University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2016 *Paper* 119

Strengthening Instrumental Variables through Weighting

Douglas Lehmann* Yun Li†
Rajiv Saran‡ Yi Li**

*The University Of Michigan, lehmannd@umich.edu
†University of Michigan School of Public Health, yunlisph@umich.edu
‡University of Michigan School of Public Health, rsaran@med.umich.edu
**University of Michigan School of Public Health, yili@med.umich.edu

http://biostats.bepress.com/umichbiostat/paper119

Strengthening Instrumental Variables through Weighting

Douglas Lehmann, Yun Li, Rajiv Saran, and Yi Li

Abstract

Instrumental variable (IV) methods are widely used to deal with the issue of unmeasured confounding and are becoming popular in health and medical research. IV models are able to obtain consistent estimates in the presence of unmeasured confounding, but rely on assumptions that are hard to verify and often criticized. An instrument is a variable that influences or encourages individuals toward a particular treatment without directly affecting the outcome. Estimates obtained using instruments with a weak influence over the treatment are known to have larger small-sample bias and to be less robust to the critical IV assumption that the instrument is randomly assigned. In this work, we propose a weighting procedure for strengthening the instrument while matching. Through simulations, weighting is shown to strengthen the instrument and improve robustness of resulting estimates. Unlike existing methods, weighting is shown to increase instrument strength without compromising match quality. We illustrate the method in a study comparing mortality between kidney dialysis patients receiving hemodialysis or peritoneal dialysis as treatment for end stage renal disease.

Strengthening Instrumental Variables Through Weighting

Douglas Lehmann[1] Yun Li[1] Rajiv Saran[2] Yi Li[1]

March 26, 2016

Abstract

Instrumental variable (IV) methods are widely used to deal with the issue of unmeasured confounding and are becoming popular in health and medical research. IV models are able to obtain consistent estimates in the presence of unmeasured confounding, but rely on assumptions that are hard to verify and often criticized. An instrument is a variable that influences or encourages individuals toward a particular treatment without directly affecting the outcome. Estimates obtained using instruments with a weak influence over the treatment are known to have larger small-sample bias and to be less robust to the critical IV assumption that the instrument is randomly assigned. In this work, we propose a weighting procedure for strengthening the instrument while matching. Through simulations, weighting is shown to strengthen the instrument and improve robustness of resulting estimates. Unlike existing methods, weighting is shown to increase instrument strength without compromising match quality. We illustrate the method in a study comparing mortality between kidney dialysis patients receiving hemodialysis or peritoneal dialysis as treatment for end stage renal disease.

1 Introduction

The randomized controlled trial (RCT) has long been considered the gold standard for obtaining treatment effects. When the treatment has been randomized to subjects it is reasonable to assume that measured and unmeasured risk factors will balance between groups, and treatment effects can be obtained through direct comparisons. While this is a major benefit of RCTs, they can be costly, and in some cases it is impossible or unethical to randomize the treatment. Observational data are a popular alternative to RCTs but come at the cost of removing control over treatment assignment from the hands of the researcher, giving rise to the possibility that treatment groups will differ in unmeasured ways that confound the relationship between treatment and outcome. Statistical methods that ignore this unmeasured confounding may give biased and misleading results (VanderWeele and Arah, 2011; Baiocchi et al., 2014). This is a primary concern in any observational study, and much research has gone into this problem.

Instrumental variable (IV) methods are widely used to deal with this issue of unmeasured confounding. These methods rely on an additional variable, termed the instrument, that influences or encourages individuals toward the treatment and only affects the outcome indirectly through its effect on treatment. In this sense, the instrument mimics randomization by randomly "assigning" individuals to different likelihoods of receiving treatment. Instruments with little influence over treatment assignment are termed weak instruments, and there are a number of problems associated with using them. Results obtained when using weak instruments suffer from greater small sample bias, and are less robust to violations of the key assumption that the instrument is randomly assigned or independent of unmeasured confounders Bound et al. (1995); Small and Rosenbaum (2008). This assumption cannot be verified and is often criticized, thus using a strong instrument is important for obtaining credible results.

The literature relating to weak instrumental variables has primarily focused on detailing the problems and limitations associated with using them. See, for example, Bound et al. (1995), Staiger and Stock (1994), Angrist et al. (1996), Small and Rosenbaum (2008), or Baiocchi et al. (2014). Variable selection methods to select a strong subset among a pool of weak instruments have been proposed (Belloni et al., 2010; Caner and Fan, 2010; Belloni et al., 2012). For working with a single weak instrument, (Baiocchi et al., 2010) proposed near-far matching, a novel method to extract a smaller study with a stronger instrument from a larger study (see also Baiocchi et al. (2012); Zubizarreta et al. (2013)). This matching-based IV methodology aims to construct pairs that are "near" on covariates but "far" in the instrument. In other words, pairs consist of subjects with similar characteristics who have received substantially different amounts of encouragement toward the treatment, with a greater difference indicating a stronger instrument. This difference is increased in near-far matching using penalties to discourage pairs with similar instrument values, while allowing a certain number of individuals to be removed from the analysis entirely. This results in a stronger instrument across a smaller number of pairs. One limitation of near-far matching is that it may strengthen the instrument at the cost of match quality.

We propose weighted IV-matching, an alternative for strengthening the instrument within this IV-matching framework. Rather than using penalties to discourage pairs who received similar encouragement, we suggest strengthening the instrument after matches have been formed through weighting, with a pair's weight being a function of the instrument within that pair. A fundamental difference between these two techniques is the stage at which the instrument is strengthened. Weighted IV-matching strengthens the instrument after matches have been formed, allowing the matching algorithm to focus only on creating good matches with similar covariate values. Near-far matching, on the other hand, strengthens the instrument and matches on covariates simultaneously, requiring the algorithm to share priority between the two goals. This generally leads to better quality matches for weighted

IV-matching, a major benefit since failing to properly match on observed confounders may lead to bias in estimation.

