Instrumental variables: Don't throw the baby out with the bathwater

Soumerai and Koppel¹ make the case against the use of instrumental variables (IVs) in outcomes research. While they do not explicitly call for a halt to the use of all IV analyses based on observational data in health services research, they generally argue against their use. Here, we present a counterargument on several points and advocate for the careful use of IV methods. We also highlight a number of ways to improve the quality of research based on IV designs.

1 | A BRIEF REVIEW AND DEFENSE OF INSTRUMENTAL VARIABLES

Confounding by indication is a critical challenge in evaluating the effectiveness of medical interventions. It is widely understood that randomization to treatment is the best analytic technique for addressing confounding by indication. The primary alternative to randomized trials is to assume that there are no unmeasured confounders. In economics, this is often referred to as “selection on observables” to indicate that selection into the treatment must be based on observed data only. Under this assumption, the investigator assumes there are no unobservable differences between the treated and control groups. While this assumption may be plausible in some settings, in health services research, treatments are purposefully chosen and administrative datasets do not record all the reasons why a particular treatment was administered or withheld.

One alternative to the assumption of no unmeasured confounding is to identify an instrument. Instruments arise primarily in two ways. First, in randomized experiments with noncompliance, randomization to treatment arms serves as an instrument for compliance with the experimental protocol. The second type of instrument is a form of natural experiment, where some circumstance produces haphazard encouragement for units to be exposed to a treatment. As such, one can view the search for natural experiments as a search for conditions where patients are either haphazardly assigned to treatments or haphazardly encouraged to take a treatment. For example, Sanwald and Schober² use relative distance as a haphazard nudge for treatment at a hospital with a catheterization laboratory. It is this second type of instrument that is critiqued by Soumerai and Koppel.¹ Note that below, we will speak of subjects being assigned to an IV. For natural experiment IVs, the assignment to the IV is often not explicit. For example, when the IV is based on distance, assignment to the IV occurs when people select their place of residence.

In either setting, for a variable to be an instrument the following three core assumptions must hold: (a) The IV is associated with the

exposure, (b) the IV is randomly or as-if randomly assigned, and (c) the IV does not have a direct effect on the outcome.³ Assumptions (b) and (c) are often combined into a single assumption related to the instrument being uncorrelated with the outcome model error term. Using the IV framework in randomized trials makes it clear why these are two separate assumptions. When the IV is randomized, assumption (b) holds by design, but assumption (c) can clearly fail. See Baiocchi et al,⁴ [Section 3] for an example where assumption (b) holds but assumption (c) may not.

Critically, assumptions (b) and (c) cannot be directly tested with data. The fact that these assumptions cannot be tested has led to the development of a series of falsification tests which allow the investigator to probe the IV assumptions. A falsification test allows analysts to demonstrate that a key IV assumption does not fail. That is, a falsification test cannot prove an assumption holds, but it can provide decisive evidence that an assumption is likely invalid. As we outline below, falsification tests are a critical way to judge the quality of an instrument.

The primary advantage of an IV design is that, subject to its assumptions holding, an IV provides a consistent estimate of the causal effect of the exposure on the outcome even in the presence of unobserved confounding between the exposure and the outcome. While randomized designs produce the highest quality evidence, IV methods are valuable since they offer the promise of reducing bias from unobserved confounders outside of randomization. Recent research outlines the promise an IV analysis. Davies et al⁵ find IV estimates that are consistent with the results from randomized trials, while estimates based on the assumption of no unmeasured confounding contradict the results from the trials. Imbens and Rosenbaum⁶ provide a particularly lucid justification for the use of instruments: The goal is to replace the implausible assumption of selection on observables with the more plausible, though not certain, IV assumptions. This statement highlights the fact that IV assumptions are best judged relative to other feasible research designs. Moreover, while the application and use of an IV is subtle and requires considerable care, we would argue that, despite the complexities, the ability to reduce bias from unmeasured confounding ensures that IV methods should remain part of the health services research toolkit. In particular, the strongest evidence base is often built from finding similar results across different research designs.⁶ Thus, an IV design or different IV designs based on different instruments can comprise one research design used in conjunction with other designs to build a robust evidence base.

