Evaluating Comparative Effectiveness of Simultaneous Liver and Kidney Transplant versus Liver Transplant Alone using Instrumental Variables

by

Mengqi Wang

BS, Biology, Slippery Rock University of Pennsylvania, 2018

Submitted to the Graduate Faculty of the

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Mengqi Wang

It was defended on

April 20, 2020

and approved by

Ada Youk, PhD, Associate Professor, Biostatistics
Graduate School of Public Health, University of Pittsburgh

Jeanine Buchanich, PhD, Research Associate Professor, Biostatistics
Graduate School of Public Health, University of Pittsburgh

Jenna Carlson, PhD, Assistant Professor, Biostatistics
Graduate School of Public Health, University of Pittsburgh

Thesis Advisor: Douglas Landsittel, PhD, Professor, Biomedical Informatics School of Medicine, University of Pittsburgh

Evaluating Comparative Effectiveness of Simultaneous Liver and Kidney Transplant versus Liver Transplant Alone using Instrumental Variables

Mengqi Wang, MS

University of Pittsburgh, 2020

Abstract

Improving the quality of medical care often requires assessment of comparative effectiveness between treatments. Although randomized controlled trials (RCTs) are considered as the gold standard for generating evidence, they may not be feasible or ethical to conduct for some comparisons. Therefore, observational studies are required to address many research questions. However, observational data may lead to a high potential for selection bias because subjects or physicians choose their treatments, which may complicate the estimation of causal effects. As one approach to overcome these issues, instrumental variables (IVs) can be used to potentially estimate unbiased causal effects in the setting of observational comparative effectiveness research.

The goal of this thesis is reducing unmeasured confounding in an observational study to compare the effectiveness of simultaneous liver and kidney transplants (SLKT) versus liver-only transplants (LTA) in patients who were on the liver transplant wait list with dialysis. We hypothesize that SLKT could lower mortality by replacing both organs in the same operation.

A two-stage least squares (2SLS) was used to estimate causal effects. The first stage was regressing treatment on IV and covariates to determine whether IV met the assumption of strongly predicting treatment. Then, the second stage least squares analysis was performed by regressing outcome on the estimated treatment and covariates. This analysis used several strategies for

formulating the IV based on geographic region, with similar results. Although our IV met the necessary assumptions, results did not show a significant causal relationship between treatment and mortality.

Findings of this thesis are significant to public health because more than ten thousand patients in the US are on the liver transplant waiting list. While performing both a kidney and liver transplant in these patients may save lives, we are not aware of any other studies that evaluated this problem using IVs or other approaches that potentially account for unmeasured confounding. By evaluating the causal effects of the different transplant approaches, physician and patients can make more informed decisions. The information may also be important for organ allocation strategies nationally.

Table of Contents

List of Tables

List of Figures

Preface

First, I would like to express my gratitude to my thesis advisor, Dr. Douglas Landsittel, for all his guidance, encouragement, and support throughout my thesis project. His advice and guidance have not only helped me understand the novel statistical method and complete this thesis smoothly, but also showed me the progress of the complete research and effective organization skill that benefit my future work.

Also, I would like to thank Dr. Ada Youk, Dr. Jeanine Buchanich and Dr. Jenna Carlson for being as my thesis committee and for taking time to give my feedback and suggestion during this process, including both writing and presenting of the thesis. Given by their help, my thesis went through smoothly and in the timely manner. I would also like to thank Dr. Ramon Bataller and Dr. Carlos Fernandez from the Division of Hepatology at the University of Pittsburgh Medical Center, for allowing us to use the data and providing valuable guiding input on the clinical problems of interest.

Additionally, I would like to thank all the faculty in the Biostatistics department at the University of Pittsburgh. I learned comprehensive and systematic statistical knowledge and gained valuable experience about team work, organization skills, and critical thinking. These harvests are so invaluable that will affect my future life.

Last, but not least, I would like to thank my family for a lifetime of support. For their continuous love, support and encouragement, I grow up as who I am and where I am today. Their love and support are the source of my power and motivation. I would also to thank extremely supportive classmates and friends, I could not have done it without them.

1.0 Introduction

To improve the quality of medical care, it is important to assess the effectiveness of one treatment or intervention compared to an alternative one. More specifically, comparative effectiveness research (CER) seeks to compare two or more known-to-be efficacious treatments or interventions, in terms of harms and benefits, for populations or subgroups, to generate evidence that compares effectiveness of interventions in pragmatic settings (Sox, 2009). CER uses both observational studies and randomized controlled trials (RCTs) to generate evidence for decision-making in medical care.

RCTs are studies where participants are assigned at random to one of two medical treatments (Spieth et al., 2016). For CER, the two groups should be different interventions (rather than a treatment versus a placebo, or no intervention at all). Because the treatment assignment is completely random, the observed difference in the outcome variable can be attributed to either a true difference, or to chance alone. In comparison, observational studies may yield differences in the outcome due to differences in subject characteristics between two treatment groups. In other words, confounding factors (which are variables associated with both the treatment and the outcome) may create bias in the results of observational studies.

In general, rigorously designed RCTs (when they are feasible and can be done in pragmatic setting) provide the best design for avoiding biases (Grossman, 2005). Results of RCTs are therefore often considered as the highest-quality evidence in clinical research as the randomization (if done correctly) eliminates potential for confounding. Results can be analyzed with common statistical tests (e.g. unadjusted tests or regression models that account for factors used in the

randomization strategy). Statistically significant results (i.e. p-values) are then readily interpretable as the probability of the observed result or more extreme occurring by chance alone.

However, a randomized controlled trial is not always feasible to conduct. Common challenges of RCTs include cost and time required to design and conduct the study (Grossman, 2005). Recruiting and retaining a generalizable study cohort is also difficult to accomplish in practice. Specifically, participants in RCTs tend to be healthier and fall into higher socio-demographic categories. Further, for some treatments, randomization may not be ethical, or physicians may feel strongly about the best treatment for their patients, and thus refuse to randomize treatment assignment. Gaining informed consent may also be impractical. Therefore, some research cannot be performed as an RCT.

Given these limitations of RCTs, observational studies may be more practical for many research questions. An observational study is a type of non-experimental research study where investigators simply observe the intervention (or use already-collected data), subsequent outcomes, and potentially confounding or predictive factors (Lu, 2009). In most cases, observational studies are more practically feasible, offer longer follow-up time, and are generalizable to a wider population.

While there are many benefits to using observational data for CER, they also lead to a high potential for selection bias because subjects or physicians choose their treatment. Subjects who, for instance, choose a more experimental treatment, are likely to have different characteristics from subjects who choose a standard treatment. If those factors, which differ between treatment groups, are related to both the outcome and treatment assignment, the resulting treatment effect (using standard statistical methods) will be biased (i.e. have a systematic error favoring one group over the other). Causal effects in the observational study may therefore be difficult to estimate. To

potentially overcome these issues, statisticians, and researchers from other disciplines (e.g. econometrics) have developed different approaches to causal inference that better address confounding to consistently estimate the causal effect. In contrast, the usual (e.g. linear or generalized linear) regression model tends to estimate only associations rather than causal effects.

There are many frameworks for causality, such as potential outcomes, and Campbell's framework (Lewis, 2019). The potential outcomes framework, which is also known as the Rubin causal model or the Neyman-Rubin potential outcomes model (Rubin, 2005), is described in section 1.2. In contrast to Rubin's approach to causal inference, Campbell's framework pays more attention to distinguishing factors that might make internal validity implausible (West and Thoemmes, 2010). For purposes of this thesis, we will use the potential outcomes framework, as it represents the most common framework for causality. Essentially, the potential outcomes framework defines the causal effect as the difference in outcomes between two hypothetical scenarios: one where the subject is exposed to the first treatment and one where they are exposed to the second treatment (or are unexposed) (Rubin, 2005). In practice, one of these outcomes is observed, and the other is the counterfactual. The average causal effect is then defined as the expected difference in potential outcomes.

While the individual causal effect is not directly estimable, there are numerous statistical approaches aimed at estimating the average causal effect over a population in an unbiased manner (or at least in a consistent, i.e. asymptotically unbiased, manner). Tools such as directed acyclic graphs (DAGs) may also be used to graphically represent potential causal relationships among covariates, and thus assist with developing the model. Statistical approaches fall into two main categories: 1) methods that use measured confounders (such as propensity score; see section 1.3), and 2) methods that attempt to emulate randomization through some instrument that predicts

treatment but is (conditional on the treatment effect) independent of the outcome. This type of instrumental variable approach attempts to account for unmeasured confounding (see section 1.4). This thesis specifically applies instrumental variables to estimate an unbiased causal effect in the setting of observational comparative effectiveness research. More specifically, the goal of this thesis is to compare the effectiveness of simultaneous liver and kidney transplants (SLKT) versus liver-only transplants (LTA) in improving survival among patients who are on dialysis. We hypothesize that SLKT will lead to a better survival rate than LTA.

1.1 Background on Kidney and Liver Transplants

There are approximately 15% of adults in the United States that have chronic kidney diseases (CDC, 2019). Many factors cause problems in the kidneys like diabetes, high blood pressure, genetic diseases, cancer, or certain medicines and illegal drug use. When conditions become severe to the extent that the patient's kidneys do not work any longer, two common treatment options can be chosen. One is dialysis, which is using a filtering machine or special fluid to help patients to filter waste out of their bodies. Dialysis does not change other functions of the kidneys; it just keeps patients alive. The alternative treatment is a kidney transplant, which is a surgery using a healthy kidney from donors to replace the non-functioning or diseased kidneys. About 430,000 Americans with kidney failure rely on regular blood-filtering dialysis treatments to survive. More than 100,000 U.S patients are waiting for kidney transplants, but only around 19,000 of those will get a kidney transplant each year (UNOS, 2019).