We illustrate these methods with a comparison of hemodialysis (HD) and peritoneal dialysis (PD) on six-month mortality among patients with end stage renal disease (ESRD) using data from the United States Renal Data System (USRDS). PD has several benefits over HD, including cost benefits, an improved quality of life, and the preservation of residual renal function (Marrón et al., 2008; Tam, 2009; Goodlad and Brown, 2013). Despite this, PD remains underutilized in the United States (Jiwakanon et al., 2010). One explanation for this may be a lack of consensus regarding the effect of PD on patient survival. An RCT to investigate this question was stopped early due to insufficient enrollment (Korevaar et al., 2003). Many observational studies suggest that PD is associated with decreased mortality though results are often conflicting (Heaf et al., 2002; Vonesh et al., 2006; Weinhandl et al., 2010; Mehrotra et al., 2011; Kim et al., 2014; Kumar et al., 2014). Complicating the issue is a strong selection bias, with PD patients tending to be younger and healthier than HD patients. Studies have dealt with this issue by measuring and controlling for important confounders, but to our knowledge none have addressed the possibility of unmeasured confounding that likely remains. We define PD as the treatment and consider a binary outcome for six-month survival. The focus on six-month survival is to study the influence of initial dialysis modality on early mortality, which tends to be high for dialysis patients. Studying early mortality can provide guidance for selecting the initial dialysis modality in order to reduce this early mortality. See, for example, Noordzij and Jager (2012), Sinnakirouchenan and Holley (2011), Heaf et al. (2002).

A possible instrument in the data is the mean PD usage at the facility level. Instruments based on mean treatment usage in a geographic region, facility, or other group are often called preference-based instruments (Brookhart and Schneeweiss, 2007; Li et al., 2015), because it is believed that these groups may have preferences that at least partially override both

measured and unmeasured patient characteristics when making treatment decisions. In other words, facilities with high PD usage are more likely to "encourage" their patients towards PD than those with low usage. Preference-based instruments are among the most commonly used instruments in practice (Garabedian et al., 2014), and methods to improve upon them may have broad applications.

The remainder of this article is organized as follows. In section 2 we outline the proposed weighted IV-matching procedure and briefly compare it to near-far matching. Inference and sensitivity are discussed in section 3. The finite sample performance of these methods are compared in section 4 through simulation, and they are illustrated with a data analysis in section 5. We conclude with a discussion in section 6.

2 Weighted IV-Matching

We begin with an outline of the IV-matching framework (Baiocchi et al., 2010, 2012) and then propose weighted IV-matching for strengthening the instrument within this framework. We briefly compare weighted IV-matching with near-far matching and highlight key differences.

With a preference-based instrument, two rounds of matching are implemented (Baiocchi et al., 2012). In the context of our motivating data example, an optimal non-bipartite matching algorithm first pairs facilities (Derigs, 1988; Lu et al., 2011). After facilities have been paired, the instrument is dichotomized into encouraging and unencouraging. This is done by comparing instrument values within each facility pair and considering the facility with the higher value to be an encouraging facility and the other to be an unencouraging facility. An optimal bipartite matching algorithm then pairs patients at the PD encouraging facility with patients in the other. This results in I pairs of two subjects with similar patient and facility characteristics that received different levels of encouragement toward PD. Instrument strength can be assessed by the average difference, or separation, of this

encouragement across pairs. For example, the instrument is considered stronger in a study in which the average encouraged and unencouraged subjects were treated at facilities with 85% and 30% treatment usage compared to one with average treatment usage of 60% and 45%.

Creating a stronger instrument in this framework is thus equivalent to increasing this separation. We propose increasing this separation by assigning more weight to pairs more likely to be influenced by the instrument. Specifically, we propose weighting by the probability that the encouraged subject receives the treatment while the unencouraged subject receives the control. This can be thought of as the probability that a pair "complies" with encouragement, and giving more weight to pairs more likely to comply creates a stronger instrument across all pairs. Without loss of generality, assume subject j in pair i was treated at the encouraging facility and subject j' at the unencouraging facility, with $Z_{ij} = 1$ indicating encouragement and $Z_{ij'} = 0$ indicating unencouragement. Let D_{ij} indicate treatment received. The weight for pair i is then calculated as

$$w_i = P(D_{ij} = 1|Z_{ij} = 1)P(D_{ij'} = 0|Z_{ij'} = 0). \tag{1}$$

Similar to separation of the instrument, this probability is a measure of instrument strength, though rather than an average across all pairs it is a measure of the influence the instrument within pair i. A stronger instrument is created when more weight is given to pairs in which the instrument has more influence over treatment. This has the effect of redistributing the data in a way to highlight "good" pairs that are more influenced by the instrument and increases separation of the instrument in the process.

In practice, the probabilities in equation (1) are unlikely to be known but will need to be estimated. Using facility level mean PD usage as the instrument, $P(D_{ij} = 1|Z_{ij} = 1)$ is estimated by the mean PD usage at the encouraging facility, while $P(D_{ij'} = 0|Z_{ij'} = 0)$ is

estimated with one minus the mean PD usage at the unencouraging facility. Weights can be standardized to maintain the effective sample size and statistical power.

The near-far matching procedure of Baiocchi et al. (2010, 2012) forces separation of the instrument in the matching process. This is done in the first round by adding a penalty to the distance measure between facilities whose instrument values are within a certain threshold, and allowing a certain number to be removed. This requires the matching algorithm to pair facilities with similar covariates and enforce separation of encouragement simultaneously, and generates an implicit tradeoff. A large penalty will dominate the distance used to reflect similarity on covariates, thereby increasing instrument separation but at the expense of match quality, whereas a small penalty may get overshadowed by the covariate distance, leading to better matches, but with less separation. Removing a number of facilities serves to alleviate some of the damage to match quality, though it may not be entirely preserved since the algorithm is still sharing priority between creating good matches and enforcing instrument separation.

A fundamental difference between weighted IV-matching and near-far matching is the stage in which the instrument is strengthened. Weighted IV-matching strengthens the instrument after matches have been formed, allowing the matching algorithm to focus only on creating good matches with similar covariate values. Near-far matching, on the other hand, strengthens the instrument in the matching process, forcing the algorithm to balance creating good matches and enforcing separation of the instrument. This difference highlights a theme that we will see when comparing the performance of these two techniques; in a tradeoff between match quality and instrument strength, weighted IV-matching tends to favor match quality while near-far matching tends to favor instrument strength. Strength in either of these areas has implications on the resulting analysis.

