Biometrics 64, 1–22 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Two Robust Tools for Inference about Causal Effects with Invalid Instruments

Hyunseung Kang^{1,*,J}, Youjin Lee^{2,**,J}, T. Tony Cai^{3,***} and Dylan S. Small^{3,****,J}

[1]: Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin, U.S.A.

[2]: Center for Causal Inference, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.

[3]: Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.

[J]: Denotes equal contribution.

**email:* hyunseung@stat.wisc.edu

***email:* youjin.lee@pennmedicine.upenn.edu

****email:* tcai@wharton.upenn.edu

*****email:* dsmall@wharton.upenn.edu

Summary: Instrumental variables have been widely used to estimate the causal effect of a treatment on an outcome. Existing confidence intervals for causal effects based on instrumental variables assume that all of the putative instrumental variables are valid; a valid instrumental variable is a variable that affects the outcome only by affecting the treatment and is not related to unmeasured confounders. However, in practice, some of the putative instrumental variables are likely to be invalid. This paper presents two tools to conduct valid inference and tests in the presence of invalid instruments. First, we propose a simple and general approach to construct confidence intervals based on taking unions of well-known confidence intervals. Second, we propose a novel test for the null causal effect based on a collider bias. Our two proposals outperform traditional instrumental variable confidence intervals when invalid instruments are present and can also be used as a sensitivity analysis when there is concern that instrumental variables assumptions are violated. The new approach is applied to a Mendelian randomization study on the causal effect of low-density lipoprotein on globulin levels.

Key words: Anderson-Rubin test; Confidence interval; Invalid instrument; Instrumental variable; Likelihood ratio test; Weak instrument.

1. Introduction

Instrumental variables have been a popular method to estimate the causal effect of a treatment, exposure, or policy on an outcome when unmeasured confounding is present (Hernán and Robins, 2006; Baiocchi et al., 2014). Informally speaking, the method relies on having instruments that are (A1) related to the exposure, (A2) only affect the outcome by affecting the exposure (no direct effect), and (A3) are not related to unmeasured confounders that affect the exposure and the outcome; see Section 2.2 for details. Unfortunately, in many applications, practitioners are unsure if all of the candidate instruments satisfy these assumptions. For example, in Mendelian randomization (Davey Smith and Ebrahim, 2003, 2004; Lawlor et al., 2008), a subfield of genetic epidemiology, the candidate instruments are genetic variants which are associated with the exposure. But, these instruments may have direct effects on the outcome, an effect known as pleiotropy, and violate (A2) (Solovieff et al., 2013). Or, the genetic instruments may be in linkage disequilibrium and violate (A3). A similar problem arises in economics where some instruments may not be exogenous, where exogeneity combines assumptions (A2) and (A3) (Murray, 2006; Conley et al., 2012).

Violation of (A1), known as the weak instrument problem, has been studied in detail; see Stock et al. (2002) for a survey. In contrast, the literature on violations of (A2) and (A3), known as the invalid instrument problem (Murray, 2006), is limited. Andrews (1999) and Andrews and Lu (2001) considered selecting valid instruments within context of the generalized method of moments (Hansen, 1982), but not inferring treatment effects after selection of valid instruments. Liao (2013) and Cheng and Liao (2015) considered estimating treatment effects when there is, a priori, a known, specified set of valid instruments and another set of instruments which may not be valid. Conley et al. (2012) proposed different approaches to assess violations of (A2) and (A3). Small (2007) considered a sensitivity analysis to assess violations of (A2) and (A3). Our work is most closely related to the recent works by Kolesár et al. (2015), Kang et al. (2016), Bowden et al. (2015), Guo et al. (2018), Windmeijer et al. (2018), Windmeijer et al. (2019), and Zhao et al. (2020). Kolesár et al. (2015) and Bowden et al. (2015) considered the case when the instruments violate (A2) and proposed an orthogonality condition where the instruments' effects on the exposure are orthogonal to their effects on the outcome. Kang et al. (2016) considered violations of (A2) and (A3) based on imposing an upper bound on the number of invalid instruments among candidate instruments, without knowing which instruments are invalid a priori or without assuming the aforementioned orthogonality condition. Windmeijer et al. (2019), under a similar setting as Kang et al. (2016), discussed consistent selection of the invalid instruments and proposed a median-Lasso estimator that is consistent when less than 50% of candidate instruments are invalid. In addition, Guo et al. (2018) proposed sequential hard thresholding to select strong and valid instruments and provided a confidence interval for the treatment effect. Windmeijer et al. (2018) used multiple confidence intervals to select a set of valid instruments and to construct a confidence interval for the treatment effect.

Instead of first selecting a set of valid or invalid instruments and subsequently testing the treatment effect, our paper directly focuses on testing the treatment effect with invalid instruments by proposing two methods. First, we propose a simple and general confidence interval procedure based on taking unions of well-known confidence intervals and show that it achieves correct coverage rates in the presence of invalid instruments. Second, we propose new tests for the null hypothesis of no treatment effect in the presence of multiple invalid instruments by leveraging different properties of the model or the implied directed acyclic graph; the nulls of these tests only depend on the number of valid instruments. We remark that our two methods can also be interpreted as a sensitivity analysis. In particular, the usual instrumental variable analysis makes the assumption that all instrumental variables are valid. Our methods allow one to relax this assumption and see how sensitive the results are by varying the parameter $\bar{s}$ that indicates the number of invalid instruments and observing how the proposed inferential quantities change from $\bar{s} = 1$ (i.e. no invalid instruments) to $\bar{s} = L$ (i.e. at most $L - 1$ instruments are invalid). We conclude by demonstrating our methods in both synthetic and real datasets.

2. Setup

2.1 *Notation*

We use the potential outcomes notation (Neyman, 1923; Rubin, 1974) for instruments laid out in Holland (1988). Specifically, let there be L candidate instruments and n individuals in the sample. Let $Y_i^{(d,\boldsymbol{z})}$ be the potential outcome if individual i had exposure d, a scalar value, and instruments $\boldsymbol{z}$, an L dimensional vector. Let $D_i^{(\boldsymbol{z})}$ be the potential exposure if individual i had instruments $\boldsymbol{z}$. For each individual, we observe the outcome Y_i, the exposure, D_i, and instruments $\boldsymbol{Z}_{i\cdot}$. We denote $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)$ and $\boldsymbol{D}_n = (D_1, \ldots, D_n)$ to be vectors of n observations. Finally, we denote $\boldsymbol{Z}_n$ to be an n by L matrix where row i consists of $\boldsymbol{Z}_{i\cdot}^T$ and $\boldsymbol{Z}_n$ is assumed to have full rank.