Even though the liver is the only organ than can regenerate itself, around thirty million Americans are affected by liver disease. Liver diseases are one of the leading causes of death

among adults between the ages of 25 and 64 in the US. Liver diseases can be genetic or caused by many factors, such as obesity, viruses, drug and alcohol use. It is life-threatening when the liver is losing or has lost its functions and a liver transplant is necessary. More than ten thousand patients in the United States are on the liver transplant waiting list, but around eight thousand liver transplant surgeries are performed every year. As a result, thousands of patients die each year while on the list (UNOS, 2019).

For patients who have both liver and kidney diseases and are on the liver transplant waiting list, there are two common interventions, simultaneous liver-kidney transplantation (SLKT), where patients receive both liver and kidney transplantations at the same time, and liver transplantation alone (LTA), where patients only consent to receive liver transplantation, and thus continue using dialysis as the treatment for kidney disease. An RCT is not suitable for this situation because a kidney is not always available, especially since organs are often allocated to patients on a regional basis. To conduct an observational comparison of SLKT versus LTA, this study will use existing databases and employ causal inference techniques. Although comparative effectiveness of these methods has previously been studied, past publications (with the exception of another thesis that used propensity score-based methods) have relied on standard statistical measures of association.

1.2 Causal Inference

Causal inference is the process of estimating the effect of a treatment or exposure on the potential outcomes. Rubin (2005) defines a causal effect as a comparison of potential outcomes, including the outcome they would have had if they had received one treatment and the outcome

they would have had if they had received another treatment (Figure 1). The potential outcomes are then either observed (for the outcome corresponding to the observed treatment) or counterfactual (for the outcome corresponding to the alternative treatment). If a given subject i had received the treatment A, then the outcome $Y_i^{(A)}$ is observed and outcome $Y_i^{(B)}$ is counterfactual and vice versa. The causal effect is the comparison between $Y_i^{(A)}$ and $Y_i^{(B)}$. In this study, the causal effect is the comparison of survival times, in the form of the hazard ratio, between the outcome caused by SLKT, $Y_i^{(SLKT)}$, and the outcome caused by LTA, $Y_i^{(LTA)}$.

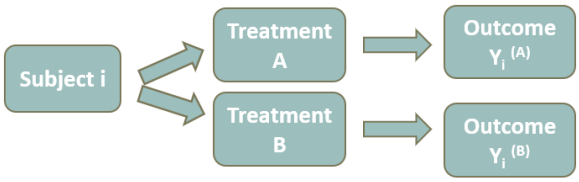


Figure 1. Explanation of causal effect

There are some assumptions for causal inference. First, choices of targeted subjects should use covariate histories as criteria; the covariate history should be measured up to but not beyond assessment of treatment received. Researchers should describe the population that gave rise to the effect estimates. The timing of the outcome assessment should also be relative to the initiation and duration of the treatment.

As mentioned above, even though observational studies tend to be more generalizable and can be used for larger data sets, the challenge of causal inference is that we cannot collect outcome data under the counterfactual case. Study participants that get one treatment may systematically differ from those who get the other treatment. Therefore, differences in outcomes may be a result of pre-existing differences, not of the treatment itself. A confounder (Figure 2) is associated with both the treatment of interest and the outcome of interest, but not in the causal path between

treatment and outcome. In terms of general research practice, it is essential to document approaches thoroughly, including pre-specification of the hypotheses and data analysis plans to make results reproducible.

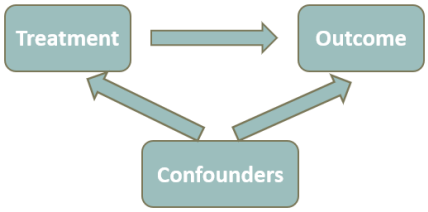


Figure 2. Relationships between treatment, outcome and confounders

There are many possible methods to handle the bias caused by lack of randomization in causal inference, such as propensity score-based methods, instrumental variable methods, g-estimation, marginal structural models, and structural nested mean models. Other than propensity score-based methods and instrumental variable methods, these three new approaches for causal inference are suitable for causal effect of a time-dependent exposure that time-dependent covariates may be confounders or intermediate variables (Robins, 2000). Because this study instead focuses on a point intervention (transplant, which is administered at a single time point), analyses were restricted to propensity score methods (in a past study) and instrumental variables. In section 1.3, the preliminary study of the propensity score method is reviewed, and the instrumental variable method is introduced in more detail in section 1.4.

1.3 Preliminary Study --- Propensity Score Method

Propensity score-based methods are another set of useful approaches for causal inference in the setting of observational data. The propensity score is the probability of an individual receiving the treatment condition (vs. the comparison), given a set of observed covariates (Rosenbaum and Rubin, 1983). The propensity score is then used, through either matching, stratifying, or weighting, to pseudo-randomize the same in an effort to make the treatment and comparison groups as similar as possible with respect to the observed covariates (Figure 3). When the causal question is addressed using non-randomized data, the assignment mechanism such as logistic regression or machine learning methods can be used to model the probability of selecting a given treatment. This predicted probability, or propensity score, is then used to match, stratify or weight the original data to 'pseudo-randomized' the data (i.e. to create a new data set that better emulates a randomized trial). A standard statistical approach, or outcomes model, is applied to the pseudo-randomized data to estimate the treatment effect.

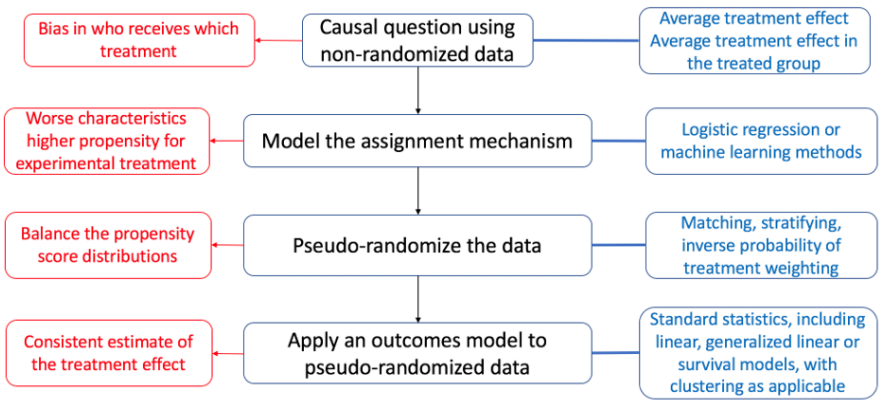


Figure 3. Schematic diagram for main idea of propensity scores

In a previous study (Srivastava, 2019), propensity-score (PS) based methods were used to estimate the causal effect, defined as the expected difference in potential outcomes. PS-based

methods refer to a series of approaches including calculating the PS, using the PS to define pseudo populations, and defining the outcomes model to estimate treatment effects. Logistic regression, classification trees, and a random forest model were three approaches used to model the treatment assignment mechanism. Three pseudo populations were created from each of the resulting PS distribution: 1:1 propensity score matching, stratification into quintiles, and inverse probability of treatment weighting (IPTW). A cox proportional hazard model was then fit for each pseudo population to estimate the treatment effect. The random forest model showed a significant estimate of treatment effectiveness with matched pseudo populations, but there was no significance when using stratified pseudo populations. Both the classification tree and random forest models led to significant estimates in IPTW pseudo populations.

In conclusion, these varied results indicated that the assignment mechanism, the approach for forming the pseudo population, and the choice of outcomes model, all can significantly influence results of PS-based methods and estimates of treatment effect. However, it is not always clear how to optimally model the treatment assignment mechanism or form the pseudo population because propensity score-based approaches are a multi-step process, not a single method, and each step needs to be carefully selected. In addition, results may be sensitive to unmeasured confounding. Although there are several methods for assessing sensitivity to unmeasured confounding, we are not aware of any such methods that apply to survival analysis.

1.4 Instrumental Variable Method

The instrumental variable (IV) method is a common approach to evaluate the causal effects of treatments when unmeasured confounding presents a serious concern in an observational study

(Baiocchi et al., 2014). More specifically, when potential confounders are not measured as part of the dataset, even causal inference methods based on propensity scores cannot consistently estimate (i.e. estimate in an asymptotically unbiased manner) the causal effect. The IV method depends on specifying a variable that is strongly predictive of treatment but conditionally independent of the outcomes. The IV is used to identify the unobserved correlation between X and Y (Figure 4).

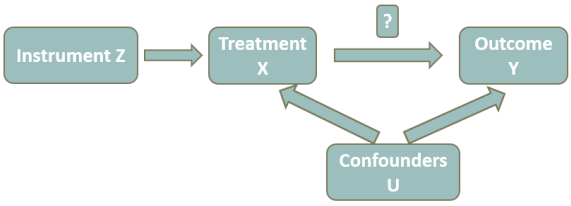


Figure 4. Path diagram of regression analysis within instrumental variable Z

Valid IVs should satisfy three features (Baiocchi et al., 2014). First, the IV directly causes changes in the treatment only, which can be verified from the data by ensuring that changes in the IV are related to changes in the treatment. Second, the IV is not associated with variation in measured or unmeasured confounders that influence the outcome. The balance check for each covariate conditional on the IV and treatment can used to assess the second assumption. Third, IVs do not cause changes in the outcome variable directly, rather, the IV only indirectly influences the outcome of interest through its effect on the treatment. The standard logistic regression can be used to test whether there is any relationship between the outcome and the IV.

