University of Louisville

ThinkIR: The University of Louisville's Institutional Repository

Electronic Theses and Dissertations

5-2014

Statistical methods for assessing treatment effects for observational studies.

Kristopher C. Gardner 1984-
University of Louisville

Follow this and additional works at: <https://ir.library.louisville.edu/etd>

Part of the [Bioinformatics Commons](#), and the [Biostatistics Commons](#)

Recommended Citation

Gardner, Kristopher C. 1984-, "Statistical methods for assessing treatment effects for observational studies." (2014). *Electronic Theses and Dissertations*. Paper 479.
<https://doi.org/10.18297/etd/479>

This Master's Thesis is brought to you for free and open access by ThinkIR: The University of Louisville's Institutional Repository. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of ThinkIR: The University of Louisville's Institutional Repository. This title appears here courtesy of the author, who has retained all other copyrights. For more information, please contact thinkir@louisville.edu.

STATISTICAL METHODS FOR ASSESSING TREATMENT EFFECTS
FOR OBSERVATIONAL STUDIES

By

Kristopher C. Gardner
B.A., University of Kentucky, 2007
M.S. University of Louisville, 2014

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

Master of Sciences

Department of Biostatistics
University of Louisville
Louisville, KY

May 2014

STATISTICAL METHODS FOR ASSESSING TREATMENT EFFECTS
FOR OBSERVATIONAL STUDIES

By

Kristopher C. Gardner
B.A., University of Kentucky, 2007
M.S., University of Louisville, 2014

A Thesis Approved on

April 17, 2014

By the following Thesis Committee:

Dr. Maiying Kong

Dr. Dongfeng Wu

Dr. Maxwell Boakye

ACKNOWLEDGEMENTS

I would first and foremost like to thank my adviser, Dr. Maiying Kong, not only for her direction and assistance in working on this project, but for helping me select a topic which I found fascinating and enjoyable to study. At least a small thanks is also due to her husband, Dr. Changbing Hu, for preparing me in my undergraduate pre-requisite courses for the Master's program. I thank him and Dr. Kong for their encouragement and friendship. I also extend my thanks to Dr. Boakye for providing the MarketScan data used for this project and for serving on the thesis committee.

I would like to thank my professors, who fielded what I am sure they considered at times to be far too many question, and (usually) did so with great patience. I would like to thank my classmates, in particular Jennifer and Hemant, without whose help and friendship I would have been overwhelmed. I would like to thank Beatrice Ugiliweneza for sharing her expertise with the dataset and SAS and without whose help I certainly could not have done this project. Thanks also to Paul, for helping me think through numerous data manipulation problems until I could arrive at workable answers.

I would like to thank my family, without whose help caring for Madelyn Joy I certainly could not have finished this project in a reasonable amount of time. And I would like to thank Abigail, whom I ensure will have her husband back once the craziness of completing this thesis is finished. I love you.

ABSTRACT

STATISTICAL METHODS FOR ASSESSING TREATMENT EFFECTS FOR OBSERVATIONAL STUDIES

Kristopher C. Gardner

April 17, 2014

Though randomized clinical (RCTs) trials are the gold standard for comparing treatments, they are often infeasible or exclude clinically important subjects, or generally represent an idealized medical setting rather than real practice. Observational data provide an opportunity to study practice-based evidence, but also present challenges for analysis. Traditional statistical methods which are suitable for RCTs may be inadequate for the observational studies. In this project, four of the most popular statistical methods for observational studies: ANCOVA, propensity score matching, regression with the propensity score as a covariate, and instrumental variables (IV) are investigated through application to MarketScan insurance claims data. Each of these methods is used to compare BMP versus autograft spinal surgeries for the outcomes length of stay, complications, and cost. Recommendations are made as to when each particular method may or may not be the optimal choice.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
INTRODUCTION.	1
ANCOVA METHOD	9
PROPENSITY SCORE METHODS.....	13
INSTRUMENTAL VARIABLES METHOD.....	22
DATA ANALYSIS.....	29
DISCUSSION	52
REFERENCES	55
APPENDIX	60
CURRICULUM VITA	67

CHAPTER I

INTRODUCTION

When there are multiple treatment methods for a common condition, it is natural to want to compare the treatments directly. Undoubtedly, the best way to accomplish this is through randomized clinical trials (RCTs). However, the costs can be prohibitive – in money, time, the availability of applicable participants, etc. – which has led researchers to turn to alternatives to RCTs, one of most common of which is comparative effectiveness research (CER). CER has been defined by the Institute of Medicine committee as “the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. The purpose of CER is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels”^[1, 2, 3]. CER requires the development, expansion, and use of a variety of data sources and methods to conduct timely and relevant research and disseminate the results in a form that is quickly usable by clinicians, patients, policy makers, and health plans and other payers^[4]. The data sources could be observational data (e.g. administrative claims data, electronic medical records, registries, or other clinical cohorts) or randomized clinical trials^[5].

RCTs are the most rigorous method of generating comparative effectiveness evidence and are considered the gold standard for evaluating the efficacy of a treatment.

RCTs incorporate randomization of test subjects: an essential component to many of the assumptions underlying traditional statistical analyses. When subjects are randomly assigned to two treatment groups, and when the sample size is large, the two treatment groups are balanced on all potential covariates, both observed and unobserved. This means that there are not fundamental underlying differences between the two treatment groups that are causing differences in treatment response, and the observed difference in response is instead due only to the treatment assignment. Randomization is the only way to effectively achieve balance of both observed and unobserved covariates between treatment groups. Traditional methods of analysis for RCTs, such as Student t-tests, or the Wilcoxon-Mann test if the outcome variable is not normally distributed, are well studied and well understood.

However, many RCTs exclude clinically relevant patient subgroups (as defined by age, sex, race, ethnicity, and comorbid conditions), commonly used comparator interventions, important patient outcomes (such as quality of life and longer-term effects), and non-expert providers. These exclusions diminish the relevance of the trial results for some important clinical and policy decisions^[5]. On the other hand, the observational data, such as administrative claims, electronic medical records, and registries, provide opportunities to study the practice-based evidence^[6]. Unlike with RCTs, the potential covariates (both observed and unobserved) could be unbalanced between treatment groups, or could even contain no overlap at all. Thus, any observed difference in outcomes between the groups could be the result of the treatments, or it could be the result of

differing covariate measures. To make valid estimates of the treatment effect, appropriate statistical methods must be used to adjust for the potential imbalance in the covariates^[7, 8]. In this project, it is my aim to examine four of the most popular statistical methods used in observational studies: the standard analysis of covariance (ANCOVA) approach, the propensity score-based regression method, the propensity score-based matching method, and the use of instrumental variables (IV). Each method has its own assumptions and is applicable under certain settings. I wish to highlight the pros and cons of each method, compare these methods via data analyses on MarketScan administrative claims data, and provide guidelines for which method may be optimal for a given CER study.

Without loss of generality, let us assume we are interested in comparing the outcome between a treatment group and a control group. The treatment effect, conceptually, is the difference in outcome a subject experiences when given treatment versus that he experiences given control. The two most common measures of treatment effect are population average treatment effect (ATE) and average treatment effect among the treated (ATT). To describe the treatment effect rigorously, I borrow the “potential outcome” notation used by Rosenbaum and Rubin^[9], Rubin^[8], Little and Rubin^[10], and Hirano, Imbens, and Ridder^[11]. Suppose a random sample of size N subjects, indexed by $i=1, \dots, N$, are drawn randomly from a large population. The i^{th} subject has the outcome $Y_i(1)$ if the subject receives the treatment, and has the outcome $Y_i(0)$ if the subject receives the control. $Y_i(0)$ and $Y_i(1)$ are called the “potential outcomes” for i^{th} subject, where either $Y_i(1)$ or $Y_i(0)$, depending on the treatment the subject received, is observed. In addition, each subject has a vector of characteristics, referred to as covariates or pre-treatment

variables, denoted by X_i . Rubin^[8] called the covariates and the potential outcomes, say $(X_i, Y_i(1), Y_i(0))$, the science, which is not affected by how or whether we try to learn about it, whether by completely randomized experiment, randomized blocks designs, or observational studies. The treatment effect for i^{th} subject would be defined as

$$\tau_i \equiv E[Y_i(1) - Y_i(0)]. \quad (1)$$

The population average treatment effect (ATE) is defined as:

$$\tau_{ate} \equiv E[Y(1) - Y(0)], \quad (2)$$

where $E(\cdot)$ denotes expectation in the population. Let T_i denote the treatment assignment for i^{th} subject ($i=1, \dots, N$), $T_i = 1$ if the subject receives treatment, and $T_i = 0$ if the subject receives control. The average treatment effect among the treated (ATT) is defined as

$$\tau_{att} \equiv E[Y(1) - Y(0) | T = 1]. \quad (3)$$

The expression for ATE and ATT are very straightforward theoretically, but in practice it is impossible to observe both $Y_i(1)$ and $Y_i(0)$ for the same subject. When subjects are randomly assigned to treatment and control groups, the treatment assignment is independent of the outcome variables, which implies that ATT and ATE are equal. However, in observational studies, ATT may be greater than ATE, if for example, people who elect treatment do so because they are more likely to benefit from the treatment than those who elect not to receive treatment. This is known as selection bias, and in medical settings, it is probably reasonable to assume that selection bias is taking place, since a doctor would reasonably be expected to recommend a procedure to a patient if he thinks that patient is a particularly good candidate for the procedure and will respond well. Conversely, the doctor may not recommend the procedure to a lesser candidate, since the benefit may not be

worth the cost. Thus, in observational studies from medical settings, it is likely that the ATT will be greater than the ATE, because the treatment group is comprised of a greater proportion of “ideal” candidates than is found in the general population.

Although the definition for ATE and ATT are straightforward, it is impossible to estimate these quantities without making speculations about the outcome one treatment would have had on a subject which, in fact, did not receive such treatment. Using the notation of Angrist and Pischke^[12], in general, the average treatment effect is estimated by $E(Y(1)|X, T = 1) - E(Y(0)|X, T = 0)$. Note that

$$\begin{aligned} & E(Y(1)|X, T = 1) - E(Y(0)|X, T = 0) \\ &= E\{(Y(1) - Y(0))|X, T = 1\} + E(Y(0)|X, T = 1) - E(Y(0)|X, T = 0). \end{aligned} \quad (4)$$

$E(Y(1)|X, T = 1) - E(Y(0)|X, T = 0)$ could be estimated by the observed difference of outcomes in treated group and control group, the first term in the right hand of Equation (4), $E\{(Y(1) - Y(0))|X, T = 1\}$, is the average treatment effect on the treated, and the second term in the right hand of Equation (4), $E(Y(0)|X, T = 1) - E(Y(0)|X, T = 0)$, is the selection bias, describing the outcome difference and the treatment group and the control group if the treatment group had received the control. In the case that there is not selection bias, the observed difference of outcomes between treated group and control group is an unbiased estimate for ATT^[12]. In RCTs, the potential outcomes, say $(Y(0), Y(1))$, are independent of the treatment assignment, thus the selection bias is zero. For non-randomized studies, Rosenbaum and Rubin^[9] coined the “ignorable treatment assignment” assumption:

$$(Y(0), Y(1)) \perp T | X. \quad (5)$$

This assumption says that the treatment assignment is independent of the potential outcome variables $(Y(0), Y(1))$. This assumption is also referred to as the “unconfoundedness assumption” according to Imbens^[13], and the “conditional independence assumption”^[12]. Under this assumption, the selection bias is zero, and

$$E\{(Y(1) - Y(0))|X, T = 1\} = E\{(Y(1) - Y(0))|X\},$$

which implies that the ATT and ATE are the same. The “overlap” assumption^[13], regarding the joint distribution of treatments and covariates, says that

$$0 < Pr(T = 1|X) < 1. \quad (6)$$

Under the “overlap” assumption, for a given covariate $X = x$, there are some subjects assigned to treatment group, and there are some subjects assigned to control group. The “overlap” assumption and the “unconfoundedness assumption”^[13] altogether are referred to as “strongly ignorable treatment assignment” assumption as expressed by Rosenbaum and Rubin^[9], which are the key assumptions for valid inferences for treatment effect based on the ANCOVA approach or the propensity score-based regression/matching methods.