For a subset $A \subseteq \{1, \ldots, L\}$, denote its cardinality $c(A)$ and its complement A^C. Let $\boldsymbol{Z}_A$ be an n by $c(A)$ matrix of instruments where the columns of $\boldsymbol{Z}_A$ are from the set A. Similarly, for any L dimensional vector $\boldsymbol{\pi}$, let $\boldsymbol{\pi}_A$ be a subvector of $\boldsymbol{\pi}$ with elements determined by A.

2.2 *Model and Definition of Valid Instruments*

For two possible values of the exposure d', d and instruments $\boldsymbol{z}', \boldsymbol{z}$, we assume the following potential outcomes model

$$Y_i^{(d',\boldsymbol{z}')} - Y_i^{(d,\boldsymbol{z})} = (\boldsymbol{z}' - \boldsymbol{z})^T \boldsymbol{\phi}^* + (d' - d)\beta^*, \quad E\{Y_i^{(0,0)} \mid \boldsymbol{Z}_{i\cdot}\} = \boldsymbol{Z}_{i\cdot}^T \boldsymbol{\psi}^* \tag{1}$$

where $\boldsymbol{\phi}^*, \boldsymbol{\psi}^*$, and β^* are unknown parameters. The parameter β^* represents the causal parameter of interest, the causal effect (divided by $d' - d$) of changing the exposure from d' to d on the outcome. The parameter $\boldsymbol{\phi}^*$ represents violation of (A2), the direct effect of the instruments on the outcome. If (A2) holds, then $\boldsymbol{\phi}^* = 0$. The parameter $\boldsymbol{\psi}^*$ represents violation of (A3), the presence of unmeasured confounding between the instruments and the outcome. If (A3) holds, then $\boldsymbol{\psi}^* = 0$.

Let $\boldsymbol{\pi}^* = \boldsymbol{\phi}^* + \boldsymbol{\psi}^*$ and $\epsilon_i = Y_i^{(0,0)} - E\{Y_i^{(0,0)} \mid \boldsymbol{Z}_{i\cdot}\}$. When we combine equation (1) along with the definition of ϵ_i, we arrive at the observed data model

$$Y_i = \boldsymbol{Z}_{i\cdot}^T \boldsymbol{\pi}^* + D_i \beta^* + \epsilon_i, \quad E(\epsilon_i \mid \boldsymbol{Z}_{i\cdot}) = 0 \tag{2}$$

The observed model is also known as an under-identified single-equation linear model in econometrics (page 83 of Wooldridge (2010)). The observed model can have exogenous covariates, say $\boldsymbol{X}_{i\cdot}$, including an intercept term, and the Frisch-Waugh-Lovell Theorem allows us to reduce the model with covariates to model (2) (Davidson and MacKinnon, 1993). The parameter $\boldsymbol{\pi}^*$ in the observed data model (2) combines both violation of (A2), represented by $\boldsymbol{\phi}^*$, and violation of (A3), represented by $\boldsymbol{\psi}^*$. If both (A2) and (A3) are satisfied, then $\boldsymbol{\phi}^* = \boldsymbol{\psi}^* = \boldsymbol{0}$ and $\boldsymbol{\pi}^* = \boldsymbol{0}$. Hence, the value of $\boldsymbol{\pi}^*$ captures whether instruments are valid or invalid. Definition 1 formalizes this idea.

Definition 1: Suppose there are L instruments along with models (1)–(2). We say that instrument $j = 1, \ldots, L$ is valid (i.e. satisfy (A2) and (A3)) if $\pi_j^* = 0$ and invalid if $\pi_j^* \neq 0$.

When there is only one instrument, $L = 1$, Definition 1 of a valid instrument is identical to the definition of a valid instrument in Holland (1988). Specifically, assumption (A2) (i.e. the exclusion restriction) which implies $Y_i^{(d,\boldsymbol{z})} = Y_i^{(d,\boldsymbol{z}')}$ for all $d, \boldsymbol{z}, \boldsymbol{z}'$, is equivalent to $\boldsymbol{\phi}^* = 0$ and assumption (A3) (i.e. no unmeasured confounding) which means $Y_i^{(d,\boldsymbol{z})}$ and $D_i^{(\boldsymbol{z})}$ are independent of $\boldsymbol{Z}_{i\cdot}$ for all d and $\boldsymbol{z}$, is equivalent to $\boldsymbol{\psi}^* = \boldsymbol{0}$, implying $\boldsymbol{\pi}^* = \boldsymbol{\phi}^* + \boldsymbol{\psi}^* = \boldsymbol{0}$. Definition 1 is also a special case of the definition of a valid instrument in Angrist et al. (1996) where our setup assumes a model with additive, linear, and constant treatment effect β^*. Hence, when multiple instruments are present, our models (1)–(2) and Definition 1 can be viewed as a generalization of the definition of valid instruments in Holland (1988). Web Appendix B contains additional discussions of (1)–(2) and Definition 1.

Given the models above, the present paper focuses on testing β^* in the presence of invalid instruments, or formally

$$H_0 : \beta^* = \beta_0 \tag{3}$$

for some value of β_0 when $\boldsymbol{\pi}^*$ may not equal to 0. Specifically, let $s^* = 0, \ldots, L - 1$ be the number of invalid instruments and let $\bar{s}$ be an upper bound on s^* plus one, $s^* < \bar{s}$. Let $v^* = L - s^*$ be the number of valid instruments. We assume that there is at least one valid instrument, even if we don't know which among the L instruments are valid; see Kang et al. (2016) for a similar setup. Also, in Mendelian randomization where instruments are genetic, the setup represents a way for genetic epidemiologists to impose prior beliefs about instruments' validity. For example, based on the investigator's expertise and prior genome wide association studies, the investigator may provide an upper bound $\bar{s}$, with a small $\bar{s}$ representing an investigator's strong confidence that most of the L candidate instruments are valid and a large $\bar{s}$ representing an investigator's weak confidence in the candidate instruments' validity.

