models. First, we used inverse probability of treatment weighting (Hirano & Imbens, 2001; Robins, Hernán, & Brumback, 2000). This method utilizes patient-specific weights defined as the inverse of the estimated probability of treatment for the treatment that patient received. For a patient who received treatment, this weight is the inverse of the predicted probability generated from the treatment models. For a patient in the comparison group, this weight is the inverse of one minus the model-based predicted probability. Second, we used 1:1 greedy matching (Rosenbaum & Rubin, 1985a), where treated patients (randomly ordered) were matched one at a time to comparison patients. Third, we used full matching (Rosenbaum, 1991), where treated patients were matched to comparison patients in sets where the number of treated and comparison patients could vary and the overall distance between matched sets was minimized. Both of these matches were made using the linear predictor from the treatment models with calipers equal to 0.2 SD of that quantity (Austin, 2011), based on the values of the linear predictor within the entire sample.

Estimation Methods

Methods that are appropriate to estimate the treatment effect and its standard error differ by propensity score method used. The list of different methods we use is shown in **Table 3.1**. For 1:1 matching, the treatment effect can be estimated by taking the difference between the averages of the matched patients from each set, as:

$$\Delta_{1:1} = \frac{1}{N_1}\sum_{i=1}^{N} A_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - A_i)Y_i = \frac{1}{N_1}\sum_{i=1}^{N}\{A_i Y_i - (1 - A_i)Y_i\}$$

where N_1 is the number of treated patients who were matched and N_0 is the number of comparison patients who were matched. For any 1:1 matching, of course, $N_1 = N_0$. The standard error of this estimate can be estimated in two ways, treating the observations as pooled and treating the observations as paired. For continuous outcomes, this is the difference between the standard error estimated when using a pooled *t*-test or a paired *t*-test. The group differences will be identical, but the standard error estimates will differ. Instead of using *t*-tests in this study, however, we will use two general linear regression models for the outcome, which will give the same results. Using regression models allows for flexibility in estimation by letting us include other factors in addition to treatment as independent variables. These other factors may be used to correct residual imbalance between groups or may account for factors that were not present at the treatment decision. For this study, both models will include the treatment indicator as the only independent variable. The first outcome model will ignore the matched sets [labeled in the results tables as “GLM”] and will report the usual regression standard errors. The second model will account for matched sets by using generalized estimating equation (GEE) methods (Liang & Zeger, 1986) where the clusters are matched sets and an exchangeable working correlation matrix is specified [labeled "GEE"]. For 1:1 matching, the specification of the working correlation is not critical, since all structures are equivalent.

Adjustments in estimation are necessary when full matching is used, since full matching results in matched sets that can have differing numbers of treated and comparison patients. Unweighted estimates from a *t*-test or basic linear model will be incorrect. Abadie & Imbens (2012) defines the treatment effect estimate from full matching results as

$$\Delta_{ABAD} = \frac{1}{N_1}\sum_{i=1}^{N} A_i\left(Y_i - \frac{1}{M_i}\sum_{j\in\mathcal{H}(i)} Y_j\right)$$

with its variance as

$$\hat{\sigma}^2_{ABAD} = \frac{1}{N_1 - 1}\sum_{i=1}^{N} A_i\left(Y_i - \frac{1}{M_i}\sum_{j\in\mathcal{H}(i)} Y_j - \Delta_{ABAD}\right)^2$$

where N_1 is the number of treated patients who were matched; M_i is the number of patients matched to patient i; and $\mathcal{H}(i)$ contains the indices of the M_i patients matched to patient i. For matched sets that contain one treated patient and multiple comparison patients, the right-hand term in Δ_{ABAD} is the difference between the outcome of the treated patient and the average of the outcomes of the matched comparison patients. For matched sets that contain multiple treated patients and one comparison patient, the right-hand term in Δ_{ABAD} is just the difference between one of those matched treated patients and the comparison patient, meaning the comparison patient will be represented in the overall average multiple times. We present these results labeled as "Abadie".

Similar to the Abadie estimator, Hansen (2004) proposes weighting entire matched sets by the number of treated patients in the set. This is referred to as ETT (effect of treatment on the treated) weighting. Although not explicitly defined, the idea behind ETT weighting is to first calculate the within-matched-set differences in the outcome by study group, and then calculating the weighted average of those differences across all sets. It is easier to make this calculation by assigning patient-level weights to each matched record. Having weights is also useful when using methods that require patient-level data, such as regression models. For matched set m with $m = 1, \dots, M$, let $\mathcal{H}_m$ contain the indices for the $N_{1,m}$ treated patients and the $N_{0,m}$ comparison patients in the set. Consider a two-stage weight. The first stage weight is a patient-level weight. Within each set, equally weight each of the patients in each study group, as:

$$w_{1i} = \begin{cases} N_{0,m}^{-1}, & A_i = 0, i \in \mathcal{H}_m \\ N_{1,m}^{-1}, & A_i = 1, i \in \mathcal{H}_m \end{cases}$$

The second stage weight is a set-level weight reflecting the number of treated patients in the set:

$$w_{2i} = N_{1,m} \text{ for } i \in \mathcal{H}_m$$

The combined weight is therefore:

$$w_i = w_{1i} \times w_{2i} = \begin{cases} \dfrac{N_{1,m}}{N_{0,m}}, & A_i = 0, i \in \mathcal{H}_m \\ 1, & A_i = 1 \end{cases}$$

The treatment effect estimate based on these weights is given by:

$$\Delta_{ETT} = \frac{1}{N_1} \sum_{i=1}^{N} \{A_i w_i Y_i - (1 - A_i) w_i Y_i\}$$

We used three different regression models to estimate the treatment effect and its standard error for the results of full matching. The first two models used ETT-weighted data and included a general linear model without any variance correction [labeled “GLM (ETT)”] and a general linear model with variance correction through the estimation of robust standard errors [labeled “GLM (ETT, robust)”]. These robust standard errors were estimated using

GEE methods with the matched set as the cluster and an independence working correlation structure. The third model we used also utilized GEE methods, but was run on unweighted data, included matched set as the cluster, and specified an exchangeable working correlation structure [labeled “GEE”].

For inverse probability of treatment weighting methods, the basic definition of the treatment effect is given by

$$\widehat{\Delta}_{IPTW} = \hat{\mu}_{1,IPTW} - \hat{\mu}_{0,IPTW}$$

where

$$\hat{\mu}_{1,IPTW} = \left(\sum_{i=1}^{N} \frac{A_i}{\hat{e}_i}\right)^{-1} \sum_{i=1}^{N} \frac{A_i Y_i}{\hat{e}_i}$$

and

$$\hat{\mu}_{0,IPTW} = \left(\sum_{i=1}^{N} \frac{1 - A_i}{1 - \hat{e}_i}\right)^{-1} \sum_{i=1}^{N} \frac{(1 - A_i) Y_i}{1 - \hat{e}_i}$$

Its standard error can be estimated as the square root of large-sample variance, given by

$$\hat{\sigma}^2_{IPTW} = n^{-2} \sum_{n=1}^{N} \hat{I}_{IPTW,i}{}^{2}$$

where

$$\hat{I}_{IPTW,i} = \frac{A_i\big(Y_i - \hat{\mu}_{1,IPTW}\big)}{\hat{e}_i} - \frac{(1 - A_i)\big(Y_i - \hat{\mu}_{0,IPTW}\big)}{1 - \hat{e}_i} - (Z_i - \hat{e}_i)\widehat{\boldsymbol{H}}'_{\beta}\widehat{\boldsymbol{E}}^{-1}_{\beta\beta}\boldsymbol{X}_i$$

$$\widehat{\boldsymbol{H}}_{\beta} = n^{-1} \sum_{i=1}^{N} \left\{\frac{A_i\big(Y_i - \hat{\mu}_{1,IPTW}\big)(1 - \hat{e}_i)}{\hat{e}_i} - \frac{(1 - A_i)\big(Y_i - \hat{\mu}_{0,IPTW}\big)\hat{e}_i}{1 - \hat{e}_i}\right\} \boldsymbol{X}_i$$

and

$$\widehat{\boldsymbol{E}}^{-1}_{\beta\beta} = n^{-1} \sum_{i=1}^{N} \hat{e}_i (1 - \hat{e}_i) \boldsymbol{X}_i \boldsymbol{X}'_{\boldsymbol{i}}$$

We refer to this as the Lunceford method [labeled as such in the results], because their manuscript (Lunceford & Davidian, 2004) was one of the first to present a formula for the large-sample variance of the estimate.

The regression-based method typically used to estimate $\widehat{\Delta}_{IPTW}$ and approximate $\hat{\sigma}_{IPTW}^2$ is a generalized linear model on the weighted data with robust standard errors. For our study, we estimated these robust standard errors in two ways. The first estimated standard errors using GEE methods with the patient as the (single-member) grouping variable and an independence working correlation structure [labeled "GLM (robust, patient)"]. The second estimated standard errors using GEE methods with provider as the grouping variable and an independence working correlation structure [labeled "GLM (robust, provider)"]. It may seem appropriate to researchers to incorporate provider-level correlations at this stage to correct the standard errors for the grouping of patients within providers. We include this specification to see how it affects results. For comparison, we also present results from a general linear model without any post-hoc correction of the standard errors [labeled "GLM"]. We do not present results from regression models estimated using GEE methods with provider as the grouping variable and an exchangeable working correlation structure because the results would be incorrect. IPTW creates appropriately weighted pseudo-populations in which the patient characteristics are balanced between study groups, and GEE methods with non-diagonal working correlation structures would disrupt this weighting.

Finally, we used doubly robust (DR) estimation (Robins, Rotnitzky & Zhao, 1994), which is an extension of IPTW methods that augment the IPTW estimates with predicted values from outcomes models. As long as either the treatment model or the outcome model is correctly specified, doubly robust methods give consistent results. These methods may be appropriate for situations like ours where providers have effects on both treatment assignment and outcomes. For example, if we used a provider-specific treatment model along with a pooled outcome model, results should be correct. Because we used both pooled and provider-specific treatment models, we also estimate DR results using pooled and provider-specific outcome models [labeled "Doubly robust (pooled)" and "Doubly robust (provider-specific)"]. In the formulas below, $\widehat{\boldsymbol{\alpha}}_1$ is the vector of parameter estimates associated with patient characteristics, $\boldsymbol{X}_i$, from an outcome regression model based solely on data from treated patients. These parameter estimates can be applied to each patient to estimate their predicted response, $m_1(\boldsymbol{X}_i, \widehat{\boldsymbol{\alpha}}_1)$, had they received treatment. Similarly, $\widehat{\boldsymbol{\alpha}}_0$ is the vector of parameter estimates associated with patient characteristics from an outcome regression model based solely on data from the comparison patients. Applying these parameter estimates to each patient's covariate vector yields their predicted response, $m_0(\boldsymbol{X}_i, \widehat{\boldsymbol{\alpha}}_0)$, had they not received treatment. The DR estimate of treatment effect is

$$\widehat{\Delta}_{DR} = \hat{\mu}_{1,DR} - \hat{\mu}_{0,DR}$$

where

$$\hat{\mu}_{1,DR} = n^{-1} \sum_{i=1}^{N} \frac{A_i Y_i - (A_i - \hat{e}_i) m_1(\boldsymbol{X}_i, \widehat{\boldsymbol{\alpha}}_1)}{\hat{e}_i}$$

and

$$\hat{\mu}_{0,DR} = n^{-1} \sum_{i=1}^{N} \frac{(1 - A_i) Y_i + (A_i - \hat{e}_i) m_0(\boldsymbol{X}_i, \widehat{\boldsymbol{\alpha}}_0)}{1 - \hat{e}_i}$$

The standard error is the square root of the large-sample variance, given by

$$\hat{\sigma}_{DR}^2 = n^{-2} \sum_{n=1}^{N} \hat{I}_{DR,i}{}^2$$

where

$$\hat{I}_{DR,i} = \frac{A_i Y_i - (A_i - \hat{e}_i) m_1(\boldsymbol{X}_i, \widehat{\boldsymbol{\alpha}}_1)}{\hat{e}_i} - \frac{(1 - A_i) Y_i + (A_i - \hat{e}_i) m_0(\boldsymbol{X}_i, \widehat{\boldsymbol{\alpha}}_0)}{1 - \hat{e}_i} - \widehat{\Delta}_{DR}$$

Metrics

For each simulation scenario (including the comparison data) and for each combination of propensity score method, treatment model and estimation method, we calculated the following metrics: Mean treatment estimate, bias, mean squared error, mean width of the 95% confidence interval based on the standard error of the estimate, and coverage probability. These were defined as:

Mean $= S^{-1} \sum_{s=1}^{S} \widehat{\Delta}_s = \overline{\Delta}$

Bias $= \overline{\Delta} - \Delta_{TRT}$

Variance $= (S-1)^{-1} \sum_{s=1}^{S} [\widehat{\Delta}_s - \overline{\Delta}]^2$

Mean squared error $= S^{-1} \sum_{s=1}^{S} [\widehat{\Delta}_s - \Delta_{TRT}]^2$

Mean width of estimated 95% CI = $S^{-1} \sum_{s=1}^{S} 2 \cdot 1.96 \widehat{\sigma}_s$

where *S* is the number of simulated data sets, Δ_{TRT} is the true treatment effect, $\widehat{\Delta}_s$ is the estimated treatment effect for data set *s* and $\hat{\sigma}_s$ is the estimated standard error of the treatment effect for data set *s*. The coverage probability is the proportion of the S simulated data sets for which Δ_{TRT} falls within the estimated 95% confidence interval.

All simulations were conducted in SAS version 9.3 (SAS Institute Inc., Cary, North Carolina). Sample SAS code for estimating propensity scores, calculating weights, performing matching, and estimating treatment effects and standard errors is shown in Appendix 1.

Simulation Study Results

Results from propensity score matching applied to the data generated without provider effects are shown in **Table 3.2**. These results demonstrate how well different methods perform before introducing provider effects into both the treatment assignment and outcome processes. All combinations of matching methods and estimation methods shown resulted in unbiased estimates of treatment effect, but only three had coverage probabilities close to the nominal value: (1) Full matching + GEE methods, (2) full matching + ETT-weighted GLM with robust standard errors, and (3) greedy matching + GEE methods. The use of methods without any variance adjustment—weighted GLM applied to full matching results or unweighted GLM applied to greedy matching results—resulted in confidence intervals that were too wide. We also found that confidence intervals based on the Abadie estimators were too narrow.

Results from inverse probability of treatment weighting methods applied to the data generated without provider effect are shown in **Table 3.3**. Again, all treatment effect estimates were unbiased. The confidence intervals for the doubly robust estimator and Lunceford estimator had appropriate coverage. The GLM without variance adjustment resulted in confidence intervals that were too wide, while the use of robust standard errors overcompensated and resulted in confidence intervals that were too narrow.

In all the results that follow, the data used were generated with provider effects on treatment assignment. The mean treatment effect estimates and mean squared errors from estimation methods applied to propensity score matching results are shown in **Table 3.4**. None of the methods used exhibited substantial bias, with the average estimates all within 3% of the true value. For unstratified matches made using the linear predictor from the pooled treatment model, mean squared error increased as the outcome ICC increased. This was not observed when stratified matching was used or when a provider-specific treatment model was used.

The mean width of nominal 95% confidence intervals and the Monte Carlo coverage probabilities for estimation methods applied to propensity score matching results are shown in **Table 3.5**. When the outcome ICC was 0, the coverage probabilities associated with unstratified matches based on the pooled treatment model were similar to those found for data without any provider effects, where the use of GEE methods or robust standard errors performed best. Coverage probabilities declined noticeably when the outcome ICC was non-zero. Even a weak outcome ICC (0.05) dropped coverage probabilities with these methods to about 85%. Coverage probabilities associated with stratified matching and a pooled treatment model were highly variable. Stratified full matching and GEE methods resulted in especially poor coverage, although this may have been due somewhat to the bias of the actual effect estimate.

When a provider-specific treatment model was used along with propensity score matching methods, the coverage probabilities were closer to nominal for all estimation methods. Full matching and ETT-weighted GLMs with robust standard errors performed very well whether the matching was stratified or unstratified. Full matching and GEE methods performed best for unstratified matches, however, while greedy matching and GEE methods performed best for stratified matches.

To summarize, when propensity score matching was used in data situations where patients were clustered within providers and there were independent provider-level effects on both treatment and outcome, the following combinations of methods performed best: (1)

Unstratified full matching and ETT-weighted GLM with robust standard errors, (2) unstratified full matching and GEE methods, (3) stratified full matching and ETT-weighted GLM with robust standard errors, and (4) stratified greedy matching and GEE methods. All of these matches were made on the linear predictor from a provider-specific treatment model. Use of a pooled treatment model did not lead to estimators which had appropriate coverage when the outcome ICC was greater than zero.

The mean treatment effect estimates and mean squared errors from estimation methods applied to inverse probability of treatment weighted data are shown in **Table 3.6**. As with matching, none of these estimators exhibit substantial bias, with all mean treatment estimates within 2% of the true value; and higher outcome ICC values led to higher MSE for most estimation methods when a pooled treatment model was used. The exception here was the doubly robust estimator that used a provider-specific outcome model. The doubly robust estimators consistently had the lowest mean squared errors among all methods, controlling for outcome ICC. And, in general, there was more error associated with methods based on the provider-specific treatment model and the pooled treatment model.

The mean width of nominal 95% confidence intervals and the Monte Carlo coverage probabilities for estimation methods applied to inverse probability of treatment weighted data are shown in **Table 3.7**. When a pooled treatment model was used, there were no methods that consistently achieved nominal coverage. When the outcome ICC was 0, all methods except the doubly robust methods had confidence intervals that were too wide. But when the outcome ICC was high (0.25) all estimation methods except the GLM with provider-level robust standard errors led to confidence intervals that were too narrow. In fact, the use of provider-level robust errors substantially overcorrected the standard errors, leading to confidence intervals that were twice, or more, as wide as those from other methods.

For IPTW methods using weights based on a provider-specific treatment model, coverage probabilities were much closer to nominal across all values of the outcome ICC. The exception was when a GLM was used without any variance correction. As above, the doubly robust estimators had the smallest confidence interval width coupled with, perhaps, the most appropriate coverage probabilities.

To summarize, when inverse probability of treatment weighting methods were used in situations where patients were clustered within providers and there were independent provider-level effects on both treatment and outcome, the following estimation methods performed best: (1) GLM with patient-level robust standard errors, (2) GLM with provider-level robust standard errors, (3) the Lunceford estimator, and (4) either doubly robust method. All methods that made use of weights based on treatment probabilities from a pooled model led to estimators with undercoverage at high ICC values for the outcome.

Clinical example

To demonstrate how different estimation methods associated with propensity score matching and inverse probability of treatment weighting may lead to difference variance estimates in practice, we revisit an analysis of the effect of different inotropes on short-term mortality. Milrinone and dobutamine are both inotropic agents that are used to increase the cardiac output of patients with decompensated heart failure (Coons, McGraw, & Murali, 2011), but the mechanism of action for each differs. (Dopamine is also in this class of medications, but its use in the population we studied, described below, was negligible.) Very few studies have compared milrinone and dobutamine directly. A few small studies found no difference in the effect of each medication on improvement of clinical symptoms (Karlsberg et al., 1996; Aranda et al., 2003; Yamani et al., 2001), but these studies were too small to evaluate clinical endpoints like short-term or long-term mortality. Abraham, et al. (2005) used data from the Acute Decompensated Heart Failure Registry (ADHERE) to examine over 5000 patients who received inotropes. After adjustment, they found significantly lower in-hospital mortality associated with milrinone compared to dobutamine [odds ratio (95% confidence interval) = 0.81 (0.65, 0.97)]. Their statistical methods differ from ours in that they did not account for clustering of patients within hospitals and they included the estimated propensity score as a covariate in a regression analysis, instead of using propensity score matching or inverse probability of treatment weighting.

For this analysis, we also used data from ADHERE registry (Adams, et al., 2005). This registry included heart failure hospitalizations from over 300 hospitals, so in this analysis providers are hospitals. Between 2001 and 2004, the registry collected information about the timing of specific medications received. We included only patients who received either milrinone or dobutamine within the first 48 hours of admission. This differs from the population used by Abraham, et al. (2005), which included inotrope receipt at any point in the hospitalization. We used this definition of exposure to identify patients receiving inotropes as an initial therapy, excluding those patients who may have received inotropes as a response to worsening heart failure later in a hospital stay. For the purposes of this analysis, we considered patients receiving milrinone as the treatment group and patients receiving dobutamine as the comparison group. The outcome of interest was in-hospital mortality, as recorded in the registry.

As above, we estimated each patient's probability of treatment (milrinone) using two different regression models. The first treatment model was a pooled model, ignoring potential hospital effects. The second treatment model was hospital-specific, estimated using a hierarchical logistic regression with hospital-specific random effects around the mean intercept. Predictors included in the treatment model were those that were believed to impact both medication selection and in-hospital mortality. These are shown in **Table 3.8**.