Analysis of covariance (ANCOVA) is the fundamental method of analysis in both randomized and non-randomized studies – including CER – as it is possible to pair with propensity score analysis and is the necessary underpinning for IV analysis, or ANCOVA can stand alone as the method of analysis in an observational study. ANCOVA is very well established, dating back to Fisher’s use in a 1932 study^[14]. The method is valued for its ease of implementation, straightforward interpretation, and widespread understanding of its applications, qualities, and drawbacks. However, when there are a large number of

covariates, one may run into the high-dimensional problem. In such cases, the propensity score based methods can easily address the high-dimensional issue.

The propensity score is defined as the conditional probability of assignment to treatment group, given the covariates,

$$e(x) = Pr(T = 1|X = x). \quad (8)$$

Rosenbaum and Rubin^[9] showed that under the “strongly ignorable treatment assignment” assumption, the difference between treatment and control means at each propensity score is an unbiased estimate of the treatment effect at that value, and consequently pair matching, subclassification and covariance adjustment on the propensity score can produce unbiased estimates of the average treatment effect. Propensity scores contain information about all observed covariates, and propensity scores have become a very popular way to balance covariates between treatment groups in observational studies, and thus produce unbiased estimates of the treatment effects. In addition, the propensity score can be used in a model as a replacement for a vector of all the individual background covariates^[16].

When there are a very large number of covariates, this can greatly simplify the model and avoid the high dimensionality problems. Either alone or with select covariates, the simpler model using propensity score rather than all covariates may allow for more reliable goodness-of-fit tests on the model^[16]. Additionally, there is no concern with over-parameterization of the model when using propensity scores, as there is with a regression using all covariates individually^[16]. While the propensity score can stand alone, it can also be combined with a vector of what are deemed the most important or influential covariates in a regression model.

While propensity score analysis can be very effective in eliminating bias due to imbalance in measured covariates, there could be significant confounding due to imbalance in unmeasured or unobserved covariates, which may make the “unconfoundedness” and “overlap” assumptions invalid. In these cases, instrumental variables (IV) for treatment assignment – if they can be found – may be applied to deal with this confounding and reduce the bias of estimates of the treatment effect. The instrumental variable Z (if one exists) should be one that is correlated with the treatment assignment variable but uncorrelated with the error term. A two-stage least squares (2SLS or TSLS) method is then generally applied to estimate the treatment effect. The first stage involves regressing the instrument onto the treatment X , resulting in an estimate of $\hat{X}$ by OLS. Then $\hat{X}$ is regressed onto the outcome variable Y in the second stage, yielding an unbiased estimate of the treatment effect.

I will apply the ANCOVA method, the propensity score method, and the instrumental variables method to compare “the effectiveness and the cost-effectiveness of two relatively new highly popular spinal technologies: BMP [bone morphogenetic proteins] vs. autograft for spinal fusion for degenerative disease” using the MarketScan insurance claims dataset, which is a database of Medicare, Medicaid, and commercial insurance claims over the years 2004-2009. The outcome variables are length of hospital stay, complications, in-hospital cost of the procedures, and total 1-year follow-up outpatient cost for procedures.

CHAPTER II

ANALYSIS OF COVARIANCE

Analysis of covariance (ANCOVA) is a fundamental statistical method of analysis for both randomized and nonrandomized studies. It is a regression analysis which includes parameter estimates for observed covariates in addition to the treatment effects we are ultimately interested in estimating. ANCOVA for two treatment groups can be written as

$$Y_i = \mu + \tau T_i + (x_i - \bar{x})'\beta + \varepsilon_i. \quad (9)$$

Y_i is the outcome measure for the i th subject, μ is the mean for the control group, τ is the average difference between the treatment group and the control group, x_i is a measurement of the observed covariates for each experimental unit, with $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, T_i is the treatment indicator variable, and ε_i is the error term, which is usually assumed to follow a normal distribution with mean zero and variance σ^2 . β is a parameter that removes the influence of the measured covariates from the error term. It is assumed in this model that the covariates properly influence Y_i in a linear fashion^[15]. By including a measure of the covariate in the regression, confounding is reduced and more precise estimates of the true treatment effects are obtained, even where randomization was employed^[15]. Of course, where randomization is not employed, as is the case with observational studies, covariance adjustment is essential for producing unbiased estimates of the treatment effect.

As stated, ANCOVA may be used to improve the precision of estimates of the treatment effect in randomized experiments. This works because randomization does not *ensure* complete balance of covariates between treatment groups, and if measurements of observed covariates are taken before treatment, ANCOVA will adjust for any imbalance of the observed covariates. Randomization is still useful because it is the only way to balance unobserved covariates. According to Cochran, the improvement in precision “depends primarily on the size of the correlation coefficient ρ between y and x on experimental units that receive the same treatment”^[15]. The actual value of the gain is approximately given by

$$\sigma_v^2(1 - \rho^2) \left\{ 1 + \frac{1}{f_e - 2} \right\} \quad (10)$$

where σ_v^2 is the observed error variance when ANCOVA is not used, f_e is the error degrees of freedom, and ρ is the correlation between x and y .

For nonrandomized experiments, ANCOVA is used to remove bias caused by imbalance in covariates. However, there are underlying assumptions for ANCOVA which should be met in order to draw valid statistical inference. The underlying assumptions are that (1) the treatment and regression effects must be additive, (2) the residuals must be independent and identically distributed following a normal distribution with mean zero, and (3) that the researcher specify the correct regression relationship, i.e., that the true relationship be linear, that it be quadratic, etc.^[15] if it is specified as such in the model. Culpepper and Aguinis^[17] note another assumption which is that (4) the covariate measurements must be taken without error. If there are measurement errors on a covariate, this covariate is referred to as a *fallible covariate*. The presence of fallible

covariates may cause a number of problems. Model diagnostics should be checked to ensure that these assumptions are not violated.

The ANCOVA method does have drawbacks. The primary drawback is that ANCOVA can do nothing to assuage the problems of unmeasured covariates. Randomization may balance unmeasured covariates so that the inference based on ANCOVA will still be valid. However, for nonrandomized studies unmeasured covariates may not be balanced. Strong theoretical expertise of the researcher guiding a study may allow for the identification of important potential unmeasured covariates.

A second drawback arises when different treatment groups have, as a whole, very different values of the measured covariates. In this case, the regression equation involves some degree of extrapolation in the range of covariate values between the various groups. This is problematic because the regression may not appropriately model this region which was unpopulated or sparsely populated in the study sample. The efficiency of the test of the difference in the estimated treatment effects suffers regardless^[15].

A third drawback results from the use of fallible covariates. Fallible covariates don't present difficulties simply in decision making. Culpepper and Aguinis state that "controlling for fallible covariates leads to biased treatment effects," and that the presence of fallible covariates can greatly inflate Type I error rates^[17]. They studied three techniques for handling fallible covariates in addition to OLS when fallible covariates are present: the errors-in-variables (EIV) method, Lord's method, and Raaijmakers and Pieter's (R&P) method. The greatest bias was produced when naïvely proceeding with ANCOVA despite knowledge of fallible covariates, suggesting that if fallible covariates are known or

suspected, a researcher should not proceed with the ANCOVA method. In short, they determined that the EIV method is optimal among these choices based on an investigation of Type I error rates, statistical power, and estimate bias^[17].

ANCOVA has become a ubiquitous method of statistical control and analysis because its implementation is straightforward and simple and it is comparatively easy to interpret the results. And when the assumptions are met, it is an effective tool for reducing bias due to measured confounders. But the drawbacks highlight the need to rely on some theoretical expertise and then proceed with diligence even if the underlying assumptions appear valid.

CHAPTER III

PROPENSITY SCORE METHODS

The use of propensity scores has become popular in observational studies. According to Austin^[18], the propensity score exists for all studies, whether they are randomized or merely observational. The true propensity score is known in a randomized trial; however, it is not known in observational studies and must be estimated from the data. Recall that the propensity score is defined as the conditional probability of an individual receiving treatment given that individual's observed covariates, that is

$$e(x) = \Pr(T = 1|X = x) \quad (11)$$

The propensity scores can be estimated using logistic regression of the following form, as indicated by D'Agostino and Austin^[16, 18]:

$$\log \frac{e(x)}{1 - e(x)} = \log \frac{\Pr(T = 1|X = x)}{1 - \Pr(T = 1|X = x)} = \alpha + X'\beta. \quad (12)$$

Since a subject is in either the treatment group or the control group, we obtain the estimated parameters α and β based on the treatment received and the subject's covariates, and we obtain a propensity score for each subject. We can then use the propensity score in different ways to obtain a valid estimate for the treatment effect. In this section, I outline four popular propensity score based methods: (1) matching based on the propensity score, (2) regression using propensity score as a covariate, (3) stratification based on the propensity score, and (4) inverse probability of treatment weighting (IPTW).

3.1 Matching methods

Matching is a technique for statistical analysis in which subjects from different exposure groups are paired together. Each pair is then treated as though it represents a single trial of the experiment or comparison in question. Subjects can be paired one-to-one, one-to-many, or many-to-many, and with or without replacement. Typically, subjects who are the closest matches for each other based on their covariates are paired, which is known as greedy matching, and each subject is removed from consideration for future pairings, which is matching without replacement. If a single subject is used in multiple pairings, this is matching with replacement. Another method is optimal matching, in which the total distance between covariate values for all matched pairs is minimized.

Greedy matching without replacement: The simplest technique is to match cases with the nearest available control based simply on the propensity score. Once a subject has been matched, he is removed from consideration for future matches. This is greedy matching without replacement based on the propensity score.

Mahalanobis metric matching: A more complex method is Mahalanobis metric matching including the logit of the estimated propensity score as a covariate. Mahalanobis matching involves calculating the distance between a case and all the controls by the following formula:

$$d(i, j) = (u - v)^T C^{-1} (u - v) \quad (13)$$

where $d(i, j)$ is the distance between case i and control subject j , u and v are the values of the variables they are being matched on, and C is the sample covariance matrix of the

matching variables from the full set of controls^[16]. Again, once a case has been matched to the nearest control, both are removed from consideration for future matches.