In the absence of prior belief about $\bar{s}$, the setup using the additional parameter $\bar{s}$ can also be viewed as a sensitivity analysis common in causal inference. In particular, similar to the sensitivity analysis presented in Rosenbaum (2002), we can treat $\bar{s}$ as the sensitivity parameter and vary from $\bar{s} = 1$ to $\bar{s} = L$ where $\bar{s} = 1$ represents the traditional case where all instruments satisfy (A2) and (A3) and $\bar{s} = L$ represents the worst case where at most $L - 1$ instruments may violate (A2) and (A3). For each $\bar{s}$, we can construct confidence intervals from our two proposed methods below and observe how increasing violations of instrumental variables assumptions through increasing $\bar{s}$ impact the resulting conclusions about β^*. Also, similar to a typical sensitivity analysis, we can find the smallest $\bar{s}$ that retains the null hypothesis of no causal effect. If at $\bar{s} = L$, our methods still reject the null, then the conclusion about the causal effect β^* is insensitive to violations of (A2) and (A3).

3. Method 1: Union Confidence Interval With Invalid Instruments

Let $B^* \subset \{1, \ldots, L\}$ be the true set of invalid instruments. In the instrumental variables literature, there are many test statistics $T(\beta_0, B^*)$ of the null hypothesis in (3) if B^* is known. Some examples include the test statistic based on two-stage least squares, the Anderson-Rubin test (Anderson and Rubin, 1949), and the conditional likelihood ratio test (Moreira, 2003); see Web Appendix C for details of these test statistics and other test statistics. By the duality of hypothesis testing and confidence intervals, inverting any of the aforementioned test statistic $T(\beta_0, B^*)$ under size α provides a $1 - \alpha$ confidence interval for β^*, which we denote as $C_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n, B^*)$

$$C_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n, B^*) = \{\beta_0 \mid T(\beta_0, B^*) \leqslant q_{1-\alpha}\} \tag{4}$$

Here, $q_{1-\alpha}$ is the $1 - \alpha$ quantile of the null distribution of $T(\beta_0, B^*)$.

Unfortunately, in our setup, we do not know the true set B^* of invalid instruments, so we cannot directly use (4) to obtain confidence intervals for β^*. However, from Section 2.2, we have a constraint on the number of invalid instruments, $c(B^*) = s^* < \bar{s}$. We can use this constraint to take unions of $C_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n, B)$ over all possible subsets of instruments $B \subset \{1, \ldots, L\}$ where $c(B) < \bar{s}$ and the confidence interval with the true set of invalid instruments $C(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n, B^*)$ will be in this union. Our proposal, which we call the union method, is exactly this except we restrict the subsets B in the union to be only of size $c(B) = \bar{s} - 1$; the restriction reduces computational overhead, while maintaining coverage (see Theorem 1).

$$C_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n) = \cup_{B, c(B) = \bar{s} - 1}\{C_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n, B)\} \tag{5}$$

While the procedure is simple and general, allowing any test statistic $T(\beta_0, B)$ with proper size control, it may be conservative. Specifically, some of the subsets B may contain every invalid instrument, leading to unbiased confidence intervals (i.e. contain β^* with probability greater than or equal to $1-\alpha$) while other subsets may not have every invalid instrument, leading to biased confidence intervals. Then, taking the union of both types of confidence intervals may elongate $C_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n)$ since we only need one unbiased confidence interval to have the desired coverage level.

One way to shorten $C_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n)$ is by pretesting whether each subset B contain invalid instruments. This is similar to a procedure by Andrews (1999) to select valid IVs through sequential testing of subsets of instruments until the pre-test retained its null; note that in our setup, the goal is to test β^* and we fix $\bar{s}$, which fixes the number of pretests to evaluate. Formally, for a $1-\alpha$ confidence interval of β^*, consider the null hypothesis that B^C, for $c(B^C) \geqslant 2$, contains only valid instruments, $H_0 : \boldsymbol{\pi}^*_{B_C} = 0$. Suppose $S(B)$ is a test statistic for this null with level $\alpha_s < \alpha$ and $q_{1-\alpha_s}$ is the $1-\alpha_s$ quantile of the null distribution of $S(B)$. One well-known example of $S(B)$ is the Sargan test (Sargan, 1958); see Web Appendix C for details. Let $\alpha_t = \alpha - \alpha_s$ be the confidence level for $C_{1-\alpha_t}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n, B)$. Then, a $1-\alpha$ confidence interval for β^* that incorporates the pretest $S(B)$ is

$$C'_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_b, \boldsymbol{Z}_n) = \cup_B \{C_{1-\alpha_t}(\boldsymbol{Y}_n, \boldsymbol{D}_b, \boldsymbol{Z}_n, B) \mid c(B) = \bar{s} - 1, S(B) \leqslant q_{1-\alpha_s}\} \qquad (6)$$

For example, if the desired confidence level for β^* is 95% so that $\alpha = 0.05$, we can run the pretest $S(B)$ at $\alpha_s = 0.01$ level and compute $C_{1-\alpha_t}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n, B)$ at the $\alpha_t = 0.04$ level. Theorem 1 shows that both $C_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n)$ and $C'_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n)$ achieve the desired $1-\alpha$ coverage of β^* in the presence of invalid instruments.

THEOREM 1: *Suppose model* (2) *holds and $s^* < \bar{s}$. Given $\alpha \in (0,1)$, consider any test statistic $T(\beta_0, B)$ with the property that for any $B^* \subseteq B$, $T(\beta_0, B)$ has size at most α under the null hypothesis in* (3)*. Then, $C_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n)$ in* (5) *has at least $1-\alpha$ coverage of β^*.*

Additionally, for any pretest $S(B)$ where $c(B^C) \geqslant 2$ and $S(B)$ has the correct size under the null hypothesis $H_0 : \boldsymbol{\pi}^*_{B^C} = 0$, $C'_{1-\alpha}(\boldsymbol{Y}_n, \boldsymbol{D}_n, \boldsymbol{Z}_n)$ has at least $1 - \alpha$ coverage.

The proof is in Web Appendix A. A caveat to our approach is computational feasibility, especially if L and $\bar{s}$ are moderately large. But, in economic applications, the number of strong and plausibly valid instruments rarely exceeds $L = 20$, and in some, but not all, Mendelian randomization studies, after linkage disequilibrium clumping and p-value thresholding, L remains small; in both cases, our procedure is computational tractable. Finally, while many tests satisfy the requirements for Theorem 1, some tests will be better than others where "better" can be defined in terms of statistical power or length of the confidence interval. In Web Appendix D, we characterize the power of common tests in the IV literature when invalid instruments are present and we show that under additional assumptions, the Anderson-Rubin test tends to have better power than the test based on two-stage least squares when invalid instruments are present.