After estimating these models, we used the methods similar to those described above to estimate the treatment effect and associated confidence interval of milrinone, compared to dobutamine, on in-hospital mortality. The methods for this example differed from those used in the simulation since the research question required methods appropriate for dichotomous outcomes. Any regression models utilized—those with and without robust standard errors, including those that used GEE methods—were specified as generalized linear models with a log link and a binary error distribution. This specification leads to the direct estimation of relative risk, which is important for any models that utilize weighting, as relative risks are collapsible, unlike odds ratios. The IPTW and doubly robust relative risk estimators are

$$\widehat{RR}_{IPTW} = \frac{\hat{\mu}_{1,IPTW}}{\hat{\mu}_{0,IPTW}}$$

and

$$\widehat{RR}_{DR} = \frac{\hat{\mu}_{1,DR}}{\hat{\mu}_{0,DR}}$$

respectively, given same $\hat{\mu}_{1,IPTW}$, $\hat{\mu}_{0,IPTW}$, $\hat{\mu}_{1,DR}$, and $\hat{\mu}_{0,DR}$ calculated above. The Lunceford large-sample variance estimator for the IPTW treatment effect is specific to linear outcomes (and risk differences), but we used a similar estimator from Williamson, Forbes & White (2014) for the variance of the log-relative risk, given as

$$\hat{\sigma}^2_{RR,IPTW} = n^{-1}\left[\hat{V}_{un} - \hat{\mathbf{v}}'\left(2\widehat{\mathbf{M}}_1 - \widehat{\mathbf{M}}_2\right)\hat{\mathbf{v}}\right]$$

where

$$\hat{V}_{un} = \frac{1}{n\,\widehat{w}_1^2\hat{\mu}_{1,IPTW}^2}\sum_{i=1}^{n}\frac{A_i\left(Y_i - \hat{\mu}_{1,IPTW}\right)^2}{\hat{e}_i{}^2} + \frac{1}{n\,\widehat{w}_0^2\hat{\mu}_{0,IPTW}^2}\sum_{i=1}^{n}\frac{(1-A_i)\left(Y_i - \hat{\mu}_{0,IPTW}\right)^2}{(1-\hat{e}_i)^2}$$

$$\hat{\mathbf{v}} = \frac{1}{n\,\widehat{w}_1\hat{\mu}_{1,IPTW}}\sum_{i=1}^{n}\frac{\mathbf{x}_i A_i\left(Y_i - \hat{\mu}_{1,IPTW}\right)(1-\hat{e}_i)}{\hat{e}_i}$$

$$+\frac{1}{n\,\widehat{w}_0\hat{\mu}_{0,IPTW}}\sum_{i=1}^{n}\frac{\mathbf{x}_i(1-A_i)\left(Y_i - \hat{\mu}_{0,IPTW}\right)\hat{e}_i}{1-\hat{e}_i}$$

$$\widehat{\mathbf{M}}_1 = \left(n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i{}'\,\hat{e}_i(1-\hat{e}_i)\right)^{-1}$$

$$\widehat{\mathbf{M}}_2 = \widehat{\mathbf{M}}_1\left(n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i{}'\,(A_i-\hat{e}_i)^2\right)\widehat{\mathbf{M}}_1$$

$$\widehat{w}_1 = n^{-1}\sum_{i=1}^{n}\frac{A_i}{\hat{e}_i}$$

and

$$\widehat{w}_0 = n^{-1}\sum_{i=1}^{n}\frac{1-A_i}{1-\hat{e}_i}$$

The large-sample variance of the log of the doubly robust relative risk estimator was taken from a SAS macro by Funk, Westreich, Weisen, & Davidian (2010), as:

$$\hat{\sigma}^2_{RR,DR} = n^{-1}\left[\frac{s_{1,DR}^2}{\hat{\mu}_{1,DR}^2} + \frac{s_{0,DR}^2}{\hat{\mu}_{0,DR}^2} - \frac{2s_{10,DR}}{\hat{\mu}_{1,DR}\,\hat{\mu}_{0,DR}}\right]$$

where $s_{1,DR}^2$ and $s_{0,DR}^2$ are the estimated variances of the patient-level components of $\hat{\mu}_{1,DR}^2$ and $\hat{\mu}_{0,DR}^2$, say d_{1i} and d_{0i}, respectively, given by

$$d_{1i} = \frac{A_i Y_i - (A_i - \hat{e}_i) m_1(\boldsymbol{X}_i, \widehat{\boldsymbol{\alpha}}_1)}{\hat{e}_i}$$

and

$$d_{0i} = \frac{(1 - A_i) Y_i + (A_i - \hat{e}_i) m_0(\boldsymbol{X}_i, \widehat{\boldsymbol{\alpha}}_0)}{1 - \hat{e}_i}$$

and $s_{10,DR}$ is the estimated covariance of these values. The only method from the simulation study without an analogue for dichotomous data was the Abadie estimator.

Clinical Example Results

In the ADHERE-HF data, 6112 patients with heart failure from 81 hospitals received either milrinone or dobutamine within 48 hours of hospital admission. The milrinone treatment rate was 38%, but ranged from 5% to 85% at different hospitals. The intraclass correlation for this treatment was 0.33. The in-hospital mortality outcome rate was just under 10%. This ranged from 0% to 32% across hospitals and had an associated intraclass correlation of 0.05. The correlation between the hospital effects on treatment and outcome was neglible (<0.02). These intraclass correlation coefficients were estimated as $\rho = \frac{\sigma_u^2}{\sigma_u^2 + \pi^2/3}$, where σ_u^2 was the estimated variance of the hospital random effects about the overall intercept in a hierarchical logistic model (Rodrıguez & Elo, 2003). Characteristics of each study group are shown in Table 3.8. While there were some statistically significant differences between the groups with respect to demographics and initial symptoms (e.g. rales, edema, and congestion), there were many similarities in medical history and initial lab results. The observed relative risk of the effect of milrinone vs. dobutamine on mortality was 0.76.

The relative risks and confidence intervals estimated by inverse probability of treatment weighting methods are shown in **Table 3.9**. All estimates of relative risk are near 0.92 and none were found to be significantly different from null (1.0). These estimates did not vary by the specification of the treatment model, although we did see slightly wider confidence intervals for most estimation methods that used weights based on the hospital-specific treatment model compared to weights based on the pooled treatment model. The relative risk of 0.94 estimated by doubly robust methods was the result most unlike the others. Given that this estimate was based on a pooled treatment model and a pooled outcome model, both of which ignore hospital effects that are known to exist, it is probably safe to assume this result is incorrect.

As expected based on the simulation results, the confidence intervals from the weighted model without any variance correction [labeled GLM] were the narrowest among all the IPTW methods. Also as expected, the widest confidence intervals were those associated with hospital-level robust standard errors [GLM (robust, provider)]. Among the other methods with more appropriate standard error estimates, the Williamson estimator had the smallest confidence interval. Theoretically, the doubly robust methods should have the smallest confidence intervals (Lunceford & Davidian, 2004), but these were based on approximations and not more precise large-sample formulas.

The relative risks and confidence intervals estimated by propensity score matching methods are shown in **Table 3.10**. There is more variability in the treatment effect estimates among these methods than there was among the IPTW methods, although almost all estimates are still not significantly different from 1.0. This variability may be related to the proportion of records that were able to be matched by each method. Only the unstratified full matches were able to assign all patients to matched sets. Because of the calipers employed in the matching, about 70% of the dobutamine patients and about 90% of the milrinone patients were able to be matched when full matching was performed within hospitals. Unstratified greedy matching based on the pooled treatment model was able to match 96% of milrinone patients to 59% of dobutamine patients. When a hospital-specific treatment model was used, these dropped to 67% and 41%, respectively. For greedy matches performed within hospitals, fewer than 60% of milrinone patients were matched and only 35% of dobutamine patients were matched. Incomplete matching can result in biased treatment effect estimates (Rosenbaum & Rubin, 1985), especially if the true effect differs by estimated propensity score, which may be happening here.

The proportion of patients matched does not explain why the estimates from GEE methods applied to unstratified full matching results were consistently stronger in favor of milrinone than estimates from the other estimation methods based on the same matches. Differences were less noticeable between results from GEE methods and other methods applied to stratified full matching results. More exploration is needed to explain these results.

For the two methods that seemed to result in estimators with the appropriate coverage based on the simulation work—ETT-weighted regressions with robust standard errors based from full matching results and GEE methods applied to greedy matching results, where both matches were based on hospital-specific treatment model results—the width of the estimated confidence intervals were similar.

Discussion

We sought to understand how well different propensity score methods performed when patients were clustered, as by provider, and there were cluster-level effects on both treatment and outcomes. Unlike previous work, we only examined situations where the cluster effects on treatment and outcome were uncorrelated, since this is a scenario that is more likely to be encountered in clinical research than a scenario in which these effects are correlated. Also, because we knew, from prior work, that treatment effect estimates from most propensity score methods would be unbiased, we were keen to investigate the quality of inference for these methods. To do so, we applied multiple treatment effect estimation methods that calculated standard errors of the estimate in different ways.

The estimation methods that consistently led to appropriate standard errors for and confidence intervals about the treatment effect estimate were those that started by estimating a provider-specific propensity score treatment model. Inverse probability of treatment weighting methods apply weights to each patient based on the patient-specific probabilities estimated by this model. The following IPTW estimation methods led to confidence intervals having the proper coverage: A GLM with patient-level or provider-level robust standard errors; the Lunceford estimator; and doubly robust methods.

The only IPTW-based method that did not lead to accurate confidence intervals was the GLM without any sort of variance correction. It is worth noting that because of how the weights are created, the sum of the weights within each study group is roughly the total number of patients in the entire sample. This means the total effectively sample size for the weighted data is twice the original sample size. This extra sample size is irrelevant for the Lunceford estimator or for the doubly robust estimators. Even the use of GEE methods to calculate robust standard errors handles these weights correctly. But maybe this extra sample size is what led to artificially small standard errors from the basic GLM. To test this, we performed post-hoc simulations based on standardized weights (Robins, Hernán, & Brumback, 2000). Standardized weights are created by scaling the weights for patients in each study group by the observed proportion of patients in that study group, yielding a new set of weights that sum to the original sample size. We found that the use of standardized weights was not enough to inflate the standard errors up to where they should have been when using a basic GLM. Confidence intervals were still too narrow.

The use of doubly robust estimators in this situation should be appealing. They are relatively easy to calculate and resulted in the narrowest confidence intervals of all methods that achieved nominal coverage of the estimated 95% confidence intervals. These estimators may appeal to researchers who do not want to estimate a provider-specific treatment effect, since it is possible to model the treatment process as a provider-specific process while modeling the outcome process as a pooled process.

Propensity score matching methods that performed best were those that used the linear predictors from a provider-specific treatment model to match treated patients to comparison patients for analysis. The combinations of matching and estimation methods that were most effective were: Unstratified full matching and either GEE methods or ETT-weighted GLM with robust standard errors; stratified full matching and ETT-weighted GLM with robust standard errors; and stratified greedy matching and GEE methods.

We found that estimation methods that were conditional on the matched sets outperformed those that ignored the matched sets. In this way, our results agree with Austin (2008) and others that argue for the use of conditional methods. While Shafer and Kang

(2008) are correct that there is no overt dependence between matched treated and comparison patients, the idea that their outcomes may be correlated is not so far-fetched. In our simulation, patients who were more likely to receive the treatment were also more likely to have poor outcomes.

It was reassuring to see that full matching performed well in our simulations. Full matching addresses a major drawback associated with 1:1 matching methods—the problem of discarded data. Full matching methods utilize all records in a sample, whereas the maximum possible proportion of records matched with 1:1 greedy (or optimal) matching is twice the rate of the smaller study group. In a sample with a 20% treatment rate, this leaves over half of the sample unused when estimating the treatment effect. Trying to match more comparison records to treated records through the use of *m*:1 fixed ratio matching is often ineffective. It has been shown that as *m* increases, substantial imbalance between study groups can be introduced (Hansen, 2004). The issue of discarded data may not be of utmost importance when comparing a treatment group to an untreated comparison group, but it is potentially very important for comparative effectiveness research when both study groups include actively treated patients. Head-to-head comparisons of medication, of dosage, or of different treatment modalities are examples where this occurs. Full matching may be the most appropriate matching method for these questions, since it is the only way to ensure complete matching of patients in both study groups.

It is worth mentioning that we did briefly explore optimal matching methods in our simulation study, but found an unusual situation. While optimal matching was able to match more treated patients to comparison patients than greedy matching, these extra matches were made at the far end of the allowable matching range (i.e. near the size of the calipers used).

And because of the data generation process used, small changes in the linear predictor were so highly associated with changes in outcome that there was bias in the effect estimates due to these additional matched sets. In practice we would not expect this. Rather, we would expect results from optimal matching that are quite close to those from greedy matching.

One set of methods available for non-parametric estimation of confidence intervals around an estimated treatment effect that we did not explore is bootstrap methods (Efron & Tibshirani, 1994). This has been advocated for use with inverse probability of treatment weighting methods (Curtis, Hammill, Eisenstein, Kramer & Anstrom, 2007) although it is time consuming and can be resource intensive. For propensity score matching, the proper use of bootstrap methods is less clear. Austin and Small (2014) presented a bootstrap method for matching that he found to generate appropriate confidence intervals. But Abadie and Imbens (2008) have argued strongly that bootstrap estimates, in general, are inappropriate when applied to matching methods.

In general, there are a few ways to consider addressing clustering of patients within provider in an analysis that uses propensity score methods. First, clustering can be addressed by including provider-specific factors in the treatment model. This, as noted above, most consistently led to accurate confidence intervals and the correct inference. Both doubly robust estimators worked well, for example, because for our simulated data a provider-specific treatment model was the correct specification of that model. Second, clustering can be addressed in the application of the propensity score methods. This is a method specific to matching. We found that if provider was ignored in the treatment model, but used for stratified matching, the confidence intervals were often too narrow, sometimes considerably so. However, if provider was incorporated into the treatment model, it turns out that matching within provider was not absolutely necessary for the calculation of appropriate standard errors. Third, clustering can be addressed during the estimation of the treatment effect. This is a method specific to inverse probability of treatment weighting methods, since there is no practical way to include provider in regression models when propensity score matching is used and the matched sets themselves need to be accounted for. Unfortunately, regression methods using inverse probability of treatment weighted data based on a pooled treatment model did not yield accurate confidence intervals by simply requesting provider-level robust standard errors. It's likely, by the way, that the provider-level robust standard errors performed as well as the patient-level robust standard errors when a provider-specific treatment model was used because weights based on that treatment model minimized or eliminated the correlation between provider and treatment. In short, as found in prior research (Austin, Grootendorst, & Anderson, 2007; Brookhart et al., 2006) when cluster is a confounding factor between treatment and outcome, even if it does not have the kind of correlated effect that would yield bias of the treatment estimate, it seems that it should be included in the treatment model like any other confounder.

Table 3.1 - List of Estimation Methods Utilized

Label	Weight	Estimation Method
Associated with propensity score matching		
GLM	None	General linear model
GLM (ETT)	ETT	General linear model
GLM (ETT, robust)	ETT	General linear model with robust standard errors via GEE methods with matched set-level independence working correlation structure
GEE	None	GEE methods with matched set-level exchangeable working correlation structure
Abadie	--	Abadie method: Δ_{ABAD}, $\hat{\sigma}^2_{ABAD}$
Associated with inverse probability of treatment weighting		
GLM	↑ IPTW ↓	General linear model
GLM (robust, patient)		General linear model with robust standard errors via GEE methods with patient-level independence working correlation structure
GLM (robust, provider)		General linear model with robust standard errors via GEE methods with provider-level independence working correlation structure
Lunceford		Lunceford method: $\widehat{\Delta}_{IPTW}$, $\hat{\sigma}^2_{IPTW}$
Doubly robust (pooled)		Doubly robust method: $\widehat{\Delta}_{DR}$, $\hat{\sigma}^2_{DR}$ $\{$ Pooled outcome model
Doubly robust (provider-specific)		Provider-specific outcome model

Abbreviations: GLM = Generalized linear model; ETT = Effect of treatment on the treated; GEE = Generalized estimating equations

Table 3.2 - Simulation results for propensity score matching methods on data generated without provider effects, by match type and estimation method. True treatment difference = 2.0.

Match type	Estimation method	Mean Treatment Estimate	Mean Squared Error (x1000)	Mean Width of 95% CI	Coverage Probability
Full	GLM (ETT)	1.998	2.37	0.315	99.9
	GLM (ETT, robust)	1.998	2.37	0.194	95.1
	GEE	1.997	1.66	0.169	95.8
	Abadie	1.998	2.37	0.164	90.4
Greedy	GLM	1.991	2.09	0.406	100.0
	GEE	1.991	2.09	0.185	95.3

Abbreviations: CI = Confidence interval; GLM = Generalized linear model; ETT = Effect of treatment on the treated; GEE = Generalized estimating equations

Table 3.3 - Simulation results for inverse probability of treatment weighting methods on data generated without provider effects, by match type and estimation method. True treatment difference = 2.0.

Estimation method	Mean treatment estimate	Mean Squared Error (x1000)	Mean width of 95% CI	Coverage probability
GLM	1.998	2.37	0.315	99.9
GLM (robust, patient)	1.998	2.37	0.164	90.4
Lunceford	1.998	2.37	0.194	95.1
Doubly robust (pooled)	1.991	2.09	0.185	95.3

Abbreviations: CI = Confidence interval; GLM = Generalized linear model

Table 3.4 - Simulation results for propensity score matching methods. Mean estimated treatment difference and mean squared error, by match type and estimation method. True treatment difference = 2.0.

		Mean Treatment Estimate			Mean Squared Error (x1000)		
Match type	Estimation method	ICC 0.00	ICC 0.05	ICC 0.25	ICC 0.00	ICC 0.05	ICC 0.25
		Pooled treatment model					
Unstratified							
Full	GLM (ETT)	1.999	2.000	2.001	1.03	2.14	7.15
	GLM (ETT, robust)	1.999	2.000	2.001	1.03	2.14	7.15
	GEE	1.999	2.001	2.001	0.71	1.73	6.86
	Abadie	1.999	2.000	2.001	1.03	2.14	7.15
Greedy	GLM	1.992	1.992	1.991	1.00	2.10	7.56
	GEE	1.992	1.992	1.991	1.00	2.10	7.56
Stratified							
Full	GLM (ETT)	1.973	1.971	1.971	2.78	2.99	2.95
	GLM (ETT, robust)	1.973	1.971	1.971	2.78	2.99	2.95
	GEE	1.943	1.941	1.942	5.15	5.64	5.57
	Abadie	1.973	1.971	1.971	2.78	2.99	2.95
Greedy	GLM	1.968	1.969	1.966	2.64	2.75	2.88
	GEE	1.968	1.969	1.966	2.64	2.75	2.88
		Provider-specific treatment model					
Unstratified							
Full	GLM (ETT)	2.041	2.042	2.045	5.99	5.92	6.54
	GLM (ETT, robust)	2.041	2.042	2.045	5.99	5.92	6.54
	GEE	2.028	2.024	2.017	3.26	3.49	3.96
	Abadie	2.041	2.042	2.045	5.99	5.92	6.54
Greedy	GLM	2.024	2.023	2.019	9.36	9.63	10.76
	GEE	2.024	2.023	2.019	9.36	9.63	10.76
Stratified							
Full	GLM (ETT)	1.983	1.983	1.983	1.77	1.69	1.88
	GLM (ETT, robust)	1.983	1.983	1.983	1.77	1.69	1.88
	GEE	1.966	1.966	1.965	2.33	2.47	2.64
	Abadie	1.983	1.983	1.983	1.77	1.69	1.88
Greedy	GLM	1.985	1.986	1.985	1.33	1.32	1.40
	GEE	1.985	1.986	1.985	1.33	1.32	1.40

Abbreviations: ICC = Intraclass correlation coefficient; GLM = Generalized linear model; ETT = Effect of treatment on the treated; GEE = Generalized estimating equations

Table 3.5 - Simulation results for propensity score matching methods. Mean width of 95% confidence interval and coverage probability, by match type and estimation method.