Mahalanobis metric matching within calipers: A third matching method is Mahalanobis metric matching within calipers based on the propensity score. The first step is to identify which control subjects are within a predetermined difference in the estimated propensity score (the caliper width) from the case to be matched. Mahalanobis distances from the case subject are then calculated for each of these potential control matches, the nearest neighbor is matched, and the two are removed from consideration for future matches. The caliper size can be determined by the investigator, though Rosenbaum and Rubin advise using a caliper that is “a quarter of a standard deviation of the logit of the propensity score.”^[9, 16] Austin suggests that the optimal caliper width (if variance of the logit of the propensity score is the same in treated and control groups) is a fifth of the standard deviation, and that this caliper width will eliminate 99% of the bias due to measured covariates^[18].

Kernel Matching: A fourth matching method, known as Kernel Matching, is described by Becker and Ichino^[19]. Succinctly, all treated subjects are matched to a weighted average of all the control subjects where the weights are inversely proportional to the distance between the propensity score estimates for the treated versus control subjects. The estimator (of average treatment effect among the treated) based on kernel matching is given by

$$\tau^K = \frac{1}{N^T} \sum_{i \in T} \left\{ Y_i^T - \frac{\sum_{j \in C} Y_j^C G\left(\frac{e_j - e_i}{h_n}\right)}{\sum_{k \in C} G\left(\frac{e_k - e_i}{h_n}\right)} \right\} \quad (14)$$

where T is the set of treated subjects, C is the set of control subjects, e_i and e_j are the propensity scores for treatment subject i and control subject j , Y_i^T and Y_j^C are the observed outcomes for treatment subject i and control subject j , N^T is the number of subjects in the treated group, h_n is a bandwidth parameter, and $G(\cdot)$ is a kernel function.

There is some disagreement about how the variance, and thus the analysis, of matched pair designs should be handled. Austin^[18] argues that while some researchers believe matched pairs can be treated as independent observations, he does not believe this is a good approach for propensity score matched samples. His reasoning is that since matched pairs have similar values of the propensity score (by definition), their covariates come from the same distribution, and baseline covariates are also related to outcomes. Austin argues for the use of a paired t-test to analyze continuous outcome data or McNemar's test to analyze dichotomous outcome data^[18].

3.2 Stratification methods

Stratification is in some sense a generalization of matching methods. Instead of cases and controls being matched directly, strata – defined by the covariates to match on – are populated by all subjects, both cases and controls, who happen to fall within each stratum based on the values of their covariates. The number of strata can depend on factors such as the total sample size, whether the cases and controls tend to cluster around specific values of covariates or not, or attempting to balance the number of subjects in each stratum while ensuring that those subjects are also as similar as possible in their covariate

balance. We consider stratification as a generalization of matching because using the maximum possible number of strata (which contain both a case and a control) would essentially result in matched pairs – each pair being its own stratum. For any stratification scheme, the investigator must determine how he wants to determine the stratification boundaries – whether they will be based on the estimated propensity scores for only the cases, for only the controls, or for the combined group of cases and controls^[16].

Stratification methods may still suffer the same problems as matching, namely that a large number of covariates will make it very difficult to find both cases and controls that will fall into the same strata, since the number of strata will grow exponentially with the number of covariates^[16]. However, Rosenbaum and Rubin showed that the propensity score is a balancing score, and it contains information about all the observed covariates that need to be balanced. Thus, stratification based on the propensity score is straightforward and does not suffer from the high dimensionality problem mentioned above. Rosenbaum and Rubin contend that stratification on the propensity score effectively balances all covariates that are used to estimate it, and that using five strata is often sufficient to remove over 90% of the bias in the covariates^[9, 16].

3.3 Regression models using the propensity score

To explain how propensity scores are useful in a regression adjustment (aka covariance adjustment), consider a regression model of the estimation of the treatment effect τ as described by D'Agostino^[16].

$$\hat{\tau} = (\bar{Y}_t - \bar{Y}_c) - (\bar{X}_t - \bar{X}_c) \beta \quad (15)$$

Y represents the response variable. The first term on the right hand side of the equation represents the average treatment effect absent any influence from the covariates, that is, the average difference in response between the treated and control groups. β is the regression estimate of the responses of the two groups on the covariates. The propensity score makes this estimation much simpler because, again, the propensity score is a single variable whereas the set of covariates for each group would likely otherwise be a vector with many variables.

A second method for regression adjustment is to include the propensity score estimates along with a small subset of the most important covariates in the regression model. D'Agostino claims that this method is analogous to the Mahalanobis metric matching method within calipers discussed previously. D'Agostino also warns against using regression adjustment if the variance in the treated group is very different than the variance in the control group, and that it's probably best to consider other methods under this scenario^[2].

Regression techniques must be approached with caution, since an underlying relationship between explanatory variables and response must be assumed, and departures from the assumed relationship can lead to biased results.

3.4 Inverse Probability of Treatment Weighting (IPTW)

IPTW “uses weights based on the estimated propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment. The use of IPTW is similar to the use of survey sampling weights

that are used to weight survey samples so that they are representative of specific populations.”^[18] The weights are defined as follows:

$$w_i = \frac{T_i}{e_i} + \frac{(1 - T_i)}{1 - e_i} \quad (16)$$

where T_i is an indicator variable for treatment versus control for subject i , and e_i is the propensity score for subject i . This weight is equal to the inverse of the probability of receiving the treatment that the subject actually received. If a subject has a very low probability of receiving the treatment he actually received, then the weights may be inaccurate, and a few proposed solutions are mentioned by Austin^[18].

The IPTW method can be used to estimate the treatment effect by the following formula:

$$\frac{1}{n} \sum_{i=1}^n \frac{T_i Y_i}{e_i} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - T_i) Y_i}{1 - e_i} \quad (17)$$

where Y_i is the outcome variable measure for subject i , and n is the number of subjects^[18, 20].

A comparison of these four propensity score methods was conducted by Austin^[18], who concluded that IPTW was essentially equal or inferior to propensity score matching methods. Also, he notes that IPTW and regression adjustment may be more sensitive to whether the propensity score has been accurately estimated as compared to matching or stratification^[18]. Numerous studies have shown that stratification on the propensity score is inferior to matching, and since “increasing the number of strata used should result in improved bias reduction,”^[18] it makes sense that stratification will not perform as well as matching, which essentially creates as many strata as are possible from the data by virtue of

each matched-pair being its own comparison stratum. Additionally, Lunceford and Davidian showed that the stratification method produces more biased estimates than other methods, and that this bias actually increases with sample size^[18, 20]. Becker and Ichino note that, in use on their data set, stratification gave slightly different results than matching techniques, though overall, results were close for all propensity score methods they investigated^[19].

Propensity score methods do have a few drawbacks for the naïve researcher. Senn, Graff, and Caputo^[21] contend that stratifying on the propensity score will produce an unbiased estimate of the treatment effect, but that it will have a greater variance than the OLS estimator in cases where linear regression is an appropriate model. They also state that including covariates in the propensity score estimation which are irrelevant to treatment response increase the variance of the estimator as well^[21]. This finding calls into question the idea that all covariates can be included indiscriminately in the propensity score estimation model and that we need not be worried about over-parameterization in that step of the propensity score analysis as argued by D'Agostino. Senn, et al, admit that some knowledge is required on the part of the analyst as to whether covariates are known to have a stronger relationship with treatment assignment or with outcome, and if the former, then propensity score analysis can be useful, though knowledge is still required about which covariates should then be used to estimate the propensity score. In any case, they conclude that “although popular, propensity score stratification is not to be considered as best practice among propensity score-based approaches”^[21]. Becker and Ichino caution that, for matching within calipers, using small caliper widths will produce stronger matches, but at

the cost of total number of matched pairs, and also at the cost of generalizability of the results to the larger population^[19].

Another issue for propensity score analyses is the presence of missing values for covariates that are used to estimate the propensity score. Matching solves this problem, in a way, in that subjects with missing values can be excluded from matches. The main downside is a reduced sample of matched pairs, as long as missing values are missing at random. Also, clinical trials where subjects drop out before completion present a similar problem. Solutions to these are areas of current research^[16]. Lastly, if the propensity score is going to be used in a regression adjustment model, Rosenbaum and Rubin warn that “covariance adjustment cannot be relied upon to perform well unless the linear discriminant is highly correlated with the propensity score”^[9]. This mirrors the previously mentioned warning issued by D’Agostino suggesting that covariance adjustment is not indicated if the treated and control groups have variances that differ greatly.

Taking these various findings into consideration, this study will focus on the methods of matching and covariance adjustment using the propensity scores.

CHAPTER IV

INSTRUMENTAL VARIABLES

Before discussing the application of instrumental variables (IV) methods, it is helpful to explain two terms typically used in econometrics, *endogeneity* and *exogeneity*, which are central to the theoretical conception of IV analysis. An *endogenous* variable refers to a variable that is correlated with the error term; the term endogenous means “determined within the system,”^[14] and an ordinary least squares (OLS) regression using an endogenous explanatory variable will result in biased estimates since the explanatory variable and the error are related. An *exogenous* variable is one that is not related to the error term. The term exogenous means “determined outside the system.” A regression using an exogenous explanatory variable will capture only the effect that the explanatory variable has on the outcome, and the estimate of this effect will be free of bias.

However, an unbiased estimate of the effect of an exogenous variable is still possible if an instrument (z) can be found^[22]. It is important that changes in z be associated with changes in the endogenous variable (x), but that changes in z itself do not directly effect changes in the outcome (y), except via the indirect route of being associated with changes in x . This is explained in the following path diagram

where u represents the error term. Note that x and u are correlated, and thus OLS will produce biased estimates of the effect of x on y . However, if we can find an instrument z which is related to x but not to the error term (i.e. z is exogenous), we can obtain an unbiased estimate of the effect of x on y ^[22]. This unbiased estimate is obtained by the two-stage least squares (2SLS) method.

To explain how 2SLS is carried out, I adapt the notation given by Brookhart, Rassen, and Schneeweiss for the two stages of regression^[23].

$$X = \alpha_0 + \alpha_1 Z + \alpha_2 C + \varepsilon_1 \quad (18)$$

$$Y = \beta_0 + \beta_1 X + \beta_2 C + \varepsilon_2. \quad (19)$$

Here, the treatment effect of treatment X is denoted in equation (19) by β_1 . C is a vector of measured covariates, ε_1 and ε_2 are error terms, Z is an instrument, and Y is the outcome. In order for $\hat{\beta}_1$ to be an unbiased estimator in ordinary least squares, the treatment exposure must be uncorrelated with the error term ε_2 . If treatment X and error ε_2 are correlated, then the error term is capturing the effect of unobserved confounders and an OLS estimate of β_1 will be biased proportional to the degree of that correlation. Therefore, an instrument Z is found which isolates the effect of X on Y by virtue of being correlated with X but uncorrelated with ε_2 . To do that, first estimate $\hat{\alpha}_0$, $\hat{\alpha}_1$, and $\hat{\alpha}_2$ using OLS regression for equation (18) above. This yields $\hat{X}$, which is plugged in to equation (19) in place of X to obtain $\hat{\beta}_1$ using OLS. This method is generalizable to multiple endogenous regressors,

where a separate first-stage regression is run to find each individual estimator $\hat{X}_i$, where $i = 1, \dots, n$ for n different treatments, and each of these estimates is then plugged into the second stage equation for the outcome^[23].

