Published in final edited form as:

Health Econ. 2014 April ; 23(4): 462–472. doi:10.1002/hec.2926.

CAN WE MAKE SMART CHOICES BETWEEN OLS AND CONTAMINATED IV METHODS?

Anirban Basu^{a,c,*} and Kwun Chuen Gary Chan^{a,b}

^aDepartment of Health Services, Pharmacy and Economics, University of Washington, Seattle, WA, USA

^bDepartment of Biostatistics and Health, University of Washington, Seattle, WA, USA

^cThe National Bureau of Economic Research, Cambridge, MA, USA

Abstract

In the outcomes research and comparative effectiveness research literature, there are strong cautionary tales on the use of instrumental variables (IVs) that may influence the newly initiated to shun this premier tool for casual inference without properly weighing their advantages. It has been recommended that IV methods should be avoided if the instrument is not econometrically perfect. The fact that IVs can produce better results than naïve regression, even in nonideal circumstances, remains underappreciated. In this paper, we propose a diagnostic criterion and related software that can be used by an applied researcher to determine the plausible superiority of IV over an ordinary least squares (OLS) estimator, which does not address the endogeneity of a covariate in question. Given a reasonable lower bound for the bias arising out of an OLS estimator, the researcher can use our proposed diagnostic tool to confirm whether the IV at hand can produce a better estimate (i.e., with lower mean square error) of the true effect parameter than the OLS, without knowing the true level of contamination in the IV.

Keywords

OLS; instrumental variable; bias; contamination; diagnostic

1. INTRODUCTION

In the context of health outcomes research and comparative effectiveness research, there appears to be a revitalization of interests in highlighting or in studying the biases that arise when an instrumental variable (IV) estimator is applied to estimate the unbiased (causal) effect of an endogenous variable on outcomes (Penrod *et al.*, 2009; Newman *et al.*, 2012; Crown *et al.*, 2011). A variable (e.g., a treatment indicator) is denoted as endogenous when other factors (confounders) that are unobserved to the analyst of the data influence both the variable in question and also the outcomes. It is well known that a naïve estimator, such as

an ordinary least squares (OLS) regression, produces a biased estimate for the effect of the endogenous variable on outcomes.

An IV is a special variable that is used to overcome the bias of the naïve estimators. An IV only directly affects the endogenous variable in question but does not affect outcomes in any other way. Therefore, the portion of variance of the endogenous variable that is predicted by the IV (which by definition would not be influenced by confounders) can be used to study the casual effect of the variable on outcomes. However, IV estimators may also produce bias when instruments are not perfectly orthogonal to the confounders (i.e., they are contaminated) and/or are poor predictors of treatment or the endogenous variable (i.e., they are weak). Sometimes, biases from IV estimation can supersede biases from naïve estimation.

There is a long history in the economics literature that has discussed these concerns (see Murray, 2006 for a summary), yet IV methods remain to be one of the most important tools to address hidden biases in observational data (Stock and Trebbi, 2003). Perhaps, the persistence of the use of the IV estimator is because careful uses of these methods can help generate causal effects outside artificial randomization. Nevertheless, there is a lack of diagnostic criteria that can guide applied researchers to conditions under which an IV at hand may generate better estimates of casual effects than a naïve estimator. This inconvenience is compounded by the recent comparative effectiveness research method literature that conveys a much stronger cautionary note for the IV methods than is necessary. For example, Crown *et al.* (2011) present a set of simulations that highlight the degrees of bias in IV estimators as affected by instrument strength, instrument contamination, and sample size. Their main results state ‘Notably, the simulations indicate a greater potential for inferential error when using IV than OLS in all but the most ideal circumstances’ and conclude that only under the most ideal circumstances are the IV methods likely to produce estimates with less estimation error than OLS. Statements such as these can be quite misleading to the applied researchers who may consequently fail to properly weigh the benefits of an IV approach with its biases. For example, in an empirical situation, where the anticipated bias from OLS is large, it should make one more likely to accept less than ideal IVs.

Our goal in this discussion would be to come up with a diagnostic criterion that can be used by an applied researcher to determine the plausible superiority of IV over an OLS estimator.¹ We propose a novel phase diagram (and provide a software to construct it) that can be used to assess the question whether the mean squared error (MSE) for estimating the effect of an endogenous variable with IVs will be larger than that from OLS given the data at hand. This can be accomplished by anticipating a reasonable lower bound of the OLS bias and without knowing the true level of contamination in the IV. Knowledge about a reasonable lower bound of the bias that arises from an OLS estimator is usually driven by the substantive knowledge of the applied area, comparison with randomized trials, or the amount of selection present on the observed explanatory variables (Altonji *et al.*, 2005).

¹We will refrain from more involved discussions regarding nonlinear functional forms (Terza *et al.* 2008) and heterogeneous treatment effects (Heckman 1997; Basu *et al.*, 2007).