		Mean width of 95% CI			Coverage probability		
		ICC	ICC	ICC	ICC	ICC	ICC
Match type	Estimation method	0.00	0.05	0.25	0.00	0.05	0.25
			Pooled treatment model				
Unstratified							
Full	GLM (ETT)	0.232	0.233	0.237	100.0	98.4	84.2
	GLM (ETT, robust)	0.135	0.138	0.151	96.8	85.8	63.3
	GEE	0.117	0.119	0.131	97.2	83.9	57.2
	Abadie	0.111	0.114	0.125	92.2	78.3	53.0
Greedy	GLM	0.283	0.284	0.290	100.0	99.6	90.4
	GEE	0.125	0.128	0.140	95.5	83.8	58.0
Stratified							
Full	GLM (ETT)	0.232	0.233	0.238	97.0	95.9	97.1
	GLM (ETT, robust)	0.156	0.156	0.156	87.4	86.1	87.1
	GEE	0.125	0.125	0.125	56.8	56.6	55.0
	Abadie	0.114	0.115	0.114	75.1	73.0	73.1
Greedy	GLM	0.316	0.318	0.325	99.5	99.3	99.3
	GEE	0.139	0.139	0.139	84.0	84.5	81.9
			Provider-specific treatment model				
Unstratified							
Full	GLM (ETT)	0.232	0.233	0.238	86.7	87.1	85.5
	GLM (ETT, robust)	0.321	0.323	0.333	97.2	97.2	95.8
	GEE	0.244	0.246	0.252	97.9	97.5	96.4
	Abadie	0.220	0.221	0.227	84.6	85.4	83.6
Greedy	GLM	0.315	0.316	0.323	89.3	89.0	88.5
	GEE	0.268	0.270	0.276	84.1	83.0	82.2
Stratified							
Full	GLM (ETT)	0.230	0.231	0.236	99.4	99.8	99.1
	GLM (ETT, robust)	0.156	0.156	0.156	93.8	94.3	93.1
	GEE	0.122	0.122	0.122	79.1	77.9	75.7
	Abadie	0.112	0.112	0.112	82.4	81.5	79.7
Greedy	GLM	0.316	0.318	0.324	100.0	100.0	100.0
	GEE	0.136	0.136	0.136	94.5	93.9	92.4

Abbreviations: CI = Confidence interval; ICC = Intraclass correlation coefficient; GLM = Generalized linear model; ETT = Effect of treatment on the treated; GEE = Generalized estimating equations

Table 3.6 - Simulation results for inverse probability of treatment weighting methods. Mean estimated treatment difference and mean squared error, by estimation method. True treatment difference = 2.0.

	Mean Treatment Estimate			Mean Squared Error (x1000)		
Estimation method	ICC 0.00	ICC 0.05	ICC 0.25	ICC 0.00	ICC 0.05	ICC 0.25
Pooled treatment model						
GLM	1.983	1.986	1.983	2.37	3.58	8.84
GLM (robust, patient)	1.983	1.986	1.983	2.37	3.58	8.84
GLM (robust, provider)	1.983	1.986	1.983	2.37	3.58	8.84
Lunceford	1.983	1.986	1.983	2.37	3.58	8.84
Doubly robust (pooled)	2.000	1.999	2.000	0.68	1.77	7.02
Doubly robust (provider-specific)	2.000	2.000	2.000	0.68	0.90	1.00
Provider-specific treatment model						
GLM	1.964	1.970	1.962	7.23	6.78	7.26
GLM (robust, patient)	1.964	1.970	1.962	7.23	6.78	7.26
GLM (robust, provider)	1.964	1.970	1.962	7.23	6.78	7.26
Lunceford	1.964	1.970	1.962	7.23	6.78	7.26
Doubly robust (pooled)	2.001	2.001	2.000	1.05	1.12	1.27
Doubly robust (provider-specific)	2.001	2.001	2.000	1.05	1.12	1.20

Abbreviations: ICC = Intraclass correlation coefficient; GLM = Generalized linear model

Table 3.7 - Simulation results for inverse probability of treatment weighting methods. Mean width of 95% confidence interval and coverage probability, by estimation method.

	Mean width of 95% CI			Coverage probability		
	ICC	ICC	ICC	ICC	ICC	ICC
Estimation method	0.00	0.05	0.25	0.00	0.05	0.25
Pooled treatment model						
GLM	0.241	0.242	0.246	98.5	95.8	80.9
GLM (robust, patient)	0.328	0.330	0.333	99.9	99.2	91.9
GLM (robust, provider)	0.637	0.650	0.696	100.0	100.0	99.9
Lunceford	0.303	0.306	0.308	99.8	98.8	88.6
Doubly robust (pooled)	0.103	0.106	0.118	95.8	79.3	52.1
Doubly robust (provider-specific)	0.103	0.103	0.103	95.6	91.7	89.7
Provider-specific treatment model						
GLM	0.241	0.242	0.246	84.0	85.7	85.4
GLM (robust, patient)	0.426	0.431	0.428	97.9	98.7	98.3
GLM (robust, provider)	0.403	0.410	0.405	97.5	98.5	97.8
Lunceford	0.402	0.408	0.404	96.9	98.0	97.3
Doubly robust (pooled)	0.126	0.130	0.145	94.3	95.0	96.3
Doubly robust (provider-specific)	0.126	0.125	0.123	94.3	93.9	93.0

Abbreviations: CI = Confidence interval; ICC = Intraclass correlation coefficient; GLM = Generalized linear model

Table 3.8 - Characteristics of ADHERE-HF patients receiving milrinone and dobutamine

Variable	Dobutamine (N = 3794)	Milrinone (N = 2318)	Standardized Difference, %
Demographics			
Age (years), Mean (SD)	69.4 (13.8)	65.7 (14.3)	26.3
Gender, Male	2,445 (64.4%)	1,595 (68.8%)	9.3
Race			8.7
White	2,788 (73.5%)	1,612 (69.5%)	
Black	683 (18.0%)	478 (20.6%)	
Other/unknown	323 (8.5%)	228 (9.8%)	
Medical History			
Anemia	1,970 (51.9%)	1,207 (52.1%)	0.3
Atrial fibrillation	1,358 (35.8%)	822 (35.5%)	0.7
Coronary artery disease	2,555 (67.3%)	1,496 (64.5%)	5.9
Chronic renal insufficiency	1,559 (41.1%)	912 (39.3%)	3.6
Chronic obstructive pulmonary disease/Asthma	1,178 (31.0%)	645 (27.8%)	7.1
Diabetes mellitus	1,617 (42.6%)	979 (42.2%)	0.8
Hyperlipidemia	1,481 (39.0%)	874 (37.7%)	2.7
Hypertension	2,317 (61.1%)	1,377 (59.4%)	3.4
Prior myocardial infarction	1,514 (39.9%)	812 (35.0%)	10.1
Peripheral vascular disease	759 (20.0%)	427 (18.4%)	4.0
Prior stroke/Transient ischemic attack	564 (14.9%)	373 (16.1%)	3.4
Current smoker	460 (12.1%)	278 (12.0%)	0.4
Pacemaker, any	1,256 (33.1%)	821 (35.4%)	4.9
Implantable cardioverter defibrillator	790 (20.8%)	616 (26.6%)	13.6
Initial evaluation			
Fatigue	1,580 (41.6%)	1,020 (44.0%)	4.8
Rales	2,439 (64.3%)	1,336 (57.6%)	13.7
Edema	2,571 (67.8%)	1,409 (60.8%)	14.6
Congestion	2,294 (60.5%)	1,266 (54.6%)	11.9
Ejection fraction			15.7
< 40	2,797 (73.7%)	1,859 (80.2%)	
40	589 (15.5%)	257 (11.1%)	
Unknown	408 (10.8%)	202 (8.7%)	
Initial vital signs and laboratory results			
Systolic blood pressure (mmHg), Mean (SD)	120.1 (28.5)	119.5 (26.3)	2.1
BUN (mg/dL), Mean (SD)	43.3 (27.1)	40.5 (25.8)	10.6
Serum sodium (mmol/L), Mean (SD)	136.4 (5.3)	136.6 (5.0)	4.4
Hemoglobin (g/dL), Mean (SD)	12.6 (2.5)	12.5 (2.4)	2.2

Values for Dobutamine and Milrinone groups presented as N (%) unless otherwise specified
Abbreviation: SD = Standard deviation

Table 3.9 - Estimated relative risk and 95% confidence interval from inverse probability of treatment weighting methods for association between milrinone (vs. dobutamine) and in-hospital mortality.

Estimation method	Relative Risk (95% CI)
Pooled treatment model	
GLM	0.92 (0.83, 1.03)
GLM (robust, patient)	0.92 (0.78, 1.09)
GLM (robust, provider)	0.92 (0.73, 1.17)
Williamson	0.92 (0.79, 1.08)
Doubly robust (pooled)	0.94 (0.79, 1.12)
Doubly robust (provider-specific)	0.91 (0.77, 1.08)
Provider-specific treatment model	
GLM	0.92 (0.83, 1.03)
GLM (robust, patient)	0.92 (0.74, 1.15)
GLM (robust, provider)	0.92 (0.71, 1.19)
Williamson	0.92 (0.79, 1.07)
Doubly robust (pooled)	0.93 (0.76, 1.14)
Doubly robust (provider-specific)	0.92 (0.75, 1.12)

Abbreviations: CI = Confidence interval; GLM = Generalized linear model

Table 3.10 - Estimated relative risk and 95% confidence interval from propensity score matching methods for association between milrinone (vs. dobutamine) and in-hospital mortality.

Match type	Estimation method	Relative Risk (95% CI)
	Pooled treatment model	
Unstratified		
Full	GLM (ETT)	0.90 (0.75, 1.09)
	GLM (ETT, robust)	0.90 (0.75, 1.10)
	GEE	0.81 (0.69, 0.95)
Greedy	GLM	0.94 (0.78, 1.14)
	GEE	0.94 (0.78, 1.13)
Stratified		
Full	GLM (ETT)	0.86 (0.71, 1.04)
	GLM (ETT, robust)	0.86 (0.66, 1.11)
	GEE	0.82 (0.69, 0.97)
Greedy	GLM	0.85 (0.67, 1.07)
	GEE	0.85 (0.67, 1.07)
	Provider-specific treatment model	
Unstratified		
Full	GLM (ETT)	0.86 (0.72, 1.04)
	GLM (ETT, robust)	0.86 (0.65, 1.14)
	GEE	0.77 (0.66, 0.91)
Greedy	GLM	0.90 (0.73, 1.11)
	GEE	0.90 (0.73, 1.11)
Stratified		
Full	GLM (ETT)	0.89 (0.74, 1.09)
	GLM (ETT, robust)	0.89 (0.66, 1.21)
	GEE	0.91 (0.75, 1.09)
Greedy	GLM	0.89 (0.69, 1.13)
	GEE	0.89 (0.70, 1.13)

Abbreviations: CI = Confidence interval; GLM = Generalized linear model; ETT = Effect of treatment on the treated; GEE = Generalized estimating equations

CHAPTER 4

A SAS MACRO FOR OPTIMAL MATCHING AND FULL MATCHING ON PROPENSITY SCORES

Introduction

The use of propensity score matching methods to balance covariates between a treated group of patients and a comparison group of patients in clinical and epidemiological research is widespread (Stürmer et al., 2006). Greedy matching (Rosenbaum & Rubin, 1985a), a method in which patients from each study group are matched one at a time and without reconsideration, is the most frequently used matching method in these studies (Austin, 2008). This preponderance of greedy matching is potentially unwarranted, given that other, often superior, matching methods have been described.

Optimal matching (Rosenbaum, 1989) and full matching (Rosenbaum, 1991) are two such methods proposed as alternatives to greedy matching. Unlike greedy matching, which seeks closely matched pairs without regard to the overall distance between matched sets, both optimal matching and full matching allow matches to be reconsidered in order to minimize the total distance between matched sets of treated and comparison patients. The difference between optimal matching and full matching is in the make-up of the matched sets. Optimal matching, like greedy matching, is a type of fixed ratio matching, where all resulting matched sets contain the same number of patients from each study group, with 1:1 matches most common. With fixed ratio matching, many records remain unmatched. Full matching addresses this limitation by allowing flexibility in the make-up of the matched sets—multiple treatment records can be matched to a single comparison record and vice versa—which results in the utilization of all records in both study groups.

When these matching methods have been compared, full matching has been shown to produce closer matched sets and better covariate balance between study groups than either optimal or greedy 1:1 matching (Ming & Rosenbaum, 2000; Gu & Rosenbaum, 1993). Others have demonstrated that trying to utilize more control patient data by performing *m*:1 matches introduces greater imbalances in the matched sets (Hansen, 2004) compared to 1:1 matching. When matching on a single variable, greedy matching and optimal matching often produce similar results, especially if there are a substantial number of control patients available for matching to each treated patient. When the number of controls per treated patient is low, optimal matching produces better matches (Gu & Rosenbaum, 1993).

The scarcity of optimal matching and full matching applications in the literature is likely due to fact that, as optimization problems, they are difficult to implement. Rosenbaum (1989 & 1991) demonstrated that one way to approach both optimal matching and full matching is by leveraging the theory and mechanics of network flow problems. By casting these matching problems as network flow problems, linear programming solvers can be employed to find the minimum distance between the two sets of records (Tardos & Kleinberg, 2006). Submitting data to these solvers requires understanding how to appropriately specify and program the nodes and arcs of a network problem. And recovering the matched sets from the solution require understanding how to identify connected records in the resulting network (Tardos & Kleinberg, 2006). Neither are trivial. And while SAS macros exist for greedy matching, full-featured macros for optimal matching and full matching do not exist (Bergstralh & Kosanke, 2003).

In this paper, we present a SAS macro for implementing both optimal matching and full matching on a scalar variable, such as the propensity score, using optimization tools found within SAS/OR, SAS's operations research software. We also review the methods and mechanics of matching and we will demonstrate the use of the macro with data from a clinical application.

Matching Methods

The propensity score was introduced by Rubin and Rosenbaum in the mid-1980s as a balancing score to be used for balancing measured characteristics between two study groups (Rosenbaum & Rubin, 1983). [Note that in comparative effectiveness research, these two study groups may both be treatment groups, treated with different modalities or strengths of the same therapy. For the purposes of this manuscript and for simplicity, we will refer to the two study groups as the treatment group and the comparison group; and we will refer to the members of these groups as patients.] The propensity score is the probability of treatment associated with each study patient. Propensity scores are usually estimated with a logistic regression having the treatment indicator as the dependent variable and factors that may confound the relationship between treatment and outcome as the predictors.

Different propensity score-based methods utilize these estimated probabilities in different ways to achieve balance, but a very common method is simply matching each patient in the treatment group to one or more patients in the comparison group having similar values on the propensity score or some function of the propensity score (e.g. linear predictor). Matching on the propensity score is a way to perform a multivariate match using a single variable, and Rubin and Rosenbaum (1985a) showed that such matching results in sets of patients from each study group that had similar distributions on measured confounders. This, in turn, allows for a direct assessment of the outcomes in each group without further adjustment for confounders. The resulting treatment effect estimate was shown to be a consistent estimator of the population-level treatment effect (Rosenbaum & Rubin, 1983).

There are many different ways to match records between two groups on a single variable like the propensity score. The simplest way to do this is to perform greedy matching, which is also sometimes referred to as nearest neighbor matching (Rosenbaum & Rubin, 1985). Implementation of greedy matching is easily programmable as a series of steps. The general algorithm can be described as follows. First, randomly order patients in the treatment group and work through them one-by-one. For each treated patient, find the comparison patient with the closest match on the matching variable and output both records as a matched set. Continue until all treated patients have been matched. This match is "greedy" in that once a patient from the comparison group is assigned, that assignment is not revisited. Most greedy matches are specified as 1:1 matches, resulting in two patients per matched set. Alternatively, researchers may specify *m*:1 fixed ratio matches. In this case, the above procedure is repeated for all patients in the treatment group and all remaining unmatched patients in the comparison group. This continues until *m* matches have been made for each treated patient.

One drawback with greedy matching is that it does not consider the total distance on the matching variables across all matched sets of treated and comparison patients. Optimal matching addresses this drawback. Optimal matching methods, in general, assign every patient in the treated group to a different patient in the comparison group simultaneously. It uses the total distance on the matching variable across all matched sets as the minimization criteria. Matching assignments are reconsidered until it is believed that the optimal solution, with the lowest possible total distance, has been found. As with greedy matching, most optimal matches are 1:1 matches, although it is possible to create *m*:1 matches in a manner similar to what is described above. After 1:1 matches are match, another *m*-1 optimal matches would be performed using all the treated patients and the unmatched comparison patients.

An obvious limitation with both greedy matching and optimal matching is that fact that there are, by definition, many records left unmatched at the end of the process that will not contribute to the outcomes analysis. While *m*:1 matching attempts to utilize more of these unmatched records than 1:1 matching, it has been shown that requiring a fixed ratio of treated to comparison patients can result in very poor matches (Hansen, 2004) which in turn can lead to bias in the treatment effect estimate. Full matching addresses these issues. The core principle of full matching is to match all records in the treatment group to all records in the comparison without requiring a fixed ratio of treated to comparison patients in the matched sets. In other words, matched set treated-to-comparison ratios are flexible and can be *m*:1 or 1:*m*, with almost no limit on the size of *m* in either direction. As with optimal matching, full matching uses the total distance on the matching variable across all matched sets as the minimization criteria and reconsiders assignments until an optimal solution has been found.

The complexity of simultaneous assignment and optimization in both optimal matching and full matching make them less amenable to programming as a series of discrete steps. Rosenbaum (1989 & 1991) recognized that each could be thought of as a minimum-cost network flow problem, a standard optimization problem (Tardos & Kleinberg, 2006). Once recast as an optimization problem, it is possible to use existing linear programming solvers to perform optimal matching and full matching. Setting up the necessary data structures and constraints for these solvers, however, can be a challenge.

Network flow problems are described using nodes and links. Links are also often referred to as arcs or edges. **Figure 4.1**, Panel A shows the general set-up for a matching problem as a network flow problem. Consider all the treated patients (1–5) in the center left-hand column, T, and all the comparison patients (a–e) in the center right-hand column, C. Each patient is a node in the network. All nodes in the treated column have directed links into all nodes in the comparison column. By adding a source node, α, to the left of the treated patients, a sink node, β, to the right of the comparisons patients and all appropriately directed links, we can imagine "flow" moving from left to right across the network. Using costs assigned to the links between patients in the treatment group and patients in the comparison group, it is possible to find the lowest cost way to move a specified amount of flow (in units) from the source node through each study group to the sink node. The resulting paths represent the final links in the matching. Figure 4.1 shows example paths for solved 1:1 optimal matching (Panel B) and solved full matching (Panel C). Note in Panel C that the full matching here has resulted in two 1:1 matches, a single 1:2 match, and a single 2:1 match. This imbalance is handled in the eventual statistical analysis.

Before the network flow problem can be solved, constraints need to be specified, including the supply of flow to put through the network, the flow costs of the links to and from the source and sink nodes, and flow capacities for all links. The first panel of **Table 4.1** lists these constraints for both optimal matching and full matching performed without stratification. The total network flow is determined by the input quantity of supply node. The sink node balances this quantity to ensure that the entire flow moves across the network. No other nodes (i.e. none of the patient nodes) initiate or terminate flow. The total units of network flow for optimal matching is the size of the smaller of the treated or comparison groups. This will result in all the patients of the smaller group being matched, but, as expected, only some patients within the larger group being matched. Total flow for full matching is, at a minimum, the size of the larger study group, which ensures all patients in both study groups get matched. For full matching, the total flow will equal the size of the larger study group only if all matched sets include only one member of the smaller study group and one or more members of the larger study group. Additional flow will be required if any of the matched sets include one member of the larger study group matched to multiple members of the smaller study group. For both unstratified optimal and full matching, the cost of a single unit of flow associated with each link from a treated patient to a comparison patient is the difference in the matching variable between those patients. In propensity score matching, this might be the difference in the linear predictors from the treatment model for each pair of patients. The capacity limits on these links are always [0, 1], with the minimum flow through the link as 0 units, or no flow, and the maximum flow as 1 unit. In other words, each link between patients can be used either once or not at all. For full matching, the link capacities out of the source node and into the sink node are all [1, ∞], meaning each of those links must be used once, but can be used multiple times. This enables both 1:*m* and *m*:1 matching. For optimal matching, the link capacities from the source node and to the sink nodes differ. For the smaller study group, the appropriate links will require 1 unit of flow across them, resulting in every patient being used. For the larger study group, the appropriate links have capacity [0, 1]. Only the patients associated with the lowest costs will be used from this group.

For stratified matches, some of these quantities change and some do not. See **Figure 4.2**, Panel A for the network set-up of a stratified matching problem. The difference from an unstratified match is that there are no links available from treated patients in one stratum to comparison patients in another. This has effects on the constraints of the problem, shown in the second panel of Table 4.1. The total flow through the network is now determined by the sum of the sizes of the smaller study group, for optimal matching, or larger study group, for full matching, within each stratum. This will only lead to a different amount of flow compared to the unstratified match of the same type if some strata have a treatment proportion under 0.5, while others have a treatment proportion greater than 0.5. Link capacities between treated and comparison patients are again based on the difference in the matching variable between pairs, with the additional note that this only applies to patients in the same stratum. Patients from different strata are not linked. One final difference has to do with the link capacities for links between the source node and the treated patients and between the comparison patients and the sink node. These capacities are now based on the sizes of the study groups within each strata. An example solution for a stratified optimal match is shown in Figure 4.2, Panel B. Note that only 4 of the 5 patients in each group are able to be matched, due to the size of each study group within each strata. An example solution for a stratified full match is shown in Figure 4.2, Panel C. Here, all patients are still matched.