IV methods are also applicable to dichotomous treatments and instruments, where the IV estimator is known as the Wald estimator and is defined as follows^[23]:

$$\beta_{IV} = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[X|Z = 1] - E[X|Z = 0]}. \quad (20)$$

Here Y is the outcome, X is the treatment, Z is the instrument, and β measures the treatment effect. “The numerator of this estimator is the ITT [intention-to-treat] estimate – i.e., the effect of the instrument on the outcome measures as a risk difference. The denominator is the difference in treatment rates between levels of the instruments...and is a measure of compliance”^[23]. This estimator is more useful than the standard ITT estimator (which equals the numerator of the Wald estimator) when there is significant non-compliance in the treatment groups, which causes bias towards the null in the ITT estimator^[23].

There are two essential assumptions underlying the validity of a potential instrument. The first is that the instrument is correlated with the explanatory variable; that is, $\text{Cov}(Z, X) \neq 0$. This assumption is essential because it is this relationship which allows us to use Z to determine the extent of the effect that X has on the outcome. The second assumption is that the instrument is unrelated to the error term, i.e., $\text{Cov}(Z, \varepsilon_2) = 0$. This is fundamentally the most important aspect for a valid instrument, since this assumption is what allows for the unbiasedness of the resulting parameter estimates. We cannot test the second assumption because we do not have an unbiased estimator of the true error

term^[14], and therefore the selection of an instrument which satisfies the second assumption must be conducted on theoretical principles. The first assumption can be easily tested based on the first stage regression of Z onto X in equation (18). If $\hat{\alpha}_1 = 0$ then the assumption of instrument relevance is violated.

Ideally, we would prefer to use OLS estimation of the treatment effect if possible, and only use IV analysis when we can be sure that our explanatory variables are endogenous and that this is producing significant bias in our estimates. A way to test for endogeneity of the explanatory variables is the Wu-Hausman test^[14]. Consider the following regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \beta_3 W_{2i} + \varepsilon_i \quad (21)$$

where W represents an exogenous regressor which is not an instrument, and consider Z_{1i} and Z_{2i} as instruments for X_{1i} . To check whether an IV analysis is necessary, we can compare the OLS and 2SLS estimates and determine whether the difference between these estimates is significant. If they differ significantly, we conclude that X_1 is an endogenous variable and thus we should proceed with IV methods^[14]. This is known as the Wu-Hausman test, which is conducted in the following way.

Estimate the first stage regression:

$$X_{1i} = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \alpha_3 W_{1i} + \alpha_4 W_{2i} + v_i \quad (22)$$

We know that each instrument (Z) is uncorrelated with ε_i , as are the exogenous regressors (W) by definition, and therefore X_{1i} is uncorrelated with ε_i only if v_i is uncorrelated with ε_i . To test this, we run the following regression using OLS:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \beta_3 W_{2i} + \delta_1 \hat{v}_i + error \quad (23)$$

and test whether $\delta_1 = 0$ using a standard t-test. If we reject the null hypothesis we conclude that X_1 is endogenous, since v_i and ε_i will be correlated^[14]. Since OLS is preferable to IV methods unless strong endogeneity is present, we should conduct this test and proceed with an IV analysis only if indicated.

If we have access to more instruments than endogenous variables, we can choose the best candidate and proceed, or we can use a linear combination of all IVs. The second approach is optimal according to Zulehner^[14]. Additionally, it is nice to have more IVs than endogenous explanatory variables because this situation allows us to test the validity of some of the instruments^[14]. Consider the same regression with outcome Y_i as shown above, where two potential instruments for X are Z_1 and Z_2 . Proceed to do an IV analysis using Z_1 as the instrument for X , and compute the residuals $\hat{\varepsilon}$. Now, if Z_2 is correlated with $\hat{\varepsilon}$ then Z_2 is not a valid instrument. Under the null hypothesis that all IVs are uncorrelated with ε , $nR_1^2 \sim \chi_q^2$ where n is the sample size, R_1^2 is the coefficient of determination from the first-stage regression, and q is the number of instrumental variables from outside the model minus the total number of endogenous explanatory variables. “If the test statistic exceeds the critical value we reject the null hypothesis and conclude that at least some of the IVs are not exogenous”^[14]. This is known as the Sargan test. We cannot know whether Z_1 is valid, because we simply have to assume that it is uncorrelated with $\hat{\varepsilon}$. Thus we can never truly know whether our instruments are valid, since the validity of this test rests on the assumption that Z_1 is indeed uncorrelated with the error term. However, the test can still be useful because if both (or all) instruments were chosen using similar logic, the lack of validity of one of them ought to warn against the validity of the other(s) as well^[14].

Likewise, if the test does not indicate that the extra instruments are invalid, it is at least some evidence of their validity as long as the theoretical reasoning for the exogeneity of Z_1 is solid. In essence, this is a test that can confirm that an instrument or instruments may be invalid, though our only guide for assuming true validity is theoretical reasoning.

The use of IV methods is fraught with difficulties and dangers. Aside from the difficulty of identifying a potential instrument, we must be wary about candidates that have only a weak correlation with the explanatory variable to be instrumented, which in practice encompasses the majority of candidates, unfortunately. These are called weak instruments. It is well recognized that using weak instruments is likely to produce estimates with large standard errors^[24]. Additionally, weak instruments produce estimates that can be highly biased^[14], which obviates the rationale for using IV analysis in the first place.

While the practical difficulty of finding valid instruments is well known, due to the reliance on theoretical knowledge and expertise of the study subject matter; Bound, Jaeger, and Baker^[24] also contend that the identification of suitable IVs may be more difficult than even traditionally thought, since even theoretically sound instruments can turn out to do a poor job. They also warn “that even researchers working with very large data sets need to be more concerned about the finite-sample properties of IV estimators”^[24]. These warnings were issued in their research on a well-known IV analysis conducted by Angrist and Krueger in 1991. Bound, et al, found that many of the findings reported in that seminal work may be dubious despite the use of very large sample sizes.

Bound, et al, also detail another drawback to IV methods, which is that the parameter estimates are actually biased for finite sample sizes. Denoting the IV parameter

estimate by $\hat{\beta}_{iv}$, they say “ $\hat{\beta}_{iv}$ is biased in the direction of the expectation of the OLS estimator of β . The magnitude of this bias depends on both the sample size...and the multiple correlation between the instruments and the endogenous explanatory variable”^[24]. As such, even very large samples will produce biased estimates (though bias does reduce with sample size), and unless Z is perfectly correlated with X (which never happens), using Z to estimate $\hat{X}$ in the first stage regression introduces a bias that is inversely proportional to their degree of correlation. The degree of bias of $\hat{\beta}_{iv}$ relative to $\hat{\beta}_{ols}$ can be estimated by $1/F$ ^[24]. This fact can give a rough guide as to whether an IV analysis provides improved estimates as compared to traditional OLS, at least.

Finally, Bound, Jaeger, and Baker report that the “common practice of adding interaction terms as excluded instruments” may increase the bias of IV parameter estimates, even though this practice reduces the standard error of the estimates^[24].

CHAPTER V

COMPARISONS OF CER METHODS AND SPINAL PROCEDURES

In this section I will compare two competing spinal procedures: Bone morphogenetic proteins (BMP) versus autograft for degenerative disc disease, by using the four common statistical methods discussed previously. Both procedures will be compared on the outcomes of length of stay, cost, and complications. I give a brief description of these procedures and the importance of comparisons between them below, then give some recommendations as far as which CER analysis methods are optimal for conducting these comparisons, and draw final conclusions.

A spinal autograft is a procedure in which a segment of bone is removed from a different section of the patient's body (usually a rib or the iliac crest) and grafted to the spine in order to achieve spinal fusion. Bone morphogenetic proteins are a family of growth factors that stimulate bone formation, and are a recent innovation (FDA approved in 2002) that represents an alternative treatment to autografts for spinal fusion. Since autografts can have a number of drawbacks including morbidity, infection, and chronic pain at the site from which the graft bone was harvested; and may also be eliminated as an option for some patients, such as those requiring very large sections of bone for the graft or those who need repeat surgery and therefore do not have available bone for an autograft, BMP has become an important alternative, representing a quarter of all fusion procedures by 2006. While BMP achieves higher success rates of spinal fusion, it has also been associated with an

increased risk of complications compared to iliac crest autografts. Although BMP is more expensive initially, it is believed that the upfront cost may be more than offset via decreased utilization of health resources in the years after the operation. Findings have been mixed. This is the reason I will investigate this question.

The data for these comparisons comes from MarketScan insurance claims data, which tracks inpatient and outpatient health resource utilization for claims from Medicare, Medicaid, and commercial insurance in the United States. We have information for patient services rendered in the years 2000-2009. The covariate measures include patient age, sex, Charlson comorbidity index (CCI), insurance type, and year of operation. The outcome variables are length of hospital stay (LOS), cost, and complications. ANCOVA and propensity score analyses were applied to compare the two procedures, and use of the county of residence as an instrumental variable for IV analysis was explored.

Before analyzing the effectiveness of the various CER methods in reducing bias, we should know how the covariate balance looks between the groups without any adjustment. Table 1 shows the covariate breakdown between the BMP and autograft comparison groups. Variables with means expressed as percentages represent the proportion of patients who fit that variable description. For example, 63.3% of autograft patients had commercial insurance.

Table 1: Summary statistics for autograft and BMP patients

	Autograft Patients (n=1308)			BMP Patients (n=649)		
	Mean	SD	95% CI	Mean	SD	95% CI
Age	57.34	13.34	(56.73, 58.18)	51.88	12.89	(50.88, 52.87)
Sex (F)	61.31%	--	(58.61, 63.96)	68.41%	--	(64.68, 71.98)
Year	2004.00	1.76	(2003.90, 2004.09)	2004.71	1.34	(2004.61, 2004.82)
Charlson	0.19	0.54	(0.16, 0.22)	0.32	0.65	(0.27, 0.37)
Commercial	63.30%	--	(60.62, 65.92)	54.24%	--	(50.32, 58.12)
Medicaid	4.43%	--	(3.38, 5.69)	36.06%	--	(32.35, 39.88)
Medicare	32.26%	--	(29.73, 34.87)	9.71%	--	(7.54, 12.25)
LOS	4.62	10.55	(4.04, 5.19)	3.96	2.49	(3.76, 4.15)
In-Hosp. Cost	\$30,354	29,919	(28,731, 31,977)	\$40,872	29,337	(38,611, 43,134)
Outp. Cost	\$16,575	21,497	(15,409, 17,741)	\$19,873	19,395	(18,378, 21,368)
Comp. Rate	13.61%	--	(11.80, 15.59)	15.72%	--	(13.00, 18.75)
Total Comps.	0.11	0.41	(0.09, 0.13)	0.14	0.48	(0.10, 0.18)

Confidence intervals for those variables were determined by the exact binomial method.