Our discussions will not focus on very weak instruments (although the diagnostics proposed will extend to instruments of any strength). The biases arising from the use of weak instruments are well documented in the literature (Bound *et al.*, 1995; Staiger and Stock, 1997). In current practice, the strength of instrument is a testable assumption, and most researchers who are familiar with the weak instrument literature would not use an instrument that is a very weak predictor of the endogenous variable (e.g., a treatment). The question then becomes, given that one has a strong instrument, how much contamination is acceptable that can still steer us towards that correct inference about average treatment effects compared with inferences based on naïve regression methods such as OLS.

2. STRUCTURAL MODELS

In what follows, we will use the following notations consistently throughout:

σ_A^2 = variance of a stochastic variable A . $\sigma_A = \sqrt{\sigma_A^2}$ = standard deviation of A .

σ_{AB} = covariance between two stochastic variables A and B .

ρ_{AB} = correlation between two stochastic variables A and B .

$\sigma_{AB.C}$ or $\rho_{AB.C}$ = partial covariance or correlation between A and B partial over C . It is the covariance or correlation between A and B , when the effects of C have been removed from both A and B .

Let us consider the effect of a covariate X on outcome Y . The underlying structural model is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (1)$$

where β_1 is the true casual effect of X on Y and ε represents a stochastic error such that $E(\varepsilon|X) = 0$ and $\text{Var}(Y|X) = \text{Var}(\varepsilon) = \sigma_\varepsilon^2$. An OLS regression of Y on X , however, will produce a biased estimate of β_1 as long as X and error ε are correlated, that is, $\sigma_{X\varepsilon} \neq 0$. The asymptotic OLS estimator will converge to (Bound *et al.*, 1995)

$$\text{plim} \hat{\beta}_{\text{OLS}} \rightarrow \beta_1 + \frac{\sigma_{X\varepsilon}}{\sigma_X^2}, \quad (2)$$

where the $\hat{\beta}_{\text{OLS}}$ signifies that the parameter is estimated from data at hand. Note that the asymptotic bias of the OLS estimator (the second term in (2)) is not zero as $\sigma_{X\varepsilon} \neq 0$.

In contrast, an IV is often used to address the bias in the OLS estimator. An IV, denoted by Z , is a variable that is correlated with X but, theoretically, is independent of ε .² That is, Z neither belongs in (1) directly nor is correlated with any variable that should have been in (1) if properly specified. An IV estimator is often implemented as a two-step process: it first generates the projection of X on Z denoted by $\hat{X}$ and then studies the effect of this projection on Y :

²Note that an instrumental variable (IV) that is identical to X , that is, $Z = X$, reproduces the ordinary least squares (OLS) estimate.

$$Y = \beta_0 + \beta_1 \hat{X} + \varepsilon, \quad (3)$$

where $\hat{X} = \alpha_0 + \alpha_1 Z$. Following the asymptotic results on the OLS estimator, the asymptotic IV estimator converges to (Bound *et al.*, 1995)

$$\text{plim} \hat{\beta}_{IV} \rightarrow \beta_1 + \frac{\sigma_{X\varepsilon}}{\sigma_X^2}, \quad (4)$$

where $\sigma_X^2 = \rho_{XZ}^2 \sigma_Z^2$ and ρ_{XZ} is the correlation between X and Z . If Z is a valid IV, then $\sigma_{X\varepsilon} = 0$ even though $\sigma_{X\varepsilon} \neq 0$; the asymptotic bias for the IV estimator is zero. However, a contaminated IV, which implies that the IV is not independent of ε , will generate a biased estimate for β_1 . If $E(Z\varepsilon) \neq 0$, it implies that $\sigma_{X\varepsilon} \neq 0$.

If we are only interested in comparing asymptotic biases and select the estimator with a lower asymptotic bias for the true parameter β_1 , we can compare the squared estimator error³:

$$\text{plim}(\hat{\beta}_{IV} - \beta_1)^2 \text{ versus } \text{plim}(\hat{\beta}_{OLS} - \beta_1)^2 \Rightarrow \left(\frac{\sigma_{X\varepsilon}}{\sigma_X^2} \right)^2 \text{ versus } \left(\frac{\sigma_{X\varepsilon}}{\sigma_X^2} \right)^2 \quad (5)$$

However, most real applications face the challenge of having finite samples, where, in addition to the comparison earlier, one must also account for the sampling error of the IV versus the OLS estimator. That is, even if the asymptotic bias of an IV estimator may be lower, a drastic increase in its sampling variability (either due to small sample size or weak instruments) may not justify the use of the IV estimator. Therefore, to select between an IV estimator versus a naïve OLS estimator in any empirical application, one should compare the potential finite sample bias of an IV estimator with an OLS estimator by using the MSE criteria, where:

$$\text{MSE} = \text{Asymp. Bias}^2 + \text{Sampling Variability}$$

and the comparison is between

$$\left(\frac{\sigma_{X\varepsilon}}{\sigma_X^2} \right)^2 + \frac{1}{n} \left(\frac{\sigma_\varepsilon^2}{\sigma_X^2} \right) \text{ versus } \left(\frac{\sigma_{X\varepsilon}}{\sigma_X^2} \right)^2 + \frac{1}{n} \left(\frac{\sigma_\varepsilon^2}{\sigma_X^2} \right). \quad (6)$$