There are a few common variations on the matches described above. One important variation on all matching methods is the use of calipers. In caliper matching, only records within a specified distance of each other are considered eligible to be matched. Calipers are useful for preventing the matching of records deemed to be too far apart. Unrestricted matches may result in too many disparate matches, which in turn can affect the balance on confounding factors between matched groups. The width of calipers to be used can be set by the researcher, but at least one simulation study led to a recommendation of calipers equal to 0.2 standard deviations of the matching variable (Austin, 2011). While some researchers recommend trimming when the distribution of the matching variables do not entirely overlap (Stürmer, Rothman, Avorn & Glynn, 2010), the use of calipers may be an alternative, if the width of the caliper is set at the largest difference desired. At one point, Rosenbaum (1989) actually called optimal matching within calipers his "method of choice". He also pointed out that greedy matching, compared to optimal matching, was less likely to result in a complete matching of treated patients when calipers were used. As a network problem, implementing caliper matching is as simple as removing links between treated patients and comparison patients for whom the difference in the matching variable is too large (**Figure 4.3**, Panel A).

Another variation on full matching, proposed by Hansen (2004), was to limit the maximum size of the matched sets through the use of ratio caps. Typical full matches do not constrain the number of treated patients matched to a comparison patient, or vice versa. It is possible to have matched sets from an unconstrained full match that contain an extreme number of treated or control patients—e.g. a set with a 100:1 treated to comparison ratio.

Hansen notes that constraining the ratios in the final matched set should have some advantages in the precision of estimates based on the matched data. He introduced the idea of thinning and thickening ratio caps that, when applied to the observed treatment odds, define the minimum and maximum ratios allowable for matching. He defined these as follows. Given observed Group 1:Group 2 treatment odds, d_{OBS}, a thickening cap of *u* (> 1) and a thinning cap of *l* (< 1), the maximum matching ratio, d_{MAX}, is given by:

$$d_{\mathrm{MAX}} = \begin{cases} \lceil ud_{\mathrm{OBS}} \rceil : 1, & ud_{\mathrm{OBS}} > 1 \\ 1 : \lfloor (ud_{\mathrm{OBS}})^{-1} \rfloor, & ud_{\mathrm{OBS}} \leq 1 \end{cases}$$

And the minimum matching ratio, d_{MIN}, is given by:

$$d_{\mathrm{MIN}} = \begin{cases} \lfloor ld_{\mathrm{OBS}} \rfloor : 1, & ld_{\mathrm{OBS}} > 1 \\ 1 : \lceil (lud_{\mathrm{OBS}})^{-1} \rceil, & ld_{\mathrm{OBS}} \leq 1 \end{cases}$$

These ratio limits can be implemented through manipulation of the minimum and maximum values for the link capacity associated with links out of the source node to the treated patients and associated with links from the comparison patients into the sink node. This involves replacing the ∞ found in the capacity limits shown in Table 4.1 with the finite number determined by the formulas above.

Table 4.2 shows what different values of the minimum and maximum allowable matching ratios would be for a given treatment rate and for different combinations of thinning and thickening caps. For example, assuming Group 1 is the treatment group, a treatment proportion of 25% is the same as an observed treated odds of 1:3. A thinning cap of 0.2 yields a minimum matching ratio of 1:15 and a thickening cap of 5 yields a maximum matching ratio of 2:1. While the sets of caps shown in this table are symmetric, they need not be in practice. Hansen (2004) gives some guidance on how to choose these caps.

Once the problem is fully specified, with or without calipers and with or without ratio caps, it can be solved using optimization algorithms. SAS solves general minimum cost network flow problems using the primary network simplex algorithm developed by Ahuja, Magnanti, and Orlin (1993). Because optimal matching results in 1:1 matches, it is actually a special case of minimum cost network flow problem called a linear assignment problem. While linear assignment problems can be solved with the simplex algorithm, it may be more efficiently solved by algorithms developed especially for such problems. SAS employs one developed by Jonker and Volgenant (1987).

The solution to a minimum cost network flow problem is a list of all links utilized to move the specified flow across the network. For matching, we are not interested in links from the source node to the treated patients or from the comparison patients to the sink node, and instead we need look only at the links from the treated patients to the comparison patients. Given the list of links, the next requirement is to identify connected components, defined as the set of nodes that are reachable from one another through the determined links. Identifying matched sets allows us to take account for them in the outcomes analysis. This is essential for full matching, since the size and composition of the matched sets can vary. Identifying connected components from optimal matches is simple. The two patients associated with a link are a connected component. It is not possible for other patients to be reachable by those patients because the matched sets only include one patient from each study group. In full matching, on the other hand, multiple treated patients can be matched to a single comparison patient, and vice versa. In this case, we need to use a depth-first search to traverse the network solution to identify all records in each matched set (Tardos & Kleinberg, 2006).

SAS Macro Details

The fmatch.sas macro presented here is available from http://people.duke.edu/~hammill/software. The macro utilizes optimization procedures available in SAS/OR, the operations research software available from SAS Institute, to solve the network problems described above. This macro transforms the input dataset into the format required by SAS/OR and appropriately specifies the constraints for the requested match type and match options. The macro does not estimate a propensity score, but instead requires that it, or any other quantity to be used for matching, is pre-calculated and available on the data set. There is some other, minor data preparation that must occur before calling the macro.

The macro requires users to specify an input dataset, the matching variable, the study group variable, and the record-level identifier. The matching variable must be numeric and is typically the estimated propensity score or the linear predictor from the treatment model. The study group variable typically reflects assigned treatment and must be a 2-level numeric variable. The first level of this variable, sorted numerically, defines Group 1 and the other defines Group 2. In this macro, the ordering of this variable is unimportant. A reverse-coded group variable will yield identical results. The record-level identifier can contain either numeric or character values, but must uniquely identify the records in the input dataset.

The default matching method used by the macro is full matching, but users may request an optimal match instead. Additionally, users may specify a stratification variable in the macro call, to request that matching occur within strata. For both unstratified and stratified full matching, the macro uses the network simplex algorithm described above. For unstratified optimal matching without calipers (option described below), the macro uses the linear assignment algorithm provided by SAS. For unstratified optimal matching with calipers and for all stratified optimal matches, however, the macro uses the network simplex algorithm. For stratified optimal matches, this is done because of the possibility that the number of matches to be made is less than the linear assignment algorithm expects (see Table 4.1). Regardless of the problem set-up, the linear assignment algorithm always expects $\min(N_t, N_c)$ matches to be made. For stratified matches where some strata have $N_{t,h} > N_{c,h}$ while others have $N_{t,h} < N_{c,h}$, the actual number of matches to be made will be less than $\min(N_t, N_c)$ and the linear assignment algorithm will report the solution to be, incorrectly, infeasible. For unstratified optimal matching with calipers, it may not be possible to make $\min(N_t, N_c)$ matches due to the caliper-limited number of potential links between patients in each study group. Figure 4.3, Panel A shows how this might happen. The algorithm would expect to make 5 matched sets where only 4 are possible. In this case (Figure 4.3, Panel B), extra arcs are created by the macro from all treated patients to an "excess" node and from that "excess" node to all comparison patients. The weight of these links are set high such that they are only used as a last resort to accommodate flow that would otherwise not be able to flow through the network. While this leaves some patients unmatched who we would otherwise expect to be matched, it avoids the optimization algorithms from reporting the problem as infeasible. The links utilized to direct flow for the 4 resulting matched sets will still reflect minimum cost.

Other macro options that can be specified are calipers and ratio caps, neither of which are used by default. To request calipers matching, users must specify the width of the calipers. This width can either be given directly or given as a multiple of the standard deviation. To constrain the size of the matched sets within full matching, users can specify the thinning or thickening ratio caps, discussed above, to be used. Thinning caps must be < 1 and thickening caps must be > 1. It should be noted that the thinning and thickening caps are applied to the observed Group 1:Group 2 odds. For stratified matches, these odds and ratio limits are determined within each stratum.

Two output datasets are created for the user. The link-level output dataset includes one record for every pair of linked records and contains the distance on the matching variables between each pair. The main, patient-level output dataset, includes one patient per record, along with a variable to identify the set number to which that patient belongs. Each record also includes the number of treated patients in the set, the number of comparison patients in the set, and the total number of patients in the set, which allows for the creation of individual or set weights as desired. The patient-level dataset is easily merged back to the source data. Both of these datasets include the stratification variable, if one was used.

A sample macro call is shown in Figure 4.5. Samples of the two output datasets created are shown in Figure 4.6 and 4.7. The full code for the macro is shown in Appendix 2.

Clinical Example

To demonstrate this macro, we used data from the Acute Decompensated Heart Failure National Registry (ADHERE) Core 1 registry (Adams, et al., 2005), which included patients hospitalized with acute decompensated heart failure from 2001 to 2003. The exposure of interest is receipt of high-dose intravenous loop diuretics. As the original study (Peacock, et al., 2009) noted, the optimal dose of diuretics is uncertain and some safety concerns had been raised about high doses of diuretics. For this analysis, a high dose of intravenous loop diuretics was defined as ≥160 mg within the first 24 hours of medication initiation. The comparison group comprised those who received <160 mg. There were 43,434 patients within 236 hospitals in the study population. Of these patients, 9,469 (21.8%) were treated with a high dose of diuretics. The hospital-level proportion of patients received high-dose diuretics varied substantially, ranging from under 2% to almost 56%.

We estimated a propensity score treatment model and used the linear predictor from this model to match patients in the high dose group with patients in the low dose group under multiple matching specifications. The treatment model was specified as a hierarchical logistic regression model, allowing for hospital-level random intercepts. Independent variables included in this model (shown in **Table 4.3**) were those deemed to be factors that could potentially confound the relationship between diuretic dose and in-hospital mortality, the outcome studied in the manuscript referenced above.

We used the fmatch.sas macro to perform optimal matching and full matching. For comparison, we also performed greedy matching. For optimal matches, we performed both an unrestricted match and a match with calipers. For full matching, we performed an unrestricted match as well as a match with thinning/thickening ratio caps and a match with calipers. For greedy matching, we performed 1:1 and 2:1 (treatment:comparison) matches with and without calipers. For all caliper matches, the width of the caliper was set to 0.2 standard deviations (SD) of the linear predictor, as recommended by Austin (2011). Of special interest was the impact of stratification on each of these matches. Therefore, all matches were performed both unstratified and stratified by hospital.

Several metrics were used to compare these different matches. First, we calculated the total distance between all matched patients from the treatment and comparison groups on the matching variable, the linear predictor from the treatment model. We also summarized, for matches that did not use calipers, the proportion of matches for which this distance was over 0.2SD of the matching variable. We calculated the percentage of patients in each of the treatment and comparison groups that were matched. We report the amount of computing time each match took to complete. And we calculated the standardized difference between the matched study groups for selected variables.

Clinical Example Results

As an objective measure of match quality, the total distance between all matched pairs of records is useful. Among the unstratified, unrestricted matches, **Table 4.4** shows how well full matching performs compared to all other types of matches. Even though full matching utilizes all records in both the treatment and comparison groups, its total distance is substantially lower than similarly unrestricted optimal and greedy 1:1 matches, both of which leave over 70% of the comparison patients unmatched. The proportion of matches after full matching with a distance over 0.2SD is also minimal, whereas both optimal and greedy matches result in about 5% of matches that exceed that distance. While greedy matching was the quickest method to finalize the matches, as might be expected, full matching was about 8 times quicker to finish than optimal matching. [Actual computing times will vary by specific hardware available, but the trends we observe should hold.]

Among stratified, unrestricted matches, full matching does not display quite the advantages reported above. In this case, the cost of matching all comparison patients to treated patients within hospitals comes at a cost reflected in the total distance metric. And all three matches—full, optimal, and greedy (1:1)—result in nearly 9% of matched pairs having a distance greater than 0.2SD on the linear predictor. There was no time advantage for any matching method when stratification was used. The limited number of potential matches resulted in very quick processing times even for those that required the use of optimization algorithms.

There are a few reasons for the differences in performance observed between unstratified and stratified full matching. The first is the often limited number of treated patients within a given hospital, which is associated with the total number of patients at that hospital. The second, which is related, is the lack of common support across the range of estimated propensity scores within each hospital. Consider an extreme example. When there is only 1 treated patient at a hospital, all comparison patients will be matched to that patient in 1 large matched set, by default. It is nearly certain that some of those comparison patients will have a value of the linear predictor that is far from that of the treated patient. When there are a large number of treated patients at a hospital (or when the match is unstratified) it is more likely that there will be patients from both study groups at all levels of the estimated propensity score. **Figure 4.4**, Panel A confirms this notion. When there are small numbers of treated patients at a hospital (i.e. <20), the proportion of matched sets created by unrestricted full matching that have a large difference on the matching variable is high. As the number of treated patients increases, this proportion decreases. For large hospitals with sufficient numbers of treated patients (Figure 4.4, Panel B) the proportion of matches having a large difference is low.

It may be advisable, therefore, to either only pursue stratified full matching for large hospitals with sufficient numbers of treated patients or to perform stratified full matching after disallowing distant matches. One way to do the latter is through trimming (Stürmer, 2010), which reduces the data in both study groups to the region of common support. Trimming is done by removing treated patients from the data who have values of the estimated propensity score outside the range found for the comparison patients, and vice versa. We do not perform trimming here, but instead use calipers to restrict potential matches. By defining calipers, the maximum distance acceptable for a match, one has still effectively trimmed the data. Note that both trimming and calipers can be used for stratified and unstratified matches.

Table 4.4 shows the effect of the use of calipers on the different types of matches. Because the unrestricted, unstratified full match did not result in many distant matches, the use of calipers did not generate much improvement in the total distance. It did substantially speed up the computing, however. The results were similar for unstratified optimal matching—a slight improvement in total distance with a noticeable speed improvement. For both of these matches, only a handful of patients that were able to be matched without the caliper restrictions were not able to be matched with them—1 patient remained unmatched by full matching and 10 patients remained unmatched by optimal matching. When calipers were used in conjunction with unstratified greedy matching, the total distance dropped substantially, but at the expense of the number of matched patients. Over 300 treated patients that were matchable when using optimal matching were not able to be matched by greedy matching. It is worrisome when this many patients in the treated group are not able to be matched to comparison patients, as incomplete matching of treated patients has been shown to be a source of bias when estimating the treatment effect (Rosenbaum & Rubin, 1985). This pattern was seen among stratified matches as well. The use of calipers yielded lower total distance and lower numbers of matched patients. Again, the decline in number of matched patients was greatest for greedy matching.

The application of ratio caps to full matching, to restrict the size of matched sets, did not result in close matches. Intuitively, this makes sense. When the matching ratios are limited, excess comparison patients that would have made close matches to a particular treated patient must be reassigned, at a higher cost, to another treated patient. Whatever gain these ratio caps may offer for precision of estimates based on the matched sets may be offset by the imbalance they induce in the matched sets. For unstratified full matching, 0.5x thinning and 2x thickening caps led to a total distance that was orders of magnitude greater than the unrestricted match. Additionally, nearly a quarter of all matches had a distance greater than 0.2SD. Less restrictive 0.2x thinning and 5x thickening caps still resulted in noticeable gains in both total distance and proportion of distant matches, compared to the unrestricted match. In addition, these restricted matches took 5 to 10 times longer to complete. Some increase in total distance was seen for the stratified full matches, but the increase was not as large as for the unstratified full matches.

The minimum and maximum study group ratios in the matched sets from full matching are shown in **Table 4.5**. The effect of ratio caps is immediately noticeable. The minimum and maximum treatment-to-comparison ratios without caps are 1:108 and 11:1. With 0.5/2 thinning/thickening caps, these fall to 1:8 and 1:1. With 0.2/5 caps, these fall to 1:18 and 2:1. Reductions in these extremes are not seen for the stratified matches, since these ratios were adjusted at the hospital level and there were hospitals with treatment rates less than 2%.

The use of 2:1 fixed ratio greedy matching in this dataset also resulted in poor matches. Nearly half of the matches made without stratification and over 36% of matches made with stratification had a distance over 0.2SD of the matching variable. Applying calipers reduced the total distance, but did not yield complete matches. Neither of the stratified 2:1 greedy matches nor the unstratified 2:1 greedy match with calipers resulted in 2 comparison patients for each treated patient. For these matches, between 15% and 30% of the matched sets only had 1 comparison patient. For matches performed within hospitals, this should have been expected, since many hospitals had treatment rates greater than 33%, the maximum that will support a 2:1 match. When calipers were used, the pool of potential matches shrank further.

To see the practical effects of each matching method on the distribution of potential confounders in the matched groups, **Table 4.6** shows the standardized differences for selected covariates that were particularly unbalanced between study groups in the complete, unmatched data. The rule of thumb typically used is that a difference of 10% or greater indicates substantial imbalance (Rosenbaum & Rubin, 1985a-match), although well-balanced study groups will typically exhibit much lower values. Note for the unstratified, unrestricted full match that none of the standardized differences shown are greater than 1.4%, for example. Even for the stratified, unrestricted full match, which led to a higher total distance between matched pairs, all standard differences are below 3.0%. The use of calipers with full matching did little to change these values, but applying the very restrictive 0.5/2 ratio caps resulted in noticeable increases in the standardized differences for several variables. The standardized difference for blood urea nitrogen (BUN) was 8.8% for the stratified full match with these caps applied.

The standardized differences associated with both unstratified optimal matches and both unstratified greedy 1:1 matches were very low. The unstratified 2:1 greedy matches led to substantial imbalance, however. Perhaps reflecting the relatively higher proportion of matched sets with differences on the matching variable over 0.2SD, the stratified, unrestricted optimal and greedy matches had a few variables with standardized differences between the study groups over 5%. The use of calipers in these cases did work to reduce these differences. As with the unstratified 2:1 greedy matches, the attempt to match multiple comparison patients to each treated patient within hospitals led to unacceptable imbalances.

It is interesting to note that even though nearly 9% of the matched pairs within the stratified full match had a large (>0.2SD) difference on the matching variable, the standardized differences did not seem to reflect this in the same way the optimal and matches, with a similar proportion of distant matches, did. The reason for this has to do with the full matching results themselves. A comparison patient with a distant match to a treated patient within a matched set may be but one of 10 or 20 comparison patients matched to that same treated patient; and because the values for all of these comparison patients are averaged, the contribution of any single comparison patient to the standardized difference is only 5% to 10% what it would be if they were the only comparison patient matched to that treated patient. In other words, there is greater tolerance of imperfect matches in full matching than there is in optimal or greedy matching.

Discussion

Full matching has been shown to have certain advantages over both greedy matching and optimal matching, not the least of which is the incorporation of all patients into the final matched sets. The macro presented here makes the implementation of full matching methods possible. Using calipers to restrict the possible links full matching can make is also recommended. In addition to reducing the computing time necessary to arrive at a solution, the use of calipers results in only minimal, if any, loss of patients in the final matched sets. If there is a need to perform stratified matching, it is also recommended that researchers perform full matching with calipers. The use of calipers in this situation is, perhaps, more essential, if the data contain small hospitals or hospitals with insufficient numbers of treated patients.

Data set size is likely to be the factor that most limits the usefulness of this macro. The clinical example above had a fairly substantial sample size of over 40,000 patients and ran without problems. Of course, the maximum size of the data set size that can be processed by the macro is determined more directly by the computing environment in which it is run. Specifically, the available memory is critical. SAS/OR optimization procedures perform tasks in memory, and the memory workpace needs of these procedures expand at non-linear rate with respect to the sample size.

The variable size of the matched sets from full matching requires attention during analysis, through either the use of appropriate cluster-level weights or cluster-level conditional statistical methods. Analyses that completely ignore the clustering by matched sets will not yield correct results. It has been recommended that the matched sets be weighted by the number of treated patients in the set. As for most other matches, an outcomes analysis with this weighting will yield the average treatment effect among the treated.

The use of greedy matching within analyses that use propensity scores has been prevalent primarily because it is an easy matching method to implement. With a macro, like the one we present here, available to perform optimal and full matching, maybe this trend can change. If researchers would like to perform propensity score matching, the use of full matching methods should be encouraged.

Figure 4.1 - Network representation of unstratified matching. Panel A shows the nodes and available links. Nodes include a flow source (α), a flow sink (β), treated patients (1–5) and comparison patients (a–e). Gray lines indicate potential links between nodes. Panel B shows an example solution for optimal matching. Panel C shows an example solution for full matching.