The first seven variables are the covariate measures, and we can see from Table 1 that the BMP and autograft patients are not at all balanced on the covariates, as the confidence intervals for the covariate values do not overlap at all. The last five variables are outcome measures, and we see that the two treatment groups differ significantly on two of the five outcomes if we don't make any adjustments for covariate imbalance. Since the two treatment groups differ significantly on every covariate measure, we should expect that standard ANOVA will produce biased estimates of the treatment effect of BMP versus autograft if performed on this dataset, as it does not adjust for the covariates. Thus, statistical methods which adjust for covariate imbalance should be applied. In section 5.1, the results for ANCOVA are presented, in section 5.2, propensity score based methods are presented, and in section 5.3, IV analysis using county of residence is explored.

For the comparisons on individual outcome measures, I used one of two statistics to identify overly influential observations and remove them from the comparison. For the length of stay outcome, visual inspection showed one obvious and extreme outlier, which

was confirmed by the DFBETAS statistic produced in SAS^[25]. In short, this statistic measures the change in each parameter estimate, i.e., the β values in the regression, that results from the deletion of the observation in question. For observation i , DFBETAS calculates

$$DFBETAS_j = \frac{b_j - b_{(i)j}}{s_{(i)}\sqrt{(\mathbf{X}'\mathbf{X})_{jj}}}, \quad (24)$$

where b_j is the parameter estimate with observation i included, $b_{(i)j}$ is the parameter estimate with observation i deleted, $s_{(i)}$ is the standard deviation of the estimate with observation i deleted, and $(\mathbf{X}'\mathbf{X})_{jj}$ is the (j, j) th element of $(\mathbf{X}'\mathbf{X})^{-1}$.

For all other outcome measures, I determined overly influential observations by the DFFITS statistic, also produced in SAS. “The DFFITS statistic is a scaled measure of the change in the predicted value for the i th observation and is calculated by deleting the i th observation”^[25]. It is calculated

$$DFFITS = \frac{\hat{y}_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{h_{(i)}}} \quad (25)$$

where $\hat{y}_i$ is the predicted outcome with observation i included, $\hat{y}_{(i)}$ is the predicted outcome with observation i deleted, $s_{(i)}$ is the standard deviation of the estimate with observation i deleted, and $h_{(i)}$ is the i th diagonal of the hat matrix with observation i deleted. According to the SAS User’s Guide, $2\sqrt{p/n}$ is recommended as a cutoff value for identifying influential observations, where p is the number of parameters in the model, and n is the sample size used in fitting the model^[25]. DFFITS with an absolute value greater than this cutoff may be considered overly influential.

5.1 The ANCOVA Method

5.1.1 Results for Length of Stay

To examine the length of stay (LOS), it is assumed that LOS follows a Poisson distribution, and an ANCOVA model with log-link function was applied. In the BMP versus autograft comparison group there were two extreme outliers for length of stay (74 and 367 days). With scaled DFBETA values and DFFITS values that were far, far removed from the values of all other observations, these were removed from the dataset due to their heavy influence on model fit^[25], which left 1955 observations. The results when including all covariates in the model and when including only the significant covariates in the model are shown in Table 2:

Table 2: Full and final ANCOVA model for length of stay for BMP vs. autograft

Full Model:	Estimate	SE	P-value	Final Model:	Estimate	SE	P-value
Intercept	1.44	0.096	<.0001	Intercept	1.44	0.091	<.0001
BMP	-0.02	0.027	0.464				
AGE	0.003	0.001	0.009	AGE	0.003	0.001	0.007
SEX (F)	0.03	0.023	0.204				
YEAR	-0.03	0.007	<.0001	YEAR	-0.03	0.007	<.0001
CHARLSON	0.20	0.016	<.0001	CHARLSON	0.20	0.016	<.0001
COMMERC.	-0.16	0.035	<.0001	COMMERC.	-0.16	0.035	<.0001
MEDICAID	-0.20	0.049	<.0001	MEDICAID	-0.21	0.046	<.0001

Here, and in all future references, the covariates are represented as follows: x_1 is an indicator variable for the BMP procedure, x_2 is a continuous variable for patient age in years, x_3 is an indicator variable for patient sex is female, x_4 is a continuous variable for the year of operation, where 2000 has been subtracted from the year in order to prevent this

variable from having an inordinate impact on the intercept terms in the models (since year of operation in this dataset ranged from 2001 to 2009, these years are represented simply by the numbers 1-9), x_5 is a continuous variable for the Charlson comorbidity index, x_6 is an indicator variable for commercial insurance, and x_7 is an indicator variable for Medicaid insurance. Thus, the final ANCOVA model for length of stay is the following:

$$\text{Log}(LOS) = 1.44 + 0.003(x_2) - 0.03(x_4) + 0.20(x_5) - 0.16(x_6) - 0.21(x_7).$$

The BMP and sex variables were removed from the final model sequentially because they were not significant predictors of length of stay. The interpretation of this model is as follows. Consider the typical patient in this dataset: she is 56 years old, has the operation in 2004, has a Charlson score of 0, and has commercial insurance. This patient has an expected stay of 3.77 days. Each 1-year increase in age increases expected length of stay by 0.3%, each 1-year increase in year of operation decreases expected length of stay by 3.0%, each 1-point increase in the Charlson comorbidity score increases expected length of stay by 22.1%, commercial insurance decreases length of stay by 17.4% over Medicare insurance, and Medicaid decreases expected length of stay by 23.4% over Medicare insurance, when all else is held equal. In this model, the BMP procedure is not significant, meaning there is no difference in expected length of stay between BMP and autograft patients.

5.1.2 Results for In-Hospital Cost

The second outcome variable for the comparison of spinal procedures is the cost of the procedure. This cost is also broken down into initial in-hospital cost and 1-year follow-up outpatient costs. Since cost cannot be negative, a common practice is to log transform

the cost variable before fitting a linear model, because the transformed cost variable more closely follows a normal distribution. I will first investigate inpatient costs. This is a measure of payments made to the hospital and the principal physician for the inpatient admission associated with the procedure. There were 62 outliers for inpatient cost in the dataset, with 14 autograft and 48 BMP patients removed, due to DFFITS statistics with absolute values greater than $2\sqrt{p/n}$, which is 0.1196 for this sample. The ANCOVA approach yields the following model for inpatient cost:

Table 3: Full and final model for inpatient cost for BMP vs. autograft

Full Model:	Estimate	SE	P-value	Final Model:	Estimate	SE	P-value
Intercept	9.42	0.188	<.0001	Intercept	9.40	0.045	<.0001
BMP	0.46	0.054	<.0001	BMP	0.45	0.053	<.0001
AGE	0.0007	0.002	0.763				
SEX (F)	0.009	0.047	0.851				
YEAR	-0.02	0.014	0.094				
CHARLSON	0.06	0.042	0.126				
COMMERC.	0.84	0.073	<.0001	COMMERC.	0.82	0.053	<.0001
MEDICAID	-0.16	0.106	0.130	MEDICAID	-0.17	0.085	0.046

$$\text{Log}(\text{Initial Cost}) = 9.40 + 0.45(x_1) + 0.82(x_6) - 0.17(x_7)$$

Sex, age, Charlson, and year of operation variables were removed sequentially from the model due to lack of significance at the $\alpha=0.05$ level, though Charlson and year were close to this cutoff. Once taking covariates into account, the BMP procedure is more expensive than autograft in initial cost by 56.83%, which aligns with the results of previous findings as well as theory. For a patient with commercial insurance, BMP would cost

\$43,045 for the procedure and associated inpatient stay, while autograft would cost only \$27,447 according to this model.

5.1.3 Results for 1-Year Follow-up Outpatient Cost

Of the 1957 patients to receive either the BMP or autograft procedure in our dataset, 1815 of them combined for 347,030 total outpatient services for which they were billed in the one-year period following their operations, while 142 did not incur outpatient costs at all over the 1-year follow-up period. These 142 patients would all be regarded as influential observations by the DFFITS statistic, but I have not removed them. Therefore, all 1957 patients remained in the sample, and the ANCOVA method yields:

Table 4: Full and final model for 1-year outpatient cost for BMP vs. autograft

Full Model:	Estimate	SE	P-value	Final Model:	Estimate	SE	P-value
Intercept	7.14	0.793	<.0001	Intercept	6.70	0.750	<.0001
BMP	0.57	0.234	0.015	BMP	0.58	0.231	0.012
AGE	0.02	0.010	0.066	AGE	0.02	0.010	0.026
SEX (F)	0.25	0.205	0.214				
YEAR	-0.04	0.059	0.479				
CHARLSON	0.28	0.173	0.106				
COMMERC.	-0.14	0.318	0.658	COMMERC.	-0.11	0.317	0.724
MEDICAID	0.97	0.433	0.025	MEDICAID	1.09	0.427	0.011

$$\text{Log}(\text{Outpatient Cost}) = 6.70 + 0.58(x_1) + 0.02(x_2) - 0.11(x_6) + 1.09(x_7)$$

The ANCOVA model shows that the BMP procedure is more expensive than autograft for 1-year outpatient costs as well, countering the theory and previous findings that BMP is less expensive in terms of follow-up care. However, due to the possibility of patients requiring re-operations, it is possible that the BMP procedure will be less expensive

in total cost over longer follow-up timeframes compared to autograft if autograft patients have a higher rate of requiring re-operations, since re-operations would not show up in initial cost or in outpatient costs.

5.1.4 Results for Complication Rate

Complications were measured over 30 days and 90 days, and in terms of rate of complications and total number of complications. I will examine the 30-day and 90-day rate of complications for BMP versus autograft patients using logistic regression since the outcome is binary. ANCOVA produces the following full and final models:

Table 5: Full and final model for 30-day complication rate for BMP vs. autograft

Full Model:	Estimate	SE	P-value	Final Model:	Estimate	SE	P-value
Intercept	-2.84	0.450	<.0001	Intercept	-2.65	0.296	<.0001
BMP	0.04	0.079	0.649				
AGE	0.02	0.007	0.018	AGE	0.01	0.005	0.013
SEX (F)	-0.01	0.069	0.875				
YEAR	0.003	0.041	0.941				
CHARLSON	0.46	0.094	<.0001	CHARLSON	0.50	0.091	<.0001
COMMERC.	-0.07	0.101	0.484				
MEDICAID	0.22	0.144	0.134				

$$\text{Log}[\text{Pr}(30 \text{ Day Complications})] = -2.65 + 0.01(x_2) + 0.50(x_5),$$

This model indicates that each 1-unit increase in a patient's age increases the chance of complications by 1%, and each 1-unit increase in the Charlson score increases the chance of complications by 64.87% within the 30 day timeframe, for fixed values of the other variable. For example, a 56 year old patient with a Charlson score of 0 has a 12.37% chance of experiencing complications within 30 days, according to the model, while a 56 year old

patient with a Charlson score of 1 has 20.39% chance of experiencing complications, which is nearly 65% greater than the chances of a patient with a Charlson score of 0. The BMP indicator is not a significant estimator of complication rate in the full model, which suggests that BMP and autograft patients do not differ in 30-day rate of complications.