Unfortunately, this comparison is difficult to make in practice. As ε remains unobservable, $\sigma_{X\varepsilon}$ and $\sigma_{X\varepsilon}$ cannot be directly measured, and thus, the comparison of the quantities in (6) becomes difficult.⁴ Nevertheless, one can make the qualitative inference from (6) that as $\sigma_{X\varepsilon}$

³Note the squared error is a better metric to compare the two estimators. Comparing mean errors, without squaring, produces ambiguous results. That is, $E(\hat{\beta}_{IV} - \beta_1) = E(\hat{\beta}_{OLS} - \beta_1)$ favors OLS only if $E(\hat{\beta}_{OLS} - \beta_1) = 0$. The same criteria favors IV if $E(\hat{\beta}_{OLS} - \beta_1) < 0$.

increases, one can have a greater margin of contamination in the IV (i.e., larger $\sigma_{X\varepsilon}$) and still have the IV estimator produce a better estimate of the true effect than OLS.

3. A NOVEL APPROACH TO COMPARE OLS VERSUS CONTAMINATED INSTRUMENTAL VARIABLE ESTIMATORS

Although there is no way one can eliminate the latent nature of ε with the data at hand, we present a new way in which such a comparison can be made simpler. Specifically, the goal is to eliminate one of the two unmeasured covariances in (5) and present the comparison in term of one unmeasured covariance and all other parameters that can be estimated readily from the data. With this new formulation, an analyst can easily detect if the IV at hand would produce better results than OLS if the analyst has some prior knowledge about the minimum bias that an OLS estimator can generate.⁵

Our approach expresses the correlation between Z and ε , $\rho_{Z\varepsilon}$, in terms of partial correlation coefficient $\rho_{Z\varepsilon \cdot X}$, where the effect of X is removed. Even though $\rho_{Z\varepsilon}$ is not directly observed in the data, $\rho_{Z\varepsilon \cdot X}$ is, as $\rho_{Z\varepsilon \cdot X} = \rho_{ZY \cdot X}$. The derivation for this equality is shown as follows:

Let us assume, without loss of generality, that each variable is normed to have mean of zero. Then,

$$\begin{aligned}\rho_{ZY \cdot X} &= E\{ZY|X\} / (\sigma_{Z \cdot X} \sigma_{Y \cdot X}) \\ &= E\{Z(X\beta + \varepsilon)|X\} / (\sigma_{Z \cdot X} \sigma_{Y \cdot X}) \text{ (replacing } Y \text{ from the model in (1))} \\ &= E\{Z\varepsilon|X\} / (\sigma_{Z \cdot X} \sigma_{\varepsilon \cdot X}) = \rho_{Z\varepsilon \cdot X},\end{aligned}\quad (7)$$

where the second to the last equality follows from $E\{ZX|X\} = 0$ and $\sigma_{Y \cdot X} = \sigma_{\varepsilon \cdot X} = \sigma_{\varepsilon}$ in (1).

Therefore, with the standard notation of a partial correlation,

$$\rho_{ZY \cdot X} = \rho_{Z\varepsilon \cdot X} = \frac{\rho_{Z\varepsilon} - \rho_{ZX}\rho_{X\varepsilon}}{\sqrt{(1 - \rho_{ZX}^2)}\sqrt{(1 - \rho_{X\varepsilon}^2)}} \Rightarrow \rho_{Z\varepsilon} = \rho_{ZY \cdot X} \left(\sqrt{(1 - \rho_{ZX}^2)}\sqrt{(1 - \rho_{X\varepsilon}^2)} \right) + \rho_{ZX}\rho_{X\varepsilon}. \quad (8)$$

Equation (8), expresses one of the latent parameter, $\rho_{Z\varepsilon}$, in terms of the other latent parameter, $\rho_{X\varepsilon}$, and other parameters that are easily computed with the data at hand. Then, coming back to our original comparison in (6), an IV estimator will produce higher MSE compared to OLS if

$$\left(\frac{\sigma_{\hat{X}\varepsilon}}{\sigma_{\hat{X}}^2} \right)^2 + \frac{1}{n} \left(\frac{\sigma_{\varepsilon}^2}{\sigma_{\hat{X}}^2} \right) - \left(\frac{\sigma_{X\varepsilon}}{\sigma_X^2} \right)^2 - \frac{1}{n} \left(\frac{\sigma_{\varepsilon}^2}{\sigma_X^2} \right) \geq 0. \Rightarrow \frac{\sigma_X^2}{\sigma_{\varepsilon}^2} \cdot \left(\frac{\sigma_{\hat{X}\varepsilon}^2}{\sigma_{\hat{X}}^2 \cdot \sigma_{\hat{X}}^2} \right) + \frac{1}{n} \left(\frac{\sigma_X^2}{\sigma_{\hat{X}}^2} \right) - \left(\frac{\sigma_{X\varepsilon}^2}{\sigma_X^2 \cdot \sigma_{\varepsilon}^2} \right) - \frac{1}{n} \geq 0$$

after multiplying left-hand side with $\sigma_X^2/\sigma_{\varepsilon}^2$, and $\sigma_X^2/\sigma_{\varepsilon}^2 > 0$.