4. Method 2: Tests for No Effect With Invalid Instruments

4.1 *The Collider Bias Test*

We introduce a new test statistic to test the null hypothesis of no treatment effect when invalid instruments are possibly present, i.e. test $H_0 : \beta^* = 0$ in equation (3). Broadly speaking, the new test is based on recognizing a collider bias in a directed acyclic graph when the null hypothesis of no effect does not hold and there is at least one valid instrument among L instruments. To illustrate, consider Figure 1 with two mutually independent instruments.

[Figure 1 about here.]

In Figure 1 (a) where both instruments are valid or in Figures 1 (b) and (c) where one of the two instruments is invalid, D is a collider, but Y is not, thanks to the lack of an edge between D and Y under the null hypothesis of no effect. Then, by the property of (non)-colliders (see Cole et al. (2009) for one illustration), Z_1 and Z_2 must be conditionally independent on Y under $H_0 : \beta^* = 0$, i.e. $Z_1 \perp\!\!\!\perp Z_2 | Y$. Critically, the conditional independence does not require us knowing which instrument is invalid or valid a priori; it doesn't require knowing which of the three graphs generated the data, so long as at least one instrument is valid.

The intuition above generalizes to more than two instruments. Formally, let $\boldsymbol{\Sigma} = (\sigma_{jk}) \in \mathbb{R}^{(L+1)\times(L+1)}$ be the covariance matrix of the instrument-outcome pair $(\boldsymbol{Z}^T, Y)$ where $\boldsymbol{Z}^T = (Z_1, \ldots, Z_L) \in \mathbb{R}^L$. Similar to the case with two instruments, if the L instruments are mutually independent of each other, conditioning on Y does not induce a collider bias between a valid instrument Z_j and any other $L - 1$ candidate instruments under the null causal effect $H_0 : \beta^* = 0$, regardless of whether these $L - 1$ candidate instruments are valid or not. The theorem below formalizes this observation and translates the null hypothesis of no effect into a collection of (conditional) independence tests.

Theorem 2: *Suppose there is at least one valid instrument among L candidate instruments. Then, the null of no treatment effect $H_0 : \beta^* = 0$ and having at least one valid instrument that is not correlated with any of other candidates is equivalent to the null of*

$$H_0 : \prod_{j=1,2,\ldots,L} \sigma_{j,L+1} = 0. \tag{7}$$

The proof is in Web Appendix A; see Drton (2006) for another example. While the results above formally rely on independence between instruments, it is possible to have dependent instruments at the expense of having a more complex null hypothesis in (7). Also, from a Mendelian randomization standpoint, we can enforce instruments to be independent of each other by choosing SNPs that are far apart from each other in genetic distance. Finally, our result differs from a recent work by Marden et al. (2018) who also proposed to use collider bias, but to test the presence of selection bias using a single instrument.

To test the null in (7), we adapt the test statistic proposed by Drton (2006) and Drton (2009) which is based on a likelihood ratio test for Gaussian graphical models that allow for some singularity constraints. Specifically, consider the following model

$$\begin{aligned} \boldsymbol{Z}_{i\cdot} &\sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_Z), \quad \boldsymbol{\Sigma}_Z = \text{diag}(\upsilon_1^2, \upsilon_2^2, \cdots, \upsilon_L^2) \\ D_i &= \boldsymbol{Z}_{i\cdot}^T \boldsymbol{\gamma}^* + \xi_i \\ Y_i &= \boldsymbol{Z}_{i\cdot}^T \boldsymbol{\pi}^* + D_i \beta^* + \epsilon_i, \quad E(\epsilon_i, \xi_i | \boldsymbol{Z}_{i\cdot}) = 0 \\ \begin{pmatrix} \epsilon_i \\ \xi_i \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_2^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}\right] \end{aligned} \tag{8}$$

The setup in (8) is a special case of model (2) with the additional assumptions that (i) D_i is linearly associated to $\boldsymbol{Z}_{i\cdot}$, (ii) the error terms are bivariate i.i.d. Normal with an unknown covariance matrix, and (iii) $\boldsymbol{\Sigma}_Z$ is diagonal. These additional assumptions are used to derive the asymptotic null distribution of the proposed test in (9) below; they are not needed to establish the relationships between the null hypotheses in Theorem 2.

Let $\boldsymbol{S}^{(L)} = (s_{jk}) \in \mathbb{R}^{(L+1)\times(L+1)}$ be the sample covariance of $(\boldsymbol{Z}_n, \boldsymbol{Y}_n)$. We propose to test (7) by computing the smallest determinant of sub-matrices of the estimated covariance matrix $\boldsymbol{S}^{(L)}$, i.e.

$$\lambda_n = \min_{j=1,2,\ldots,L} \left(n \log \left(\frac{s_{jj} \det(\boldsymbol{S}^{(L)}_{-j,-j})}{\det(\boldsymbol{S}^{(L)})} \right) \right) \tag{9}$$

We call λ_n the collider bias test and Theorem 3 shows the limiting null distribution of λ_n.

Theorem 3: *Let $\boldsymbol{W} = (W_{jk})$ be a $L \times L$ symmetric matrix where each entry is an independent χ_1^2 random variable. Let $V^* \subset \{1, 2, \ldots, L\}$ be a set of valid instruments among L instruments and $v^* = c(V^*)$. Under model (8) and the null hypothesis of no effect, $H_0 : \beta^* = 0$, the collider bias test statistic λ_n in (9) converges to the minimum of χ_L^2-distributed random variables, which we denote as $\underline{\chi}^2_{L,v^*}$*

$$\lambda_n \overset{n\rightarrow\infty}{\longrightarrow} \min_{j \in V^*} \left(\sum_{k=1}^{L} W_{jk} \right) := \underline{\chi}^2_{L,v^*}. \tag{10}$$

For any size $\alpha \in (0, 1)$, we can use the asymptotic null distribution of λ_n in Theorem 3 to obtain a critical value $\underline{\chi}^2_{L,v^*,1-\alpha}$, which is the $1 - \alpha$ quantile of the $\underline{\chi}^2_{L,v^*}$ distribution. We would reject the null of no effect if the observed λ_n exceeds the critical value. Theorem 3 also shows that the asymptotic null distribution of the collider bias test λ_n does not depend on the exact set of valid instruments V^*; it only depends on the number of valid instruments v^*. Finally, for a fixed α, as the number of valid instruments v^* increases, the critical value becomes smaller. In other words, by allowing a greater number of invalid instruments into our test statistic, we push the critical value farther away from zero and make it harder to reject the null hypothesis of no effect.