The rate of complications over a 90-day follow-up period shows nearly identical results to the 30-day period. The procedure variable is not significant in either timeframe, and only age and Charlson score are significant in the ANCOVA model, with identical parameter estimates and a slightly different intercept. Again, the conclusion is that BMP patients do not differ from autograft patients in the rate of complications over 90 days post surgery.

5.1.5 Results for Total Number of Complications

Though the BMP and autograft treatment groups do not differ in the rate of complications post surgery, perhaps they do differ in the total number of complications they experience. We were able to extract information for the following types of post-surgery complications: renal, cardiac, neurological, deep vein thrombosis or pulmonary embolism, pulmonary complications, infection, or wound complications. This means that any given patient could experience a maximum of 7 different complications as a result of the procedure. Since complications are a count variable, I assume an underlying Poisson distribution and use a log-link function in a general linear model. In the 30-day follow-up period, adjusting for covariates yields the following ANCOVA model:

Table 6: Full and final model for total number of complications for BMP vs. autograft

Full Model:	Estimate	SE	P-value	Final Model:	Estimate	SE	P-value
Intercept	-3.41	0.548	<.0001	Intercept	-2.66	0.189	<.0001
BMP	0.07	0.157	0.647				
AGE	0.01	0.007	0.110				
SEX (F)	0.02	0.141	0.887				
YEAR	0.06	0.040	0.131	YEAR	0.08	0.038	0.047
CHARLSON	0.50	0.066	<.0001	CHARLSON	0.54	0.061	<.0001
COMMERC.	0.18	0.213	0.389				
MEDICAID	0.33	0.268	0.224				

$$\text{Log}(\text{Total Comps}) = -2.66 + 0.08(x_4) + 0.54(x_5)$$

This model suggests that only a patient's Charlson comorbidity index and the year of operation significantly affect the total number of complications in the 30 days post surgery. Age was nearly significant, but not at the $\alpha=0.05$ level. Interestingly, this model suggests that procedures performed in more recent years have higher numbers of complications over a 30-day follow-up than procedures performed in the earlier years of the sample.

The 90-day follow-up period again provides very similar results. The same variables are significant for the final model as they were for the 30-day follow-up, and the parameter estimates are also very similar. All the findings mentioned for complication rate and total number of complications suggest it doesn't matter whether a 30-day or a 90-day follow-up period is chosen as far as discerning the differences between the BMP and autograft treatments.

5.2 Matching on the Estimated Propensity Scores

Recall that BMP versus autograft patients were significantly different on all five measured covariates in the study: age, sex, year of operation, Charlson comorbidity score, and insurance type. Before examining the results of propensity score methods, let's first examine the extent to which propensity score matching is useful in reducing imbalance on the measured covariates in the dataset. I conducted an optimal one-to-one matching of BMP patients with autograft patients using the *vmatch* macro^[26] from the Mayo Clinic. According to its authors, "The *vmatch* macro matches each of N cases with a minimum of "a" controls to a maximum of "b" controls from a total pool of M controls. The *vmatch* macro uses the case-control "distance matrix" as input. This matrix has one row per case and one column per potential control. Each cell entry is the distance, D_{ij} , between the i-th case and the j-th potential control. Output includes the assignments of cases to controls and summaries of the matching efficacy"^[26]. *vmatch* requires the use of the *nobs* macro^[27] and the *dist* macro^[28] in order to run. Using a caliper distance of 1/5 the standard deviation of the estimated propensity scores as recommended by Austin^[18], optimal one-to-one matching resulted in 463 of the 649 BMP patients being matched to autograft patients as controls. The covariate breakdown after matching is described in the Table 7 below.

Table 7: BMP vs. autograft covariate balance after matching on propensity score

	Autograft Patients (n=463)			BMP Patients (n=463)		
	Mean	SD	95% CI	Mean	SD	95% CI
Age	54.33	11.63	(53.27, 55.39)	53.32	12.34	(52.20, 54.45)
Sex (F)	64.36%	--	(59.81, 68.73)	66.31%	--	(61.80, 70.60)
Year	2004.65	1.30	(2004.53, 2004.77)	2004.63	1.25	(2004.52, 2004.75)
Charlson	0.23	0.55	(0.18, 0.28)	0.26	0.60	(0.20, 0.31)
Commercial	75.81%	--	(71.64, 79.64)	76.03%	--	(71.87, 79.85)
Medicaid	10.58%	--	(7.93, 13.75)	10.37%	--	(7.74, 13.51)
Medicare	13.61%	--	(10.62, 17.07)	13.61%	--	(10.62, 17.07)

When comparing to the covariate breakdown given at the beginning of section 5, we can see that one-to-one matching on the propensity score has eliminated nearly all of the imbalance in observed covariates between the two groups. There is a great deal of overlap in the confidence intervals for all of the observed covariates now, whereas before matching on the propensity score there was no overlap in any of them.

Heeding the argument by Austin^[18] that matched pairs should not be considered independent, comparisons between the two treatment groups will be conducted using a paired t-test for continuous outcome variables or McNemar's test for dichotomous outcome variables.

Results for Matching on the Estimated Propensity Scores

Table 8 summarizes the results of matched pair analysis based on the estimated propensity scores for all of the outcome variables of interest.

Table 8: Results for all outcomes based on matched samples

	Autograft Patients		BMP Patients		P-value for Diff in Means
	Mean	SE	Mean	SE	
Length of Stay	4.74	17.00	3.98	2.56	0.332
In-Hospital Cost	35,809	30,401	43,505	28,217	<.0001
1-Year Outp. Cost	21,132	29,157	19,947	19,061	0.432
30-Day Comp. Rate	13.17%	--	14.69%	--	0.518
90-Day Comp. Rate	13.61%	--	16.41%	--	0.245
30-Day Total Comps.	0.11	0.38	0.13	0.46	0.593
90-Day Total Comps.	0.12	0.39	0.14	0.47	0.457

For the length of stay outcome, a paired t-test shows that the difference in means is 0.77 days, with BMP having the shorter average stay, with a 95% confidence interval of (-0.78, 2.32). It is of note that the autograft patient with an in-hospital length of stay of 367 days was matched to a BMP patient based on their covariate measures, and virtually the

entire difference in means is due to that one extreme outlier. If this patient had not been matched, the 462 other matched pairs would have combined for 1835 total days among BMP patients and 1831 total days among autograft patients, which would make the difference in their means virtually zero. Even with this observation in the sample, the difference in means is not significant.

The ANCOVA method showed that the BMP procedure is not a significant predictor of length of stay. The propensity score matching method mirrored that finding, and excluding the outlier patient, the difference in means from the paired t-test followed a normal distribution extremely closely, which is the primary assumption for the validity of the paired t-test.

For in-hospital cost, a paired t-test shows the difference in mean in-hospital cost to be \$7,697, with autograft patients having a lower mean cost, with a 95% confidence interval of (3,941, 11,452), so the result is significant. The difference in means approximates a normal distribution rather well. Again, this agrees with the ANCOVA result that showed the BMP procedure to be a significant factor in in-hospital cost.

The difference in mean 1-year outpatient cost post surgery is \$1,185, with BMP patients having a lower mean cost, and a 95% confidence interval of (-1,776, 4,146), meaning the result is not significant. This is in stark contrast to the results obtained via the ANCOVA method, which had BMP patients with a significantly higher outpatient cost over a 1-year follow-up period post surgery.

For 30-day complication rate, McNemar's test returns a statistic value of 0.4188, which from a chi-square distribution with 1 degree of freedom, has a p-value of 0.52 for

values this extreme or more extreme. Thus, BMP does not differ significantly from autograft in 30-day complication rate. This can also be seen from the odds ratio of 1.13 for autograft versus BMP patients, with a 95% confidence interval of (0.78, 1.65). These findings match the ANCOVA results that show that BMP does not differ significantly from autograft in 30-day complication rate.

As before, similar results are found for the 90-day complication rate. McNemar's test returns a statistic value of 1.35. From a chi-square distribution with 1 degree of freedom, this has a p-value of 0.24 for values this extreme or more extreme. The odds ratio for autograft versus BMP patients is 1.25, with a 95% confidence interval of (0.87, 1.79), and thus BMP does not significantly differ from autograft in 90-day complication rate.

The mean difference in total number of complications for 30-day follow-up is 0.015, with autograft patients having fewer complications on average, with a 95% confidence interval of (-0.04, 0.07). Thus, there is not a significant difference in total complications 30 days post surgery between BMP and autograft patients. The mean difference in this case follows a normal distribution with very small variance.

The 90-day follow-up period shows similar results: the mean difference in total complications is 0.022 with autograft patients having fewer complications, with a 95% confidence interval of (-0.04, 0.08). Again, the mean difference follows a normal distribution with small variance.

5.3 Using Propensity Score as a Covariate in Regression

This method is analogous to the ANCOVA method presented in section 5.1, but now the estimated propensity score replaces the vector of the five individual covariates in the

regression model. As before, outliers were removed based on DFBETAS or DFFITS statistics before running the regressions.

5.3.1 Results for Length of Stay

Two outliers were removed as their DFBETAS statistics values were more than double any of the other observations in the sample, leaving 1955 patients in the comparison. The model shows that the propensity score is highly significant (p-value < .0001) but the BMP procedure indicator variable is not significant (p-value = 0.45). Thus we would conclude that covariate measures do affect length of stay but that BMP does not differ from autograft in expected length of stay. This result is in agreement with both the ANCOVA method and the method of matching on the propensity score.

5.3.2 Results for In-Hospital Cost

76 outliers for in-hospital cost were removed due to a DFFITS statistic with an absolute value greater than 0.0639, leaving 1881 patients in the sample. The model shows that the BMP variable is highly significant (p-value < .0001), and that the expected cost for BMP patients is 53.72% more than for autograft. The propensity score was not significant in this model, indicating that the covariates do not significantly influence inpatient cost. Again, this matches the results of both the ANCOVA and the propensity score matching methods suggesting that BMP does differ from autograft significantly in in-hospital cost, with BMP being more expensive.

5.3.3 Results for 1-Year Follow-up Outpatient Cost

For outpatient cost, as with ANCOVA, there were 142 patients with no outpatient costs at all over the 1-year follow-up, and all would have been labeled outliers by the DFFITS

statistic, but I did not remove them. As a result, all 1957 patients remained in the sample. The model shows that both BMP (p-value = 0.01) and the propensity score (p-value = 0.03) have a significant effect on outpatient cost, with BMP patients having expected costs 78% greater than autograft patients. This finding is in agreement with the ANCOVA results, but in stark contrast to the propensity score matching method, which found no effect for BMP versus autograft on 1-year outpatient cost.

5.3.4 Results for Complication Rate

For the 30-day complication rate, neither BMP nor the propensity score is a significant predictor of development of complications post surgery. This is also true for the 90-day complication rate. Both of these findings are in agreement with the ANCOVA and propensity score matching methods which suggest that BMP does not differ significantly from autograft in rate of complications.

5.3.5 Results for Total Number of Complications

For the total number of complications 30 days post surgery, BMP is not a significant predictor of the number of complications, but the propensity score is (p-value = 0.007). The 90-day total number of complications has the same results. ANCOVA and propensity score matching methods also showed that BMP does not differ significantly from autograft in terms of the total number of complications.