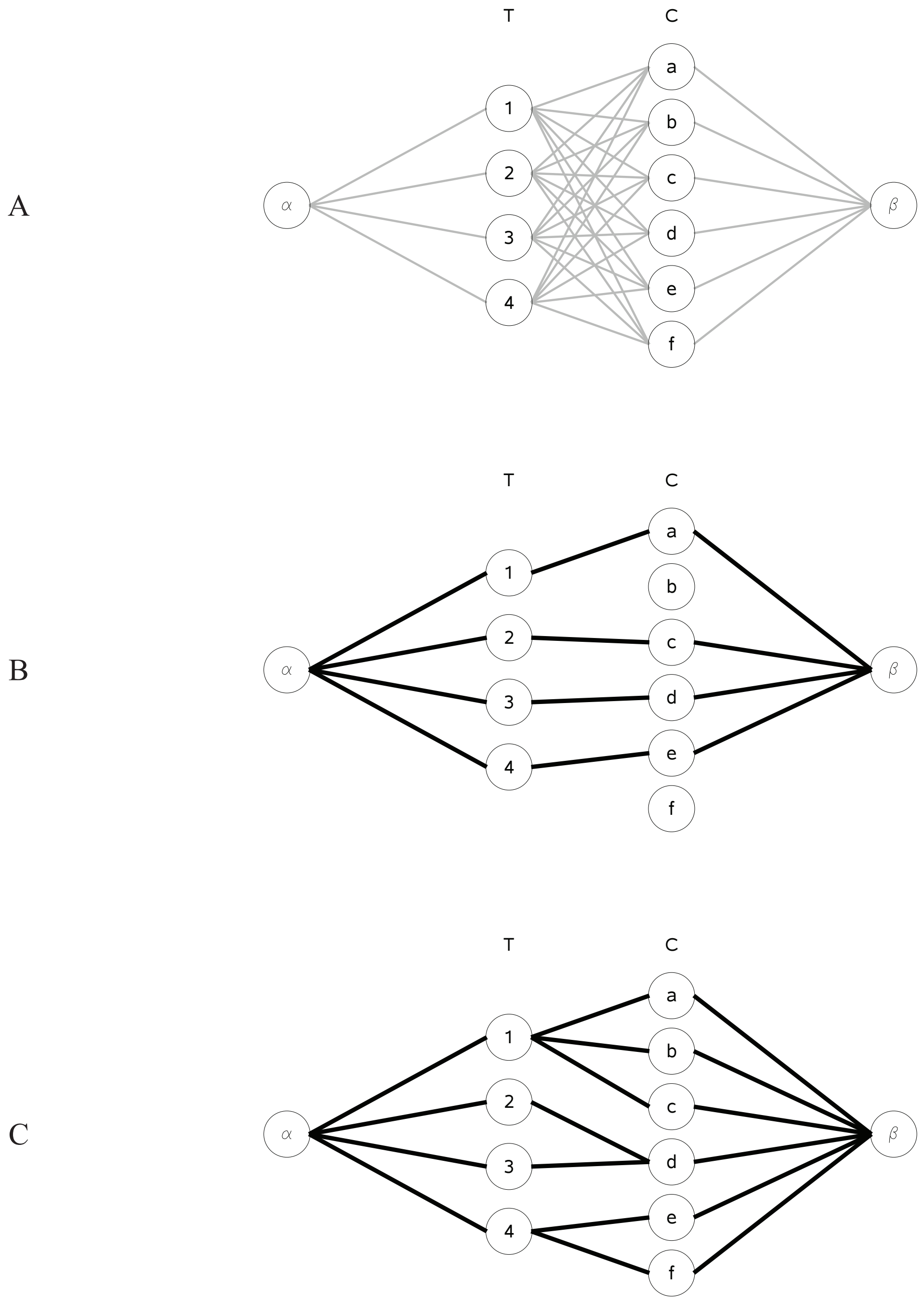


Figure 4.2 - Network representation of stratified matching. Panel A shows the nodes and available links. Nodes include a flow source (α), a flow sink (β), treated patients (1–5) and comparison patients (a–e). Gray lines indicate potential links between nodes. Panel B shows an example solution for optimal matching. Panel C shows an example solution for full matching.

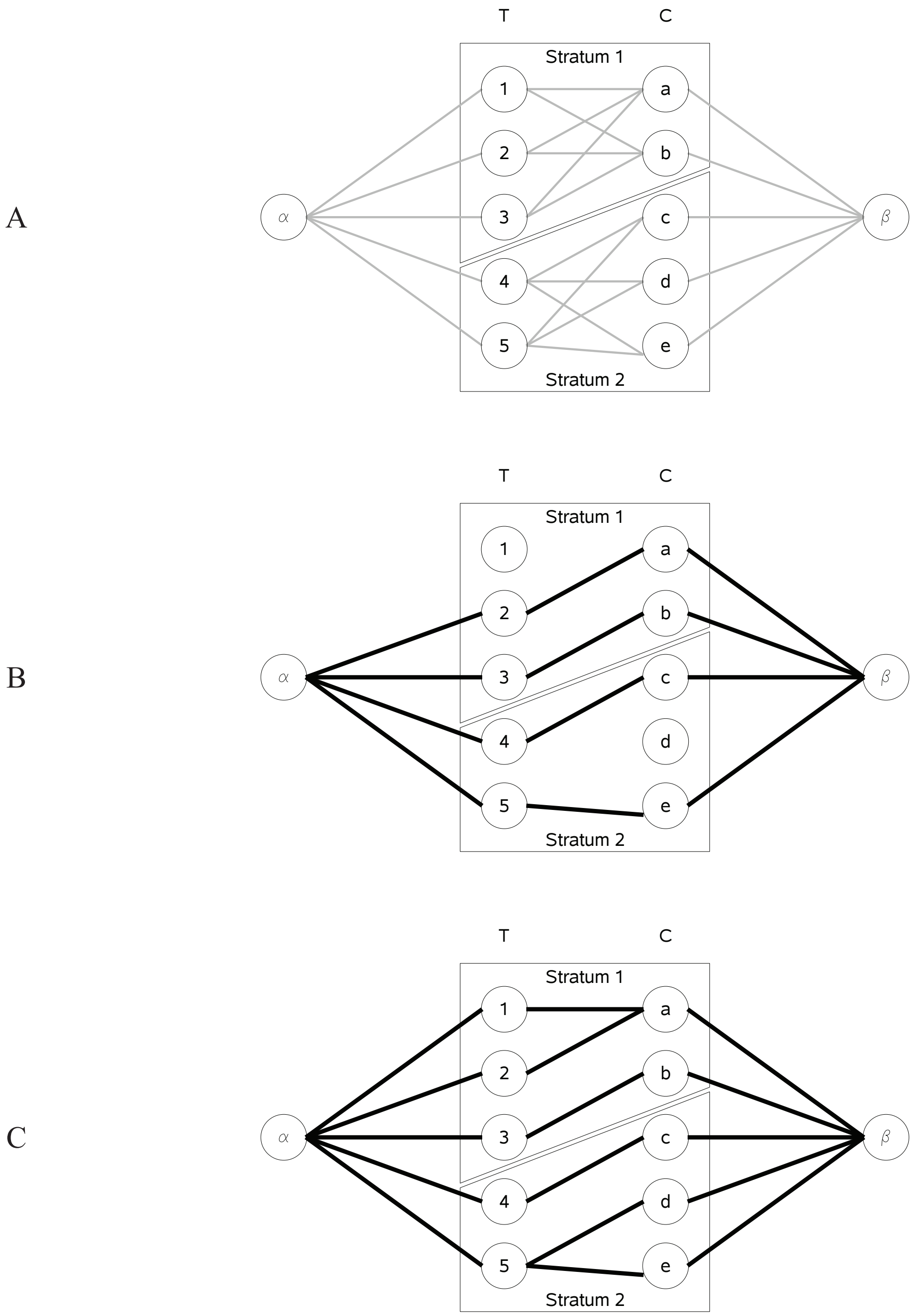


Figure 4.3 - Network representation of matching with calipers. Panel A shows the nodes and available links. Nodes include a flow source (α), a flow sink (β), treated patients (1–5) and comparison patients (a–e). Gray lines indicate potential links between nodes. All potential links shown between patients in the treatment (T) and comparison (C) groups are those with distance less than the given caliper width. Panel B shows an example solution for optimal matching with calipers, which required the use of an excess node (ε).

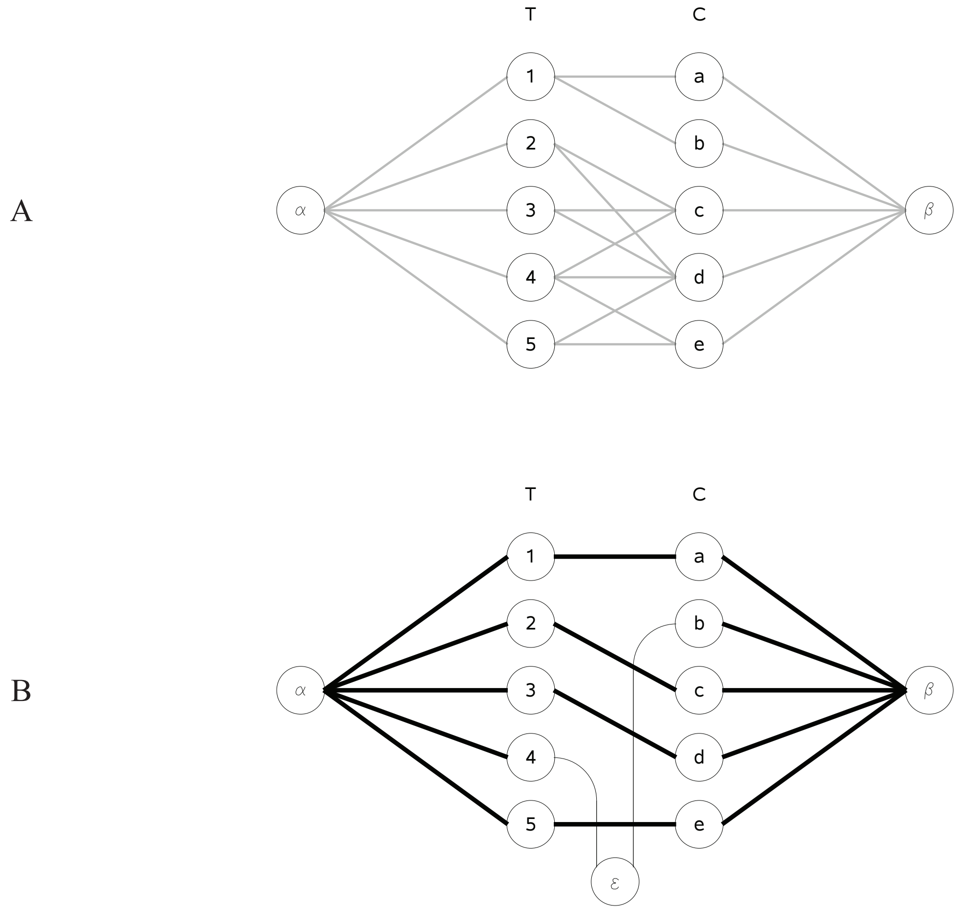


Figure 4.4 - Proportion of differences on the matching variable greater than 0.2SD within matched sets at each site. Panel A shows these proportions by number of treated patients at the site. Panel B shows these proportions by number of total patients at the site, with sites additionally categorized by number of treated patients.

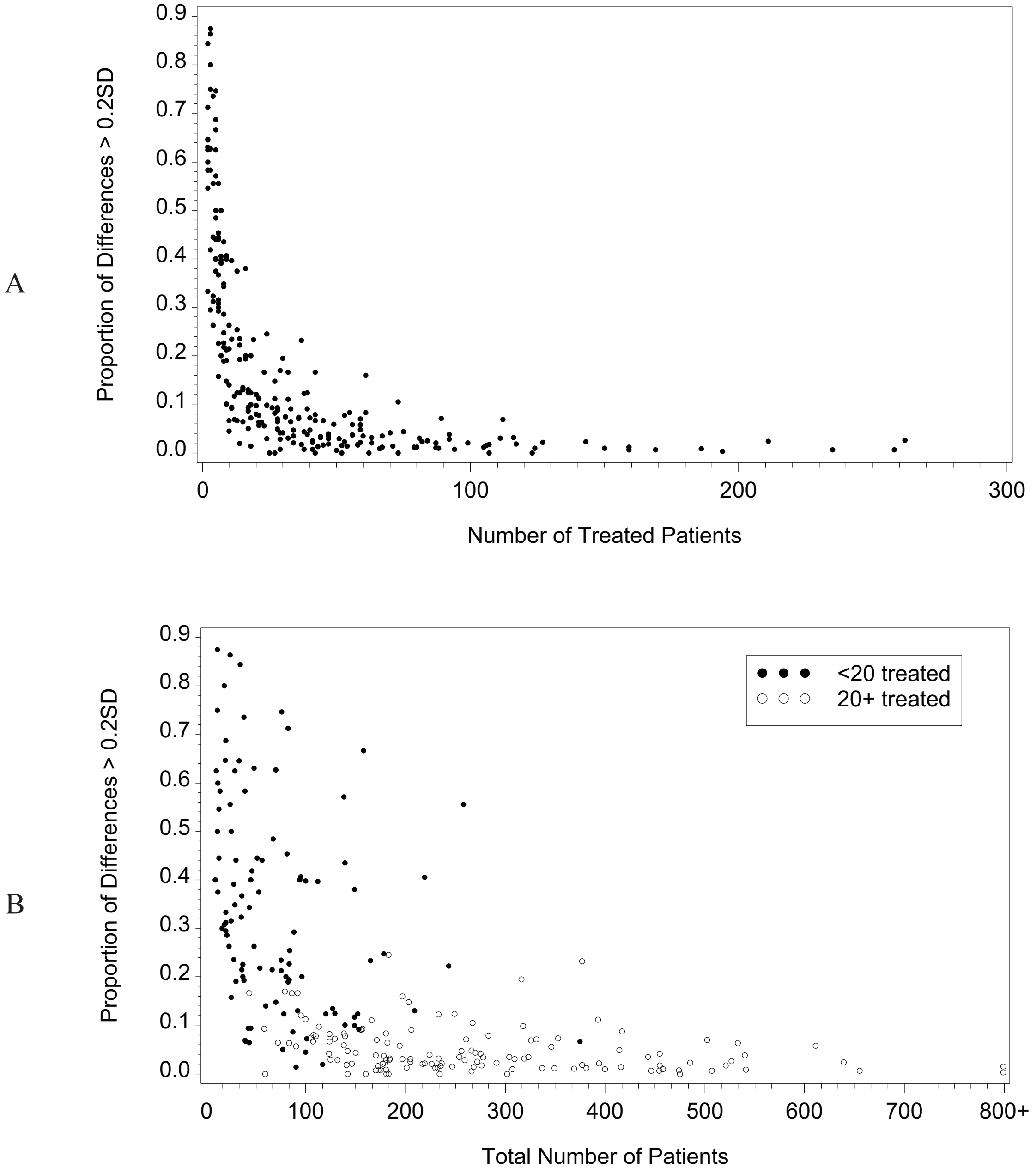


Figure 4.5 - Sample macro call for the full matching macro, FMATCH.SAS

```
* Macro available at:
* http://www.duke.edu/~hammill/software.html
;
%include “fmatch.sas”;
%macro fmatch(
    INDS           = mydata,
    MATCHVAR       = ps,
    GROUPVAR       = trt,
    IDVAR          = ptid,
    STRATVAR       = site,
    CALIPER        = 0.5,
    CALIPER_TYPE   = sd,
    RATIO_MAX      = 10,
    RATIO_MIN      = 0.1,
    MATCHTYPE      = full,
    OUTLINKS       = matchlinks,
    OUTDS          = matchrecs
);
```

Figure 4.6 - Sample records from the link-level output dataset generated by the full matching macro, FMATCH.SAS

```
GRP1    GRP2   STRATA      DIST
A153    Q224      1      0.05424
K197    Q224      1      0.49399
B171    Q224      1      0.06481
G136    M268      1      0.44148
L146    M268      1      0.42249
Q249    M268      1      1.17942
K193    M268      1      0.80213
H126    M268      1      0.38412
W139    E137      2      0.15742
W139    R016      2      0.00820
               ...
```

Figure 4.7 - Sample records from the record-level output dataset generated by the full matching macro, FMATCH.SAS

```
STUDY_ID    TRT   STRATA    SETNUM    SET_N1    SET_N2    SET_N
  A153       0       1         1         3         1        4
  B171       0       1         1         3         1        4
  K197       0       1         1         3         1        4
  Q224       1       1         1         3         1        4
  G136       0       1         2         5         1        6
  H126       0       1         2         5         1        6
  K193       0       1         2         5         1        6
  L146       0       1         2         5         1        6
  M268       1       1         2         5         1        6
  Q249       0       1         2         5         1        6
  E137       1       2         4         1         2        3
  R016       1       2         4         1         2        3
  W139       0       2         4         1         2        3
                              ...
```

Table 4.1. - Network characteristics associated with optimal matching and full matching, performed with and without stratification

	Unstratified		Stratified	
Network Characteristic	Optimal Matching	Full Matching	Optimal Matching	Full Matching
Total Network Flow (in units)	$\min(N_t, N_c)$	$\geq \max(N_t, N_c)$	$\sum_{h=1}^{N_h} \min\left(N_{t,h}, N_{c,h}\right)$	$\geq \sum_{h=1}^{N_h} \max\left(N_{t,h}, N_{c,h}\right)$
Link Costs				
$T_i \rightarrow C_j$	$\lvert M_i - M_j \rvert$		$\begin{cases} \lvert M_i - M_j \rvert, & T_i, C_j \text{ same stratum} \\ \text{no link}, & T_i, C_j \text{ different stratum} \end{cases}$	
$\alpha \rightarrow T_i$	0		0	
$C_j \rightarrow \beta$	0		0	
Link Capacity [Min, Max]				
$T_i \rightarrow C_j$	$[0, 1]$		$[0, 1]$	
$\alpha \rightarrow T_i$	$\begin{cases} [1, 1], & N_t \leq N_c \\ [0, 1], & N_t > N_c \end{cases}$	$[1, \infty]$	$\begin{cases} [1, 1], & N_{t,h} \leq N_{c,h} \\ [0, 1], & N_{t,h} > N_{c,h} \end{cases}$	$[1, \infty]$
$C_j \rightarrow \beta$	$\begin{cases} [0, 1], & N_t \leq N_c \\ [1, 1], & N_t > N_c \end{cases}$	$[1, \infty]$	$\begin{cases} [0, 1], & N_{t,h} \leq N_{c,h} \\ [1, 1], & N_{t,h} > N_{c,h} \end{cases}$	$[1, \infty]$

Where

N_t = # of treated patients

$N_{t,h}$ = # of treated patients in strata *h*

N_c = # of comparison patients

$N_{c,h}$ = # of comparison patients in strata *h*

N_h = # of strata

T_i = Treated patient *i*

M_i = Value of the matching variable for treated patient *i*

C_j = Comparison patient *j*

M_j = Value of the matching variable for comparison patient *j*

α = source node

β = Sink node

Table 4.2 - Ratio bounds for full matching associated with different combinations of thinning caps, thickening caps and treatment group proportions

Group 1	Observed	Thinning	Thickening	Group 1:Group 2 Bounds for Matching	
Proportion	Group 1:Group 2	Cap	Cap	Minimum	Maximum
0.05	1:19	0.8	1.25	1:24	1:15
		0.5	2	1:38	1:9
		0.2	5	1:95	1:3
		0.1	10	1:190	1:1
0.25	1:3	0.8	1.25	1:4	1:2
		0.5	2	1:6	1:1
		0.2	5	1:15	2:1
		0.1	10	1:30	4:1
0.40	1:1.5	0.8	1.25	1:2	1:1
		0.5	2	1:3	2:1
		0.2	5	1:8	4:1
		0.1	10	1:15	7:1
0.50	1:1	0.8	1.25	1:2	2:1
		0.5	2	1:2	2:1
		0.2	5	1:5	5:1
		0.1	10	1:10	10:1
0.60	1.5:1	0.8	1.25	1:1	2:1
		0.5	2	1:2	3:1
		0.2	5	1:4	8:1
		0.1	10	1:7	15:1
0.75	3:1	0.8	1.25	2:1	4:1
		0.5	2	1:1	6:1
		0.2	5	1:2	15:1
		0.1	10	1:4	30:1
0.95	19:1	0.8	1.25	15:1	24:1
		0.5	2	9:1	38:1
		0.2	5	3:1	95:1
		0.1	10	1:1	190:1

Table 4.3 - Patient characteristics included as independent variables in the treatment model

Category	Characteristic
Demographics	Age (years)
	Gender (Male, Female)
	Race (White, Black, Other/unknown)
Medical History	Anemia
	Atrial fibrillation
	Chronic renal insufficiency
	Chronic obstructive pulmonary disorder
	Coronary artery disease
	Diabetes mellitus
	Hypercholesterolemia
	Hypertension
	Peripheral vascular disease
	Prior myocardial infarction
	Smoker (current)
Medical devices in place	Pacemaker
	Implantable cardioverter defibrillator
Initial examination	Fatigue
	Rales
	Edema
	Congestion
	Ejection fraction
	Systolic blood pressure
	Blood urea nitrogen (mg/dL)
	Serum sodium (mEq/L)
	Hemoglobin (g/dL)