⁴Note that ε represents the true error in the structural model and not the residuals that are computed after OLS estimation that has been forced by OLS to be orthogonal to x .

⁵An experienced analyst can draw the same conclusions by directly comparing the IV and OLS estimates. What I present here is a different diagnostic criterion to make this comparison.

By replacing $\sigma_{\hat{X}}^2/\sigma_X^2=\rho_{XZ}^2$ in the first and the second terms, $\sigma_{\hat{X}\epsilon}^2/(\sigma_\epsilon^2 \cdot \sigma_X^2)=\rho_{Z\epsilon}^2$ in the first term and $\sigma_{X\epsilon}^2/(\sigma_X^2 \cdot \sigma_\epsilon^2)=\rho_{X\epsilon}^2$ in the third term, we have

$$\begin{aligned} &\Rightarrow \left(\frac{\rho_{Z\epsilon}^2}{\rho_{XZ}^2} \right) \\ &+ \frac{1}{n} \left(\frac{1}{\rho_{XZ}^2} \right) \\ &-(\rho_{X\epsilon}^2) - \frac{1}{n} \geq 0 \Rightarrow \frac{\rho_{Z\epsilon}^2 + 1/n}{\rho_{XZ}^2} - (\rho_{X\epsilon}^2 + 1/n) \geq 0 \Rightarrow \left(\rho_{ZY.X} \left(\sqrt{(1-\rho_{ZX}^2)} \sqrt{(1-\rho_{X\epsilon}^2)} \right) + \rho_{ZX} \rho_{X\epsilon} \right)^2 \\ &+ (1/n) \\ &- (\rho_{ZX} \rho_{X\epsilon})^2 \\ &- (\rho_{XZ}^2/n) \geq 0 \end{aligned}$$

(replacing $\rho_{Z\epsilon}$ from (8))

$$\Rightarrow \rho_{ZY.X}^2 ((1-\rho_{ZX}^2)(1-\rho_{X\epsilon}^2)) + 2\rho_{ZY.X} \rho_{ZX} \rho_{X\epsilon} \left(\sqrt{(1-\rho_{ZX}^2)} \sqrt{(1-\rho_{X\epsilon}^2)} \right) + \frac{1}{n} (1-\rho_{XZ}^2) \geq 0. \quad (9)$$

Equation (9) shows the conditions under which different phases of superiority for IV occur.

If the sample size is not very large and the instrument is rather weak (i.e., ρ_{XZ}^2 is small), the last term may dominate and thereby result in a higher MSE for the IV estimator. However, in applications with strong instruments and decent sample size, a nonideal instrument (i.e., with contamination) may still produce lower MSE than an OLS estimator if residual (unobserved) confounding is strong. To illustrate this, we first define the phase boundaries (thresholds) by the solution of the quadratic equation in (7) with respect to $\rho_{ZY.X}$:

$$\frac{-\rho_{ZX} \rho_{X\epsilon} \pm \sqrt{(n\rho_{ZX}^2 \rho_{X\epsilon}^2 + \rho_{ZX}^2 - 1)/n}}{\sqrt{(1-\rho_{ZX}^2)} \sqrt{(1-\rho_{X\epsilon}^2)}}. \quad (10)$$

For given values of n and ρ_{ZX} based on the data at hand, the value of $\rho_{ZY.X}$, also determined from the data at hand, must lie *between* the phase boundaries for any hypothesized value of $\rho_{X\epsilon}$ for the IV estimator to have lower MSE than the OLS estimator. If instead the value of $\rho_{ZY.X}$ lies outside these phase boundaries, then it signifies that the IV estimator would generate higher MSE compared with a naïve OLS estimator.

Various extensions of this simple criterion are possible as explained below.

3.1. Presence of other exogenous factors

Most applications will have some other exogenous covariates available in the data. We denote this by W . Some or all of these variables may be confounders. The structural outcomes model becomes:

$$Y = \beta_0 + \beta_1 X + \beta_2 W + \varepsilon' \quad (11)$$

By standard definitions, it follows that, in parallel to equations (7) and (8): $\rho_{ZY.XW} = \rho_{Z\varepsilon.XW}$, and

$$\rho_{Z\varepsilon.W} = \rho_{ZY.XW} \left(\sqrt{(1 - \rho_{ZX.W}^2)} \sqrt{(1 - \rho_{X\varepsilon.W}^2)} \right) + \rho_{ZX.W} \rho_{X\varepsilon.W},$$

where each quantity is expressed to be conditional on W . The condition defined in (9) for the phase plot readily extends to this situation where W s are present. One only needs to calculate all the correlation parameters partial to these W s⁶:

$$\rho_{ZY.XW}^2 \left((1 - \rho_{ZX.W}^2) (1 - \rho_{X\varepsilon.W}^2) \right) + 2\rho_{ZY.X} \rho_{ZX.W} \rho_{X\varepsilon.W} \left(\sqrt{(1 - \rho_{ZX.W}^2)} \sqrt{(1 - \rho_{X\varepsilon.W}^2)} \right) + \frac{1}{n} (1 - \rho_{XZ.W}^2) \geq 0. \quad (12)$$