3 Inference and Sensitivity

3.1 Notation

We define causal effects of interest using the potential outcomes framework (Neyman, 1923; Rubin, 1974; Angrist et al., 1996). Let $Z_{ij} = 1$ if subject j in pair i is encouraged toward treatment, $Z_{ij} = 0$ otherwise. Let $D_{ij}(Z_{ij})$ indicate treatment received for subject j in pair i given their encouragement, and let $Y_{ij}(Z_{ij}, D_{ij})$ indicate mortality. $D_{ij}(Z_{ij})$ and $Y_{ij}(Z_{ij}, D_{ij})$ are referred to as a subjects "potential outcomes." For encouraged subjects, with $Z_{ij} = 1$, we observe treatment $D_{ij}(1)$ and response $Y_{ij}(1, D_{ij})$. Similarly for unencouraged subjects, we observe $D_{ij}(0)$ and response $Y_{ij}(0, D_{ij})$. Our interest lies in estimating the parameter

$$\lambda = \frac{\sum_i \sum_j \left(Y_{ij}(1, D_{ij}) - Y_{ij}(0, D_{ij})\right)}{\sum_i \sum_j \left(D_{ij}(1) - D_{ij}(0)\right)}. \tag{2}$$

This parameter is often referred to as the local average treatment effect (Imbens and Angrist, 1994; Angrist et al., 1996). In contrast to an average treatment effect, which is applicable to the entire population, the local effect is interpreted as an average treatment effect among a subgroup of the population known as "compliers." Depicted in Table 1, compliers are individuals that will take the treatment that they are encouraged to take. Unfortunately,

Table 1: Population subgroups defined by the effect of encouragement on treatment. $D(1)$ denotes the treatment a subject will receive if they are encouraged toward treatment, while $D(0)$ denotes the treatment they will receive if they are not.

		$D(1)$	
		0	1
$D(0)$	0	Never-takers	Compliers
	1	Defiers	Always-takers

since we never observe subjects under both states of encouragement, we never observe both $Y_{ij}(1, D_{ij})$ and $Y_{ij}(0, D_{ij})$ or both $D_{ij}(1)$ and $D_{ij}(0)$, and we must estimate λ from the

data. We impose the following five assumptions to aid us in estimation (Angrist et al., 1996; Baiocchi et al., 2014).

A1. Stable Unit Treatment Value Assumption (SUTVA). Often known as no interference, SUTVA requires that individuals' outcomes be unaffected by the treatment assignment of others, and will be violated if spillover effects exist between treatment and control groups. SUTVA allows us to consider a subjects potential outcomes as a function of only their treatment and encouragement, rather than treatment and encouragement assignments across the entire population.

A2. Random assignment of the instrument. The instrument is assumed to be randomly assigned, and this implies that it is independent of any unobserved confounders. It is often stated conditional on measured confounders. This assumption cannot be verified to hold, and weak instruments are especially sensitive to violations (Baiocchi et al., 2014; Bound et al., 1995; Small and Rosenbaum, 2008; Staiger and Stock, 1994).

A3. Exclusion Restriction. The instrument can only affect the outcome through its effect on treatment. This requires that $Y_{ij}(1, D_{ij} = d) = Y_{ij}(0, D_{ij} = d)$ for all i, j and for $d = 0, 1$, which cannot be verified since both potential outcomes are never observed for any individual.

A4. Nonzero association between instrument and treatment. A nonzero association between the instrument and outcome implies that $E[D_{ij}(1) - D_{ij}(0)] \neq 0$.

A5. Monotonicity. The monotonicity assumption states that there are no defiers, or subjects that always do the opposite of what they are encouraged to do, and implies that $D_{ij}(1) \geq D_{ij}(0)$ for all i, j.

Assumptions *A1* and *A2* allow for unbiased estimation of the instruments effect on the outcome and treatment, or the numerator and denominator in (2). By exclusion restriction, never-takers and always-takers do not contribute to estimation since their treatment and response values do not vary with encouragement. Monotonicity ensures that the group of defiers is empty, while a nonzero association between the instrument and treatment ensures

that the group of compliers is not empty. Thus, with the addition of *A3-A5*, λ is interpreted as the average causal effect of the treatment among the compliers. Further discussion of these assumptions can be found in Imbens and Angrist (1994), Angrist et al. (1996), or Baiocchi et al. (2014), among many others.

3.2 Estimation and Inference

Let $Y_{ij} = Z_{ij}Y(1, D_{ij})+(1-Z_{ij})Y(0, D_{ij})$ denote the observed response and $D_{ij} = Z_{ij}D_{ij}(1)+(1 - Z_{ij})D_{ij}(0)$ the observed treatment for subject j in pair i. Estimate λ as

$$\hat{\lambda} = \frac{\sum_{i=1}^{I} \hat{w}_i \sum_{j=1}^{2} [Z_{ij}Y_{ij} - (1 - Z_{ij})Y_{ij}]}{\sum_{i=1}^{I} \hat{w}_i \sum_{j=1}^{2} [Z_{ij}D_{ij} - (1 - Z_{ij})D_{ij}]}. \tag{3}$$

For inferences regarding λ, Baiocchi et al. (2010) develop an asymptotically valid test for the null hypothesis $H_0^{(\lambda)}$. $H_0^{(\lambda)}$ is true under many population distributions, and thus is a composite null hypothesis. The size of a test for a composite null is the supremum over all null hypotheses in the composite null, and a test is considered valid if it has size less than or equal to its nominal level. Using statistics

$$\begin{aligned} T(\lambda_0) &= \frac{1}{I}\sum_{i=1}^{I} \hat{w}_i \left[\sum_{j=1}^{2} Z_{ij}(Y_{ij} - \lambda_0 D_{ij}) - \sum_{j=1}^{2}(1 - Z_{ij})(Y_{ij} - \lambda_0 D_{ij})\right] \\ &= \frac{1}{I}\sum_{i=1}^{I} V_i(\lambda_0) \end{aligned}$$

and

$$S^2(\lambda_0) = \frac{1}{I(I-1)}\sum_{i=1}^{I}[V_i(\lambda_0) - T(\lambda_0)]^2$$

we can test $H_0^{(\lambda)}$ by comparing $T(\lambda_0)/S(\lambda_0)$ to a standard normal cumulative distribution for large I. Inverting this test and solving for $T(\lambda_0)/S(\lambda_0) = 0$ and $\pm$ 1.96 provides an estimate and 95% confidence interval for λ. We refer interested readers to Baiocchi et al.

(2010) for a detailed discussion of this test statistic, its distribution and related issues.