4.2 *Minimum of Wald Tests*

We also present another test proposed by the referee to test the null of no effect when the instruments are mutually independent of each other. The test is motivated by writing Y_i in (8) in its reduced-form (i.e. only as a function of Z_i), which is $Y_i = \mathbf{Z}_{i\cdot}^T \mathbf{\Gamma}^* + \xi_i \beta^* + \epsilon_i$ where $\mathbf{\Gamma}^* = (\Gamma_1^*, \ldots, \Gamma_L^*)^T = (\boldsymbol{\pi}^* + \boldsymbol{\gamma}^* \beta^*)$. Let $\hat{\Gamma}_j^*$ be the maximum likelihood estimator (MLE) of Γ_j^*. Then under the null of no treatment treatment, when instrument j is valid, the Wald test statistic $w_j = (\hat{\Gamma}^*)_j^2 / Var(\hat{\Gamma}_j^*)$ follows an asymptotic χ_1^2 distribution. Also, if there is one valid instrument and the instruments are independent of each other, the minimum of L Wald statistics $w_1, \ldots, w_L$ will still be asymptotically χ_1^2. Theorem 4 generalizes this idea and shows that the minimum of L Wald statistics is distributed as the minimum of v^* independent χ_1^2 where v^* is the number of valid instruments.

Theorem 4: *Suppose there is at least one valid instrument among L instruments and all instruments are independent of each other. Then, under the null $H_0 : \beta^* = 0$,*

$$W_{n,L} = \min_{j=1,2,\ldots,L} \left(\frac{\hat{\Gamma}_j^2}{\widehat{Var}(\hat{\Gamma}_j)} \right) \xrightarrow{d} \underline{\chi}^2_{1,v^*}, \tag{11}$$

where $\widehat{Var}(\hat{\Gamma}_j)$ is a consistent estimator of $Var(\hat{\Gamma}_j)$.

We call $W_{n,L}$ the minimum of Wald tests. Overall, in comparison to the method in Section 3, a disadvantage of both the collider bias test and the minimum of Wald tests is that they do not directly produce confidence intervals for β^*; they only produces evidence against the null hypothesis of no effect. But, the collider bias test and the minimum of Wald tests have much lower computational overhead because both do not require unions. Regardless, all the methods we presented can handle a very small proportion of valid instruments and maintain the correct size; our methods only require one valid instrument while other methods in the literature require more valid instruments.

5. Simulation Study With Invalid Instruments

We conduct a simulation study to evaluate the performance of our methods when invalid instruments are present. The simulation setup follows equation (8) with $n = 1000$ individuals, $L = 10$ candidate instruments, and each instrument is independent from each other. For each simulation setting, we generate 1000 independent replicates. We test the null causal effect $H_0 : \beta^* = 0$ and vary β^*. We change $\boldsymbol{\pi}^*$'s support from 0 to 0.1 and vary the number of invalid instruments (i.e. the number of 0.1's in $\boldsymbol{\pi}^*$) by changing the number of non-zero $\boldsymbol{\pi}^*$'s. We set $\sigma_1 = \sigma_2 = 2$, $\rho = 0.8$, and $v_j^2 = 2$ for $j = 1, \ldots, L$. We set $\gamma_j^* = \{\theta\sigma_1^2/(n\sigma_z^2)\}^{1/2}$ and consider $\theta = 25$. The parameter θ is closely related to the concentration parameter and is a measure of instrument strength (Stock et al., 2002) when only valid instruments are present. Here, $\theta = 25$ represents weak instruments; in Web Appendix E, we consider $\theta = 150$ representing strong instruments and a mixture of both strong and weak instruments.

We compare the statistical power between the union method, the collider bias test, and the minimum of Wald tests. We vary the true number of invalid instruments s^*. For the union method, we set the upper bound on s^* to be $\bar{s} = s^* + 1$, and use the Anderson-Rubin test and the conditional likelihood ratio test since they are robust to weak instruments.

[Figure 2 about here.]

Figure 2 presents the power of different methods under weak instruments. When $s^* = 1$, the union methods have similar power as the oracle method; here, the oracle methods refer to methods that know exactly which instruments are valid and invalid a priori. But, when $s^* = 5$, i.e. when 50% of instruments are invalid, the union methods have near-oracle power if the treatment effect β^* is positive, but they have smaller power than the oracle methods when β^* is negative. This asymmetric power may arise from the inflection points of the likelihood function (Kleibergen, 2007) as well as non-identifiability issues when $\pi_j^* + \beta^* \gamma_j^* = 0$. The minimum of Wald tests has less power than the union methods except when $s^* = 5$ and $\beta^* < 0$ and $s^* = 9$. Finally, the collider bias test generally has the smallest power compared to the other methods.

Tables 1 presents the coverage proportion and the median lengths of 95% confidence intervals when the instruments are weak and $\bar{s} = 5$; again, the true $\beta^* = 0$. As expected, compared to the naive methods, our method provides at least 95% coverage with invalid instruments. But, our methods are conservative and only as s^* gets close to $\bar{s}$ do we see that our methods' coverage rates reach close to 95%. We also noticed that as $\bar{s}^*$ increased, the Anderson-Rubin test produces slightly shorter intervals than the CLR test when invalid instruments are present.

[Table 1 about here.]

Web Appendix E presents additional simulation studies with different test statistics, s^*, and strength. In brief, when the instruments are strong with $\theta = 150$ and s^* is close to $\bar{s}$, the Anderson-Rubin test and the pretesting methods with two-stage least squares or conditional likelihood ratio test perform well with respect to power and length, with the Anderson-Rubin test being the simpler alternative since it doesn't use a pretest. Between the Anderson-Rubin test and the conditional likelihood ratio test, the Anderson-Rubin test dominates the conditional likelihood ratio test for $s^* > 0$. This finding differs from the advice in the weak instruments literature where the conditional likelihood ratio test generally dominates the

Anderson-Rubin test (Andrews et al., 2006; Mikusheva, 2010). Finally, when we have a mix of strong and weak invalid instruments, the results are somewhere in between the strong instrument case and the weak instrument case presented here.

Overall, the simulation studies suggest that there is no uniformly dominant test for the treatment effect across all scenarios concerning invalid instruments. The performance depends both on the number of invalid instruments and instrument strength. Nevertheless, for practice, we recommend the union method using the Anderson-Rubin test as it has decent power across various scenarios.