3.2. Presence of multiple instruments

When there are multiple instruments, one can use each instrument in a just-identified system and assess the potential validity of each instrument following (9). If all instruments are to be used simultaneously in the first-stage regression in an overidentified system, then prior to the first stage, a single composite instrument can be formed by regressing X over the vector of Z s (but not W s).⁷ For example, when there are two instruments, $Z = (Z_1, Z_2)$,

$$X = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + \nu. \quad (13)$$

The predictions from this model, $X(\hat{Z})$, can be used as a scalar IV in the main analysis, and the validity of the combination of all the instruments can be assessed by applying the calculations in (11) and (12) to this single composite IV.

3.3. Simulations

We conducted simulation studies to assess whether the diagnostic test proposed could lead to a decision that has a lower MSE. A total of 5000 independent dataset replicates were generated for each simulation scenario, and each data set contained $n = 1000$ or $10,000$ observations (Tables I and II, respectively).⁸ The data (Z, ε, X) were generated from a trivariate normal distribution with mean zero, unit variance, and correlations $\rho_{Z\varepsilon}$, ρ_{ZX} , and $\rho_{X\varepsilon}$. We considered $\rho_{ZX} = 0.10, 0.20$, and 0.40 , which corresponded to weak, moderate, and strong instruments, $\rho_{X\varepsilon} = 0.10, 0.15$, and 0.25 , which corresponded to weak, moderate, and strong confounding. We also varied the strength of the contamination of instrument $\rho_{Z\varepsilon}$ for every scenario.

⁶Note that $\rho_{X\varepsilon.W} = \rho_{X\varepsilon'}$.

⁷ W s are not included here as the goal is to form a scalar quantity that will serve as a single composite instrument. This composite instrument can then be used in the first-stage regression where W s are included.

⁸These sample sizes are most typical of health economic analyses. For example, a Medline search of all IV applications in the past 3 years revealed that 70% of them had sample size greater than 5000. In fact in comparative effectiveness research, with the emergence of more integrated data, these sample sizes are only likely to increase.

We compared the MSE of OLS and IV estimators averaging across 5000 simulations (columns (a) and (b) in Tables I and II). We compared these MSEs that are generated because of the use of one or the other estimator in all data replicates to the average MSE generated if we followed our diagnostic tests and varied the use of IV or OLS across replicate datasets (column (c)). To decide whether to use OLS or IV estimator on the basis of inequality (9) in each dataset, we estimated $\rho_{ZY.X}$ and ρ_{ZX} from that data. The partial correlation was estimated by the sample correlation of residuals from a linear regression of Z on X and a linear regression of Y on X . The correlation ρ_{ZX} was estimated by the sample correlation of Z and X . For each combination of $\rho_{Z\epsilon}$, ρ_{ZX} , and $\rho_{X\epsilon}$ values, we also report the proportion of times across 5000 replicate datasets that inequality (9) based on estimated values leads us to OLS (column (d)) and contrast it to whether OLS would be the preferred estimator if the true values were known (column (e)).

The results for $n = 1000$ and $n = 10,000$ are given in Tables I and II, respectively. The simulations showed that the choice of estimator based on inequality (9) would lead to a lower MSE than either OLS or IV (contrast column (c) to either (a) or (b)). This confirmed our belief that we can make a smart choice between OLS and contaminated IV that lead to a practical improvement in lowering MSE on the basis of sampling variations. The improvement is particularly significant when the MSE of OLS and IV are comparable. In those cases, the MSE of the mixed strategy could have a 25% reduction in MSE compared with either OLS or IV on average.

Moreover, the decision to use IV over OLS using estimated values of parameters in criterion (9) tracks well with the decision if the true values were known (column (d) versus (e)), and this comparison improves with sample size (Table I versus Table II).

Next, we evaluated the performance of estimators when the strength of instrument contamination was not fixed in advance. A random $\rho_{Z\epsilon}$ was generated in each of the 5000 simulations from a uniform (0, 0.1) distribution, whereas we fixed the strength of instrument and confounding. Sample size for each simulated data set was 10,000. We compared the MSE from estimation following decision rule (9) to MSE by using OLS or IV for all simulated data sets. The results are shown in Table III. Again, the MSE was reduced for every scenario by using inequality (9), indicating informed choices were being made between OLS and IV estimators across data replicates.