This inference procedure does not, however, provide a standard error estimate. A sandwich type variance estimate can be obtained following a procedure similar to that discussed in Lunceford and Davidian (2004) and Li and Greene (2013). Define the following estimating equations with respect to $\theta = (\mu_{Y_1}, \mu_{Y_0}, \mu_{D_1}, \mu_{D_0}, \boldsymbol{\beta}')$,

$$\mathbf{0} = \sum_{i=1}^{I}\sum_{j=1}^{2}\phi_{ij}(\theta) = \sum_{i=1}^{I}\sum_{j=1}^{2}\begin{bmatrix} w_i Z_{ij}(Y_{ij} - \mu_{Y_1}) \\ w_i(1 - Z_{ij})(Y_{ij} - \mu_{Y_0}) \\ w_i Z_{ij}(D_{ij} - \mu_{D_1}) \\ w_i(1 - Z_{ij})(D_{ij} - \mu_{D_0}) \\ S_\beta(\boldsymbol{\beta}) \end{bmatrix} \tag{4}$$

where $\mu_{Y_1} = E(w_i Z_{ij} Y_{ij})/E(w_i Z_{ij})$, $\mu_{Y_0} = E(w_i(1 - Z_{ij})Y_{ij})/E(w_i(1 - Z_{ij}))$ and similar for μ_{D_1} and μ_{D_0}. $S_\beta(\boldsymbol{\beta})$ correspond to the score equations for estimating parameters $\boldsymbol{\beta}$, often from a logistic regression, for the probabilities used in equation (1) for determining the weight. This procedure allows for simultaneous estimation of w_i and λ. We estimate $var(\hat{\theta})$ with $(2I)^{-1}\hat{A}^{-1}\hat{B}\hat{A}^{-T}$, where $\hat{A} = \sum_{i=1}^{I}\sum_{j=1}^{2}\partial\phi_{ij}(\theta)/\partial\theta|_{\theta=\hat{\theta}}$ and $\hat{B} = \sum_{i=1}^{I}\sum_{j=1}^{2}\phi_{ij}(\theta)\phi_{ij}^{T}(\theta)|_{\theta=\hat{\theta}}$. Applying the multivariate delta method with $g(\theta) = (\mu_{r_T} - \mu_{r_C})/(\mu_{d_T} - \mu_{d_C})$, an estimate of $var(\hat{\lambda})$ is obtained as $\nabla g(\theta)^T \hat{var}(\hat{\theta}) \nabla g(\theta)$. This approach does not take into account the matching process and can be expected to overestimate the variance, though it was found to perform well in simulations. In sections 4 and 5, intervals and coverage results will be based on the permutation inference procedure.

3.3 Sensitivity

An important benefit of working with stronger instruments is the increased robustness of resulting estimates to violations of the assumption that the instrument is randomly assigned

or independent of unmeasured confounders. In this section we describe a sensitivity analysis outlined in Rosenbaum (2002) and applied to IV-matching in Baiocchi et al. (2010, 2012). The goal of this sensitivity analysis is to determine how far an instrument can deviate from being randomly assigned before the qualitative results of the study are altered, with more robust results remaining consistent under larger deviations. In other words, how large would an unmeasured instrument-outcome confounder have to be to explain what appears to be a significant treatment effect?

Following Rosenbaum (2002), deviation from random assignment is quantified by assuming that within pair i matched on covariates $\mathbf{X}$, subjects j and $j^{'}$ differ in their odds of receiving encouragement by at most a factor of $\Gamma \geq 1$, where

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1-\pi_{ij'})}{\pi_{ij'}(1-\pi_{ij})} \leq \Gamma \text{ for all } i, j, j' \text{ with } X_{ij} = X_{ij'} \tag{5}$$

and $\pi_{ij} = P(Z_{ij} = 1|X_{ij})$. Under the random assignment assumption, $\pi_{ij} = \pi_{ij'}$ and $\Gamma = 1$. As the assumption becomes increasingly violated, these probabilities diverge and Γ increases.

The sensitivity analysis is conducted by using Γ in inference procedures to obtain bounds on the p-value associated with testing $H_0 : \lambda = 0$. For matched pairs, this involves comparing the sum of events in the encouraged group among discordant pairs with two binomial distributions, one with probability $p^- = \frac{1}{1+\Gamma}$ and another with probability $p^+ = \frac{\Gamma}{1+\Gamma}$. This is done for increasing values of Γ until a previously rejected H_0 becomes accepted, e.g. a significant effect is no longer significant. The maximum deviation that can be sustained is given by the largest Γ value in which the upper bound on the p-value remains less than 0.05, with larger maximum deviations indicating more robust results. When pairs are weighted, normal approximations to the binomials can be used for obtaining p-values using the weighted sum of events in the encouraged group among discordant pairs.

Rosenbaum and Silber (2009) discuss how the univariate parameter Γ can be mapped to

two components as $\Gamma = (\Delta\Lambda + 1)/(\Delta + \Lambda)$, where Λ represents the effect of an unmeasured confounder on the instrument and Δ the effect of an unmeasured confounder on the outcome. For example, an unmeasured confounder that triples the odds of receiving encouragement ($\Lambda = 3$) while doubling the odds of experiencing the event ($\Delta = 2$) is equivalent to a deviation from random assignment of size $\Gamma = 1.4$. This mapping of Γ allows the sensitivity analysis to remain simple while providing a useful interpretation of its magnitude.

4 Simulation

In this section we compare the finite sample performance of three IV-matching techniques through simulation. The standard IV-match (IVM) uses the full data and makes no attempt to strengthen the instrument, while weighted IV-matching (WIVM) and near-far matching (NFM) will create stronger instruments as described in section 2. For the NFM procedure, we add a penalty to the distance between facilities if their instruments are within a distance equal to the interquartile range of instrument values. As in Baiocchi et al. (2010), we specify a penalty function that begins at 0 and increases exponentially a pairs instrument values become closer, and allow 50% of facilities to be removed during the matching process.

4.1 Setup

One thousand datasets are generated containing $i = 1, ..., 200$ facilities with $j = 1, ..., 40$ subjects at each. Binary treatment D and binary outcome Y are randomly assigned with

$$P(D_{ij} = 1) = logit^{-1}(\gamma_i + \alpha X_{1,i} + \delta X_{2,ij} + \nu_{ij}), \tag{6}$$

$$P(Y_{ij} = 1) = logit^{-1}(\beta D_{ij} + \alpha X_{1,i} + \delta X_{2,ij} + \epsilon_{ij}). \tag{7}$$

$\gamma_i \sim N(0,1)$ represents a facility effect. Standard normal covariates $X_{1,i}$ and $X_{2,ij}$ represent observed confounders and are used for matching. $X_{1,i}$ is a facility level confounder and $X_{2,ij}$ is a patient level confounder. Coefficients α, δ, and β represent the effects of X_1, X_2, and D, respectively. Unobserved confounding is created by generating $(\nu_{ij}, \epsilon_{ij})$ as bivariate normal with correlation $\rho = .75$. The proportion of treated individuals at each facilities serves as the instrument.

To obtain the "true" local average treatment effect that we wish to estimate, or λ in (2), we need counterfactual treatments and responses for every individual. These are not easily obtained under the current setup, since γ, not encouragement, is in equation (6). Furthermore, we do not know which counterfactual state an individual will be considered to have been observed in until after matching, since subjects are determined to have been observed in an encouraging or unencouraging facility by comparing instrument values within pairs. Despite this caveat, suitable counterfactuals can be obtained in the following way.