Angrist and Pischke (2009) used an analysis of Vietnam War veterans as an example to explain why the IV could work to estimate the causal effect even without measuring the confounders. They indicated that the military participation (X) might lower lifetime income (Y) because of the sort of psychological effects of going to war. However, the military participation is

likely correlated with unobserved factor such as, preference of office work, academic attitude, value of money for choosing jobs, etc., which also detrimentally impact lifetime income. Then the effect from OLS will be greater than the true causal effect of military participation on lifetime income. This study used draft eligibility as the IV to address the causal effect of military participation because there is a highly positive correlation between military participation and draft eligibility.

Subjects were grouped into three categories: eligible for the draft and went to war, not eligible for the draft and went to war, and did not go to war. The average of lifetime income for individuals who participated in war was much lower than the lifetime income for individuals who did not participate in war, but the effect of military participation on lifetime income was overstated. In fact, the individuals who were not eligible for the draft, but still participated in the war, were likely to earn less during their lifetime. After excluding these individuals, the comparison of the lifetime income between individuals who were eligible and also participated in the war and individuals who did not participate in war estimated the true causal effect of military participation on the lifetime income because there are no other reason to suggest these two types individuals are different.

There are limitations to consider when using IV methods (Baiocchi et al., 2014). More specifically, an IV can be categorized as a strong or weak IV, based on its association with treatment selection. The weak IV, even if it is valid, could cause high variance in the causal estimate, and thus produce misleading inferences from two stage least squares. Trade-offs between an IV estimate with a large variance and a conventional estimate with possible bias need to be considered when deciding on the best method. IV methods are the most helpful for the analysis when unmeasured confounding is a major concern in the study and there exists a valid IV that is

strongly associated with choice of treatments. IV methods could be considered as secondary analyses or sensitivity analyses when unmeasured confounding is not a serious issue in the study or a valid and strong IV is not available. In this thesis, we attempt to find a valid IV that can compare the effectiveness of SLKT and LTA to control unmeasured confounding in the observational study.

1.5 Statement of Problem

In this study, we attempt to determine the causal effect of SLKT versus LTA in patients who were on the liver transplant wait list with dialysis between 2006-2016, with a minimum of one year of follow-up. More specifically, the goal of this thesis is to conduct an IV analysis to reduce the potential for unmeasured confounding that could bias the comparison.The results from this study could provide important information for future changes to organ allocation when both liver and kidney diseases are present.

2.0 Method

The analysis was conducted in four steps. First, the data set was managed by omitting missing values and patients without one-year follow-up. The outcome of interest was specified as survival time less than one year, and the primary covariate of interest was the treatment groups of either LTA or SLKT. Second, suitable covariates were chosen as the IVs according to the aforementioned criteria for defining IVs. Third, other covariates of interest were introduced to the model and risk difference estimates for mortality were compared. Last, assumptions of IV methods were assessed to see whether they were violated in this study. This study aims to determine the causal effect of SLKT versus LTA for reducing the risk of death by performing IV methods to reduce unmeasured confounding. The study has substantial importance for organ allocation in kidney and liver diseases.

2.1 Data Sources

Patient information was collected from the United Network for Organ Sharing (UNOS), which is the platform to gather, analyze and publish all Organ Procurement and Transplantation Network (OPTN) data. The original dataset included 2,587 observations and 23 covariates. Based on input from clinical investigators, ten covariates were retained for this study: date of kidney transplant; three continuous covariates: BMI, recipient age and patient survival time in days; three binary variables: sex (male/female), transplant type (LTA/SLKT) and death status (yes/no); and three categorical variables: diabetes (none, type I, type II), regions where listed or transplanted

(Region 1-11), and race (black/not black) included. After omitting 114 patients with missing values and 508 patients without one-year follow-up, there were 1,965 observations in the final analytic sample.

2.2 Standard Logistic Regression Analysis

A logistic regression model is a generalized linear model typically used to model the relationship between a binary outcome and treatment or exposure of interest. Logistic regression can address questions such as whether, and to what degree, an exposure or treatment is associated with the outcome. Logistic regression requires a binary outcome, with independently and identically distributed data. While variables in the model may be correlated, their association on the outcome may be impossible to disentangle if any variables are approximately equal to a linear combination of the other variables (e.g. multicollinearity). Also, while logistic regression makes no assumptions about the distribution of the variables in the model, the link function used in the model does assume that each of the continuous variables are linearly related to the log odds of the outcome.

While each of these assumptions may be reasonable for our study, the logistic model still assesses only associations rather than causation. Specifically, the logistic model estimates the ratio in the log-odds given the observed covariate values, which is not the causal effect (i.e. expected difference, or log-odds, between potential outcomes). Thus, the results of the logistic model are unlikely to represent an unbiased estimate of the causal effect. In addition, the logistic model, as is the case for any other regression model, depends on the observed covariates and cannot adjust for unmeasured confounding.

2.3 Instrumental Variable Analysis

2.3.1 Assumptions of IV Methods

As mentioned in section 1.4, an instrument is the variable that directly affects the treatment, only affects the outcome through the effect on the treatment, and is not associated with unmeasured confounding conditional on covariates. Thus, there are three basic assumptions of IV methods in general: the relevance assumption, the exclusion restriction, and the independence assumption (Baiocchi et al., 2010). The relevance assumption indicates that the IV, Z, directly have a causal effect on the treatment, X. In other words, IV is associated with the treatment received. We assumed that patients in certain regions are encouraged to have an SLKT. In this case, we assume that the patient is encouraged to have SLKT ($E(X^{SLKT}|U) > E(X^{LTA}|U)$). The exclusion restriction requires that Z affects the outcome Y only through X. Figure 4 depicts this relationship, Y^{SLKT} is the potential outcome for the patient that received SLKT; Y^{LTA} is the potential outcome for the patient that received LTA, there is no direct outcome from the IV. The independence assumption describes that there is no confounding between the effect of Z on Y, which is exogenous. If the variable is uncorrelated with the unmeasured confounding, then this variable is exogenous. When the instrument and treatment are binary, the independence assumption can also be written as "Z is independent of (X^1, X^0, $Y^{1,1}$, $Y^{1,0}$, $Y^{0,1}$, $Y^{0,0}$) |U"; 0 and 1 in X indicates two categories of IV, and in Y indicates combinations of IV and treatment (Baiocchi et al., 2010).

Because the instrument and treatment are binary variables in this thesis, there are two extensions of the assumptions. One is called stable unit treatment value assumption (SUTVA), the other is the monotonicity assumption (Baiocchi et al., 2010). SUTVA suggests that the treatment only affects the patient who takes that treatment, and this treatment will not have other versions

that cause differences in the outcome. To be specific, the patient who accepts SLKT will not influence the next patient who will receive either SLKT or LTA. Also, SLKT treatment will only generate one outcome; there are no other outcomes for SLKT treatment.

Introducing the monotonicity assumption requires definitions for the following terms: Complier, Never-taker, Always-taker, and Defier. These are the four subgroups of patients or subjects for which we observe both actual assignment and the counterfactual assignment (Table 1). Always-takers are the subjects who are always treated no matter under which instrumental assignment. Alternatively, Never-takers are the subjects who are never treated regardless of instrumental assignments. Compliers are the subjects with treatments following instrumental assignments. Defiers are the subjects with treatments opposite of instrumental assignments, which is essentially related to the monotonicity assumption. In clinical trials, the effect homogeneity assumption is never achieved because it is biologically implausible to keep the effect of the treatment X on the outcome Y constant among patients. Furthermore, we cannot identify these four subgroups in real world situations because we can only observe the treatment under the actual assignment; the counterfactual assignment will never be observed.

Table 1. Four subgroups defined by combinations between actual and counterfactual assignments and exposure

	Z = 0	
	X = 0	X = 1
Z = 1		
X = 0	Never-takers	Defiers
X = 1	Compliers	Always-takers

Alternatively, the more feasible assumption of monotonicity only requires that there are no Defiers (Lousdal, 2018), i.e. "no subject i with $D_i^1 = 0$, $D_i^0 = 1$ (Baiocchi et al., 2010). Always-takers and Never-takers have no causal effect of the instrument on the treatment because they have

constant values of treatments regardless of instrumental assignments. If we assume no Defiers exist, Compliers are the only subgroup having effect of the instrument Z on the treatment X. Then the Complier average causal effect (CACE) or the local average treatment effect (LATE), $E(Y_i^1 - Y_i^0 \mid \text{Compliers})$, can be estimated to infer the treatment effects.

2.3.2 Sources for IVs

Based on the assumptions above, finding a strong valid IV is vital in IV analysis. Randomized assignment in trials would be a desirable instrument because the CACE could then be estimated (Lousdal, 2018). However, in observational studies, a variety of factors, both measured and/or unmeasured, may affect a patient's treatment selection and/or the treatment assigned by their physician. Because of this, natural features are commonly chosen as the instrument variable (e.g. geographical variation, physical distance to facility, physician's preference, genetic factors, or timing variables) (Lousdal, 2018).

In this thesis, the region is one variable that could be used as described above and was chosen as the IV. The United States was partitioned into 11 regions (as shown in Appendix A). To simplify the analysis, the region variable was dichotomized using two approaches. In one approach, the regions were divided into high proportion of SLKT group and low proportion of SLKT group. If the proportion of SLKT in the region was higher than 40%, then this region was regarded as the high proportion of SLKT group, and vice versa. The other approach was to only keep the two regions with extreme proportions of SLKT. The region with the highest proportion of SLKT was regarded as yes for IV predictive SLKT, and the region with the lowest proportion of SLKT was regarded as no for IV predictive SLKT. Based on these two classifications of regions, there were two series of IV analyses.