6. Data Analysis: Mendelian Randomization in the Framingham Heart Study

We use our proposed methods to study the effect of low-density lipoprotein (LDL-C) on globulin levels among individuals in the Framingham Heart Study (FHS) Offspring Cohort. Over several decades, the FHS has been one of the most popular epidemiologic cohort studies to identify risk factors for CVD, and recently, Mendelian randomization has been used to uncover causal relationships in the presence of unmeasured confounding (Smith et al., 2014; Mendelson et al., 2017). Traditional Mendelian randomization requires every instrument to be valid in order to test for a treatment effect, a tall order for many studies. Our proposed methods relax this requirement and allow some of the instruments to be invalid.

For the main analysis, we selected nine SNPs that are known to be significantly associated with LDL-C measured in mg/dL (Kathiresan et al., 2007; Ma et al., 2010; Smith et al., 2014) and are located in different chromosomes or in linkage equilibrium; the latter ensures that candidate instruments are, to the best extent possible, statistically independent each other. Our outcome Y is continuous globulin levels (g/L) measured at Offspring Cohort Exam 2. Globulins are known to have an important role in liver function, clotting, and immune system. We also use subjects' age and sex as covariates $\boldsymbol{X}$. Web Appendix F has additional details about the data.

To mitigate concerns for cryptic relatedness (Voight and Pritchard, 2005; Astle and Balding, 2009), which means genetic variants and phenotypes may share the same correlation structure due to familial relationships among the subjects, we selected one subject at random from each family in the Offspring Cohort linked by sibling or marriage relationships; note that there is no parent-child relationship within the Offspring Cohort. By doing so, we retain only 60% of the subjects ($n = 1445$ for the main analysis and $n = 1,480$ for the secondary analysis) from $n = 2,499$ with no missing data. This may reduce power of our analysis (Pierce et al., 2010), but possibly lead to a more valid analysis of the true effect (Lee and Ogburn, 2019).

Table 2 summarizes results from testing the null of no effect using our procedures; we vary $\bar{s}$ from 1 to 8. The test statistic from the minimum of Wald tests is $W_{L,n} = 0.0134$ and the collider bias test statistic is $\lambda_n = 9.6571$. As $\bar{s}$ varies, the critical values of the minimum of Wald tests and the collider bias test become larger and the null becomes harder to reject. We see that the collider bias test is able to reject the null of no effect even if there are at most 6 invalid instruments ($\bar{s} = 7$); in contrast, the minimum of Wald tests is not able to reject the null of no effect even with no invalid instruments. Among the union methods, the Anderson-Rubin test is able to reject the null of no effect when there is at most one invalid instrument and the two-stage least squares method is able to reject the null when there is no invalid instruments; the conditional likelihood ratio test failed to reject the null even if there are no invalid instruments.

[Table 2 about here.]

The empirical results concerning the collider bias test are contrary to what we found in the simulation study where the collider bias test had less power than the minimum of Wald tests. This may be due to a number of reasons, but we found evidence of correlations between some pairs of nine instruments, possibly due to cryptic relationships or kinships that may be present even after considering at most one subject from each family in our data analysis. In particular, when we ran a regression between four strongest SNPs (rs11591147, rs646776, rs2228671, rs2075650) where one of the four SNPs is regressed on others, we find that one of the three regression coefficients had p-values less than or equal to 0.05. Given this, we suspect that the collider bias test may be more sensitive to the presence of correlations than the minimum of Wald tests so it rejects the null (7) even when small correlations between instruments are present. To partially verify this, Web Appendix F conducts additional analysis where we change the candidate instruments from nine to four instruments. These four instruments were selected based on strength plus weak correlation amongst each other. With four instruments, we found that we can reject the null with more invalid instruments present. Also, matching the simulation study, the collider bias test cannot reject the null when there is more than one invalid instrument.

Table 3 shows the 95% confidence intervals based on the union method as we vary $\bar{s}$ varies from 1 to 3. First, we see that when $\bar{s} = 1$ and we don't allow any invalid instruments, the AR and the Sargan-based methods return empty intervals, suggesting that invalid instruments among the candidate set of instruments are present. As $\bar{s}$ moves away from 1 we see that most confidence intervals indicate a positive effect of LDL-C on globulin levels. Also, similar to Table 2, the confidence intervals from the Anderson-Rubin test and the two-stage least squares test with a pre-test do not cover 0 even if there is at most one invalid instrument. Overall, the data analysis shows evidence of a positive causal effect between LDL-C and globulin levels and this conclusion is robust so long as there are a few invalid instruments.

[Table 3 about here.]

7. Discussion

This paper proposes two methods to conduct inference for the treatment effect when instruments are possibly invalid. The first method is a simple modification of pre-existing methods in instrumental variables to construct robust confidence intervals for the causal effect. The second method is based on testing the null hypothesis of no causal effect and produces valid inference so long as there is at least one valid instrument. We show through numerical experiments and data analysis how our proposed methods can be used to strengthen conclusions about the presence and direction of a causal effect using instrumental variables when invalid instruments may be present.

Acknowledgements

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No.N01-HC-25195 and HHSN268201500001I). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Funding for CARe genotyping was provided by NHLBI Contract N01-HC-65226. The research of Hyunseung Kang was supported in part by NSF Grant DMS-1811414. The research of Tony Cai was supported in part by NSF Grants DMS-1712735 and DMS-2015259 and NIH grants R01-GM129781 and R01-GM123056.

References

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* **20,** 46–63.

Andrews, D. W. and Lu, B. (2001). Consistent model and moment selection procedures for gmm estimation with application to dynamic panel data models. *Journal of Econometrics* **101,** 123–164.

Andrews, D. W. K. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica* **67,** 543–563.

Andrews, D. W. K., Moreira, M. J., and Stock, J. H. (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* **74,** 715–752.

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.

Astle, W. and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science* **24,** 451–471.

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

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International Journal of Epidemiology* **44,** 512–525.

Cheng, X. and Liao, Z. (2015). Select the valid and relevant moments: An information-based lasso for gmm with many moments. *Journal of Econometrics* **186,** 443–464.

Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., and Poole, C. (2009). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* **39,** 417–420.

Conley, T. G., Hansen, C. B., and Rossi, P. E. (2012). Plausibly exogenous. *Review of Economics and Statistics* **94,** 260–272.

Davey Smith, G. and Ebrahim, S. (2003). 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32,** 1–22.