Table 4.4 - Metrics describing match results, by matching method

Matching Method	Total Distance	% Matches with Distance > 0.2 SD	% Treated Patients Matched	% Comparison Patients Matched	Computing Time Used (minutes)
		Unstratified matches			
Full matching					
Unrestricted*	47.0	0.1	100	100	29
0.2SD calipers	40.0	--	100	99.9	2
0.5 / 2 ratio caps**	7278.2	24.6	100	100	336
0.2 / 5 ratio caps	1487.2	6.8	100	100	155
Optimal matching					
Unrestricted	293.5	4.9	100	27.9	232
0.2SD calipers	273.6	--	99.9	27.8	9
Greedy matching					
1:1 match, unrestricted	297.9	5.0	100	27.9	< 1
1:1 match, 0.2SD calipers	20.5	--	96.7	27.0	< 1
2:1 match, unrestricted	3994.5	49.3	100	55.8	< 1
2:1 match, 0.2SD calipers	106.9	--	96.8	46.3	< 1
		Matches stratified by clinical site			
Full matching					
Unrestricted	2258.4	8.9	100	100	< 1
0.2SD calipers	945.6	--	98.4	91.2	< 1
0.5 / 2 ratio caps	3413.0	15.0	100	100	< 1
0.2 / 5 ratio caps	2334.4	9.2	100	100	< 1
Optimal matching					
Unrestricted	554.3	8.8	99.8	27.8	< 1
0.2SD calipers	257.0	--	96.4	26.9	< 1
Greedy matching					
1:1 match, unrestricted	616.4	8.9	99.8	27.8	< 1
1:1 match, 0.2SD calipers	158.2	--	93.7	26.1	< 1
2:1 match, unrestricted	2807.3	36.5	99.8	51.3	< 1
2:1 match, 0.2SD calipers	328.9	--	93.7	44.7	< 1

* Unrestricted matches are those without calipers or (for full matching) thinning and thickening ratio caps

** Thinning, thickening ratio caps applied to observed treated:comparison ratio

Abbreviation: SD = Standard deviation

Table 4.5 - Minimum and maximum treated-to-comparison ratios in matched sets resulting from full matching

Full Matching Specification	Minimum Treated:Comparison	Maximum Treated:Comparison
Unstratified		
Unrestricted*	1:108	11:1
0.2SD calipers	1:108	11:1
0.5 / 2 ratio caps**	1:8	1:1
0.2 / 5 ratio caps	1:18	2:1
Stratified		
Unrestricted	1:142	10:1
0.2SD calipers	1:100	10:1
0.5 / 2 ratio caps	1:127	3:1
0.2 / 5 ratio caps	1:142	5:1

* Unrestricted matches are those without calipers or thinning / thickening ratio caps

** Thinning, thickening ratio caps applied to observed treated:comparison ratio

Abbreviation: SD = Standard deviation

Table 4.6 - Standardized differences for selected covariates between matched treatment and control groups, by matching method used

Matching Method	Anemia	Coronary Artery Disease	Chronic Renal Insufficiency	Diabetes Mellitus	Edema	Age	Blood Urea Nitrogen	Hemoglobin
Observed	14.0	8.4	25.5	24.3	25.9	21.8	27.4	14.8
Unstratified matches								
Full matching								
Unrestricted*	0.9	1.4	1.2	0.9	1.1	0.3	1.3	0.4
0.2SD calipers	1.0	1.3	1.3	0.7	1.1	0.4	1.3	0.2
0.5 / 2 ratio caps**	1.6	0.4	3.2	2.9	2.3	2.8	5.5	3.4
0.2 / 5 ratio caps	0.1	1.7	0.6	0.2	0.4	0.8	0.3	0.8
Optimal matching								
Unrestricted	0.7	1.3	2.4	0.7	1.4	0.1	4.0	1.8
0.2SD calipers	0.6	1.4	2.2	0.8	1.5	0.2	3.7	1.6
Greedy matching								
1:1 match, unrestricted	0.5	1.2	2.2	0.5	1.5	0.1	3.7	1.4
1:1 match, 0.2SD calipers	0.4	1.5	0.1	0.3	1.9	0.6	0.1	0.8
2:1 match, unrestricted	5.4	3.6	11.1	7.6	6.4	8.0	14.6	5.1
2:1 match, 0.2SD calipers	3.2	1.6	6.5	4.6	3.6	4.5	8.2	2.9
Matches stratified by clinical site								
Full matching								
Unrestricted	2.2	0.9	1.1	0.8	0.6	0.5	2.9	0.9
0.2SD calipers	2.8	1.2	0.9	0.7	0.2	0.6	3.3	1.2
0.5 / 2 ratio caps	0.7	0.8	3.8	2.8	3.7	3.4	8.8	1.8
0.2 / 5 ratio caps	2.1	0.8	1.2	0.7	0.7	0.8	3.2	1.2
Optimal matching								
Unrestricted	1.5	1.1	6.3	3.6	1.6	1.5	8.9	1.9
0.2SD calipers	0.1	0.1	3.6	1.7	0.1	0.7	4.6	0.7
Greedy matching								
1:1 match, unrestricted	1.5	1.2	6.2	4.3	1.5	1.5	9.1	2.0
1:1 match, 0.2SD calipers	0.5	0.3	2.0	0.9	0.9	1.3	2.6	0.2
2:1 match, unrestricted	6.7	6.1	14.7	11.7	11.4	10.8	18.5	6.3
2:1 match, 0.2SD calipers	2.0	2.3	6.8	4.7	3.7	3.4	9.5	2.5

* Unrestricted matches are those without calipers or (for full matching) thinning and thickening ratio caps

** Thinning, thickening ratio caps applied to observed treated:comparison ratio

Standardized differences presented as % of SD

Abbreviation: SD = Standard deviation

CHAPTER 5

CONCLUSION

For research questions regarding the real-world effectiveness and safety of medical therapies and devices, researchers must often rely on observational data. Unlike controlled clinical trials, the assignment of treatment to patients in routine medical practice is not randomized. One class of methods used extensively by researchers to address this selection problem is propensity score methods. The role of the healthcare provider has not typically been accounted for when propensity score methods are employed, despite the fact that provider, by imparting an effect on both patient-level treatment assignment and patient-level outcomes, is a potential confounding factor.

When a healthcare provider has measurable impacts on both a patient's treatment assignment and their downstream outcomes, simulation results demonstrated that not accounting for these provider effects could lead to biased estimates of treatment effect when using propensity score methods. This was true specifically when a provider's direct effect on treatment was correlated with their effect on outcome; a situation that occurs when providers having better patient outcomes use therapies at higher (or lower) rates than other providers. Propensity score methods that incorporated provider were able to control this error. Even when provider effects on treatment and outcome were uncorrelated, it was still important to account for provider in the propensity score treatment model. Failure to do so resulted in confidence intervals around the estimated treatment effect that were either substantially too wide or too narrow, depending on the estimation methods used.

It may be that these methods are applicable only to a specific subset of clinical research questions. Researchers should take the time to understand if the question at hand could involve confounding by provider. One question to ask is whether or not there is a reasonable expectation that provider has an effect on treatment and outcome. It is often easier to expect that providers exhibit differential treatment propensities. Some providers may have a preference for certain therapies as first-line therapies and others as second-line, while other providers prefer the opposite. Some providers are more likely to incorporate new therapies or techniques more quickly than others. As a result, there is often a noticeable distribution of treatment rates among providers. Considering whether or not providers have an effect on outcome can be more difficult, as this can vary by the outcome, the time horizon, the treatment setting, etc. In general, short-term outcomes associated with direct treatment (e.g. surgical procedure) or prolonged care (e.g. hospital-based care) are more likely to result in stronger provider effects on outcomes. As a preliminary step in data analysis, the presence of provider effects on treatment and outcome can be checked using hierarchical models and provider-specific random intercept terms. The extent of variability of the random intercepts for both treatment and outcome is a guide to the strength of these effects. To prevent estimating the treatment effect prematurely, we recommend modeling the outcome, at this stage, using just the comparison group (or just a single study group, if both include treated patients).

Researchers should also check to see that the distribution of provider treatment rates is not extremely bimodal. If one group of providers has a very high treatment rate and another group has a very low treatment rate, there may not be enough within-provider variation to allow for appropriate comparisons when adjusting for provider in the analysis. In fact, if there are too many providers who exclusively treat or exclusively don't treat patients with the treatment of interest, methods that control for provider are not even possible. Ideally, the distribution of provider treatment rates should be somewhat normally distributed. Data may include too many providers who exclusively (or nearly exclusively) treat or don't treat patients for a few reasons. First, it may be that provider, in the data, reflects the practice of a single physician. Any single physician is more likely than a group of physicians to have strong preferences for a particular therapy or course of treatment. When provider reflects a group of physicians in a practice or in a hospital, there is usually variability in the preferences across those physicians. Second, it may be there are many providers who seem to exclusively treat (or don't treat) because the number of patients per physician in the data is very small. If there is interest in incorporating provider into propensity score methods, a relatively large number of patients per provider is desirable.

Finally, as with any propensity score analysis, researchers need to ensure that they have the data necessary to fully characterize the treatment assignment. A critical assumption of all these methods is that there is no unmeasured confounding. Whether or not provider is included as a factor in the propensity score treatment model or not, data that lacks important confounders will lead to biased treatment estimates.

Once a researcher is comfortable with the idea that their clinical question may involve confounding by provider, estimation of the treatment effect still needs to be done properly. After including provider in the treatment model, there are a few directions the analysis can go. If there is interest in using inverse probability of treatment weighting methods, strong consideration should be given to the Lunceford estimator and to doubly robust methods. If there is interest in using propensity score matching, consideration should be given to full matching. Patients within the resulting matched sets either need to be appropriately weighted for analysis or the use of GEE methods is required.

Full matching may not be familiar to most analysts. A criticism of typical 1:1 propensity score matching, whether stratified by provider or not, is that the data from many patients are not utilized in the outcomes analysis. Full matching addresses this issue by optimally assigning all treated patients and all comparison patients into variably-sized matched sets. The result is closer matches between study groups than those obtained by other matching methods. For comparative effectiveness research, where head-to-head comparisons of therapies involve two study groups that both include treated patients, full matching should be considered as a primary matching method. Full matching is not currently utilized frequently because it is difficult to implement. A macro to perform full matching by leveraging SAS optimization procedures was presented.

There are a few obvious extensions of this work that need to be explored. The first is how well propensity score methods that incorporate provider perform when the outcome of interest in dichotomous. We showed that these methods are appropriate when the outcome is continuous and the treatment effect reflects a simple difference between mean group outcomes. Quantities of interest for binary data include risk differences, risk ratios, and odds ratio. It may also be important to examine the situation where treatment models are entirely provider-specific. We simulated data scenarios where providers had baseline levels of treatment that were higher or lower than average, but we did not otherwise alter the treatment assignment mechanism by provider. This may not reflect reality, but it’s not clear if provider

-specific intercepts in a treatment model are sufficient to produce correct treatment effect estimates or if the treatment model would need to more closely reflect the each provider's treatment process.

APPENDIX 1

MISCELLANEOUS SAS CODE FOR ESTIMATING PROPENSITY SCORES, CALCULATING WEIGHTS, PERFORMING MATCHING, AND ESTIMATING APPROPRIATE TREATMENT EFFECT ESTIMATES

```
/*****************************************************************
* Generated data set information                                  *
*   Data set name = DS                                            *
*   Variables:                                                    *
*     A = Treatment (0/1)                                         *
*     Y = Outcome (continuous)                                    *
*     X1 = Covariate (0/1)                                        *
*     X2 = Covariate (continuous)                                 *
*     X3 = Covariate (continuous)                                 *
*     X4 = Covariate (continuous)                                 *
*     IDX = Patient ID variable                                   *
*     SITE = Cluster ID variable                                  *
*                                                                 *
* For greedy matching, there is a macro (gmatch.sas) at           *
*    http://people.duke.edu/~hammill/software                     *
* that can be used for matching on a single variable like PS.     *
 *****************************************************************/

/*****************************************************************
* Pooled propensity score treatment model
*  - Propensity score (PS1) and linear predictor (XB1) saved back
*    onto input dataset DS
 *****************************************************************/
proc logistic descending data=ds;
    model a = x1 x2 x3 x4;
    output out=ds pred=ps1 xbeta=xb1;
run;

/*****************************************************************
* Cluster-specific propensity score treatment model #1
*  - Random effects (intercept only) for cluster
*    [random slopes can be added, if desired]
*  - Propensity score (PS2) and linear predictor (XB2) saved back
*    onto input dataset A
*  - Sometimes you need to specify a less stringent ABSPCONV value
*    (PROC GLIMMIX statement option) than the default of 1E-8 for
*    this model to converge
 *****************************************************************/
proc glimmix data=ds abspconv=1e-4;
    class site;
    model a = x1 x2 x3 x4 / link=logit dist=bin s;
    nloptions maxiter = 50;
    random intercept / subject=site;
    output out=ds pred(ilink)=ps2 pred=xb2;
run;
```

```
/****************************************************************
* Cluster-specific propensity score treatment model #2
*  - Fixed effects (intercept only) for cluster
*    [cluster-specific slopes can be created by adding
*     interactions with the cluster variable, if desired]
*  - Propensity score (PS3) and linear predictor (XB3) saved back
*    onto input dataset A
 ****************************************************************/
proc logistic descending data=ds;
    class site;
    model a = site x1 x2 x3 x4;
    output out=ds pred=ps3 xbeta=xb3;
run;

/****************************************************************
* Create inverse probability of treatment weights (W1, W2, W3)
* based on estimated propensity scores (PS1, PS2, PS3).
 ****************************************************************/
data ds;
    set ds;

    * Keep estimated probabilities of treatment (A = 1) for Lunceford
    * estimator, doubly robust estimators, etc.
    ;
    e1 = ps1;
    e2 = ps2;
    e3 = ps3;

    * When A = 0, need to flip b/c we need the probability of receiving
    * the treatment actually received
    ;
    if a = 0 then do;
        ps1 = 1 - ps1;
        ps2 = 1 - ps2;
        ps3 = 1 - ps3;
    end;

    w1 = 1 / ps1;
    w2 = 1 / ps2;
    w3 = 1 / ps3;
run;

/****************************************************************
* Create matched study groups using 1:1 unstratified greedy
* matching with calipers = 0.2SD
 ****************************************************************/
%gmatch(
    inds = ds,
    matchvar = xb1,
    groupvar = a,
    idvar = idx,
    caliper = 0.2,
    caliper_type = SD,
    randseed = 20150603,
    outds = matched1
)
```

```
/****************************************************************
* Create matched study groups using 1:1 cluster-stratified greedy
* matching with calipers = 0.2SD
 ****************************************************************/
%gmatch(
    inds = ds,
    matchvar = xb1,
    groupvar = a,
    idvar = idx,
    stratvar = site,
    caliper = 0.2,
    caliper_type = SD,
    randseed = 20150603,
    outds = matched2
)

/****************************************************************
* Create matched study groups using variable ratio unstratified
* full matching with calipers = 0.2SD
 ****************************************************************/
%fmatch(
    inds = ds,
    matchvar = xb1,
    groupvar = a,
    idvar = idx,
    caliper = 0.2,
    caliper_type = SD,
    matchtype = FULL,
    outds = matched3
)

/****************************************************************
* Create matched study groups using variable ratio cluster-
* stratified full matching with calipers = 0.2SD
 ****************************************************************/
%fmatch(
    inds = ds,
    matchvar = xb1,
    groupvar = a,
    idvar = idx,
    stratvar = site,
    caliper = 0.2,
    caliper_type = SD,
    matchtype = FULL,
    outds = matched4
)
```

```
/*****************************************************************
* Create ETT weights assuming A is coded 0/1 and "treatment" is
* when A = 1.  The FMATCH.SAS macro outputs set counts based on
* the sorted value of A, so it is essential to understand your
* data.  For full matching results based on these data, SET_N1
* (produced by the macro on the output dataset) is the count of
* comparison patients in the matched set while SET_N2 is the
* count of treated patients in the matched set.
 ****************************************************************/
data matched3;
    set matched3;

    select (a);
        when (0) SETWT = set_n2 / set_n1;
        when (1) SETWT = 1;
        * no otherwise;
    end;
run;

/*****************************************************************
* For analysis code below:
*   - Be sure to use appropriate dataset (DS, MATCHED1, etc.)
*   - Use desired weight (W1, W2, etc.) -- Note that you can
*     create a null weight (e.g. W0 = 1) for use with matched
*     data that is otherwise unweighted
*   - Some of this code may need to be put within a macro wrapper
*     b/c of %do loops, etc.
 ****************************************************************/

/*****************************************************************
* Pooled t-test
 ****************************************************************/
proc ttest data=ds;
    var y;
    class a;
    weight w1;
run;

/*****************************************************************
* Matched t-test (for 1:1 matched results)
*   - Requires transpose to flatten data within matched pairs
 ****************************************************************/
proc transpose data=ds out=dst prefix=Y;
    var y;
    id a;
    by setnum;
run;

proc ttest data=dst;
    paired y0 * y1;
run;

/*****************************************************************
* Regression models:
*   GLM (via printmle)
*   GLM w/robust SEs (subject-level)
 ****************************************************************/
proc genmod data=ds;
    class idx;
    model y = a ;
    weight w1;
    repeated subject=idx / printmle;
run;
```

```
/****************************************************************
* Regression model: GLM w/robust SEs (cluster-level)
 ****************************************************************/
proc genmod data=ds;
    class site;
    model y = a ;
    weight w1;
    repeated subject=site / type=ind;
run;

/****************************************************************
* Regression model: GEE w/cluster-level exchangeable correrlation
 ****************************************************************/
proc genmod data=ds;
    class site;
    model y = a ;
    weight w1;
    repeated subject=site / type=exch;
run;

/****************************************************************
* Regression models: ETT-weighted GLM w/robust SEs (matched-set)
*    (SETNUM output by matching macros)
 ****************************************************************/
proc genmod data=ds;
    class setnum;
    model y = a ;
    weight setwt;
    repeated subject=setnum / type=ind;
run;

/****************************************************************
* Regression model: GEE w/matched-set-level exchangeable
*    correlation (SETNUM output by matching macros)
 ****************************************************************/
proc genmod data=ds;
    class setnum;
    model y = a ;
    repeated subject=setnum / type=exch;
run;

/****************************************************************
* Lunceford estimator + standard error (using IPTW weights);
 ****************************************************************/
proc sql noprint;
    select
        sum(y * a0 / (1 - e1)) / sum(a0 / (1 - e1)) as mu0,
        sum(y * a1 / e1) / sum(a1 / e1) as mu1,
        count(*) as nall
    into
        :mu0, :mu1, :nall
    from ds;
quit;
```

```
* If using a cluster-specific weight (W2 or W3), need to create individual
* indicator variables for each cluster.  There are multiple ways to do this,
* below is one for a dataset with 50 sites where the sites are numbered 1 to
* 50.  If using pooled weight (W1), can ignore this data step.
;
data ds;
    set ds;

    array allsites(50) site1 - site50;

    %do i = 1 %to 50;
        allsites(&i) = 0;
    %end;

    allsites(site) = 1;
run;

proc iml;
    use ds;

    read all var {x1 x2 x3 x4} into xraw;
    ebbsum = j(4, 4, 0);
    m2sum = j(4, 4, 0);
    vsum1 = j(4, 1, 0);
    vsum0 = j(4, 1, 0);

    * If using a cluster-specific weight, need to replace the read
    * and init code above to incorporate the cluster indicators like:
    *
    *    read all var {
    *        %do s = 1 %to 50;
    *            site&s
    *        %end;
    *        x1 x2 x3 x4} into xraw;
    *    ebbsum = j(54, 54, 0);
    *    m2sum = j(54, 54, 0);
    *    vsum1 = j(54, 1, 0);
    *    vsum0 = j(54, 1, 0);
    *
    ;

    read all var {y} into y;
    read all var {a} into a;
    read all var {a0} into a0;
    read all var {a1} into a1;
    read all var {e1} into e;

    x = xraw`;
    isum = 0;

    diff = &mu1 - &mu0;
    v1 = (y - &mu1) # a1 / e;
    v2 = (y - &mu0) # a0 / (1 - e);
    v3pre = (a - e);

    do i = 1 to &nall;
        xxt = x[,i] * x[,i]`;
        ebbsum = ebbsum + e[i] # (1 - e[i]) # xxt;
        vsum1 = vsum1 + x[,i] # (y[i] - &mu1) # a1[i] # (1 - e[i]) / e[i];
        vsum0 = vsum0 + x[,i] # (y[i] - &mu0) # a0[i] # e[i] / (1 - e[i]);
    end;

    ebb = ebbsum / &nall;
    hb2 = (vsum1 + vsum0) / &nall;

    do i = 1 to &nall;
        toadd = v1[i] - v2[i] - v3pre[i] # (hb2` * ebb * x[,i]);
        isum = isum + toadd # toadd;
    end;

    estvar = isum / (&nall * &nall);
    se = sqrt(estvar);

    outdata = diff || se;

    create lunceford from outdata [colname={"diff" "se"}];
    append from outdata;
    close outdata;
quit;

/****************************************************************
* Doubly robust estimators + standard error (using IPTW weights)
* DR estimates based on both a pooled outcome model and a
* cluster-specific outcome model are produced.
 ****************************************************************/
data ds;
    set ds;

    if not a then do;
        y_notrt = y;
        y_trt = .;
    end;
    else do;
        y_notrt = .;
        y_trt = y;
    end;
run;

* GLM predictors for DR;
proc genmod data=ds;
    model y_notrt = x1 x2 x3 x4;
    output out=ds pred=m0_glm;
run;

proc genmod data=ds;
    model y_trt = x1 x2 x3 x4;
    output out=ds pred=m1_glm;
run;

* Mixed predictors for DR;
proc mixed data=ds;
    class site;
    model y_notrt = x1 x2 x3 x4 / outpred=ds(
        drop=alpha df lower resid stderrpred upper _level_:
        rename=(pred = m0_mix)
    );
    random intercept / subject=site;
run;

proc mixed data=ds;
    class site;
    model y_trt = x1 x2 x3 x4 / outpred=ds(
        drop=alpha df lower resid stderrpred upper _level_:
        rename=(pred = m1_mix)
    );
    random intercept / subject=site;
```

```
run;

proc sql noprint;
    select
        sum(y * a0 / (1 - &e)) / count(*) as part2_1,
        sum(y * a1 / &e) / count(*) as part1_1,
        sum((a - &e) * m0_mix / (1 - &e)) / count(*) as part2_2_mix,
        sum((a - &e) * m1_mix / &e) / count(*) as part1_2_mix,
        sum((a - &e) * m0_glm / (1 - &e)) / count(*) as part2_2_glm,
        sum((a - &e) * m1_glm / &e) / count(*) as part1_2_glm
    into
        :p2_1, :p1_1, :p2_2x, :p1_2x, :p2_2l, :p1_2l
    from ds;
quit;

data dr;
    set ds end=final;

    retain var_gee 0 var_mix 0 var_glm 0;

    tau_mix = &p1_1 - &p1_2x - &p2_1 - &p2_2x;
    tau_glm = &p1_1 - &p1_2l - &p2_1 - &p2_2l;

    piece_mix =
        (y * a1 / e1) -
        ((a - e1) * m1_mix / e1) -
        (y * a0 / (1 - e1)) -
        ((a - e1) * m0_mix / (1 - e1)) -
        tau_mix
    ;

    piece_glm =
        (y * a1 / e1) -
        ((a - e1) * m1_glm / e1) -
        (y * a0 / (1 - e1)) -
        ((a - e1) * m0_glm / (1 - e1)) -
        tau_glm
    ;

    var_mix = var_mix + (piece_mix * piece_mix);
    var_glm = var_glm + (piece_glm * piece_glm);

    if final then do;
        var_mix = var_mix / (_n_ * _n_);
        se_mix = sqrt(var_mix);

        var_glm = var_glm / (_n_ * _n_);
        se_glm = sqrt(var_glm);

        output;
    end;

    keep tau: var_: se_:;
run;
```