Consider patients treated at facilities with $\gamma_i > 0$ to be observed in the encouragement state, while those at facilities with $\gamma_i \leq 0$ to be observed in the unencouragement state. For individuals in the encouragement state, we have $D_{ij} = D_{ij}(1)$ and $Y_{ij} = Y_{ij}(1, D_{ij})$ from equations (6) and (7). For counterfactuals, sample a γ from the unencouragement group and denote it γ^*. $D_{ij}(0)$ is then obtained using equation (6) with $P(D_{ij} = 1) = P(D_{ij}(0) = 1) = logit^{-1}(\gamma_i^* + \alpha X_{1,i} + \delta X_{2,ij} + \nu_{ij})$ and $Y_{ij}(0, D_{ij})$ is obtained using equation (7) with $P(Y_{ij} = 1) = P(Y_{ij}(0, D_{ij})) = logit^{-1}(\beta D_{ij}(0) + \alpha X_{1,i} + \delta X_{2,ij} + \epsilon_{ij})$. Counterfactuals for patients observed in the unencouragement state can be obtained similarly. After obtaining $D_{ij}(1)$, $D_{ij}(0)$, $Y_{ij}(1, D_{ij})$, and $Y_{ij}(0, D_{ij})$, these are plugged into equation (2) for the true effect, λ.

4.2 Simulation Results

4.2.1 Instrument Strength

The present work is motivated by the desire to create a stronger instrument by increasing the separation of encouragement within pairs. Table 2 shows that both WIVM and NFM were able to do so, increasing the standardized difference in encouragement approximately 25% and 65%, respectively. All things being equal, the stronger instrument is preferred. Looking at match quality in the next section, however, we will see that all things are not equal.

Table 2: Separation of encouragement within pairs based on 1,000 simulations. Reported is the mean treatment usage at unencouraging facilities ($\bar{Z}_U$), encouraging facilities ($\bar{Z}_E$), and the standardized difference between them, calculated as St Diff $= 100(\bar{Z}_E - \bar{Z}_U)/\sqrt{.5(s^2_{Z_E} + s^2_{Z_U})}$ where $s^2_{Z_E}$ and $s^2_{Z_U}$ are sample variances of the instrument in each group.

	$(\bar{Z}_U, \bar{Z}_E)$	St Diff
IVM	(37%, 62%)	141
WIVM	(35%, 65%)	175
NFM	(30%, 70%)	232

4.2.2 Match Quality

Table 3 reports balance of covariates X_1 and X_2 as indicated by the standardized difference within pairs. The WIVM procedure produced consistently better covariate balance than the NFM procedure. The particularly poor balance of facility level X_1 under the NFM procedure shows that introducing penalties to the match negatively affected the ability to properly match on X_1 in the first round.

The pattern seen in Tables 2 and 3 shows a tradeoff of instrument strength and match quality between WIVM and NFM. WIVM allows the matching algorithm to focus entirely on matching on covariates, and strengthens the instrument through weighting after the matches have been formed. NFM, on the other hand, incorporates penalties into the match

Table 3: Standardized differences in covariates X_1 and X_2 within pairs. Results based on 1,000 simulations.

	(α, δ)	IVM	WIVM	NFM
X_1	$(0, 0)$	0.01	0.01	0.34
	$(0.25, 0.25)$	0.15	0.14	18.01
	$(0.50, 0.50)$	0.14	0.16	36.10
X_2	$(0, 0)$	0.01	0.02	0.10
	$(0.25, 0.25)$	0.58	0.68	1.02
	$(0.50, 0.50)$	1.35	1.57	2.10

to enforce separation of the instrument, requiring the matching algorithm to share priority between matching on covariates and strengthening the instrument. A large penalty might dominate the distance used for matching and diminish the ability to properly match on covariates. In the tradeoff between instrument strength and match quality, WIVM is willing to trade less instrument strength for higher quality matches, while NFM is willing to trade lower quality matches for a stronger instrument. In results that follow, we will see that strength or weakness in either area has important implications on inferences and sensitivity.

4.2.3 Estimation and Coverage

Table 4 presents simulation results relating to estimation and coverage of λ under increasing magnitudes of observed confounding. When α and δ are zero and matching on X_1 and X_2 is trivial, each method is nearly unbiased and maintains nominal coverage. WIVM and NFM achieved lower mean squared error than IVM, which is one benefit associated with stronger instruments (Wooldridge, 2001). As α and δ increase and matching on X_1 and X_2 becomes more important, the performance of IVM and WIVM remain mostly unchanged. NFM, on the other hand, results in increased bias and mean squared errors and low coverage rates, which can be attributed to the inability of the NFM procedure to properly match on X_1.

Table 4: Bias, mean squared error (MSE), and 95% coverage probabilities (CP) for estimation of λ based on 1,000 simulations. Bias and MSE are multiplied by 1,000. Coverage probabilities are based on confidence intervals obtained using the permutation inference procedure discussed in section 3.

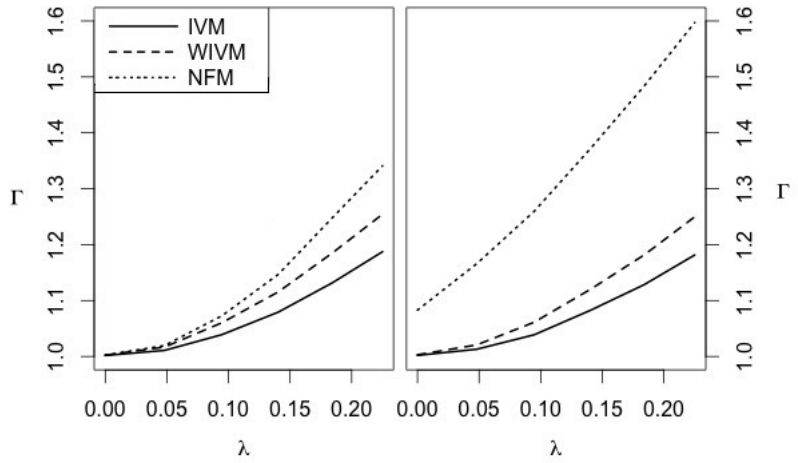
			IVM			WIVM			NFM		
(α, δ)	β	λ	Bias	MSE	CP	Bias	MSE	CP	Bias	MSE	CP
(0, 0)	0.0	0.0	4.6	2.3	94.3	4.5	1.6	93.9	2.5	1.7	94.2
	0.6	0.14	1.4	2.0	94.3	1.3	1.4	95.1	0.3	1.4	95.6
	1.0	0.23	4.6	1.9	94.6	3.7	1.4	95.2	2.5	1.4	95.0
(0.25, 0.25)	0.0	0.0	3.0	2.1	94.8	4.9	1.5	94.6	29.2	2.5	84.9
	0.6	0.14	4.9	1.9	95.2	4.9	1.5	95.2	25.8	2.3	87.6
	1.0	0.23	4.4	1.8	95.0	4.7	1.3	96.0	26.4	2.2	86.5
(0.50, 0.50)	0.0	0.0	8.7	2.4	93.6	9.7	1.7	94.2	93.9	10.5	28.3
	0.6	0.14	8.7	2.3	94.3	7.9	1.7	93.6	88.6	9.7	30.9
	1.0	0.23	4.0	2.0	93.7	3.7	1.5	93.8	78.6	7.8	38.3

4.2.4 Sensitivity

In this section we report simulation results for studying violations of random assignment of the instrument. Figure 1 presents how large a deviation from random assignment estimates are robust to, as defined by Γ. Larger values of Γ correspond with more robust results. These curves are naturally upward sloping since larger effects are more robust, all else being equal.