2.4 Two Stage Least Squares (2SLS) Regressions

The ordinary least squares (OLS) estimator is used in traditional regression analysis to infer the relationship between X and Y. The coefficient of interest, β, is a function of the covariance between X and Y and the variance of X, where $\beta = \sigma_{XY} / \sigma^2_{XY}$ (Pokropek 2016). Under this circumstance, when X is uncorrelated with U (a potential confounder), corr (X, U) = 0, we can say X is an exogenous variable. When Y is correlated with a potential confounder, U, corr (Y, U) ≠ 0, then Y is an endogenous variable. However, having an endogenous variable is not completely suitable for the IV analysis. In the IV analysis, X and Y are correlated with U, but Z is uncorrelated with U. Thus, Z is the IV that can be used to estimate β, the association between X and Y. However, the OLS estimator of β will be biased and inconsistent in this situation because part of the variance in Y is endogenous and part is exogenous.

Therefore, a two-stage least squares estimator is necessary to infer β by the 'ivreg' function in the 'ivpack' package in R. According to the assumptions in section 2.2.1, the complier average causal effect (CACE) can be written as

$$\mathrm{E}(Y_i^1 - Y_i^0 \mid \text{Compliers}) = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(X|Z = 1) - E(X|Z = 0)}, \qquad (1)$$

where we can see the IV is estimated through two stages of OLS estimations. The first stage is regressing X on Z to eliminate the endogenous part of the variation in Y. We can write the equation:

$$X = a + bZ + e, \qquad (2)$$

where X is the observed treatment, a is the intercept, b is the coefficient of the IV Z, and e is the error term. From equation (2), we can obtain the consistently estimated E(X|Z). The second stage

is to solve the exogenous part of the variation by regressing Y on the estimated E(X|Z), the equation can be written as:

$$Y = \beta_0 + \beta_1 * E(X|Z) + e \quad (3)$$

where Y is the outcome variable, β_0 is the intercept, β_1 is the coefficient of the estimated E(X|Z), which is also the inferred relationship between X and Y, and e is the error term.

Sometimes, the IV will be valid or become a strong IV only after adjusting on covariates (Baiocchi et al., 2010). After incorporating covariates into the model, the first stage of 2SLS will be regressing the treatment X on the covariates and instrument Z to obtain the estimated X, and then regressing the outcome Y on the estimated X and covariates in the second stage. In the current study, we adjusted for sex, age, BMI, diabetes, and race as covariates in the IV model that can turn region into the strong IV.

2.5 Assessing the IV Assumptions

The three basic assumptions of IV methods relevance assumption, exclusion restriction, and independence were assessed. The relevance assumption was assessed by the first stage of the 2SLS. This is the most convenient approach to test whether there is association between the instrument and the treatment. If this assumption is violated, it would be necessary to change a different variable as the instrument. Baiocchi et al. (2010) stated that IV would violate the exclusion restriction if the IV was associated with other treatments which accompanied the primary treatment of interest. In this case, besides SLKT and LTA treatments, all subjects were treated with the dialysis; therefore, our IV region is not associated with the concomitant treatment. The exclusion restriction was satisfied in the current study.

The association between the IV and the measured confounders may reflect whether the independence assumption is violated (Baiocchi et al., 2010). The imbalance of measured covariates across level of the IV and the bias ratio should be assessed to prove if there is association between the IV and observed covariates. If so, there might be some associations between IV and the unmeasured confounding.

2.6 Overview of Statistical Analyses

For each of the 11 regions, the frequency and proportion of the two treatments were calculated. The total number of deaths and the number of deaths within one-year follow-up were calculated for each region, along with the proportion of deaths. The central tendencies (i.e. mean, median, and range) of the continuous covariates BMI and age were calculated. To simplify and standardize, all covariates were dichotomized as binary variables. The mean was used as a cut point to dichotomize BMI and age into high and low categories. Diabetes was categorized as "No Diabetes", "Insulin Dependent Diabetes", "Non-Insulin Dependent Diabetes", and "Diabetes, Dependency Unknown" in the original dataset, and was dichotomized as "No Diabetes" and "Diabetes", collapsing the three diabetes categories together, in the current analysis. Self-reported race had three responses Hispanic/Latino, Black and Other, then was dichotomized as black and non-black. The binary IV is the simplest and most frequent option; thus, we dichotomized region using two strategies. The first approach was to dichotomize region into high proportion of SLKT region and low proportion of SLKT region. The second strategy was only to keep the highest and the lowest proportion of SLKT region as IV. These two strategies of dichotomizing region followed the same process of the logistic regression and IV regression analysis.

In each of the two strategies for dichotomizing region, six logistic regression models were performed. To determine whether the IV has the direct influence on the treatment, the IV region was regressed on the treatment. To investigate whether the IV has the direct effect on the outcome, the IV region was regressed on the outcome. Then four logistic regression models: outcome regress on treatment, outcome regress on treatment and region, outcome regress on treatment and other covariates, and outcome regress on treatment, region and other covariates, were built to compare with the IV analysis.

In the IV analysis, the first stage model was regressing IV and the matrix of all covariates on the treatment. The first stage models with and without IV were performed and the F test from ANOVA was used to determine whether the IV was valid and strong enough for two stage least squares inference. The proportion of compliers was estimated using the first stage logistic regression model. The two stage least squares analysis was performed by the package in R called 'ivpack' (Jiang and Small, 2014). The IV regression model was run via 'ivreg' function. The odds ratios were calculated by exponentiating coefficients of the logistic regression model. Finally, the robust standard errors with or without clustering were both calculated. The transplant center code was used as the unit of clustering.

The imbalance of measured covariates across level of the IV was calculated to assess whether there was an association between IV and the observed covariates. When measured confounders are proxies for the true confounder, an association between the IV and measured confounders could imply that there would be an association between the IV and unmeasured confounders. The probabilities of each covariate conditional on $Z = 0$ and $Z = 1$ were calculated separately, and a chi-squared test was performed to determine whether there is significant

difference for each covariate between Z = 0 and Z = 1. CrossTable function in R package ‘gmodels’ (Warnes, et al.,2018) was used to achieve above calculation.

All analyses were performed by the Software R version 3.6.1 (R Foundation for Statistical Computing).

3.0 Results

This section presents the results of the methods described above. Section 3.1 presents a summary of the data, including mortality rates by region and by transplant method. Section 3.2 gives the results for the IV approach where region was dichotomized into high and low proportions of SLKT. Section 3.3 gives the results for the IV approach using region with the two extreme proportions of SLKT.

3.1 Summary Statistics

Table 2 provides the summary statistics for frequency of treatments and total deaths for each region as well as the proportion of deaths within one year. Using the information from this table, region was dichotomized into two groups. In the first strategy, regions 3, 4, 7, 8, 10, and 11 had more than 40% of SLKT; these regions were then defined as yes for the IV for using SLKT; regions 1, 2, 5, 6, and 9 had lower than 40% of SLKT and were defined as the IV for using LTA (i.e. not using SLKT).

The first N column in Table 2 displays the percentage of patients in each region relative to the entire sample; the N columns under LTA and SLKT indicated the percentages of patients using each treatment in each region. The death columns under LTA and SLKT shows the percentages of total deaths for each treatment separately in each region. Also, the last total death indicated the percentage of patients who had survival time less than one year. As shown in Table 2, region 5 has the highest percentage of the patients (25.4% patients of the total). However, region 5 has the

lowest percentage of SLKT usage (29%). Region 7 has around 16% patients in the total sample, and 53% of patients had SLKT, which is the highest values of the SLKT compared to other regions. In the second strategy, region 5 and region 7 were used as the lowest and highest percentages of SLKT, respectively. Table 2 also describes the mortality rates by region. The mortality for each treatment in each region varies from 20% to 40%. The highest mortality is using LTA in region 2 and region 10, both of which are greater than 37%. The lowest mortality is also using LTA in region 6, which is less than 22%. Comparing mortality between LTA and SLKT shows that regions 2, 5, 8 have the similar mortalities with either treatment. Region 2 has the relatively higher mortality and region 5 has the relatively lower mortality using either treatment. In region 9, 21.5% of patients survived less than one year after transplantations, which was the highest percentage. In region 6 and region 11, less than 10% of patients survived less than one year after transplantations.