APPENDIX 2

SAS CODE FOR THE FULL MATCHING MACRO, FMATCH.SAS

```
/***************************************************************************
| Program: FMATCH.SAS                                                       |
| Purpose: Perform full matching or optimal matching on a single scalar     |
|          variable between 2 study groups                                  |
| Author:  Brad Hammill                                                     |
| Date:    2015Jan01                                                        |
| Output:  Two datasets named by the &OUTLINKS and &OUTDS macro variables   |
|          that contain link-level results and record-level results         |
|                                                                           |
| Modifications:                                                            |
 ***************************************************************************/

/***************************************************************************
| INDS           Input dataset *REQUIRED*, no default                       |
| MATCHVAR       Matching variable *REQUIRED*, no default                   |
|                  The variable on which to match members of Group 1 to     |
|                  to Group 2 (see GROUPVAR)                                |
| GROUPVAR       Group variable *REQUIRED*, no default                      |
|                  This variable defines study groups                       |
|                  Only 2 non-missing levels allowed                        |
|                  The first level (sorted alphanumeric) of this variable   |
|                  defines Group 1.  The second level defines Group 2.      |
| IDVAR          Record-level ID variable *REQUIRED*, no default            |
|                  This variable must be unique across records              |
| STRATVAR       Stratification variable, default = NONE                    |
|                  This variable defines the strata within which matching   |
|                  occurs, if desired                                       |
| CALIPER        Caliper width, default = NONE                              |
|                  This variable defines the maximum distance allowable for |
|                  matching on MATCHVAR between groups. Assumed to be an    |
|                  absolute value unless noted as a multiple of the observed|
|                  SD in CALIPER_TYPE option below                          |
| CALIPER_TYPE   Type of caliper to apply, ABS (default) | SD               |
|                  ABSolute calipers match records within the distance noted|
|                  by the CALIPER option.                                   |
|                  SD calipers match records within a multiple (noted with  |
|                  the CALIPER option) of the observed standard deviation   |
|                  of the MATCHVAR.                                         |
|                  Ex: To match within 0.2SD of the MATCHVAR, specify       |
|                      CALIPER=0.2 and CALIPER_TYPE=SD                      |
| RATIO_MIN      Thinning cap, if specified must be < 1                     |
| RATIO_MAX      Thickening cap, if specified must be > 1                   |
|                  Both are applied to the observed ratio of Group 1 to     |
|                  to Group 2 records.  Constrains the matching ratio to be |
|                  within some multiple of the observed ratio.              |
|                  [See: Hansen, JASA 99:467 pp 609-618]                    |
|                  Default is NONE, which leads to an uncontrained match.   |
|                  To apply limits, specify both RATIO_MIN and RATIO_MAX    |
|                  options and ensure that RATIO_MAX > 1 and RATIO_MIN < 1. |
| MATCHTYPE      Type of match to make: FULL (default) or OPT               |
|                  FULL matching ensures that all records in Group 1 are    |
|                  matched to all records in Group 2 (within calipers, if   |
|                  specified).  Matched set sizes can vary.                 |
|                  OPTimal matching performs optimal 1:1 matching (within   |
|                  calipers, if specified).                                 |
| OUTLINKS       Output dataset for links, default = _OUTLINKS              |
|                  This is a link-level dataset that includes record IDs    |
|                  matched from each group and distance between records.    |
|                  Dataset variables:                                       |
|                    GRP1 = Record ID for Group 1 record                    |
|                    GRP2 = Record ID for Group 2 record                    |
|                    [STRATVAR] = Strata, if specified for link             |
|                    DIST = Distance between MATCHVAR values                |
| OUTDS          Output dataset for records, default = _OUTMATCH           |
|                  This is a record-level dataset that includes record ID,  |
|                  matched set number, and weights for analysis.            |
|                  Dataset variables:                                       |
|                    [IDVAR] = Record ID                                    |
|                    [GROUPVAR] = Group variable value                      |
|                    [STRATVAR] = Strata, if specified for link             |
|                    SETNUM = Matched set index                             |
 ***************************************************************************/

%macro fmatch(
    INDS          = ,
    MATCHVAR      = ,
    GROUPVAR      = ,
    IDVAR         = ,
    STRATVAR      = NONE,
    CALIPER       = NONE,
    CALIPER_TYPE  = ABS,
    RATIO_MAX     = NONE,
    RATIO_MIN     = NONE,
    MATCHTYPE     = FULL,
    OUTLINKS      = _OUTLINKS,
    OUTDS         = _OUTMATCH
);

    * Local macro variables;
    %local BREAK INFINITY G1 G2 SOLVER_STATUS;
    %let BREAK = 0;
    %let INFINITY = 1E14;


    * Input parameter checks;
    data _check1;
        MERGEVAR = 1;

        SPEC_INDS = upcase("&inds");
        SPEC_MATCHVAR = upcase("&matchvar");
        SPEC_GROUPVAR = upcase("&groupvar");
        SPEC_IDVAR = upcase("&idvar");
        SPEC_STRATVAR = upcase("&stratvar");
        SPEC_CALIPER = upcase("&caliper");
        SPEC_CALIPER_TYPE = upcase("&caliper_type");
        SPEC_RATIO_MAX = upcase("&ratio_max");
        SPEC_RATIO_MIN = upcase("&ratio_min");
        SPEC_MATCHTYPE = upcase("&matchtype");
        SPEC_OUTLINKS = upcase("&outlinks");
        SPEC_OUTDS = upcase("&outds");

        ABORT1 = 0;
        PARM_NOMATCH = 0;
        PARM_NOGROUP = 0;
        PARM_NOID = 0;
        PARM_1RATIO = 0;
        PARM_HILORATIO = 0;
        PARM_NOINDS = 0;
        PARM_BADINDS = 0;
        PARM_BADMATCHVAR = 0;
```

```
PARM_BADGROUPVAR = 0;
PARM_BADIDVAR = 0;
PARM_BADSTRATVAR = 0;
PARM_BADCALIPTYPE = 0;
PARM_BADMATCHTYPE = 0;

%if %length(&matchvar) = 0 %then %do ;
    PARM_NOMATCH = 1;
    ABORT1 = 1;
    call symput("BREAK", 1);
%end;

%if %length(&groupvar) = 0 %then %do ;
    PARM_NOGROUP = 1;
    ABORT1 = 1;
    call symput("BREAK", 1);
%end;

%if %length(&idvar) = 0 %then %do ;
    PARM_NOID = 1;
    ABORT1 = 1;
    call symput("BREAK", 1);
%end;

%if &ratio_max ^= NONE | &ratio_min ^= NONE %then %do ;
    %if &ratio_min = NONE | &ratio_max = NONE %then %do;
        PARM_1RATIO = 1;
        ABORT1 = 1;
        call symput("BREAK", 1);
    %end;
    %else %do;
        if  not (&ratio_min < 1  and &ratio_max > 1) then do;
            PARM_HILORATIO = 1;
            ABORT1 = 1;
            call symput("BREAK", 1);
        end;
    %end;
%end;

%if %length(&inds) = 0 %then %do ;
    PARM_NOINDS = 1;
    ABORT1 = 1;
    call symput("BREAK", 1);
%end;

%if %sysfunc(exist(&inds)) = 0 %then %do ;
    PARM_BADINDS = 1;
    ABORT1 = 1;
    call symput("BREAK", 1);
%end;
%else %do;
    dsid = open("&inds");
    if varnum(dsid, "&matchvar") = 0 then do;
        PARM_BADMATCHVAR = 1;
        ABORT1 = 1;
        call symput("BREAK", 1);
    end;
    if varnum(dsid, "&groupvar") = 0 then do;
        PARM_BADGROUPVAR = 1;
        ABORT1 = 1;
        call symput("BREAK", 1);
    end;
    if varnum(dsid, "&idvar") = 0 then do;
```

```
                PARM_BADIDVAR = 1;
                ABORT1 = 1;
                call symput("BREAK", 1);
            end;
            if "&stratvar" ne "NONE" and varnum(dsid, "&stratvar") = 0 then do;
                PARM_BADSTRATVAR = 1;
                ABORT1 = 1;
                call symput("BREAK", 1);
            end;
            rc = close(dsid);
            drop dsid rc;
        %end;

        %if &caliper ^= NONE %then %do;
            select(upcase(substr("&caliper_type", 1, 1)));
                when ("S") call symput("caliper_type", "SD");
                when ("A") call symput("caliper_type", "ABS");
                otherwise do;
                    PARM_BADCALIPTYPE = 1;
                    ABORT1 = 1;
                    call symput("BREAK", 1);
                end;
            end;
        %end;

        select(upcase(substr("&matchtype", 1, 1)));
            when ("O") call symput("matchtype", "OPT");
            when ("F") call symput("matchtype", "FULL");
            otherwise do;
                PARM_BADMATCHTYPE = 1;
                ABORT1 = 1;
                call symput("BREAK", 1);
            end;
        end;

        if "&stratvar" = "NONE" then
            call symput("stratvar", "_ONE_");
    run;

    %if &BREAK = 0 %then %do;
        * Make working copy of input data;
        data _useds;
            set &inds;

            _ONE_ = 1;
        run;

        * Data checks and input record counts;
        proc sql noprint;
            * Counts from input DS, including how many missing key variables;
            create table _check2 as
            select
                1 as MERGEVAR,
                count(*) as N_INPUT,
                sum(missing(&matchvar)) as NMISS_MATCH,
                sum(missing(&groupvar)) as NMISS_GRP,
                sum(missing(&idvar)) as NMISS_ID,
                sum(missing(&stratvar)) as NMISS_STRAT,
                count(&idvar) as N_NOMISS_ID,
                count(distinct &idvar) as N_ID,
                count(distinct &stratvar) as N_STRATA,
                count(distinct &groupvar) as GRP_LEVELS
            from _useds
            ;
        quit;

        data _check2;
            set _check2;

            ABORT2 = 0;
            INDATA_BADGRP = 0;
            INDATA_BADIDS = 0;

            if grp_levels ne 2 then do;
                INDATA_BADGRP = 1;
                ABORT2 = 1;
                call symput("BREAK", 1);
            end;
            if n_id ne n_nomiss_id then do;
                INDATA_BADIDS = 1;
                ABORT2 = 1;
                call symput("BREAK", 1);
            end;
        run;

    %end;

    %if &BREAK = 0 %then %do;

        * Check input group counts;
        proc sql noprint;
            select distinct &groupvar into :G1 - :G2
            from _useds;

            create table _check3 as
            select
                1 as MERGEVAR,
                sum(&groupvar = &g1) as N_G1,
                sum(&groupvar = &g2) as N_G2
            from _useds;
        quit;

        * Keep useable data, add numeric index, output crosswalk;
        data
            _useds
            _xwalk(keep = _IDX &idvar &groupvar &stratvar)
        ;
            set _useds;
            where
                not missing(&matchvar) and
                not missing(&groupvar) and
                not missing(&idvar) and
                not missing(&stratvar)
            ;

            _IDX = _n_;
        run;

        * Figure caliper width, if based on SD;
        %if &caliper ^= NONE & %upcase(&caliper_type) = SD %then %do;

            %if &stratvar = _ONE_ %then %do;
                proc means data=_useds noprint;
                    var &matchvar;
                    output out=_sd STD=SD_MATCH;
                run;
            %end;
            %else %do;
                proc means data=_useds noprint;
                    var &matchvar;
                    class &stratvar;
                    output out=_rawsd std=sitesd;
                run;

                proc means data=_rawsd noprint;
                    var sitesd;
                    output out=_sd mean=SD_MATCH;
                run;

                proc delete data=_rawsd;
                run;
            %end;

            data _null_;
                set _sd;

                call symput("caliper", &caliper * sd_match);
            run;

            proc delete data=_sd;
            run;

        %end;
        %put INFO: Caliper width = &caliper ;

        proc sql;
            create table _links as
            select
                use1._idx as FROM,
                use2._idx as TO,
                use1.&stratvar as STRAT,
                abs(use1.&matchvar - use2.&matchvar) as WEIGHT,
                . as LOWER,
                1 as UPPER
            from
                _useds use1,
                _useds use2
            where
                use1.&groupvar = &g1 and
                use2.&groupvar = &g2 and
                use1.&stratvar = use2.&stratvar
                %if &caliper ^= NONE %then
                    and abs(use1.&matchvar - use2.&matchvar) < &caliper
                ;
            ;
        quit;

        proc sql;
            create table _check4 as
            select
                1 as MERGEVAR,
                count(distinct from) as USE_G1,
                count(distinct to) as USE_G2,
                count(distinct from) + count(distinct to) as USE_LIMIT,
                count(distinct STRAT) as USE_STRATA
            from _links;
        quit;

        data _check4;
```

```
        set _check4;

        ABORT3 = 0;
        LINK_NODATA = 0;

        if use_limit = 0 then do;
            LINK_NODATA = 1;
            ABORT3 = 1;
            call symput("BREAK", 1);
        end;
    run;

%end;

%if &BREAK = 0 %then %do;
    * Optimal matching without strata and without calipers can use the linear
    * assignment solver.  Otherwise, create the graph input manually for the
    * simplex solver.
    ;
    %if &caliper = NONE &
        &stratvar = _ONE_ &
        &matchtype = OPT %then %do;

        proc optnet
            loglevel = moderate
            graph_direction = directed
            data_links = _links
        ;
            linear_assignment
                out = _mcf
            ;
        run;

        data _null_;
            STATUS = scan(substr("&_OROPTNET_LAP_", 8), 1);

            call symput("SOLVER_STATUS", STATUS);
        run;
    %end;
    %else %do;
        proc sql;
            create table _stratinfo as
            select
                strat,
                count(distinct from) as N_G1,
                count(distinct to) as N_G2,
                count(distinct from) / count(distinct to) as OBSRATIO
            from _links
            group by strat
            order by strat;
        quit;

        data _stratinfo;
            set _stratinfo;

            %if &matchtype = FULL %then %do;
                REQFLOW = max(n_g1, n_g2);
                %if &ratio_max ^= NONE & &ratio_min ^= NONE %then %do;
                    U_RATIO = obsratio * &ratio_max;
                    L_RATIO = obsratio * &ratio_min;
                    format u_ratio l_ratio best5.;

                    if u_ratio > 1 then do;
                TO_MAX = ceil(u_ratio);
                FROM_MIN = 1;
            end;
            else do;
                TO_MAX = 1;
                FROM_MIN = floor(1 / u_ratio);
            end;

            if l_ratio > 1 then do;
                TO_MIN = floor(l_ratio);
                FROM_MAX = 1;
            end;
            else do;
                TO_MIN = 1;
                FROM_MAX = ceil(1 / l_ratio);
            end;
        %end;
        %else %do;
            FROM_MIN = 1;
            FROM_MAX = &INFINITY;
            TO_MIN = 1;
            TO_MAX = &INFINITY;
        %end;
    %end;
    %else %do;
        REQFLOW = min(n_g1, n_g2);

        FROM_MAX = 1;
        TO_MAX = 1;

        if n_g1 <= n_g2 then do;
            FROM_MIN = 1;
            TO_MIN = 0;
            *TO_MIN = .;
        end;
        else do;
            FROM_MIN = 0;
            *FROM_MIN = .;
            TO_MIN = 1;
        end;
    %end;

    format obsratio best5.;
run;

proc sql;
    create table _nodes as
    select
        sum(reqflow) as REQFLOW
    from _stratinfo;
quit;

data _nodes;
    set _nodes;

    * FROM / SUPPLY node;
    NODE = -1;
    WEIGHT = reqflow;
    WEIGHT2 = &infinity;
    output;

    * TO / DEMAND node;
    NODE = -2;
    WEIGHT = -&infinity;
    WEIGHT2 = 0;
    output;
run;

proc sql;
    create table _links2 as
    select
        case FROM_IND
            when 1 then -1
            when 0 then NODE
        end as FROM,
        case FROM_IND
            when 1 then NODE
            when 0 then -2
        end as TO,
        case FROM_IND
            when 1 then FROM_MIN
            when 0 then TO_MIN
        end as LOWER,
        case FROM_IND
            when 1 then FROM_MAX
            when 0 then TO_MAX
        end as UPPER,
        0 as WEIGHT,
        s.STRAT
    from
        _stratinfo s,
        (   select
                1 as FROM_IND,
                FROM as NODE,
                STRAT
            from _links
            UNION
            select
                0 as FROM_IND,
                TO as NODE,
                STRAT
            from _links
        ) l
    where
        s.strat = l.strat
    ;
quit;

%if &caliper ^= NONE &
    &matchtype = OPT %then %do;

    * Set up excess node to prevent infeasibility;
    data _links2;
        set _links2;

        output;

        if from = -1 then do;
            from = to;
            to = -9;
            weight = 10 * &caliper;
            lower = .;
            upper = 1;
            output;
        end;
        if to = -2 then do;
```

```
                to = from;
                from = -9;
                weight = 10 * &caliper;
                lower = .;
                upper = 1;
                output;
            end;
        run;
    %end;

    data _links;
        set
            _links
            _links2
        ;
    run;

    proc optnet
        loglevel = moderate
        graph_direction = directed
        data_nodes = _nodes
        data_links = _links
        out_links = _mcf
        internal_format = thin
    ;
        mincostflow
            logfreq = 1000
        ;
    run;

    data _null_;
        Status = scan(substr("&_OROPTNET_MCF_", 8), 1);

        call symput("SOLVER_STATUS", Status);
    run;

    data _mcf;
        set _mcf;
        where
            mcf_flow and
            from not in (-1, -2, -9) and
            to not in (-1, -2, -9)
        ;
    run;
%end;

proc optnet
    data_links = _mcf
    out_nodes = _connected
;
    concomp;
run;

proc sql;
    create table &outlinks as
    select
        x1.&idvar as GRP1,
        x2.&idvar as GRP2,
        %if &stratvar ^= _ONE_ %then
            x1.&stratvar,;
        l.weight as DIST
    from
        _mcf l,
            _xwalk x1,
            _xwalk x2
        where
            l.from = x1._idx and
            l.to = x2._idx
        ;

        create table &outds as
        select
            x.&idvar,
            x.&groupvar,
            %if &stratvar ^= _ONE_ %then
                x.&stratvar,;
            c.concomp as SETNUM
        from
            _connected c,
            _xwalk x
        where
            c.node = x._idx
        order by
            c.concomp,
            x.&idvar,
            x.&groupvar
        ;

        create table _stratwt as
        select
            setnum,
            sum(&groupvar = &g1) as SET_N1,
            sum(&groupvar = &g2) as SET_N2,
            count(*) as SET_N
        from &outds
        group by setnum
        order by setnum;
    quit;

    data &outds;
        merge
            &outds
            _stratwt
        ;
        by setnum;
    run;

    proc sql;
        create table _check5 as
        select
            1 as MERGEVAR,
            count(distinct GRP1) as LINK_G1,
            count(distinct GRP2) as LINK_G2,
            count(distinct GRP1) + count(distinct GRP2) as LINK_N,
            %if &stratvar ^= _ONE_ %then
                count(distinct &stratvar) as LINK_STRATA,
            ;
            sum(DIST) as TOTAL_DIST
        from &outlinks;
    quit;

%end;

* Output specifications and status;
data _null_;
    merge
    _check1
    %if %sysfunc(exist(_check2)) %then
        _check2
    ;
    %if %sysfunc(exist(_check3)) %then
        _check3
    ;
    %if %sysfunc(exist(_check4)) %then
        _check4
    ;
    %if %sysfunc(exist(_check5)) %then
        _check5
    ;
;
by mergevar;

* Macro info;
put "-----FMATCH macro called-----";
put;
put "Macro parameters specified";
put "  Input dataset:               " spec_inds;
put "  Matching variable:           " spec_matchvar;
put "  Group variable:              " spec_groupvar;
put "  ID variable:                 " spec_idvar;
put "  Stratification variable:     " spec_stratvar;
put "  Caliper width:               " spec_caliper;
if spec_caliper ne "NONE" then
put "  Caliper type:                " spec_caliper_type;
put "  Max ratio multiplier:        " spec_ratio_max;
put "  Min ratio multiplier:        " spec_ratio_min;
put "  Match type:                  " spec_matchtype;
put "  Link-level output dataset:   " spec_outlinks;
put "  Record-level output dataset: " spec_outds;
put;

* If aborted, indicate why
* Else, report results
;
if abort1 then do;
    put "ERROR: Macro aborted";
    if parm_nomatch then
        put "ERROR: No matching variable specified";
    if parm_nogroup then
        put "ERROR: No group variable specified";
    if parm_noid then
        put "ERROR: No ID variable specified";
    if parm_1ratio then
        put "ERROR: Both RATIO_MIN and RATIO_MAX need to be specified if
             one is specified";
    if parm_hiloratio then
        put "ERROR: RATIO_MIN must be < 1.0 and RATIO_MAX must be > 1.0";
    if parm_noinds then
        put "ERROR: No input dataset specified";
    if parm_badinds then
        put "ERROR: Input dataset does not exist";
    if parm_badmatchvar then
        put "ERROR: Match variable does not exist";
    if parm_badgroupvar then
        put "ERROR: Group variable does not exist";
    if parm_badidvar then
        put "ERROR: ID variable does not exist";
    if parm_badstratvar then
        put "ERROR: Stratification variable does not exist";
        if parm_badcaliptype then
            put "ERROR: Caliper type must be ABS or SD";
        if parm_badmatchtype then
            put "ERROR: Match type must be OPT or FULL";
    end;
    %if %sysfunc(exist(_check2)) %then %do;
        else if abort2 then do;
            put "ERROR: Macro aborted";
            if indata_badgrp then
                put "ERROR: Group variable has " grp_levels " levels -- 2 are
                     required";
            if indata_badids then
                put "ERROR: ID variable is not unique";
        end;
    %end;
    %if %sysfunc(exist(_check4)) %then %do;
        else if abort3 then do;
            put "ERROR: Macro aborted";
            if link_nodata then
                put "ERROR: No data to link after applying caliper (if
                     specified) and after";
                put "       removing records with missing data in key fields";
        end;
    %end;
    %if %sysfunc(exist(_check5)) %then %do;
        else do;
            USE_PCT = trim(left(put(link_n / n_input, percentn7.1)));
            G1_PCT = trim(left(put(link_g1 / n_g1, percentn7.1)));
            G2_PCT = trim(left(put(link_g2 / n_g2, percentn7.1)));
            put "Matching completed";
            put "  " LINK_N "of " N_INPUT "(" USE_PCT +(-1) ") records used for
                 matching";
            if spec_caliper ne "NONE" and use_limit ne n_input then
            put "     - Some records may have had no potential matches within
                 the caliper";
            if nmiss_match > 0 then
            put "     - " nmiss_match "record(s) missing matching variable";
            if nmiss_grp > 0 then
            put "     - " nmiss_grp "record(s) missing group variable";
            if nmiss_id > 0 then
            put "     - " nmiss_id "record(s) missing ID variable";
            %if &stratvar ^= _ONE_ %then %do;
                STRAT_PCT = trim(left(put(link_strata / n_strata,
                            percentn7.1)));
                put "  " LINK_STRATA "of " N_STRATA "(" STRAT_PCT +(-1) ")
                     strata used for matching";
            %end;
            if nmiss_strat > 0 then
            put "     - " nmiss_strat "record(s) missing stratification
                     variable";
            put "  Group 1 defined as %upcase(&groupvar) = &g1, " link_g1 "of "
                 n_g1 "(" g1_pct +(-1) ") used for matching";
            put "  Group 2 defined as %upcase(&groupvar) = &g2, " link_g2 "of "
                 n_g2 "(" g2_pct +(-1) ") used for matching";
            put "  Total distance between matched records is " total_dist;
            put "  Solver status is %trim(&solver_status) (anything other than
                 OPTIMAL may indicate a problem)";
            put;
        end;
    %end;
    put "------------------------------";
run;
```

```
    proc datasets library=work nolist;
        delete _check1;
        %if %sysfunc(exist(_check2)) %then %do; delete _check2; %end;
        %if %sysfunc(exist(_check3)) %then %do; delete _check3; %end;
        %if %sysfunc(exist(_check4)) %then %do; delete _check4; %end;
        %if %sysfunc(exist(_check5)) %then %do; delete _check5; %end;
        %if %sysfunc(exist(_stratinfo)) %then %do; delete _stratinfo; %end;
        %if %sysfunc(exist(_nodes)) %then %do; delete _nodes; %end;
        %if %sysfunc(exist(_links)) %then %do; delete _links; %end;
        %if %sysfunc(exist(_links2)) %then %do; delete _links2; %end;
        %if %sysfunc(exist(_mcf)) %then %do; delete _mcf; %end;
        %if %sysfunc(exist(_connected)) %then %do; delete _connected; %end;
        %if %sysfunc(exist(_stratwt)) %then %do; delete _stratwt; %end;
        %if %sysfunc(exist(_useds)) %then %do; delete _useds; %end;
        %if %sysfunc(exist(_xwalk)) %then %do; delete _xwalk; %end;
    run;

%mend;
```