Results in Figure 1 show that more robust results were obtained after creating stronger instruments. An interesting finding can be seen when comparing results from the left panel in Figure 1 to the right panel. As α and δ increase from 0 in the left panel to 0.5 in the right panel, results for IVM and WIVM are unchanged but those for NFM seem to improve greatly. This apparent improvement arises from the biased estimates obtained after failing to properly match on X_1. These biased estimates appear more robust than their unbiased counterparts, and cause the curve to shift left by about the size of this bias. This serves as a warning that this sensitivity analysis assumes measured confounders have been properly adjusted and we are able to obtain unbiased estimates. Γ can therefore be misleading if match quality is poor.

Figure 1: Sensitivity results based on 1,000 simulations. Lines represent the size of an unobserved bias, as quantified by Γ, that would be required to explain a significant finding. Larger values of Γ correspond with more robust estimates. Left: $(\alpha, \delta) = (0, 0)$, Right: $(\alpha, \delta) = (0.5, 0.5)$



5 Data Analysis

In this section we illustrate IV-matching (IVM), weighted IV-matching (WIVM), and near-far matching (NFM) with a study comparing mortality in the first six months between patients receiving hemodialysis (HD) or peritoneal dialysis (PD) as treatment for end stage renal disease. Complete information on 164,195 adults initiating dialysis for the first time between January 1, 2010 and December 31, 2013 was obtained from the United States Renal Data System. The analysis was restricted to patients being treated at dialysis facilities with at least ten patients that used both HD and PD during the study period. The analysis was conducted as intention-to-treat, with treatment defined as the modality prescribed at the onset of dialysis.

The instrument, facility mean PD usage, was calculated on data from 2007-2009 to avoid correlation with any patient level confounders. The instrument varied greatly across facilities, ranging from 0 to 100% with a mean of 9.8%. The correlation coefficient between a facilities

2007-2009 and 2010-2013 PD usage was 0.68.

Figure 2 and Table A1 of the appendix confirm that patients treated with PD are generally healthier than those treated with HD. On average, they are six years younger, receive more pre-ESRD care, suffer less comorbidities, and are more likely to be employed than HD patients. Additionally, facilities with higher PD usage tend to be larger, as indicated by the higher number of nurses, social workers, and hemodialysis stations. Since these factors could be related to unmeasured confounders that affect patient outcomes, it is important to control for these variables when matching.

5.1 Constructing Matches

We follow the two round matching procedure described in section 2 for constructing matches. An optimal non-bipartite match first pairs facilities. Within each of these pairs, an optimal bipartite match pairs patients from one facility with patients in the other.

For the first round facility level match, we defined the distance between facilities using a Mahalanobis distance based on the facility covariates in Figure 2. For the NFM procedure, a penalty was added to this distance if facilities instrument values were within 14% of each other (the inter-quartile range), and half of facilities were dropped from the analysis. For the second round patient level match, we matched on a prognostic score based on the patient level covariates in Figure 2. For the WIVM procedure, a weight was assigned to each pair based on equation (1), where probabilities were estimated using the instrument, facility mean PD usage from 2007-2009.

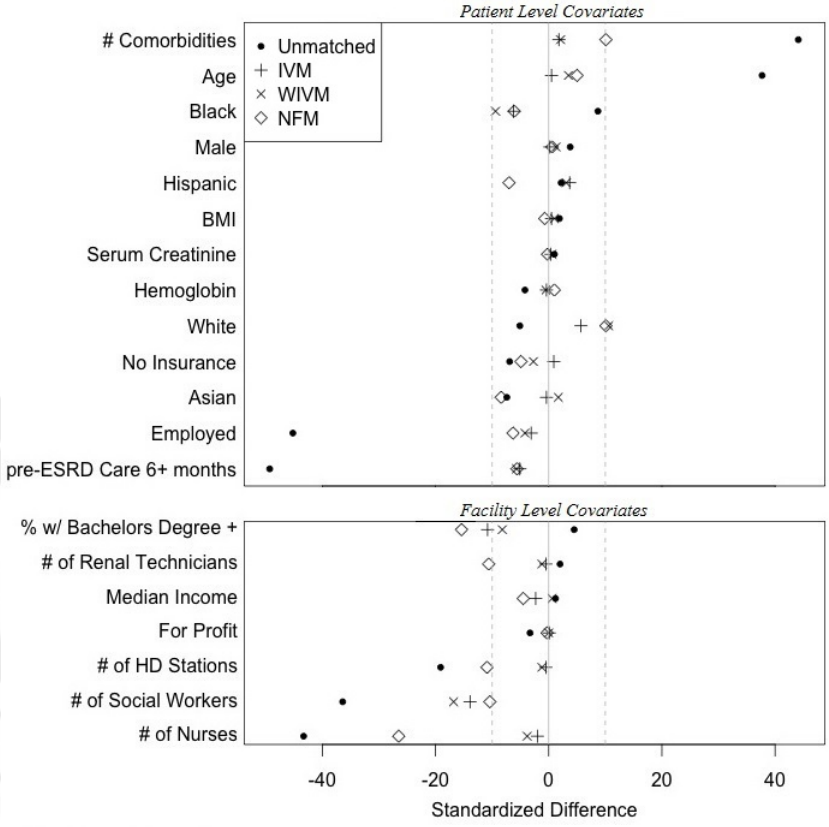
5.2 Results

Of the 164,195 patients, 128,700 were paired using the IVM and WIVM procedure, while 67,904 were paired using the NFM procedure. The average unencouraged and encouraged

patient was treated at a facility with PD usage from 2007-2009 of 4.7% and 15.3% using the IVM procedure, 6.3% and 27.8% using the WIVM procedure, and 3.8% and 25.3% using the NFM procedure. For WIVM and NFM, the increased separation corresponds with roughly a 100% increase in the standardized difference in encouragement, with neither procedure performing notably better than the other in terms of instrument strength.