Table 2. Frequency table of transplant method and death rate before and at 1 year by region

Region	N	LTA		SLKT		Death within one year
		N	Death	N	Death	
1	86 (4.4%)	53 (61.6%)	15 (28.3%)	33 (38.4%)	11 (33.3%)	11 (12.8%)
2	164 (8.3%)	107 (65.2%)	40 (37.4%)	57 (34.8%)	21 (36.8%)	28 (17.1%)
3	235 (12.0%)	138 (58.7%)	32 (23.2%)	97 (41.3%)	33 (34.0%)	29 (12.3%)
4	177 (9.0%)	102 (57.6%)	25 (24.5%)	75 (42.4%)	22 (29.3%)	30 (16.9%)
5	499 (25.4%)	354 (70.9%)	81 (22.9%)	145 (29.1%)	32 (22.1%)	58 (11.6%)
6	45 (2.3%)	28 (62.2%)	6 (21.4%)	17 (37.8%)	5 (29.4%)	4 (8.9%)
7	323 (16.4%)	152 (47.1%)	49 (32.2%)	171 (52.9%)	40 (23.4%)	39 (12.1%)
8	98 (5.0%)	54 (55.1%)	14 (25.9%)	44 (44.9%)	12 (27.3%)	10 (10.2%)
9	79 (4.0%)	55 (69.6%)	18 (32.7%)	24 (30.4%)	7 (29.2%)	17 (21.5%)
10	161 (8.2%)	93 (57.8%)	35 (37.6%)	68 (42.2%)	18 (26.5%)	28 (17.4%)
11	98 (5.0%)	50 (51.0%)	11 (22.0%)	48 (49.0%)	12 (25.0%)	7 (7.1%)
Overall	1965	1241 (60.2%)	344 (27.7%)	819 (39.8%)	228 (27.8%)	261 (13.3%)

3.2 Strategy One: High Proportion of SLKT vs Low Proportion of SLKT

3.2.1 The Logistic Regression Analysis

The treatment was regressed on the IV region, and the ANOVA test was used to calculate the p-value of this model, which was less than 0.0001, and was therefore highly significant. The

IV we chose in this strategy was significantly associated with the treatment, thus indicating that the relevance assumption was not violated. The outcome was then regressed on the IV region, to determine whether there was an association between the outcome and the IV. The p-value of the model was 0.78, so there was no evidence shown that the IV region was directly associated with the outcome. We can assume that region only affects the outcome through the exposure, which means the exclusion restriction was not violated.

Four logistic regression models were performed: outcome regressed on treatment, outcome regressed on treatment and region, outcome regressed on treatment and other covariates, and outcome regressed on treatment, region and other covariates (results not shown). Based on these four models, there was no evidence that the treatment was statistically significantly associated with the outcome. Because of this, IV analysis was performed to estimate the causal relationship between the treatment and the outcome.

3.2.2 The IV Regression Analysis

In the first approach, region was dichotomized as more than 40% SLKT and less than 40% SLKT. Shown in Table 1, regions 3, 4, 7, 8, 10, 11 had more than 40% patients who chose SLKT; for this group, the IV was defined as yes SLKT. Regions 1, 2, 5, 6, 9 had less than 40% patients who chose SLKT; for this group, the IV was defined as no SLKT. In the first-stage model, the treatment variable (i.e. the actual transplant method received) was regressed on the IV and the observed covariates. A partial F statistic was calculated to test whether the instrument had an effect in the first model. The partial F statistic was 39.8, indicating that the IV was strong enough for two stage least squares inference to be reliable. The proportion of compliers was also calculated and 13.78% of the subjects were estimated as compliers.

The two-stage least squares model was fitted by 'ivreg' function in the 'ivpack' R package. Table 3 gives the summary of non-robust standard errors. The residual standard error was 0.3385 on 1958 degrees of freedom, and the multiple R-Squared and adjusted R-Squared were 0.0086 and 0.0056, respectively. The F test was 2.208 on 6 and 1958 degrees of freedom, and the p-value was 0.040, which was statistically significant. Based on Table 3, we estimated that the effect of using SLKT (treatment) was to reduce the mortality rate for compliers by 0.029 or 28.6 patients per 1000 patients on the waiting list. The odds of death within one year for the high SLKT group is 3% lower than the odds of death within one year for the low SLKT group, but was not statistically significant.

Table 3. Non-robust standard errors of the 2SLS model

	Estimate	Std. Error	t value	P-value	OR	CI.2.5 %	CI.97.5 %
Intercept	0.014	0.077	0.18	0.85			
Treatment	-0.028	0.11	-0.25	0.81	0.97	0.78	1.21
Sex	0.0019	0.016	0.12	0.91	1.00	0.97	1.03
Diabetes	0.027	0.022	1.22	0.22	1.03	0.98	1.07
Race	0.046	0.040	1.16	0.25	1.05	0.97	1.13
BMI	0.0059	0.018	0.33	0.74	1.01	0.97	1.04
Age	0.041	0.016	2.52	0.012	1.04	1.01	1.08

* BMI indicates the body mass index.

The standard errors in Table 3 were non-robust, Table 4 shows the results of standard errors that are robust for heteroscedasticity with and without clustering. The transplant center code was chosen as the unit of clustering. Robust standard errors did not lead to any substantial change as compared to the non-robust error after accounting for clustering. In both analyses, we concluded that the treatment (SLKT or LTA) did not have a significant influence on the outcome in the IV analysis.

Table 4. Robust standard errors (RSE) with and without clustering

	RSE without clustering			RSE with clustering		
	Std. Error	t value	P-value	Std.Error	t value	P-value
Intercept	0.072	0.20	0.84	0.067	0.21	0.83
Treatment	0.11	-0.25	0.81	0.12	-0.22	0.83
Sex	0.016	0.12	0.91	0.016	0.12	0.91
Diabetes	0.021	1.26	0.21	0.020	1.35	0.18
Race	0.034	1.38	0.17	0.030	1.52	0.13
BMI	0.018	0.34	0.74	0.016	0.37	0.71
Age	0.016	2.55	0.011	0.017	2.42	0.015

3.2.3 The Association Between IV and Unmeasured Confounding

The proportion of measured covariates across levels of the IV was calculated by the Cross-Table function in R. Table 5 shows the balance of covariates between patients living in low SLKT region (Z = 0) or high SLKT region (Z = 1). All covariates seem balanced between the two types of region, except race, with high SLKT region patients being much more likely to be non-black.

Table 5. Balance of measured covariates between patients living in low SLKT region vs high SLKT region

Covariate X	P (X | Low SLKT region)	P (X | High SLKT region)	p-value
Sex (Female)	28.9%	35.5%	0.57
Diabetes (Yes)	14.5%	19.6%	0.21
Race (Non-black)	43.2%	52.6%	0.0048
BMI (>30)	23.5%	28.4%	0.45
Age (>55)	24.7%	31.4%	0.71

From Table 5, all covariates except race had p-values greater than 0.05, indicating that they were not statistically significantly different between low SLKT region and high SLKT region. Because we did not find evidence of association between the IV and any measured covariates, we assumed that IV was not associated with any unmeasured covariates either.

3.3 Strategy Two: The Highest and Lowest Proportion of SLKT

3.3.1 The Logistic Regression Analysis

This approach used the same analysis process as the previous approach. Treatment was regressed on IV region, and the ANOVA test was used to calculate the p-value of this model, which was less than 0.0001, and thus highly significant. The IV we chose in this strategy was significantly associated with treatment, proving that the relevance assumption held. Outcome was regressed on IV region to determine whether there was association between the outcome and the IV. The p-value of model was 0.85, indicating that there was no evidence that IV region was directly associated with the outcome. We can assume that region only affected the outcome through the exposure, which means that the exclusion restriction was not violated.

Four logistic regression models were performed: outcome regressed on treatment, outcome regressed on treatment and region, outcome regressed on treatment and other covariates, and outcome regressed on treatment, region and other covariates (results not shown). In these four models, treatment was a factor that had a statistically significant influence on the outcome, but the models were not statistically significant. The results from the logistic regression are biased because they do not consider the unmeasured confounding, so the IV analysis was performed to estimate the causal relationship between the treatment and the outcome.

3.3.2 The IV Regression Analysis

In the second approach, a dichotomized version of region with region 7 as the highest percentage of SLKT and region 5 as the lowest percentage of SLKT was used as our IV. Like

approach one, the treatment variable was regressed on the IV and observed covariates, which was the first stage model. The partial F statistic was calculated by the ANOVA test as 49, which is greater than 10, so that IV is considered valid for two stage least squares inference to be reliable. With this model, 23.63% of the subjects are estimated as compliers.

The two-stage least squares model was fitted with results in Table 6 gave the non-robust standard errors of the IV model. The residual standard error was 0.32 on 815 degrees of freedom. The multiple R-Squared was 0.0020, and the adjusted R-Squared was approximately 0.0. The F test was approximately 1.00 on 6 and 815 DF and the p-value was 0.43. We estimated that the effect of moving to SLKT for treatment will increase the mortality rate for compliers by 0.019 or 19 patients per 1000 subjects who are on the waiting list. As shown in Table 6, the odds of death within one year for the high SLKT group are 2% higher than the odds of death within one year for the low SLKT group, which is not statistically significant.

Table 6. Non-robust standard errors of the 2SLS squares model

	Estimate	Std. Error	t value	P-value	OR	CI.2.5 %	CI.97.5 %
Intercept	0.067	0.16	0.43	0.67			
Treatment	0.019	0.099	0.19	0.85	1.02	0.84	1.24
Sex	-0.027	0.024	-1.13	0.26	0.97	0.93	1.02
Diabetes	0.029	0.028	1.02	0.31	1.03	0.97	1.09
Race	-0.0049	0.071	-0.069	0.94	0.99	0.86	1.14
BMI	0.0058	0.024	0.24	0.81	1.00	0.96	1.05
Age	0.033	0.023	1.41	0.16	1.03	0.99	1.08

The standard errors in Table 6 were non-robust, Table 7 shows the results of standard errors that are robust for heteroscedasticity with and without clustering. The transplant center code was chosen as the unit of clustering. Robust standard errors did not change a lot comparing with the non-robust error with the addition of cluster. Therefore, we concluded that the treatment (SLKT or LTA) did not have a significant influence on the outcome in the IV analysis.