Davey Smith, G. and Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology* **33,** 30–42.

Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics.* Oxford University Press, New York.

Drton, M. (2006). Algebraic techniques for gaussian models. *arXiv preprint math/0610679* .

Drton, M. s. (2009). Likelihood ratio tests and singularities. *The Annals of Statistics* **37,** 979–1012.

Guo, Z., Kang, H., Cai, T. T., and Small, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80,** 793–815.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50,** 1029–1054.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology* **17,** 360–372.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology* **18,** 449–484.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association* **111,** 132–144.

Kathiresan, S., Manning, A. K., Demissie, S., D'agostino, R. B., Surti, A., Guiducci, C., Gianniny, L., Burtt, N. P., Melander, O., Orho-Melander, M., et al. (2007). A genome-wide association study for blood lipid phenotypes in the framingham heart study. *BMC Medical Genetics* **8,** S17.

Kleibergen, F. (2007). Generalizing weak instrument robust iv statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. *Journal of*

Econometrics **139,** 181–216.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., and Imbens, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics* **33,** 474–484.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27,** 1133–1163.

Lee, Y. and Ogburn, E. L. (2019). Network dependence and confounding by network structure lead to invalid inference. *arXiv preprint arXiv:1908.00520* .

Liao, Z. (2013). Adaptive gmm shrinkage estimation with consistent moment selection. *Econometric Theory* **29,** 857–904.

Ma, L., Yang, J., Runesha, H. B., Tanaka, T., Ferrucci, L., Bandinelli, S., and Da, Y. (2010). Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the framingham heart study data. *BMC Medical Genetics* **11,** 55.

Marden, J. R., Wang, L., Tchetgen, E. J. T., Walter, S., Glymour, M. M., and Wirth, K. E. (2018). Implementation of instrumental variable bounds for data missing not at random. *Epidemiology (Cambridge, Mass.)* **29,** 364–368.

Mendelson, M. M., Marioni, R. E., Joehanes, R., Liu, C., Hedman, Å. K., Aslibekyan, S., Demerath, E. W., Guan, W., Zhi, D., Yao, C., et al. (2017). Association of body mass index with dna methylation and gene expression in blood cells and relations to cardiometabolic disease: a mendelian randomization approach. *PLoS medicine* **14,** e1002215.

Mikusheva, A. (2010). Robust confidence sets in the presence of weak instruments. *Journal of Econometrics* **157,** 236–247.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*

71, 1027–1048.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *The Journal of Economic Perspectives* **20,** 111–132.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9 (with discussion) translated in statistical sciences. *Statistical Science* **5,** 465–472.

Pierce, B. L., Ahsan, H., and VanderWeele, T. J. (2010). Power and instrument strength requirements for mendelian randomization studies using multiple genetic variants. *International Journal of Epidemiology* **40,** 740–752.

Rosenbaum, P. R. (2002). *Observational Studies.* Springer Series in Statistics. Springer-Verlag, New York, second edition.

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

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* **26,** 393–415.

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association* **102,** 1049–1058.

Smith, J. G., Luk, K., Schulz, C.-A., Engert, J. C., Do, R., Hindy, G., Rukh, G., Dufresne, L., Almgren, P., Owens, D. S., et al. (2014). Association of low-density lipoprotein cholesterol–related genetic variants with aortic valve calcium and incident aortic stenosis. *Journal of the American Medical Association* **312,** 1764–1771.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14,** 483–495.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic*

Statistics **20,** 518–529.

Voight, B. F. and Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genet* **1,** e32.

Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association* **114,** 1339–1350.

Windmeijer, F., Liang, X., Hartwig, F., and Bowden, J. (2018). The confidence interval method for selecting valid instruments. Technical report, Technical report, University of Bristol.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* MIT press, second edition.

Zhao, Q., Wang, J., Hemani, G., Bowden, J., and Small, D. S. (2020+). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *The Annals of Statistics* .

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

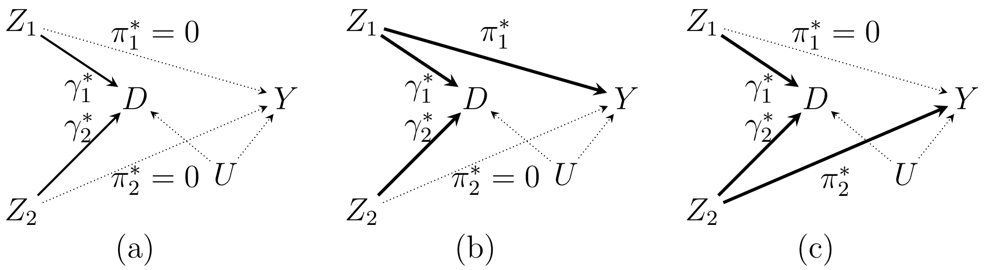

Figure 1: Causal directed acyclic graph with two candidate instruments Z_1 and Z_2 when $H_0 : \beta^* = 0$ holds. Solid lines are non-zero causal paths and dotted lines are *possibly non-zero* causal paths. A variable U indicates an unmeasured confounder between D and Y. We use γ^* and π^* to label each edge. Our setup supposes that at least one instrument is valid, i.e. $\pi_1^* \pi_2^* = 0$, without knowing which π^* is zero.