3.4. Use of diagnostic criterion in an application

We illustrate the application of the proposed diagnostic criteria in an empirical analysis that we had conducted elsewhere (Basu *et al.*, 2012). The empirical analysis was an IV analysis used to estimate the causal effect of a generic drug (X) called risperidone ($n = 24,028$) compared with other branded drugs ($n = 54,503$) on annual number of schizophrenia-related hospitalizations (Y) among patients with schizophrenia who started on either drug after 6 months of drug-clean period

The 24-state Medicaid data from 2003 to 2005 was used in this analysis. Because the data arose from Medicaid claims file, the typical risk factors (W) that were controlled for were patient demographics and comorbidities. By the nature of the data, it was expected that there

was selection bias in choosing risperidone due to patient characteristics that were not captured by these claims data. For example, underlying severity of the disease, which is most likely driving these patients back to their physicians to obtain a prescription drug, was unmeasured in the data. Not only would this severity levels affect the primary outcome measure of hospitalizations following drug receipt but could also very much influence the choice between risperidone and other drugs. By 2003–2004, a series clinical trials has shown risperidone to be the least efficacious on average (Edgell *et al.*, 2008; Volavka *et al.*, 2002; Zhao, 2004). This evidence was also later confirmed by a large clinical trial (Lieberman *et al.*, 2005). Thus, it was expected that a positive correlation existed between the unobserved severity levels and hospitalizations, whereas a negative correlation existed between severity and risperidone choice, making a naïve regression estimate biased downward.

The traditional regression estimator, which allowed for adjustment of observed factors only, produced an effect estimate of 0.03 (SE = 0.01, $p = 0.011$) on the annual number of hospitalizations per patient, implying a very small effect of risperidone on hospitalizations compared to the branded products. Because it was anticipated that this estimate was biased downward, an IV analysis was conducted.

Two IVs were used: (i) the frequency with which a patient's physician prescribed risperidone during the 6 months prior to the patient's index start date and (ii) the average rate of risperidone use in a patient's zip code in the year prior to the index date. Both variables were expected to be associated with risperidone use but not otherwise associated with a patient's risk of hospitalization (Brookhart and Schneeweiss, 2007). Both were found to be strong predictors of treatment receipt.

In this empirical setting, the question that this current paper seeks to answer is whether using these IVs would lead to lower MSE than the naïve regression estimator. To answer this question, first, we follow (13) to collapse the two IVs into a scalar IV. To do this, we regressed the risperidone indicator variable on these two IVs and used the predicted probability of risperidone use as a scalar IV, $X(\hat{Z})$, for our analyses.

Next, because there were additional risk factors (W) in the analyses, we used criterion (12) to derive the phase boundaries and use them to draw a phase plot that is shown in Figure 1. The shaded portions in the plot are the regions where the IV estimator at hand will generate lower MSE than OLS and hence is superior.

To understand this diagnostic phase plot, one must understand that we are comparing an observed quantity, $\rho_X(\hat{Z})_{Y.XW}$, on the Y-axis and an unobserved quantity, $\rho_{X\epsilon}$, on the X-axis. For any plausible level of the unobserved quantity, we can check whether the level of the observed quantity falls in the gray region of the graph. For example, in Figure 1, the estimated $\rho_X(\hat{Z})_{Y.XW}$ from our data is 0.016. With this value and the criterion in (12), the figure identifies the cutoff for $\rho_{X\epsilon} = -0.037$, beyond which the IV at hand will produce smaller MSE than OLS. That is, if the analyst contemplates that the negative correlation between the endogenous variable (risperidone) and the unmeasured confounders is at least as great as -0.04 , then one should be confident that the IV at hand will produce lower MSE

that an OLS estimator. In general, the smaller the absolute value of $\rho_X(\hat{Z})_{Y.XW}$, the better the chance that the IV estimator will be better than OLS.

Does this tell us if we should use IV in this analysis? To answer this question, we have to invoke substantive knowledge about the plausible levels of $\rho_{X\epsilon}$. We base this knowledge on the largest clinical trial in this field (Lieberman *et al.*, 2005) and find that when the relative risk for risperidone compared with olanzapine (which constituted the majority share of the branded drugs in 2003–2004) on schizophrenia-related hospitalizations is extrapolated to the average hospitalization rate in our sample, it indicates that the true effect of risperidone on annual number of schizophrenia-related hospitalizations per patient could be 0.41, if not more. If we were to believe that 0.41 is the true effect, it would indicate that $\rho_{X\epsilon}$ could be as large as -0.19 .⁹

Looking back at Figure 1, the shaded range of the graph corresponding to the $\rho_{X\epsilon}$ value of -0.19 on the X-axis not only captures the estimated $\rho_X(\hat{Z})_{Y.XW}$ of 0.016 from our data but also can possibly accept even large values of $\rho_X(\hat{Z})_{Y.XW}$ and still have an IV estimator to produce lower MSE than a naïve estimator. This implies that even if we had used an IV with slightly higher contamination that we currently have, we would have still expected it to produce lower MSE than the OLS results.

In practice, one would construct such a phase diagram with the data at hand by calculating $\rho_{ZY \cdot X}$ from the data and finding the corresponding margin of $\rho_{X\epsilon}$ where the IV at hand will produce lower MSE compared with OLS.¹⁰

4. CONCLUSIONS

In this paper, we propose a diagnostic for figuring out whether the IV(s) at hand would estimate treatment effects with lower MSE than that from a traditional regression such as OLS. The diagnostic method identifies the plausible range of correlations between the treatment and the error in the outcomes regression, where the IVs at hand would produce better results than an OLS regression. Applied researchers may find such a metric to be useful in addition to the other diagnostics and tests relevant to carry out a proper IV analysis.