5.4 Instrumental Variables Method

In previous exploratory analyses on this dataset, it was determined that patient county of employment is the only potentially suitable instrumental variable. Many other variables from the dataset were investigated, but ruled out. By way of example, physician

ID and hospital location are two variables that are commonly considered as potential IVs in medical observational studies; however, they are unsuitable for this study because in this dataset they have missing values, and the values are not missing at random. As such, using them as IVs would introduce bias. Therefore, this section will focus on the suitability of county of employment as an IV for the outcomes being studied.

5.4.1 Testing for Endogeneity of the BMP Procedure Variable

Recall that unless the explanatory variable is indeed correlated with the error term, we would prefer to use OLS estimation. Thus, the first step when considering an IV approach is to test this condition with the Wu-Hausman test. However, the Wu-Hausman test itself relies upon OLS, and is not fit to handle many-leveled categorical variables for either the potentially endogenous regressor or for the IV which will be used in its place. If both the endogenous variable and the instrument are binary, we have the Wald estimator and we can proceed. If only the endogenous variable is categorical with many levels, Lochner and Moretti^[29] have proposed a distinct test of exogeneity. But the case where the potential instrument is a many-leveled categorical variable requires further study. Some researchers proceed with the Wu-Hausman test even when either the endogenous or instrumental variable is categorical (usually binary), but this is technically not the proper purview of the Wu-Hausman test.

The difficulty of empirically determining the endogeneity of the BMP variable in this case is probably an argument against proceeding with an IV analysis, unless there is a strong theoretical reason to believe that the BMP procedure variable is endogenous. If it is not, IV

analysis is certain to be far more detrimental than helpful, given the results to be presented in the next section.

5.4.2 Testing Instrument Relevance

Since both BMP and employee county are categorical variables, we cannot take a measure of correlation, but there are a number of options for assessing the degree of association between these variables. Since the relationship between BMP and employee county cannot naturally be classified in terms of one variable “explaining” the other, I prefer to measure association between them with a *reduction in error* approach. In short, this approach attempts to measure how much knowledge one variable gives us in trying to predict the value of another variable. Symmetric λ is one statistic which uses this approach. This statistic is calculated by

$$\lambda = \frac{\sum_i r_i + \sum_j c_j - r - c}{2n - r - c}. \quad (26)$$

Consider a $2 \times J$ contingency table with BMP procedure status composing the rows $i = 1, 2$, and employee county values composing the columns $j = 1, \dots, J$, and n_{ij} is the observed count in the cell for row i and column j . Then

$$r_i = \max_j(n_{ij}),$$

$$c_j = \max_i(n_{ij}),$$

$$r = \max_j(n_{.j}),$$

$$c = \max_i(n_{i.})^{[30]}.$$

Symmetric λ will have values between 0 and 1, and it represents the reduction in error rate a prognosticator would experience by having knowledge of one variable in trying to predict the other variable. If the variables have no association whatsoever, a prognosticator will

not see any reduction in the error rate of his guesses, and $\lambda = 0$. Perfect association between the variables gives $\lambda = 1$, since the prognosticator would see 100% reduction in his rate of errors by knowing the value of one variable before trying to predict the value of the other^[31].

The symmetric lambda statistic value for association between BMP and employee county is 0.09 with a 95% confidence interval of (0.08, 0.11). Knowing the value of BMP would reduce errors in guessing employee county (or vice versa) by about 9.4% over guessing without this knowledge. This is a significant result, meaning that employee county passes the assumption of instrument relevance, though it would likely be considered a weak instrument, since this level of association is not particularly strong. As mentioned, it is rare in practice for instruments to be very strong, but with a sample of only 1663 (since employee county is missing for 294 patients in the dataset), the weakness of this instrument raises quite a red flag about proceeding with an IV analysis. Bound, Jaeger, and Baker cast doubt upon the findings of a classic Angrist and Krueger study (1991) in part due to using a weak instrument, and their sample size exceeded 300,000^[24]. Recall that the bias of an IV estimate is indirectly proportional to its correlation with the endogenous variable. They also warn that if the instrument is only weakly correlated with the endogenous variable, this can greatly exacerbate the problems that are created if the instrument is actually correlated with the error term as well^[24].

5.4.3 Hypothesizing Instrument Exogeneity

As mentioned in section 4 of this study, the exogeneity of a chosen instrument cannot be tested empirically, as we do not have an unbiased estimate of the true error. The

exogeneity of the instrument must therefore be assumed to be true based on theoretical grounds. It seems reasonable to assume that a patient's county of employment would have no effect whatsoever on his response to either the BMP or autograft procedure, and thus it seems reasonable to assume that county of employment is exogenous. We must keep in mind though, that even a very slight correlation between employee county and the error term would create highly biased estimates of the treatment effect, due to the instrument's weak correlation with the BMP procedure variable discussed above^[24].

5.5 Recommendations

Because each outcome variable represents an independent analysis, I will address each outcome individually. For the length of stay, regression and matching assumptions are all satisfied, and all three methods agreed, so choice of method is simply a matter of preference.

For in-hospital cost, the final model contains only categorical predictors in the ANCOVA method, so linearity of the relationship is not a concern. The error appears more or less normally distributed (see Figure 1 Panel A) and with constant variance (see Figure 1 Panel B), though the model fits more poorly in the region of patients with very low inpatient cost (see Figure 2). If the dataset were smaller, ANCOVA may be preferable to PS matching if matching resulted in only a small number of pairs. Overall, ANCOVA appears valid here, and all three methods produced the same conclusions. PS matching also appears valid here, as the difference in means between the two treatment groups follows a normal distribution pretty closely.

For 1-year outpatient cost, we had the odd circumstance of 142 of the 1957 patients having no outpatient costs in the 1-year follow-up, which I chose not to discard. A researcher could choose to simply discard all of these observations and proceed with ANCOVA or PS regression. However, the sample was composed of 67% autograft patients, yet autograft patients accounted for 79% of the zero outpatient cost observations. To throw these observations out is to discard pertinent information, in my opinion – autograft patients are more likely to avoid outpatient costs post surgery. If these observations remain in the sample, ANCOVA assumptions are violated (see Figure 3). Thus, if a researcher deems the outlier observations to contain pertinent information, he should proceed with the PS matching method, or another method which does not make assumptions about linearity or about normality of the error term. Unsurprisingly, ANCOVA and PS regression produced the same conclusion – that BMP differed significantly from autograft – but PS matching produced an entirely different result. The PS matching result is far more likely to be valid. Considering that the zero counts produced some outliers for the difference in cost, the difference still follows a normal distribution reasonably well.

For complication rate, model assumptions for ANCOVA methods appear to be satisfied. In the matched pairs design there were no sparse cell counts, so McNemar's test is valid here as well, allowing for PS matching. All three methods agreed with each other.

For total number of complications, linearity of relationship, normality of error, and constancy of error variance appear to be satisfied (see Figure 4 Panel A). However, the error term does not appear to be independent of the predictor (see Figure 4 Panel B). As a result, ANCOVA assumptions are violated. The difference in means of matched pairs follows

a normal distribution (with large variance), so PS matching is valid. The results of all three methods were in agreement, but PS matching is indicated in this case as the optimal choice.

CHAPTER VI

DISCUSSION

ANCOVA has its drawbacks, which is why propensity scores, IV methods, and numerous other methods have been developed in the first place. Primarily, ANCOVA produces biased estimates of the treatment effect when important covariates are omitted from the model. This is where IV methods find their application. Secondly, covariate adjustment is ineffectual when the covariates are so numerous that multiple experimental units (from each of the treatment assignments) cannot be found with the same covariate measures. This is where propensity score methods are useful. Also, as mentioned, ANCOVA depends on proper specification of the regression relationship, but this is true of all statistical models, with varying degrees of severity when this condition is violated.

Propensity score analyses do have drawbacks. For example, Senn, Graf, and Caputo^[21] find that stratifying on the propensity score will produce estimates with greater variance than least squares regression under the circumstances where both methods are unbiased, indicating that propensity score analyses should not be used simply for convenience or simplicity's sake, but only when there is strong risk of bias due to confounding. They also warn that including covariates irrelevant to the response (which researchers may wish to do if the covariates are believed to be related to assignment) reduces efficiency further. That is to say, the exemption from over-parameterization does not come without a cost. Winkelmayr and Kurth^[32] caution that "even though [propensity

scores] can balance observed baseline covariates between exposure groups, they do nothing to balance unmeasured characteristics and confounders,” (which is the role of the instrumental variables method). “Secondly, one cannot use covariates that may be affected by the exposure of interest in the model that estimates the PS.” This indicates a need for some theoretical expertise about which covariates could be potentially troublesome if included in a model estimating the propensity scores.

The drawbacks to IV analysis are rather severe. First, the selection of a suitable instrument is a difficult task that relies heavily on theoretical reasoning, and the suitability of an instrument is difficult or impossible to test empirically. Even instruments that seem highly intuitive candidates can be fatally flawed in reality. Second, as stated by Brookhart, Rassen, and Schneeweiss^[23], even when valid, “IV methods are inefficient and should not be used as a primary analysis unless unmeasured confounding is thought to be strong.” Both of these drawbacks highlight the strong level of theoretical expertise required to even consider using IV methods, first simply for the recognition of the need for IV methods and second for the identification of potential instruments. Third, IV analysis is consistent only in large samples.

Because IV analysis is useful (superior to OLS) only when unmeasured confounding is strong, and is also highly sensitive to the validity of the chosen instrument(s) along with other drawbacks, it behooves the researcher to use IV analysis only when necessary. The Wu-Hausman test is one method for determining whether the treatment is correlated with the error term by comparing OLS estimates to 2SLS estimates. If we have more instruments than we do treatment effects to estimate, we can use the Sargan test to check whether any

of the “excess” instruments are exogenous (unrelated to the error term), though this test requires some strong assumptions^[14]. Additionally, further research is needed for situations where the potential instrument is a many-leveled non-ordinal categorical variable.