Table 7. Robust standard errors (RSE) with and without clustering

	RSE without clustering			RSE with clustering		
	Std.Error	t value	P-value	Std.Error	t value	P-value
Intercept	0.15	0.45	0.65	0.14	0.48	0.63
Treatment	0.098	0.19	0.85	0.12	0.16	0.88
Sex	0.025	-1.09	0.27	0.023	-1.18	0.24
Diabetes	0.029	0.99	0.32	0.027	1.07	0.28
Race	0.075	-0.066	0.95	0.066	-0.074	0.94
BMI	0.024	0.24	0.81	0.017	0.34	0.73
Age	0.023	1.46	0.15	0.024	1.38	0.17

3.3.3 The Association Between IV and Unmeasured Confounding

The proportion of measured covariates across levels of the IV was calculated by the Cross-Table function in R. Table 8 shows the balance of covariates between patients living in low SLKT region (Z = 0) or high SLKT region (Z = 1). All covariates seem balanced between the two types of region, except race, with high SLKT region patients being much more likely to be non-black.

Table 8. Balance of measured covariates between patients living in low SLKT region vs high SLKT region

Covariate X	P (X | Low SLKT region)	P (X | High SLKT region)	p-value
Sex (Female)	38.9%	25.4%	0.87
Diabetes (Yes)	20.4%	12.2%	0.42
Race (Non-black)	60.0%	37.3%	0.0011
BMI (>30)	32.4%	18.4%	0.066
Age (>55)	33.7%	23.1%	0.35

From Table 8, all covariates except race have p-values greater than 0.05, indicating that they were not statistically significantly different between low SLKT region and high SLKT region. Because we did not find evidence of association between the IV and any measured covariates, we assumed that IV was not associated with any unmeasured covariates either.

4.0 Discussion

In a previous study, Srivastava (2019) used propensity score methods to compare the causal effectiveness of two treatments. She used three approaches (logistic regression, classification trees and random forest model) to model the treatment assignment, and three pseudo populations were created from each of the three PS distributions (1:1 matching, stratification into quintiles, and inverse probability of treatment weighting). However, most outcome models were not statistically significant with no statistically significant differences between SLKT and LTA treatment for liver transplant patients on dialysis. Even though some results show that SLKT was more effective than LTA for dialysis patients, models also displayed more overlaps in propensity score distributions, which were not well-balanced. Thus, it is unclear that which treatment assignment approaches and PS distributions are more optimal combinations to compare the causal effectiveness of two treatments.

PS-based method is a multi-step analysis, each single step in the analysis could lead to completely different results. The analyses described above were essential because they demonstrated a CER of the causal effect for two treatments only using observational data. Yet, the major limitation of PS-based analysis is unmeasured confounding. In this specific data set, the regional allocation for SLKT is inconsistent because the use of SLKT is largely region-based due to transplant organ and recipients. Also, the degree to which the region affects the outcome is hard to measure. Unmeasured confounding in this data set is a necessary issue which needed to be considered. IV analysis is commonly used to estimate causal inferences when the unmeasured confounding is a potential complexity in the study.

In the current study, an IV approach was used with geographic region dichotomized to simplify the data set. Two strategies were used to dichotomize region. In strategy one (section 3.2), the OLS regressions estimates indicated that the treatment (SLKT/LTA) did not have statistically significant association with the outcome (survival time less than one year) regardless of adjustment for other covariates. Because the unmeasured confounding was not considered in the OLS regression, the IV analysis about strategy one was performed. Overall, the first stage model of the IV analysis indicated that dichotomized region was the strong IV. In the second stage model, SLKT slightly reduced the mortality for compliers, but not statistically significantly. The odds of death within one year for patients who used SLKT were lower than the odds of death within one year for patients who used LTA.

In strategy two (section 3.3), the OLS regression estimates showed there was a statistically significant association between treatment and outcome with or without controlling for the other covariates. As the results in the OLS regression may be biased because of unmeasured confounding, the IV analysis was performed for strategy two. The dichotomized region in the strategy two was also regarded as the strong IV from the first stage model of the IV analysis. However, SLKT in the second stage of model showed slightly increasing mortality for compliers, although not statistically significant. The odds of death within one year for SLKT patients were higher than the odds of death within one year for LTA patients in the analysis of strategy two. In summary, this IV analysis did not show significant causal effect between treatment and the mortality of patients using SLKT or LTA. SLKT and LTA did not have statistically significantly different effectiveness for liver transplant patients on dialysis.

According to the balance check in both strategies, the distribution of measured confounding across levels of the IV and treatment is equivalent, except race (shown in Table 5 and Table 8),

indicating that the IV did not have any association with measured and unmeasured confounding. The independence assumption was met. However, not all assumptions can be totally assessed, but methods have been proposed to test certain parts of the assumptions. Also, even if some assumptions are not fully satisfied, e.g. the IV only has a weak association with treatment, it still could provide some useful information. The monotonicity assumption is not suitable for the transplant study. It is implausible to have defiers in the transplant patients; therefore, we assumed the monotonicity assumption was satisfied. As for SUTVA, it is hard to test in this study because the amount of donated kidney and liver is limited, and if one patient matches successfully, it will decrease the probability of other patients matching. SLKT treatment may affect LTA treatment in some way. SUTVA is not testable and may be violated in this case, which may cause some limitations. Further, sensitivity analysis can be performed to test how sensitive the results are to these violations of the IV assumptions.

Except violations of the IV assumptions, the strategies of choosing IV and dichotomizing IV could also influence results. The US was partitioned into eleven regions in the UNOS dataset; however, use of binary IVs is the most common approach in the literature. Therefore, we dichotomized the eleven regions into high versus low proportion of SLKT. Partitioning the whole country into only two levels could, however, lead to substantial variability and inaccurate results showing not statistically significant causal effects between the treatment and the outcome. Future analyses could use a continuous IV, such as the distance between matched donors and recipients, or the number of donors and recipients in the specific region. There are many possible sources of data for defining the IV. Although preference-based IVs (region, hospital or individual physician) are common, numerous other choices, such as calendar time or genetic variants may be used as

potential IVs in other scenarios. Defining multiple IVs to capture unmeasured confounding is another possible, although less common approach.

In addition to the above-described statistical considerations, the availability and complexity of transplantation data also complicate findings. The choice of SLKT versus LTA, and subsequent results of the operation, depend on many factors, such as the organ allocation policy, the availability of matching organs, risk factors that affect potential for graft rejection, and increased medical costs. The transplantation system is very complicated in the US (DeRoos et al., 2019), including comprehensive physical checks for both donors and receivers, strict screening criteria for deceased donor or living donor, and legally required consent and monitoring for both donors and recipients. The need for organ matching before transplant, the distance between donor and recipient, the feasibility of transferring the organs, and the background of patients who need both transplants, all impact the ability to select SLKT versus LTA. A number of other factors unrelated to the transplant method (e.g. potential risk for graft rejection, lifestyle and psychological factors) further complicate the results in a way that is difficult to account for in the analysis (Galletta et al., 2016). Data in transplantation are often limited, thus further limiting the utility of observational data to estimate the causal relationship and comparative effectiveness of the two treatments.

The current study is informative in three points. First, IV analysis is potentially more useful in the clinical setting than other statistical approaches when unmeasured confounding threatens our ability to make causal inferences. Second, it is important to choose a valid and strong IV in the analysis. The assumptions of IV analysis need to be assessed and hold for IVs to be useful for a given clinical problem. Last, due to the availability of organ donation and complexity of transplantation, it is difficult to use any single statistical method to compare the effectiveness of

SLKT and LTA and estimates the causal effect. The real-world challenges and complexities of the organ allocation system need to be considered in interpreting the study results. This thesis is a preliminary study of an IV analysis in the context of observational data, which yielded some potentially useful findings for informing clinical decisions for liver transplant patients on dialysis.

Appendix A Region Partition

Region 1: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Eastern Vermont

Region 2: Delaware, District of Columbia, Maryland, New Jersey, Pennsylvania, West Virginia, Northern Virginia

Region 3: Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, Puerto Rico

Region 4: Oklahoma, Texas

Region 5: Arizona, California, Nevada, New Mexico, Utah

Region 6: Alaska, Hawaii, Idaho, Montana, Oregon, Washington

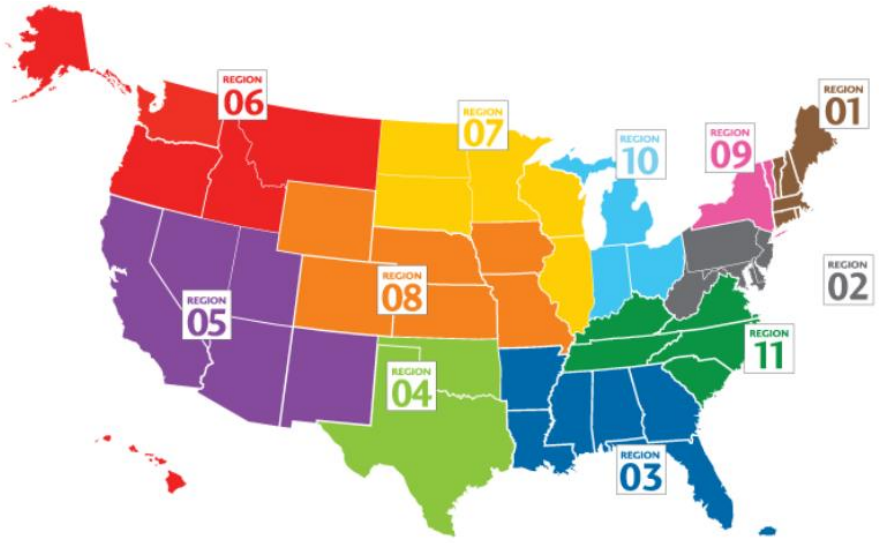
Region 7: Illinois, Minnesota, North Dakota, South Dakota, Wisconsin