ACKNOWLEDGEMENTS

The author acknowledges support from the National Institute of Health Research Grants R01MH083706, RC4CA155809, and R01CA155329. The authors thank Willard Manning and two anonymous reviewers for their comments on an earlier version of the paper and take responsibility on all errors in the paper.

REFERENCES

Altonji JG, Elder TE, Taber CR. Selection on observed and unobserved variables: assessing the effectiveness of catholic schools. *Journal of Political Economy*. 2005; 113(1):151–184.

⁹From (2), Bias = $(\beta_{OLS} - \beta) = \rho_{X\epsilon}/\sigma_X = (0.03 - 0.41) = -0.38$. Therefore, $\rho_{X\epsilon} = -0.19$, because $\sigma_X = 0.50$ in our data.

¹⁰A Stata command—ivcheck—can be downloaded from author's website that will produce this diagram for any data at hand (to be posted at <http://faculty.washington.edu/basua/software.html>).

- Basu A, Jena AB, Goldman DP, Philipson TJ, Dubois R. Incorporating patient heterogeneity into comparative effectiveness research: evidence from schizophrenia. Mimeo, University of Washington. 2012
- Bound J, Jaeger DA, Baker RM. Problems with instrumental variable estimation when correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*. 1995; 90(430):443–450.
- Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *International Journal of Biostatistics*. 2007; 3(1):14.
- Crown WH, Henk HJ, Vanness DJ. Some cautions on the use of instrumental variables estimators in outcomes research: How bias in instrumental variable estimators is affected by instrument strength, instrument contamination, and sample size. *Value in Health*. 2011; 14(8):1078–1084. [PubMed: 22152177]
- Edgell ET, Andersen SW, Johnstone BM, Dulisse B, Revicki D, Breier A. Olanzapine versus risperidone. A prospective comparison of clinical and economic outcomes in schizophrenia. *Pharmacoeconomics*. 2008; 18(6):567–579. [PubMed: 11227395]
- Lieberman JA, Stroup SS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, Keefe RSE, Davis SM, Davis CE, Lebowitz BD, Severe J, Hsiao JK. the CATIE Investigators. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *The New England Journal of Medicine*. 2005; 353(12):1209–1223. [PubMed: 16172203]
- Murray MP. Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives*. 2006; 20:111–132.
- Newman TB, Vittinghoff E, McCulloch CE. Efficacy of phototherapy for newborns with hyperbilirubinemia: a cautionary example of an instrumental variable. *Medical Decision Making*. 2012; 32(1):83–92. [PubMed: 21859678]
- Penrod JD, Goldstein NE, Deb P. When and how to use instrumental variable in palliative care research. *Journal of Palliative Medicine*. 2009; 12(5):471–474. [PubMed: 19416044]
- Staiger D, Stock JH. Instrumental variable regressions with weak instruments. *Econometrica*. 1997; 65(3):557–586.
- Stock JH, Trebbi F. Who invented instrumental variable regression? *Journal of Economic Perspectives*. 2003; 17(3):177–194.
- Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics*. 2008; 27(3):531–543. [PubMed: 18192044]
- Volavka J, Czobor P, Sheitman B, Lindenmayer J-P, Citrome L, McEvoy JP, Cooper TB, Chakos M, Lieberman JA. Clozapine, olanzapine, risperidone, and haloperidol in the treatment of patients with chronic schizophrenia and schizoaffective disorder. *The American Journal of Psychiatry*. 2002; 159:255–262. [PubMed: 11823268]
- Zhao Z. Economic outcomes associated with olanzapine versus risperidone in the treatment of uncontrolled schizophrenia. *Current Medical Research and Opinion*. 2004; 20(7):1039–1048.

X: Endogenous Variable; Z: Instrumental Variable; Y: Outcome; W: Other Exogenous Variables; MSE: Mean Squared Error

Figure 1.

Phase diagram to detect plausible superiority of instrumental variable (IV) estimator over ordinary least squares (OLS) given data.