Unless confounding is thought to be strong due to unmeasured covariates, or the sample size is small, propensity score matching seems to be the safest choice for dealing with covariate imbalance between treatment groups. Its assumptions are easily tested, and its drawbacks are rather mild. ANCOVA and propensity score regression techniques have stricter assumptions and some not so easily tested, though they may be preferable if the sample size is small and only few matched pairs can be created by matching techniques. IV methods should only be used when there is strong suspicion of endogeneity of explanatory variables, and then only provided that a valid instrument can be found. It requires very strong assumptions which may not be empirically testable at all. The drawbacks to this method are too severe for use in situations where it is not absolutely imperative to eliminate bias created by unmeasured confounders

REFERENCES

1. National Research Council. (2009). What is comparative effectiveness research?. *Initial National Priorities for Comparative Effectiveness Research*, 29-60. Washington DC: The National Academies Press. Available at http://books.nap.edu/openbook.php?record_id=12648&page=29
2. Sox HC, Greenfield S. (2009). Comparative effectiveness research: a report from the Institute of Medicine. *Annals of Internal Medicine* 151(3), 203-205. doi: 10.7326/0003-4819-151-3-200908040-00125. Retrieved from <http://annals.org/article.aspx?articleid=744633>
3. Sox HC. (2010). Defining comparative effectiveness research: the importance of getting it right. *Medical Care* 48(6), S7-S8. doi: 10.1097/MLR.0b013e3181da3709. Retrieved from [http://journals.lww.com/lww-medicalcare/Fulltext/2010/06001/Defining Comparative Effectiveness Research The.3.aspx](http://journals.lww.com/lww-medicalcare/Fulltext/2010/06001/Defining_Comparative_Effectiveness_Research_The.3.aspx)
4. Agency for Healthcare Research and Quality. What Is comparative effectiveness research. Retrieved from <http://effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/> March 16, 2014
5. Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. (2009). Good research practices for comparative effectiveness research: Analytic methods to improve causal inference for nonrandomized studies of treatment effects using secondary data sources: The ISPOR good research practices for retrospective database analysis task force report—Part III. *Value in Health* 12(8), 1062-1073. Retrieved from <http://www.ispor.org/taskforces/documents/RDPartIII.pdf>
6. Luce BR, Kramer JM, Goodman SN, Connor JT, Tunis S, Whicher D, Schwartz JS. (2009). Rethinking randomized clinical trials for comparative effectiveness research: The need for transformational change. *Annals of Internal Medicine* 151(3), 206-209. doi:

10.7326/0003-4819-151-3-200908040-00126. Retrieved from
<http://annals.org/article.aspx?articleid=744634>

7. Horn SD, Gassaway J. (2007). Practice-based evidence study design for comparative effectiveness research. *Medical Care* 45(10), S50-S57. Retrieved from
http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CDYQFjAB&url=http%3A%2F%2Fwww.researchgate.net%2Fpublication%2F5936418_Practice-based_evidence_study_design_for_comparative_effectiveness_research%2Ffile%2F0b49518b0eebc58ac.pdf&ei=360pU7zENsnCyAHgg4DgDQ&usg=AFQjCNHQPntgsxhJfOlKv-6kH4TuDEp-Xw&sig2=-pE7M-VCMTjiP9qUNfQOAg&bvm=bv.62922401,d.aWc&cad=rja
8. Rubin DB. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 100(469), 322-331. Retrieved from <http://www.stat.cmu.edu/~fienberg/Rubin/Rubin-JASA-05.pdf>
9. Rosenbaum PR, Rubin DB. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41-55. Retrieved from
<http://www.jstor.org/discover/10.2307/2335942?uid=3739664&uid=2129&uid=2134&uid=2&uid=70&uid=4&uid=3739256&sid=21102597812051>
10. Little RJ, Rubin DB. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. *Annual Review of Public Health* 21(1), 121-145. Retrieved from
http://www.bios.unc.edu/~mhudgens/sismid2013/Little_Rubin_2000.pdf
11. Hirano K, Imbens GW, Ridder G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4), 1161-1189. Retrieved from
http://scholar.harvard.edu/imbens/files/efficient_estimation_of_average_treatment_effects_using_the_estimated_propensity_score.pdf
12. Angrist JD, Pischke JS. (2009). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton, New Jersey.
13. Imbens GW. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86(1), 4-29. Retrieved from
<http://www.ics.uci.edu/~sternh/courses/265/imbensreview1994.pdf>

14. Zulehner C. Instrumental variables regression. Retrieved from <http://homepage.univie.ac.at/Christine.Zulehner/iv.pdf>
15. Cochran WG. (1957). Analysis of covariance: Its nature and uses. *Biometrics* 13(3), 261-281. Retrieved from <http://personal.psc.isr.umich.edu/yuxie-web/files/pubs/Articles/Cochran1957.pdf>
16. D'Agostino RB. (1998). Propensity score methods for bias reduction in the comparison of a treatment group to a non-randomized control group. *Statistics in Medicine* 17, 2265-2281. Retrieved from <http://www.stat.ubc.ca/~john/papers/DAgostinoSIM1998.pdf>
17. Culpepper SA, Aguinis H. Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods* 16(2), 166-178. doi: 10.1037/a0023355. Retrieved from [http://www.researchgate.net/publication/51074175_Using_analysis_of_covariance_\(ANCOVA\)_with_fallible_covariates/file/d912f5123810a431ec.pdf](http://www.researchgate.net/publication/51074175_Using_analysis_of_covariance_(ANCOVA)_with_fallible_covariates/file/d912f5123810a431ec.pdf)
18. Austin PC. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46(3), 399-424. doi: 10.1080/00273171.2011.568786. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3144483/>
19. Becker SO, Ichino A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal* 2(4), 358-377. Retrieved from <http://www.stata-journal.com/article.html?article=st0026>
20. Lunceford JK, Davidian M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 23, 2937-2960. Retrieved from <http://www4.stat.ncsu.edu/~davidian/statinmed.pdf>
21. Senn S, Graf E, Caputo A. (2007). Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in Medicine* 26(30), 5529-5544. doi: 10.1002/sim.3133. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/sim.3133/abstract>
22. Cameron CA. (2008). *ch04iv* [lecture notes from ECON 240 at UC Davis]. Retrieved from <http://cameron.econ.ucdavis.edu/e240a/ch04iv.pdf>

23. Brookhart MA, Rassen JA, Schneeweiss S. (2010). Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety* 19, 537-554. doi: 10.1002/pds.1908. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2886161/>
24. Bound J, Jaeger DA, Baker RM. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443-450. Retrieved from <http://www.jstor.org/stable/2291055>
25. SAS Institute Inc. (2014). Details: REG Procedure – Influence Statistics. Retrieved from http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_reg_sect040.htm
26. Bergstralh E, Kosanke J. (2003). *vmatch* [SAS macro]. Retrieved from <http://www.math.smith.edu/sasr/examples.php>
27. Author unknown, property of Mayo Clinic College of Medicine. (2004). *nobs* [SAS macro]. Retrieved from <http://www.math.smith.edu/sasr/examples.php>
28. Bergstralh E, Kosanke J. (2003). *dist* [SAS macro]. Retrieved from <http://www.math.smith.edu/sasr/examples.php>
29. Lochner L, Moretti E. (2014). Estimating and testing models with many treatment levels and limited instruments. *The National Bureau of Economic Research* [working paper]. Retrieved from <http://emlab.berkeley.edu/~moretti/nonlinearities.pdf>
30. SAS Institute Inc. (2014). Details: FREQ Procedure – Statistical Computations – Measures of Association. Retrieved from https://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_freq_a0000000659.htm
31. Gingrich P. (2004). Chapter 11 - Correlation. *Introductory Statistics for the Social Sciences* [text for Sociology 201 at University of Regina, Regina, Saskatchewan, Canada]. Retrieved from <http://uregina.ca/~gingrich/text.htm>

32. Winkelmayr C, Kurth T. (2004). Propensity scores: help or hype? *Nephrology Dialysis Transplantation* 19, 1671-1673. doi: 10.1093/ndt/gfh104. Retrieved from <http://ndt.oxfordjournals.org/content/19/7/1671.full>

APPENDIX

Figure 1 Panel A

Figure 1 Panel B

Figure 1: Panel A shows that the error term is approximately normally distributed for the ANCOVA model for the in-hospital cost outcome variable, and Panel B shows that the error term does not show a pattern which would indicate that error variance is non-constant. Thus, these assumptions of the ANCOVA model are not violated for analysis of in-hospital cost.

Figure 2: The Q-Q plot shows that the model fits poorly in the region where in-hospital cost is low, as the residuals in that region are much larger than expected in absolute value. The same issue can be seen in Figure 1 Panel A, where the left tail of the distribution of residuals extends out beyond what would be expected if normally distributed.

Figure 3: This graph shows that the error term is not normally distributed for the ANCOVA model of outpatient cost if zero-cost patients are not removed from the sample. If a researcher wishes to proceed with an analysis that includes these patients, he should not use ANCOVA or PS regression.

Figure 4 Panel A

Figure 4 Panel B

Figure 4: Panel A shows that total number of complications appears to follow a linear relationship with its predictors. Panel B shows that the error term does not appear to be i.i.d., and thus, ANCOVA assumptions are violated.

SAS Code: This section contains some example SAS code for conducting the ANCOVA and propensity score methods discussed in the paper, as well as SAS code for determining the symmetric lambda statistic used to test instrument relevance for the IV method.

To run ANCOVA for length of stay:

```
Proc GenMod data=BMP_Auto4 plots=all;  
  Class Sex Insurance BMP;  
  Model Days = BMP Age Sex Year2 Charlson Insurance  
            / dist=poi link=log type3;  
run;
```

The “GENMOD” procedure runs generalized linear models in SAS; the “class” statement lists the variables in the model which are categorical; BMP, Age, Sex, etc. are the explanatory variables in the model; the “dist” option allows you specify the distribution for the response variable (“poi” for Poisson); and “link” option allows you to specify the link function for that distribution (“log” for the standard log-link function). This code gives the full model for length of stay; to get the reduced model, explanatory variables were removed sequentially based on the p-value.

To get the estimated propensity scores for the BMP and autograft patients:

```
Proc logistic data=Dataset.BMP_Auto;  
  Class BMP Sex Insurance;  
  Model BMP (event='1') = Age Sex Year Insurance Charlson  
            / selection = stepwise;  
  Output out=EstPropensity Pred=PropScore;  
run;
```

The “LOGISTIC” procedure runs a logit model by default. The “selection” option allows you to choose the variable selection method for the final model. The “output” statement writes a new SAS dataset which contains whichever variables from the logistic regression that you tell SAS to write – I have told SAS to write the variable for the predicted probability with the

“Pred=” option, and named that variable PropScore. I can then use this variable PropScore to do propensity score regression, illustrated in the following code:

```
Proc GenMod data=PSLOS plots=all;  
  Class BMP;  
  Model Days = BMP PropScore / dist=poi link=log type3;  
run;
```

This code is identical to the ANCOVA code above, except that PropScore replaces the vector of covariates which were each listed individually in the ANCOVA model after BMP.

For the continuous outcomes in the propensity score matched sample, the paired t-test is conducted by:

```
Proc ttest data=Dataset.PSMatchedLOS;  
  paired BMPLOS*AutoLOS;  
run;
```

The “TTEST” procedure conducts Student’s t-test. The “paired” statement indicates that it is a matched-pairs design (not independent). The two outcomes to be compared (BMP length of stay with autograft length of stay) are joined with an asterisk.

To get the symmetric λ statistic for measuring association between categorical variables which are not necessarily explanatory:

```
Proc freq data=Dataset.BMP_Auto;  
  tables BMP*EmpCty / measures cl;  
run;
```

The “tables” statement creates the contingency table between the variables for which you wish to measure association (BMP and employee county in this case), the “measures” option provides a number of statistics that measure association, and the “cl” option provides confidence limits for those measures.

CURRICULUM VITA

NAME: Kristopher C. Gardner

ADDRESS: 2610 Broadmeade Rd.
Louisville, KY 40205

DOB: Louisville, Kentucky – September 12, 1984

EDUCATION: B.A. Foreign Language and International Economics (FLIE)
University of Kentucky
2003-2007

B.A. Spanish
University of Kentucky
2003-2007

M.S. Biostatistics
University of Louisville
2012-2014

AWARDS: National Merit Scholar
2003