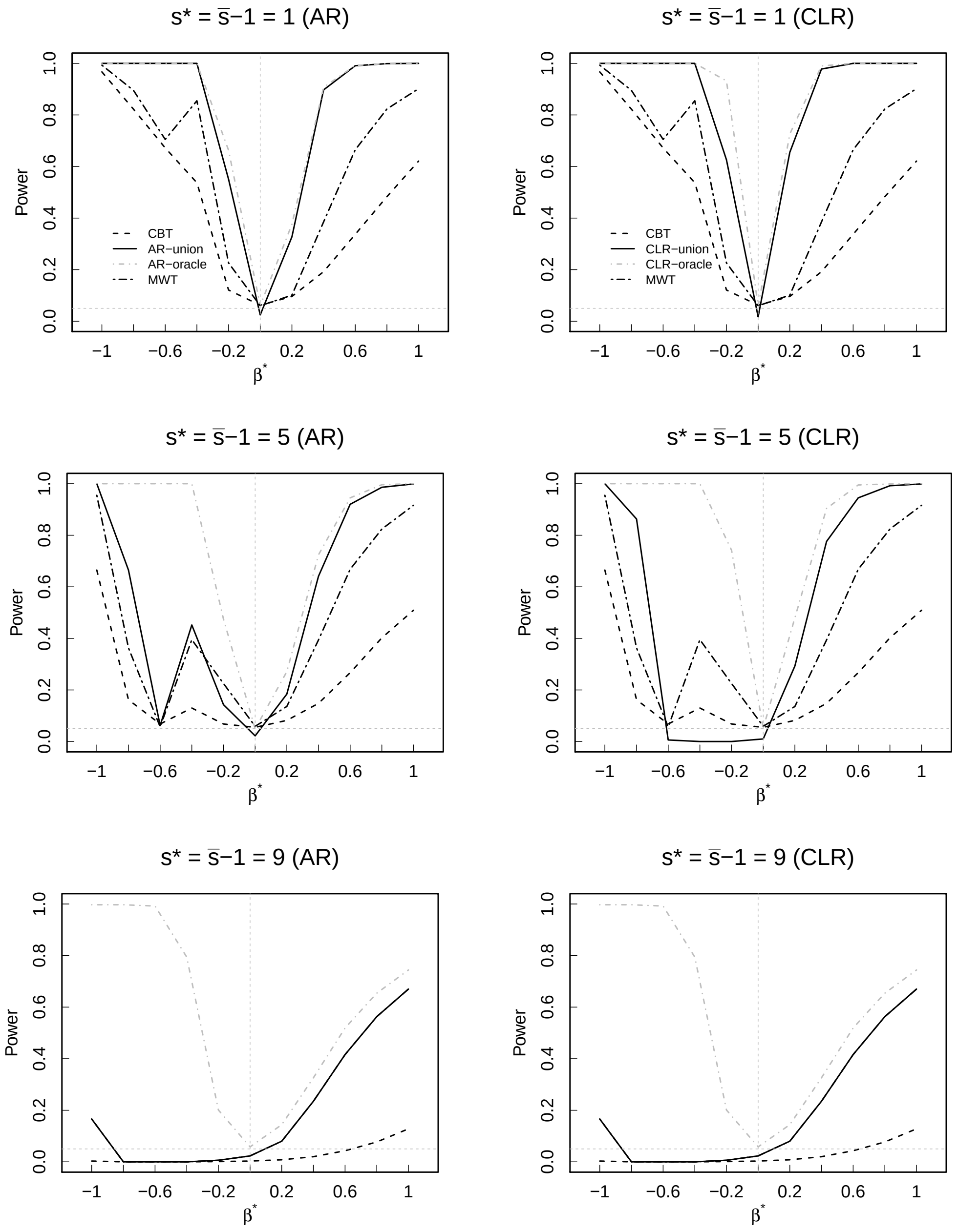

Figure 2: Power of different methods with varying numbers of invalid instruments. The instruments are weak and the Type I error is fixed at $\alpha = 0.05$. AR denotes the Anderson-Rubin test, CBT denotes the collider bias test, CLR denotes the conditional likelihood ratio test, and MWT denotes the minimum of Wald tests. The oracle tests are tests that know exactly which instruments are valid.

Table 1: Coverage rates (CR) and median length of 95% confidence intervals under weak instrumental strength and $\bar{s} = 5$. TSLS denotes two-stage least squares, AR denotes the Anderson-Rubin test, and CLR denotes the conditional likelihood ratio test.

Method	Test	$s^* = 0$		$s^* = 1$		$s^* = 2$		$s^* = 3$		$s^* = 4$	
		CR	Length	CR	Length	CR	Length	CR	Length	CR	Length
Naive	AR	93.1	0.400	44.9	0.066	11.0	0.000	1.9	0.000	0.2	0.000
	CLR	94.3	0.253	91.9	0.253	78.3	0.251	61.6	0.246	33.7	0.240
Our method	AR	100.0	0.881	100.0	0.827	100.0	0.767	99.5	0.691	97.3	0.604
	CLR	100.0	0.604	100.0	0.637	100.0	0.677	99.9	0.725	98.7	0.770
Oracle	AR	93.1	0.400	93.9	0.421	93.4	0.433	93.7	0.455	95.1	0.488
	CLR	94.3	0.253	95.0	0.268	95.2	0.285	94.5	0.306	94.5	0.332

Table 2: P-Values of testing the null Hypothesis of no effect of low-density lipoprotein on globulin levels using different methods. AR denotes the Anderson-Rubin test, CLR denotes the conditional likelihood ratio test, TSLS denotes two-stage least squares, Sargan denotes the Sargan pre-test, MWT denotes the minimum of Wald tests, and CBT denotes the collider bias test. R denotes rejecting the null of no effect and NR denotes not rejecting the null of no effect.

Test statistics	$\bar{s} = 1$	$\bar{s} = 2$	$\bar{s} = 3$	$\bar{s} = 4$	$\bar{s} = 5$	$\bar{s} = 6$	$\bar{s} = 7$	$\bar{s} = 8$
AR	R	R	NR	NR	NR	NR	NR	NR
CLR	NR	NR	NR	NR	NR	NR	NR	NR
TSLS	R	NR	NR	NR	NR	NR	NR	NR
SarganTSLS	R	R	NR	NR	NR	NR	NR	NR
MWT (p-value)	0.4226	0.4649	0.5128	0.5629	0.6189	0.6804	0.7486	0.8235
CBT (p-value)	0.0025	0.0029	0.0041	0.0082	0.0156	0.0339	0.0746	0.3796

Table 3: 95% Confidence intervals for the effect of low-density lipoprotein on globulin levels using the union method. AR denotes the Anderson-Rubin test, CLR denotes the conditional likelihood ratio test, TSLS denotes two-stage least squares, and Sargan denotes the Sargan pre-test. We vary $\bar{s}$, the total number of allowable invalid instruments from 1 to 3, with $\bar{s} = 1$ indicating no invalid instruments and $\bar{s} = 3$ indicating at most two invalid instruments. Cells with NA represent empty confidence intervals.

	$\bar{s} = 1$	$\bar{s} = 2$	$\bar{s} = 3$
AR	NA	(0.0073, 0.0646)	(-0.0208, 0.0811)
CLR	(-0.0001, 0.0508)	(-0.0172, 0.0620)	(-0.0365, 0.0742)
TSLS	(0.0012, 0.0446)	(-0.0119, 0.0553)	(-0.0265, 0.0661)
SarganTSLS	NA	(0.0067, 0.0586)	(-0.0304, 0.0698)
SarganCLR	NA	(0.0067, 0.0654)	(-0.0422, 0.0793)