Table I

Comparison of OLS and IV estimators, $n = 1000$

Pz_{β}	Strong instrument					Moderate instrument					Weak instrument				
	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)
Weak confounding	0.00	0.011	0.006	0.005	30.5	0	0.011	0.027	0.011	100	1	0.011	0.376	0.011	100
	0.01	0.011	0.007	0.006	34.4	0	0.011	0.030	0.011	100	1	0.011	0.380	0.011	100
	0.025	0.011	0.010	0.007	46.6	0	0.011	0.043	0.011	100	1	0.011	0.537	0.011	100
	0.05	0.011	0.022	0.009	76.3	1	0.011	0.092	0.011	100	1	0.011	0.908	0.011	100
	0.10	0.011	0.069	0.011	99.2	1	0.011	0.292	0.011	100	1	0.011	2.897	0.011	100
Moderate confounding	0.00	0.024	0.006	0.006	7.1	0	0.023	0.028	0.020	83.5	1	0.024	0.452	0.024	100
	0.01	0.024	0.007	0.007	8.8	0	0.024	0.028	0.020	84.6	1	0.024	0.478	0.024	100
	0.025	0.023	0.010	0.009	17.5	0	0.023	0.042	0.021	88.7	1	0.024	0.573	0.024	100
	0.05	0.025	0.022	0.015	42.9	0	0.023	0.091	0.023	95.9	1	0.023	1.899	0.023	100
	0.10	0.025	0.068	0.023	94.9	1	0.023	0.289	0.023	99.8	1	0.024	3.672	0.024	100
Strong confounding	0.00	0.063	0.006	0.006	0.3	0	0.063	0.027	0.023	22.5	0	0.063	0.677	0.059	92.8
	0.01	0.063	0.007	0.007	0.2	0	0.064	0.029	0.025	25.1	0	0.063	0.879	0.060	93.6
	0.025	0.063	0.010	0.010	0.6	0	0.064	0.041	0.031	34.5	0	0.063	1.258	0.061	95.3
	0.05	0.063	0.022	0.021	5.2	0	0.063	0.088	0.047	64.2	1	0.063	1.842	0.062	98.6
	0.10	0.063	0.068	0.052	53.7	1	0.063	0.283	0.063	97.7	1	0.063	4.476	0.063	100

(a) Average mean squared error (MSE) of ordinary least squares (OLS). (b) Average MSE of instrumental variable (IV). (c) Average MSE across OLS/IV used based on the decision from inequality (9). (d) Percentage of simulated data in which inequality (9) was satisfied (OLS is used) on the basis of estimated parameters. (e) Whether inequality (9) is satisfied with true parameters (1 = IV MSE > OLS MSE and OLS should be used).

Table II

Comparison of OLS and IV estimators, $n = 10,000$

	Strong instrument					Moderate instrument					Weak instrument					
	$p_{Z\beta}$	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)	(a)	(b)	(c)	(d)	(e)
Weak confounding	0.00	0.010	0.001	0.001	0	0	0.010	0.003	0.002	8.1	0	0.010	0.010	0.009	83.8	0
	0.01	0.010	0.001	0.001	0	0	0.010	0.005	0.004	22.1	0	0.010	0.020	0.009	91	1
	0.025	0.010	0.005	0.004	5.9	0	0.010	0.018	0.009	77.8	1	0.010	0.073	0.010	99.4	1
	0.05	0.010	0.016	0.010	88.6	1	0.010	0.065	0.010	100	1	0.010	0.267	0.010	100	1
	0.10	0.010	0.063	0.010	100	1	0.010	0.254	0.010	100	1	0.010	1.039	0.010	100	1
Moderate confounding	0.00	0.023	0.001	0.001	0	0	0.023	0.003	0.003	0.4	0	0.023	0.011	0.009	26.8	0
	0.01	0.023	0.001	0.001	0	0	0.023	0.005	0.005	2.5	0	0.023	0.019	0.013	46.8	0
	0.025	0.023	0.005	0.005	0	0	0.023	0.018	0.015	36.3	0	0.023	0.073	0.021	91.9	1
	0.05	0.023	0.017	0.017	15.6	0	0.023	0.065	0.022	98.6	1	0.023	0.262	0.023	100	1
	0.10	0.023	0.063	0.023	100	1	0.023	0.255	0.023	100	1	0.023	1.034	0.023	100	1
Strong confounding	0.00	0.063	0.001	0.001	0	0	0.063	0.003	0.003	0	0	0.063	0.010	0.010	2.4	0
	0.01	0.063	0.001	0.001	0	0	0.063	0.005	0.005	0	0	0.063	0.020	0.018	9.4	0
	0.025	0.063	0.005	0.005	0	0	0.063	0.018	0.018	0.4	0	0.063	0.072	0.048	58.5	1
	0.05	0.063	0.016	0.016	1.6	0	0.063	0.065	0.054	53.4	1	0.062	0.262	0.062	99.7	1
	0.10	0.063	0.063	0.059	51.3	1	0.063	0.253	0.063	100	1	0.063	1.039	0.063	100	1

(a) Average mean squared error (MSE) of ordinary least squares (OLS). (b) Average MSE of instrumental variable (IV). (c) Average MSE across OLS/IV used based on the decision from inequality (9). (d) Percentage of simulated data in which inequality (9) was satisfied (OLS is used) on the basis of estimated parameters. (e) Whether inequality (9) is satisfied with true parameters (1 = IV MSE > OLS MSE and OLS should be used).

Table III

Comparison of (a) ordinary least squares, (b) instrumental variable, and (c) ordinary least squares/instrumental variable based on the decision from inequality (9), $n = 10000$, when instrument contamination strength is random

	Strong instrument			Moderate instrument			Weak instrument		
	(a)	(b)	(c)	(a)	(b)	(c)	(a)	(b)	(c)
Weak confounding	0.010	0.021	0.007	0.010	0.094	0.009	0.010	0.358	0.010
Moderate confounding	0.023	0.022	0.014	0.023	0.088	0.018	0.022	0.347	0.021
Strong confounding	0.063	0.021	0.020	0.062	0.084	0.041	0.063	0.337	0.052