Region 8: Colorado, Iowa, Kansas, Missouri, Nebraska, Wyoming

Region 9: New York, Western Vermont

Region 10: Indiana, Michigan, Ohio

Region 11: Kentucky, North Carolina, South Carolina, Tennessee, Virginia



Appendix B Code

```
#Instrumental Variable methods

#install.packages("haven")
#install.packages("AER")
#install.packages("gmodels")

setwd("C:/Users/mengqi/Documents/MasterSpring2020/Capstone/Thesis")
library(haven)
data <- read_dta("editedPSdata.dta")
summary(data)
colnames(data)
dim(data)
table(data$tx_typ)

# data pre-processing
myvars<-c("tx_typ", "GENDER", "diab", "region", "race", "ctr_code",
        "bmi_tcr", "age", "death", "tx_date", "ptime")
data_omit<-data[myvars]
data_omit<-na.omit(data_omit)

# SUBSETTING DATA
# 1 yr followup
d1<-subset(data_omit, tx_date < "2015-01-01")
d1_death<-subset(d1, death=="1")
d1_alive<-subset(d1, death=="0")
# working dataset is data_omit
data_omit<-d1

remove(d1)
remove(d1_alive)
remove(d1_death)

# NEW DEATH VARIABLE
#table for deaths and survival time <= 1 year
table(data_omit$death, data_omit$ptime<=365)

#adding indicator for death within 1 year
#data_omit$death1yr<-data_omit$death
data_omit$death1yr<-ifelse(data_omit$ptime<=365,
                    ifelse(data_omit$death==1, 1, 0),0)
```

```
#checking count of subjects with death within 1 year
working<-data_omit$death1yr[data_omit$death1yr==1]
remove(working)

# SUMMARY STATISTICS
table(data_omit$death)
table(data_omit$region)

# recode region 5 as region 0 as reference group
#data_omit$region<-ifelse(data_omit$region==5, 0, data_omit$region)


#recode as factor variables
data_omit$region<-factor(data_omit$region)
data_omit$race<- factor(data_omit$race)
data_omit$diab <- factor(data_omit$diab)
data_omit$death <- factor(data_omit$death)
data_omit$death1yr <- factor(data_omit$death1yr)
data_omit$GENDER <- factor(data_omit$GENDER)
data_omit$tx_typ <- factor(data_omit$tx_typ)

# DESCRIPTIVES
summary(data_omit$bmi_tcr)
sd(data_omit$bmi_tcr)
summary(data_omit$age)
sd(data_omit$age)
table(data_omit$diab)
table(data_omit$race)

#recode diab to ordered 0=no 1=typeII 2=typeI
data_omit$diabNew<-data_omit$diab
data_omit$diabOld<-data_omit$diab

data_omit$diabNew[data_omit$diabNew==5]<-NA
data_omit$diabNew[data_omit$diabNew==998]<-NA

#remove missing (1965 obs left)
data_omit<-na.omit(data_omit)
#drop empty levels
data_omit$diabNew<-droplevels(data_omit$diabNew)
data_omit$diabNew<-ifelse(data_omit$diabNew==1, 0, 1)
data_omit$diabNew <- factor(data_omit$diabNew)
table(data_omit$diabNew)

# dichotomize bmi, age, race, and diab
data_omit$bmi_tcr <- ifelse(data_omit$bmi_tcr<30, 0, 1)
```

```
data_omit$bmi_tcr <- factor(data_omit$bmi_tcr)
data_omit$age <- ifelse(data_omit$age<55, 0,1)
data_omit$age <- factor(data_omit$age)
data_omit$race <- ifelse(data_omit$race==0 | data_omit$race==1, 1,0)
data_omit$race <- factor(data_omit$race)

library(gmodels)
CrossTable(data_omit$region,data_omit$death1yr)
table(data_omit$tx_typ, data_omit$death,data_omit$region)

# Approach 1: recode region 3,4,7,8,10,11 as yes, others as no
data_omit$region.high.SKLT<-
ifelse(data_omit$region==1|data_omit$region==2|data_omit$region==5|data_omit$region==6|da
ta_omit$region==9, 0, 1)
data_omit$region.high.SKLT<-factor(data_omit$region.high.SKLT)

# typical logistic regression
library(aod)

#Logistic regression model of treatment with IV
model.xz <- glm(tx_typ ~ region.high.SKLT, family = binomial(link = 'logit'), data =
data_omit)
model.xz
summ.xz <- summary(model.xz)
summ.xz
nmod <- glm(tx_typ~1, family = 'binomial',data = data_omit) ##"null" mod
anova(nmod, model.xz, test = 'Chisq')

#Logistic regression model of death with IV
model.yz <- glm(death1yr ~ region.high.SKLT, family = binomial(link = 'logit'), data =
data_omit)
summ.yz <- summary(model.yz)
summ.yz
nullmod <- glm(death1yr~1, family = 'binomial',data = data_omit) ##"null" mod
anova(nullmod, model.yz, test = 'Chisq')

#Logistic regression model of death with treatment
model.xy <- glm(death1yr ~ tx_typ, family = binomial(link = 'logit'), data = data_omit)
summ.xy <- summary(model.xy)
summ.xy
anova(nullmod, model.xy, test = 'Chisq')

#Logistic regression model of death with treatment and IV
model.xyz <- glm(death1yr ~ tx_typ+region.high.SKLT, family = binomial(link = 'logit'),
data = data_omit)
summ.xyz <- summary(model.xyz)
```

```
summ.xyz
anova(nullmod, model.xyz, test = 'Chisq')

#Logistic regression model of death with treatment and xmat
model1 <- glm(death1yr ~ tx_typ+age+race+GENDER+bmi_tcr+diabNew, family = binomial(link = 'logit'), data = data_omit)
summ <- summary(model1)
summ
anova(nullmod, model1, test = 'Chisq')

#Logistic regression model of death with treatment, IV and xmat
model2 <- glm(death1yr ~ tx_typ+region.high.SKLT+age+race+GENDER+bmi_tcr+diabNew, family = binomial(link = 'logit'), data = data_omit)
summ <- summary(model2)
summ
anova(nullmod, model2, test = 'Chisq')

# The IV Analysis

library(ivpack)
# y is the nx1 vector of the outcome (mortality)
# d is the nx1 vector of the treatment (1 if SKLT, 0 if LTA)
# xmat is the nxp matrix of observed covariates (e.g., gender, age, race, etc.)
# z is the IV
# (1 if regions with more SKLT, 0 regions with less SKLT)

# Fit first stage model
y <- as.numeric(data_omit$death1yr)-1
d <- as.numeric(data_omit$tx_typ)-1
vars <- c("GENDER", "diabNew", "race", "bmi_tcr", "age")
xmat <- data.matrix(data_omit[vars])
z <- as.numeric(data_omit$region.high.SKLT)-1

# Fit first stage model
first.stage.model=lm(d ~ z+xmat)
summary(first.stage.model)
# Calculate Partial F statistic for testing whether instrument has an effect
# in the first stage model
first.stage.model.without.z=lm(d ~ xmat)
summary(first.stage.model.without.z)
anova(first.stage.model.without.z,first.stage.model)

# The partial F statistic is 39.844, which is much greater than 10,
# so that IV is strong enough for two stage least squares inference to be reliable.
```

```
# Estimate proportion of compliers
first.stage.model.logistic=glm(d ~ z+xmat,family=binomial)
newdata.z1=data.frame(z=rep(1,length(z)),xmat);
expected.treatment.zequal1=predict(first.stage.model.logistic,
                        newdata=newdata.z1,type="response");
newdata.z0=data.frame(z=rep(0,length(z)),xmat)
expected.treatment.zequal0=predict(first.stage.model.logistic,
                        newdata=newdata.z0,type="response");
proportion.compliers=mean(expected.treatment.zequal1-expected.treatment.zequal0);
proportion.compliers
# 0.1378
# We estimate that 13.78% of the subjects are compliers


# Two stage least squares analysis
ivmodel=ivreg(y ~ d+xmat | z + xmat)
# This summary gives the non-robust standard errors
summiv <- summary(ivmodel)
summiv
summIVtable <- as.data.frame(summiv$coefficients)
summIVtable$OR <- exp(coef(ivmodel))
summIVtable$CI <- exp(confint(ivmodel))
write.csv(summIVtable, paste(" IV regression ",".csv",sep=""))
# We estimate that the effect of going to a SKLT is to
# reduce the mortality rate for compliers by 0.02857 or 28.57 patients per 1000
# subjects who are on the waiting list.

# Standard errors that are robust for heteroskedasticity but not clustering
robust.se(ivmodel)

# Huber-White standard errors that account for clustering due to transplant center
# and are also robust to heteroskedasticity
# ctr_code is transplant center code
cluster.robust.se(ivmodel,data_omit$ctr_code)

# Imbalance measurement
CrossTable(data_omit$GENDER,data_omit$region.high.SKLT, prop.chisq = FALSE, chisq = TRUE)
CrossTable(data_omit$diabNew,data_omit$region.high.SKLT, prop.chisq = FALSE, chisq = TRUE)
CrossTable(data_omit$race,data_omit$region.high.SKLT, prop.chisq = FALSE, chisq = TRUE)
CrossTable(data_omit$bmi_tcr,data_omit$region.high.SKLT, prop.chisq = FALSE, chisq = TRUE)
CrossTable(data_omit$age,data_omit$region.high.SKLT, prop.chisq = FALSE, chisq = TRUE)
```

```
# Approach 2: only keep region 5 as no, region 7 as yes
data_omit1 <- subset(data_omit, region== "5" | region == "7")
data_omit1$region.extreme<-ifelse(data_omit1$region==7, 1, 0)

#recode as factor variables
data_omit1$region.extreme<-factor(data_omit1$region.extreme)

#Logistic regression model of treatment with IV
model.xz <- glm(tx_typ~region.extreme, family = binomial(link = 'logit'), data = data_omit1)
model.xz
summ.xz <- summary(model.xz)
summ.xz
nmod <- glm(tx_typ~1, family = 'binomial',data = data_omit1) ##"null" mod
anova(nmod, model.xz, test = 'Chisq')

#Logistic regression model of death with IV
model.yz <- glm(death1yr ~ region.extreme, family = binomial(link = 'logit'), data = data_omit1)
summ.yz <- summary(model.yz)
summ.yz
nullmod <- glm(death1yr~1, family = 'binomial',data = data_omit1) ##"null" mod
anova(nullmod, model.yz, test = 'Chisq')

#Logistic regression model of death with treatment
model.xy <- glm(death1yr ~ tx_typ, family = binomial(link = 'logit'), data = data_omit1)
summ.xy <- summary(model.xy)
summ.xy
anova(nullmod, model.xy, test = 'Chisq')

#Logistic regression model of death with treatment and IV
model.xyz <- glm(death1yr ~ tx_typ+region.extreme, family = binomial(link = 'logit'), data = data_omit1)
summ.xyz <- summary(model.xyz)
summ.xyz
anova(nullmod, model.xyz, test = 'Chisq')

#Logistic regression model of death with treatment and xmat
model1 <- glm(death1yr ~ tx_typ+age+race+GENDER+bmi_tcr+diabNew,
        family = binomial(link = 'logit'), data = data_omit1)
summ <- summary(model1)
summ
anova(nullmod, model1, test = 'Chisq')
```

```
#Logistic regression model of death with treatment, IV and xmat
model2<-glm(death1yr~tx_typ+region.extreme+age+race+GENDER+bmi_tcr+diabNew,
          family = binomial(link = 'logit'), data = data_omit1)
summ <- summary(model2)
summ
anova(nullmod, model2, test = 'Chisq')

# The IV Analysis

y2 <- as.numeric(data_omit1$death1yr)-1
d2 <- as.numeric(data_omit1$tx_typ)-1
xmat2 <- data.matrix(data_omit1[vars])
z2 <- as.numeric(data_omit1$region.extreme)-1

# Fit first stage model
first.stage.model=lm(d2 ~ z2+xmat2)
# Calculate Partial F statistic for testing whether instrument has an effect
# in the first stage model
first.stage.model.without.z=lm(d2 ~ xmat2)
anova(first.stage.model.without.z,first.stage.model)

# The partial F statistic is 49.08, which is greater than 10,
# so that IV is valid for two stage least squares inference to be reliable.

# Estimate proportion of compliers
first.stage.model.logistic=glm(d2 ~ z2+xmat2,family=binomial)
newdata.z1=data.frame(z2=rep(1,length(z2)),xmat2);
expected.treatment.zequal1=predict(first.stage.model.logistic,
                        newdata=newdata.z1,type="response");
newdata.z0=data.frame(z2=rep(0,length(z2)),xmat2)
expected.treatment.zequal0=predict(first.stage.model.logistic,
                        newdata=newdata.z0,type="response");
proportion.compliers=mean(expected.treatment.zequal1-expected.treatment.zequal0);
proportion.compliers
# 0.2363
# We estimate that 23.63% of the subjects are compliers


# Two stage least squares analysis
ivmodel=ivreg(y2 ~ d2 + xmat2 | z2 + xmat2)
# This summary gives the non-robust standard errors
summiv <- summary(ivmodel)
summiv
summIVtable <- as.data.frame(summiv$coefficients)
summIVtable$OR <- exp(coef(ivmodel))
summIVtable$CI <- exp(confint(ivmodel))
```

```
write.csv(summIVtable, paste(" IV regression ",".csv",sep=''))

# We estimate that the effect of going to a SKLT is to
# increase the mortality rate for compliers by 0.01905 or 19.05 patients per 1000
# subjects who are on the waiting list.

# Standard errors that are robust for heteroskedasticity but not clustering
robust.se(ivmodel)

# Huber-White standard errors that account for clustering due to transplant center
# and are also robust to heteroskedasticity
# ctr_code is transplant center code
cluster.robust.se(ivmodel,data_omit1$ctr_code)

# Imbalance measurement
CrossTable(data_omit1$GENDER,data_omit1$region.extreme, prop.chisq = FALSE, chisq = TRUE)
CrossTable(data_omit1$diabNew,data_omit1$region.extreme, prop.chisq = FALSE, chisq = TRUE)
CrossTable(data_omit1$race,data_omit1$region.extreme, prop.chisq = FALSE, chisq = TRUE)
CrossTable(data_omit1$bmi_tcr,data_omit1$region.extreme, prop.chisq = FALSE, chisq = TRUE)
CrossTable(data_omit1$age,data_omit1$region.extreme, prop.chisq = FALSE, chisq = TRUE)
```