Covariate balance after matching is presented in Figure 2 as well as Table A2 of the appendix. Each method is seen to improve covariate balance on average. IVM and WIVM, however, generally resulted in better balance than NFM, particularly for facility level covariates where NFM seems to struggle. These results are similar to those seen in the simulations of section 4. Estimation results reported in table **??** indicate that PD has a protective effect

Figure 2: Covariate balance before and after matching as indicated by the standardized differences within pairs. Dashed grey lines are at ±10. Standardized differences larger than this have been suggested to represent an imbalance (Normand et al., 2001).



on mortality in the first six months. For example, $\hat{\lambda} = -0.09$ suggests that for every 100 subjects that are encouraged to switch from HD to PD, there are nine fewer deaths in the first six months. Both WIVM and NFM decreased the width of the confidence interval associated with λ compared to IVM, with NFM leading to the narrowest interval. While WIVM and NFM created equally strong instruments, WIVM ultimately led to the more robust results since NFM estimated a smaller effect. Though results appear similar for each of the three methods in this particular analysis, they could differ quite substantially in other scenarios, particularly when important group level covariates are present or difficult to adjust for.

Table 5: Estimate and 95% confidence interval for the local average treatment effect, λ, as well as sensitivity parameter Γ.

	$\hat{\lambda}$	95% CI	Γ
IVM	-0.09	(-0.14, 0.03)	1.03
WIVM	-0.09	(-0.15, -0.06)	1.09
NFM	-0.07	(-0.10, -0.04)	1.07

6 Discussion

Weak instrumental variables present many problems to an IV analysis. Of particular concern is that results obtained using weak instruments are sensitive to small unmeasured confounders affecting instrument assignment. This problem cannot be alleviated with increasing sample sizes. While we cannot verify the assumption that the instrument is independent of unmeasured confounders, working with stronger instruments increases robustness to violations of it (Bound et al., 1995; Small and Rosenbaum, 2008; Baiocchi et al., 2010).

In this article, we proposed a weighting procedure for building a stronger instrument in the IV-matching framework. The key idea is that we can redistribute the data through weighting to highlight pairs in a way that increases the overall instrument strength. The proposed weights were based on the probability that a pair complies with encouragement,

or that within that pair the encouraged subject received treatment while the unencouraged subject received control. Other weights could be considered, the only requirement being that more weight is assigned to pairs that are more influenced by the instrument. In future work we are considering the possibility of an "optimal" weight, perhaps subject to a constraint on covariate balance.

Compared with existing methods, weighting is able to build a stronger instrument without compromising match quality. This is because weights are applied to strengthen the instrument after matches have been formed, as opposed to methods that strengthen the instrument simultaneously with matching. This is a major strength of the proposed method since failing to properly match on important covariates leads to biased effect estimates and misleading sensitivity results.

Using data from the United States Renal Data System, the proposed method was illustrated in a study comparing mortality in the first six months between patients receiving hemodialysis or peritoneal dialysis as treatment for end state renal disease. The proposed weighting procedure was able to create a stronger instrument while maintaining the integrity of matches. A protective effect of peritoneal dialysis was found, suggesting that there are nine fewer deaths for every 100 patients that are encouraged to switch from hemodialysis to peritoneal dialysis.

While the the current work focused on building a stronger instrument within an IV-matching framework, the idea might not be limited to this setting. In future research, we investigate the use of weighting to increase instrument strength in more common instrumental variable procedures.

Appendix

Table A1: Summary of covariates before matching. Patient level covariates are compared across dialysis modality and facility level covariates are compared across first and fourth quartile of the PD usage.

Patient Covariates	HD	PD	St Diff
N	142,737	21,458	-
Outcome			
Death w/in 6 months	14%	4%	35.7
Covariates			
Age	64	58	37.7
Male	57%	55%	3.8
Bmi	29.6	29.5	1.9
6+ months pre-ESRD care	45%	69%	-49.3
# of comorbidities	2.4	1.9	44.1
Hemoglobin	9.9	10.6	-4.2
Serum creatinine	6.6	6.4	1.0
No insurance	7%	8%	-6.9
White	68%	71%	-5.1
Black	26%	22%	8.7
Asian	4%	5%	-7.4
Hispanic	13%	12%	2.2
Employed	9%	26%	-45.2

Facility Covariates	Q 1	Q 4	St Diff
Instrument			
PD usage	3%	30%	-208
Covariates			
For profit	85%	86%	-3.3
# of nurses	6.7	8.7	-43.3
# of technicians	8.2	8.1	2.0
# of social workers	0.8	1.1	-36.4
# of HD stations	20.3	21.9	-19.1
Median income	$51,086	$50,850	1.2
Bachelors degree +	23.7%	23.4%	4.5

Table A2: Summary of covariates after matching, by matching algorithm. U and E correspond to patients considered to have been treated at unencouraging or encouraging PD facilities

	IVM (64,350 pairs)			WIVM (64,350 pairs)			NFM (33,702 pairs)		
	U	E	St Diff	U	E	St Diff	U	E	St Diff
Instrument									
Facility % PD 2007-09	4.7%	15.3%	-96.3	6.3%	27.8%	-194.8	3.8%	25.3%	-195.2
Treatment									
PD	10.0%	16.3%	-18.7	11.5%	23.5%	-35.7	9.0%	23.8%	-44.1
Outcome									
Died w/in 6 months	11.9%	11.3%	1.7	11.7%	10.7%	3.3	11.8%	10.8%	3.1
Patient Covariates									
Age	62.8	62.7	0.5	62.6	62.1	3.5	63.0	62.2	5.0
Male	57.0%	56.9%	0.2	57.3%	56.7%	1.3	57.0%	56.7%	0.6
BMI	30.5	30.4	0.5	30.5	30.2	1.0	31.0	31.2	-0.7
6+ mos pre-ESRD care	47.8%	50.3%	-5.1	50.2%	53.1%	-5.7	50.4%	53.2%	-5.6
# of comorbidities	2.5	2.4	1.8	2.5	2.4	2.1	2.6	2.4	10.1
Hemoglobin	9.9	10.0	-0.4	9.9	10.0	-0.5	10.0	9.9	1.0
Serum Creatinine	6.6	6.5	0.4	6.7	6.5	0.6	6.5	6.5	0.2
No insurance	7.1%	6.9%	0.9	6.8%	7.5%	-2.7	6.1%	7.4%	-4.9
White	68.6%	65.9%	5.7	68.9%	63.9%	10.6	68.8%	64.1%	10.1
Black	25.3%	28.1%	-6.2	25.1%	29.2%	-9.4	26.7%	29.4%	-6.1
Asian	3.8%	3.8%	0.4	4.1%	3.8%	1.7	2.6%	4.2%	-8.4
Hispanic	13.8%	12.6%	3.7	13.4%	12.4%	3.1	9.9%	12.2%	-7.0
Employed	11.4%	12.4%	-3.1	12.3%	13.7%	-4.2	11.6%	13.6%	-6.3
Facility Covariates									
For profit	84.3	84.3	0.1	81.1	81.1	0.0	83.3	83.4	-0.3
# of nurses	9.1	9.2	-2.0	10.0	10.2	-3.8	9.2	10.7	-26.5
# of technicians	9.8	9.9	-0.5	9.7	9.8	1.2	9.0	9.7	-10.6
# of social workers	1.1	1.3	-13.9	1.1	1.4	-16.8	1.1	1.3	-10.4
# of HD stations	24.0	24.0	-0.5	24.2	24.4	-1.2	23.5	24.6	-10.9
Median income	$50,874	$51,343	-2.32	$50,618	$50,496	0.6	$50,470	$51,368	-4.5
Bachelors degree +	23.4	25.0	-10.8	24.0	25.1	-8.2	23.5	25.7	-15.4