REFERENCES

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica, 76*(6), 1537-1557.

Abadie, A., & Imbens, G. W. (2012). A Martingale Representation for Matching Estimators. *Journal of the American Statistical Association, 107*(498), 833-843.

Abraham, W. T., Adams, K. F., Fonarow, G. C., Costanzo, M. R., Berkowitz, R. L., LeJemtel, T. H., . . . Wynne, J. (2005). In-hospital mortality in patients with acute decompensated heart failure requiring intravenous vasoactive medications: an analysis from the Acute Decompensated Heart Failure National Registry (ADHERE). *Journal of the American College of Cardiology, 46*(1), 57-64.

Adams, K. F., Fonarow, G. C., Emerman, C. L., LeJemtel, T. H., Costanzo, M. R., Abraham, W. T., . . . ADHERE Scientific Advisory Committee Investigators. (2005). Characteristics and outcomes of patients hospitalized for heart failure in the United States: rationale, design, and preliminary observations from the first 100,000 cases in the Acute Decompensated Heart Failure National Registry (ADHERE). *American Heart Journal, 149*(2), 209-216.

Ahuja, R., Magnanti, T., & Orlin, J. (1993). *Network flows*. Upper Saddle River, NJ: Prentice-Hall.

Aranda, J. M., Schofield, R. S., Pauly, D. F., Cleeton, T. S., Walker, T. C., Monroe, V. S., . . . Hill, J. A. (2003). Comparison of dobutamine versus milrinone therapy in hospitalized patients awaiting cardiac transplantation: a prospective, randomized trial. *American Heart Journal, 145*(2), 324-329.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics & Data Analysis, 55*(4), 1770-1780.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine, 27*(12), 2037-2049.

Austin, P. C. (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics-Simulation and Computation, 38*(6), 1228-1234.

Austin, P. C. (2010). Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology, 172*(9), 1092-1097.

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics, 10*(2), 150-161.

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

Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine, 33*(24), 4306-4319.

Barbash, G. I., Friedman, B., Glied, S. A., & Steiner, C. A. (2014). Factors associated with adoption of robotic surgical technology in US hospitals and relationship to radical prostatectomy procedure volume. *Annals of Surgery, 259*(1), 1-6.

Bergstralh, E. K., Jon. (2003). Computerized matching of cases to controls using the greedy matching algorithm with a fixed number of controls per case. Retrieved 01/01/2015, from http://www.mayo.edu/research/documents/gmatchsas/doc-10027248.

Bernheim, S. M., Grady, J. N., Lin, Z., Wang, Y., Wang, Y., Savage, S. V., . . . Merrill, A. R. (2010). National Patterns of Risk-Standardized Mortality and Readmission for Acute Myocardial Infarction and Heart Failure Update on Publicly Reported Outcomes Measures Based on the 2010 Release. *Circulation: Cardiovascular Quality and Outcomes, 3*(5), 459-467.

Bhattacharya, J., & Vogt, W. B. (2007). Do instrumental variables belong in propensity scores? : National Bureau of Economic Research Technical Working Paper No. 343. Retrieved 01/01/2015, from http://www.nber.org/papers/t0343.

Brookhart, M. A., & Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The International Journal of Biostatistics, 3*(1), 1-25.

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

Carpenter, W. R., Reeder-Hayes, K., Bainbridge, J., Meyer, A.-M., Amos, K. D., Weiner, B. J., & Godley, P. A. (2011). The role of organizational affiliations and research networks in the diffusion of breast cancer treatment innovation. *Medical Care, 49*(2), 172-179.

Casey, M., Burlew, M., & Moscovice, I. (2010). Critical Access Hospital Year 5 Hospital Compare Participation and Quality Measure Results: *Flex Monitoring Team Policy Brief #15*. Retrieved 01/01/2015, from http://www.flexmonitoring.org/wp-content/uploads/2007/04/BriefingPaper26-HospitalCompare5.pdf.

Centers for Medicare and Medicaid Services. Hospital Compare. Retrieved 05/01/2014, from http://www.medicare.gov/hospitalcompare.

Cook, A. J., Wellman, R. D., Marsh, T. L., Tiwari, R. C., Nguyen, M. D., Russek-Cohen, E., . . . Nelson, J. C. (2012). Statistical Methods for Estimating Causal Risk Differences (PRISM). Retrieved 01/01/2015, from http://mini-sentinel.org/methods/ methods_development/details.aspx? ID=1037.

Coons, J. C., McGraw, M., & Murali, S. (2011). Pharmacotherapy for acute heart failure syndromes. *American Journal of Health-System Pharmacy, 68*(1), 21-35.

Coye, M. J., & Kell, J. (2006). How hospitals confront new technology. *Health Affairs, 25*(1), 163-173.

Curtis, L. H., Hammill, B. G., Eisenstein, E. L., Kramer, J. M., & Anstrom, K. J. (2007). Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. *Medical Care, 45*(10), S103-S107.

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

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Fisher, E. S., Bell, J., Tomek, I., Esty, A., Goodman, D., & Bronner, K. (2010). Trends and regional variation in hip, knee, and shoulder replacement. *Dartmouth Atlas Surgery Report. Hanover, NH: The Dartmouth Institute for Health Policy and Clinical Practice*.

Friedman, L. M., Furberg, C., & DeMets, D. L. (2010). *Fundamentals of clinical trials*. New York, NY: Springer.

Funk, M. J., Westreich, D., Weisen, C., & Davidian, M. (2010). Doubly robust estimation of treatment effects. *Analysis of Observational Health Care Data Using SAS. Ed. by Douglas Faries et al. Cary, NC: SAS Institute*, 85-103.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Goldman, L. E., & Dudley, R. A. (2008). United States rural hospital quality in the Hospital Compare Database—accounting for hospital characteristics. *Health Policy, 87*(1), 112-127.

Goodman, D. C., Fisher, E. S., & Chang, C.-H. (2011). After hospitalization: a Dartmouth atlas report on post-acute care for Medicare beneficiaries. *The Dartmouth Institute for Health Policy and Clinical Practice*.

Goodney, P. P., Travis, L., Lucas, F. L., Fisher, E. S., Goodman, D. C., Bronner, K. K., & Esty, A. R. (2010). Trends and regional variation in carotid revascularization. *The Dartmouth Institute for Health Policy and Clinical Practice*.

Griswold, M. E., Localio, A. R., & Mulrow, C. (2010). Propensity score adjustment with multilevel data: setting your sites on decreasing selection bias. *Annals of Internal Medicine, 152*(6), 393-395.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405-420.

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

Hernán, M. A. (2005). Invited commentary: hypothetical interventions to define causal effects—afterthought or prerequisite? *American Journal of Epidemiology, 162*(7), 618-620.

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003'by Peter Austin, Statistics in Medicine. *Statistics in Medicine, 27*(12), 2055-2061.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology, 2*(3-4), 259-278.

Hong, G. (2010). Marginal mean weighting through stratification: Adjustment for selection bias in multilevel data. *Journal of Educational and Behavioral Statistics, 35*(5), 499-531.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy. *Journal of the American Statistical Association, 101*(475), 901-910.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association, 47*(260), 663-685.

Horvitz-Lennon, M., Alegría, M., & Normand, S.-L. T. (2012). The Effect of Race-Ethnicity and Geography on Adoption of Innovations in the Treatment of Schizophrenia. *Psychiatric Services, 63*(12), 1171-1177.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., & Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models. *The American Statistician, 58*(4), 272-279.

Jonker, R., & Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing, 38*(4), 325-340.

Karlsberg, R. P., DeWood, M. A., DeMaria, A. N., Berk, M. R., Lasher, K. P., & Group, M. D. S. (1996). Comparative efficacy of short-term intravenous infusions of milrinone and dobutamine in acute congestive heart failure following acute myocardial infarction. *Clinical Cardiology, 19*(1), 21-30.

Kelcey, B. (2011). Multilevel Propensity Score Matching within and across Schools. *Society for Research on Educational Effectiveness*.

Kim, J., & Seltzer, M. (2007). Causal Inference in Multilevel Settings in Which Selection Processes Vary across Schools. CSE Technical Report 708. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Krumholz, H. M., Merrill, A. R., Schone, E. M., Schreiner, G. C., Chen, J., Bradley, E. H., . . . Straube, B. M. (2009). Patterns of hospital performance in acute myocardial infarction and heart failure 30-day mortality and readmission. *Circulation: Cardiovascular Quality and Outcomes, 2*(5), 407-413.

Li, F., Zaslavsky, A. M., & Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine, 32*(19), 3373-3387.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

Lindenauer, P. K., Bernheim, S. M., Grady, J. N., Lin, Z., Wang, Y., Wang, Y., . . . Drye, E. E. (2010). The performance of US hospitals as reflected in risk-standardized 30-day mortality and readmission rates for medicare beneficiaries with pneumonia. *Journal of Hospital Medicine, 5*(6), E12-E18.

Localio, A. R., Berlin, J. A., Ten Have, T. R., & Kimmel, S. E. (2001). Adjustments for center in multicenter studies: an overview. *Annals of Internal Medicine, 135*(2), 112-123.

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.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India, 2*(1), 49-55.

Massachusetts Department of Public Health. (2013). Adult Coronary Artery Bypass Graft Surgery in the Commonwealth of Massachusetts, Fiscal Year 2011 Report.

Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56*(1), 118-124.

Munson, J. C., Morden, N. E., Goodman, D. C., Valle, L. F., Wennberg, J. E., & Bronner, K. K. (2013). Dartmouth Atlas of Medicare Prescription Drug Use. *The Dartmouth Institute for Health Policy and Clinical Practice*.

Neuhaus, J. M., Hauck, W. W., & Kalbfleisch, J. D. (1992). The Effects of Mixture Distribution Misspecification when Fitting Mixed-Effects Logistic Models. *Biometrika, 79*(4), 755-762.

New York State Department of Health. (2012). Adult Cardiac Surgery in New York State 2008 – 2010.

Peacock, W., Costanzo, M., De Marco, T., Lopatin, M., Wynne, J., & Mills, R. (2009). Impact of intravenous loop diuretic on outcomes of patients hospitalized with acute decompensated heart failure: Insight from the ADHERE Registry. *Cardiology, 113*, 12-19.

Pennsylvania Health Care Cost Containment Council. (2013). Cardiac Surgery in Pennsylvania: Information About Hospitals and Cardiothoracic Surgeons.

Rassen, J. A., Mittleman, M. A., Glynn, R. J., Brookhart, M. A., & Schneeweiss, S. (2010). Safety and effectiveness of bivalirudin in routine care of patients undergoing percutaneous coronary intervention. *European Heart Journal, 31*(5), 561-572.

Ray, W. A., Murray, K. T., Griffin, M. R., Chung, C. P., Smalley, W. E., Hall, K., . . . Stein, C. M. (2010). Outcomes With Concurrent Use of Clopidogrel and Proton-Pump InhibitorsA Cohort Study. *Annals of Internal Medicine, 152*(6), 337-345.

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11*(5), 550-560.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*(427), 846-866.

Rodrıguez, G., & Elo, I. (2003). Intra-class correlation in random-effects models for binary data. *The Stata Journal*, 3(1), 32-46.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association, 84*(408), 1024-1032.

Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological)*, 597-610.

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. (1985a). The bias due to incomplete matching. *Biometrics*, 103-116.

Rosenbaum, P. R., & Rubin, D. B. (1985a). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38.

Ross, J. S., Cha, S. S., Epstein, A. J., Wang, Y., Bradley, E. H., Herrin, J., . . . Krumholz, H. M. (2007). Quality of care for acute myocardial infarction at urban safety-net hospitals. *Health Affairs, 26*(1), 238-248.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688-701.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods, 13*(4), 279-313.

Schneeweiss, S., Solomon, D. H., Wang, P. S., Rassen, J., & Brookhart, M. A. (2006). Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: An instrumental variable analysis. *Arthritis & Rheumatism, 54*(11), 3390-3398.

Shahian, D. M., Nordberg, P., Meyer, G. S., Blanchfield, B. B., Mort, E. A., Torchiana, D. F., & Normand, S.-L. T. (2012). Contemporary performance of US teaching and nonteaching hospitals. *Academic Medicine, 87*(6), 701-708.

Sheets, N. C., Goldin, G. H., Meyer, A.-M., Wu, Y., Chang, Y., Stürmer, T., . . . Carpenter, W. R. (2012). Intensity-modulated radiation therapy, proton therapy, or conformal radiation therapy and morbidity and disease control in localized prostate cancer. *JAMA: The Journal of the American Medical Association, 307*(15), 1611-1620.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology, 59*(5), 437.e1-437.e24.

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

Tardos, E., & Kleinberg, J. (2006). *Algorithm Design.* Reading, MA: Addison-Wesley.

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46*(3), 514-543.

Venkitachalam, L., Lei, Y., Magnuson, E. A., Chan, P. S., Stolker, J. M., Kennedy, K. F., . . . Cohen, D. J. (2011). Survival Benefit With Drug-Eluting Stents in Observational

Studies Fact or Artifact? *Circulation: Cardiovascular Quality and Outcomes, 4*(6), 587-594.

Walker, A. M. (2013). Matching on provider is risky. *Journal of Clinical Epidemiology, 66*(8), S65-S68.

Wennberg, J. E. (2002). Unwarranted variations in healthcare delivery: implications for academic medical centres. *BMJ: British Medical Journal, 325*(7370), 961-964.

Williamson, E. J., Forbes, A., & White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine, 33*(5), 721-737.

Yamani, M. H., Haji, S. A., Starling, R. C., Kelly, L., Albert, N., Knack, D. L., & Young, J. B. (2001). Comparison of dobutamine-based and milrinone-based therapy for advanced decompensated congestive heart failure: hemodynamic efficacy, clinical outcome, and economic impact. *American Heart Journal, 142*(6), 998-1002.