Bibliography

Angrist, J. and Pischke, J. "Mostly Harmless Econometrics: An Empiricist's Companion." Princeton University Press. (2009). Princeton, NJ.

Baiocchi, M., Cheng, J., and Small, D. S. "Tutorial in Biostatistics: Instrumental variable methods for causal inference." Statistics in Medicine. 33, 13. (2014): 2297-2340. DOI: 10.1002/sim.6128.

"Chronic Kidney Disease Basics," Centers for Disease Control and Prevention (CDC), accessed December 12, 2019, https://www.cdc.gov/kidneydisease/basics.html.

DeRoos, L. J., Marrero, W. J., Tapper, E. B., et al. "Estimated Association Between Organ Availability and Presumed Consent in Solid Organ Transplant." JAMA Network. 2, 10. (2019). DOI: 10.1001/jamanetworkopen.2019.12431.

Galletta, D., Lauria, I., Longobardi, T., et al. "The Complexity of Life Donor Renal Transplantation: The Role and Effectiveness of Multidisciplinary Approach in the Medical Route and Psychological Operation." Journal of Clinical Research Bioethics. 7, 5. (2016). DOI: 10.4172/2155-9627.1000289.

Grossman, J. "The randomized controlled trial gold standard, or merely standard?" Perspectives in Biology and Medicine. 48, 4. (2005): 516-534. DOI: https://doi.org/10.1353/pbm.2005.0092.

Jiang, Y. and Small, D. (2014). ivpack: Instrumental Variable Estimation. R package version 1.2. https://CRAN.R-project.prg/package=ivpack.

Lewis, M. A. "Comment on frameworks for causal inference." Journal of Family Violence. 34, 8. (2019): 711-714. DOI: https://doi.org/10.1007/s10896-019-00046-2.

Lousdal, M. L. "An introduction to instrumental variable assumptions, validation and estimation." Emerging Themes in Epidemiology. 15, 1. (2018): 1-7. DOI: https://doi.org/10.1186/s12982-018-0069-7.

Lu, C. Y. "Observational studies: a review of study designs, challenges and strategies to reduce confounding." Clinical Practice. 63, 5. (2009): 691-697. DOI: https://doi.org/10.1111/j.1742-1241.2009.02056.x.

Pokropek, A. "Introduction to instrumental variables and their application to large-scale assessment data." Large-scale Assessments in Education. 4, 4. (2016): 1-20. DOI: https://doi.org/10.1186/s40536-016-0018-2

Robins, J. M. “Marginal structural models versus structural nested models as tools for causal inference.” Statistical models in Epidemiology, the Environment, and Clinical trials. 116. (2000): 95-133. DOI: https://doi.org/10.1007/978-1-4612-1284-3_2.

Rosenbaum, P. R. and Rubin, D. B. “The central role of the propensity score in observational study for causal effects.” Biometrika. (1983): 41-45.

Rubin, D. B. “Causal inference using potential outcomes.” American Statistical Association. 100, 469. (2005): 322-331. DOI: https:// doi.org/10.1198/016214504000001880.

Sox, H. C. “Comparative effectiveness research: a report from the institute of medicine.” Ann Intern Med. 151, 3. (2009): 203-205. DOI: 10.7326/0003-4819-151-3-200908040-00125.

Spieth, M. P., Kubasch, S. A., Penzlin, I. A., et al. “Randomized controlled trials – a matter of design.” Dove Medical Press. 12. (2016): 1341-1349. DOI: 10.2147/NDT.D101938.

Srivastava, A. “Impact of the treatment assignment model on propensity score-based methods” University of Pittsburgh master’s thesis. (2019).

“Transplant Trends,” United National for Organ Sharing (UNOS), accessed December 31, 2019, https://unos.org/data/transplant-trends/.

Warnes, G. R., Bolker, B., Lumley, T., Johnson, R. C. Contributions from Randall C. Johnson are Copyright SAIC-Frederick, Inc. Funded by the Intramural Research Program, of the NIH, National Cancer Institute and Center for Cancer Research under NCI Contract NO1-CO-12400. (2018). Gmodels: Various R Programming Tools for Model Fitting. R package version 2.18.1. https://CRAN.R-project.org/package=gmodels.

West, S. G. and Thoemmes, F. “Campbell’s and Ruin’s perspectives on causal inference.” Psychol Methods. 12, 1. (2010): 18-37. DOI: 10.1037/a0015917.