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91,** 444–455.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in medicine* **33,** 2297–2340.

Baiocchi, M., Small, D., Lorch, S., and Rosenbaum, P. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association* **105,** 1285–1296.

Baiocchi, M., Small, D. S., Yang, L., Polsky, D., and Groeneveld, P. W. (2012). Near/far matching: a study design approach to instrumental variables. *Health Services and Outcomes Research Methodology* **12,** 237–253.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80,** 2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. (2010). Lasso methods for gaussian instrumental variables models. *arXiv:1012.1297* .

Bound, J., Jaeger, D. A., and Baker, R. M. (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.

Brookhart, M. A. and Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The international journal of biostatistics* **3,**.

Caner, M. and Fan, Q. (2010). The adaptive lasso method for instrumental variable selection. Technical report, Working Paper, North Carolina State University.

Derigs, U. (1988). Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research* **13,** 225–261.

Garabedian, L. F., Chu, P., Toh, S., Zaslavsky, A. M., and Soumerai, S. B. (2014). Potential bias of instrumental variable analyses for observational comparative effectiveness research-potential bias of instrumental variable analyses for observational cer. *Annals of Internal Medicine* **161,** 131–138.

Goodlad, C. and Brown, E. (2013). The role of peritoneal dialysis in modern renal replacement therapy. *Postgraduate medical journal* **89,** 584–590.

Heaf, J. G., Løkkegaard, H., and Madsen, M. (2002). Initial survival advantage of peritoneal dialysis relative to haemodialysis. *Nephrology Dialysis Transplantation* **17,** 112–117.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62,** pp. 467–475.

Jiwakanon, S., Chiu, Y.-W., Kalantar-Zadeh, K., and Mehrotra, R. (2010). Peritoneal dialysis: an underutilized modality. *Current opinion in nephrology and hypertension* **19,** 573–577.

Kim, H., Kim, K. H., Park, K., Kang, S.-W., Yoo, T.-H., Ahn, S. V., Ahn, H. S., Hann, H. J., Lee, S., Ryu, J.-H., Kim, S.-J., Kang, D.-H., Choi, K. B., and Ryu, D.-R. (2014). A population-based approach indicates an overall higher patient mortality with peritoneal dialysis compared to hemodialysis in korea. *Kidney International* **86,** 991–1000.

Korevaar, J. C., Feith, G., Dekker, F. W., van Manen, J. G., Boeschoten, E. W., Bossuyt, P. M., and T Krediet, R. (2003). Effect of starting with hemodialysis compared with

peritoneal dialysis in patients new on dialysis treatment: a randomized controlled trial. *Kidney International* **64,** 2222–2228.

Kumar, V. A., Sidell, M. A., Jones, J. P., and Vonesh, E. F. (2014). Survival of propensity matched incident peritoneal and hemodialysis patients in a united states health care system. *Kidney International* **86,** 1016–1022.

Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The International Journal of Biostatistics* **9,** 215–234.

Li, Y., Lee, Y., Wolfe, R. A., Morgenstern, H., Zhang, J., Port, F. K., and Robinson, B. M. (2015). On a preference-based instrumental variable approach in reducing unmeasured confounding-by-indication. *Statistics in Medicine* **34,** 1150–1168.

Lu, B., Greevy, R., Xu, X., and Beck, C. (2011). Optimal nonbipartite matching and its statistical applications. *The American Statistician* **65,**.

Lunceford, J. K. and 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.

Marrón, B., Remón, C., Pérez-Fontán, M., Quirós, P., and Ortíz, A. (2008). Benefits of preserving residual renal function in peritoneal dialysis. *Kidney International* **73,** S42–S51.

Mehrotra, R., Chiu, Y.-W., Kalantar-Zadeh, K., Bargman, J., and Vonesh, E. (2011). Similar outcomes with hemodialysis and peritoneal dialysis in patients with end-stage renal disease. *Archives of internal medicine* **171,** 110.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science* **5,** 463–480.

Noordzij, M. and Jager, K. (2012). Survival comparisons between haemodialysis and peritoneal dialysis. *Nephrology Dialysis Transplantation* .

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

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

Rosenbaum, P. R. and Silber, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association* **104,**.

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

Sinnakirouchenan, R. and Holley, J. L. (2011). Peritoneal dialysis versus hemodialysis: risks, benefits, and access issues. *Advances in Chronic Kidney Disease* **18,** 428–432.

Small, D. S. and Rosenbaum, P. R. (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association* **103,** 924–933.

Staiger, D. O. and Stock, J. H. (1994). Instrumental variables regression with weak instruments. *Econometrica* .

Tam, P. (2009). Peritoneal dialysis and preservation of residual renal function. *Peritoneal Dialysis International* **29,** S108–S110.

VanderWeele, T. J. and Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)* **22,** 42–52.

Vonesh, E., Snyder, J., Foley, R., and Collins, A. (2006). Mortality studies comparing peritoneal dialysis and hemodialysis: what do they tell us? *Kidney International* **70,** S3–S11.

Weinhandl, E. D., Foley, R. N., Gilbertson, D. T., Arneson, T. J., Snyder, J. J., and Collins, A. J. (2010). Propensity-matched mortality comparison of incident hemodialysis and peritoneal dialysis patients. *Journal of the American Society of Nephrology* **21,** 499–506.

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S., and Rosenbaum, P. R. (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Ann. Appl. Stat.* **7,** 25–50.