2 | A REBUTTAL OF CLAIMS

Next, we review and rebut two key claims made by Soumerai and Koppel.¹ First, they outline the possibility of confounding between the IV and the outcome as a key shortcoming in IV designs. Such confounding is a violation of assumption (b). Their argument is an elaboration of a review of IV studies in Garabedian et al⁷ that contends that many IV designs may be subject to this form confounding due to the fact that they identified possible confounders for 65 studies. Soumerai and Koppel¹ conclude, “the majority of IVs are not a reliable way to control for bias in medical effectiveness research.”

We argue that a more honest assessment of the IV design is not whether there may be some bias due to confounding, but whether the IV is more biased relative to a study that controls for observed exposure-outcome confounders under the selection on observables assumption. High-quality IVs have a haphazard element to them to reduce bias from instrument outcome confounders. For example, Keele et al⁸ use an IV based on a physician's preference for operative care in emergency admissions. This is a setting where patients are much less likely to select their physician based on the type of medical treatment he or she prefers, and physicians rotate based on preset schedules. The key idea is that assignment to a physician with a specific treatment preference is closer to random than assignment to a specific treatment.

While assignment to an IV may not be perfectly as-if random, it should be less confounded than assignment to the exposure of interest. The most commonly applied falsification test for IVs focuses on the level of bias reduction due to baseline covariates being balanced by IV status. More specifically, if the IV behaves in an as-if random fashion, our expectation is that sample means of baseline covariates should not differ significantly across levels of the IV. If the means of baseline covariates differ significantly by IV status, assumption (a) has been falsified. In the first use of distance to a medical facility as an IV, the authors demonstrated that patient severity was balanced by IV status.⁹ Ideally, baseline measures of key confounders are available for this type of analysis.

Recent research has demonstrated how investigators can formally analyze the balance associated with the IV relative to the balance that results from assuming selection on observables holds.¹⁰⁻¹² Using these methods, researchers can provide evidence on whether the IV is less confounded than the treatment of interest—that is, researcher can state whether there appears to be less bias associated with an IV analysis as compared to an analysis that uses risk adjustment. One weakness of the analysis in Sanwald and Schober² is that they present rather limited results for this type of falsification test. In general, we think it is more useful to critique specific applications of IV based on a comparative study of bias. That is, does the IV have less bias than if the study were conducted assuming selection on observables? Using existing methods, analysts can provide clear evidence on this point. For example, Keele et al⁸ present several analyses to demonstrate whether there is more bias under IV relative to risk adjustment.

Overall, we would argue that a general critique of IV based on the possibility of confounding is unhelpful, since empirical evidence can shed light on whether an instrument appears to be valid. Moreover, falsification tests can also shed light on whether a proposed instrument should be abandoned. If a proposed instrument does little to balance baseline covariates, that is one indication the IV is invalid. Like all falsification tests, balance does not prove that assumption (b) holds, but widespread imbalance can demonstrate that assumption (b) is implausible.

Next, Soumerai and Koppel¹ argue that a key limitation of using an IV design is that it is a cross-sectional data analytic technique. Specifically, based on arguments in Soumerai et al,¹³ they argue that longitudinal analyses such as interrupted time series are generally superior research designs. Research designs based on longitudinal data have some advantages; however, statistical adjustment for confounders using longitudinal data is a variation on the assumption of no unmeasured confounding. Typically, in these designs one has to assume that either the unobserved confounders are time-invariant or that there are no unobserved differences across treated and control group after adjusting for past outcomes. In general, there is little reason to think either is true in a wide set of health services research applications. See O'Neill et al¹⁴ for a review of the assumptions necessary for many longitudinal designs. Morgan and Winship,¹⁵ [ch. 11] provide a useful overview of the strong assumptions needed for interrupted time series designs. We would argue that longitudinal research designs are useful, but certainly no panacea. Moreover, in general, IV designs should have a longitudinal component. Baseline covariates should be collected before assignment to the IV to avoid bias from conditioning on a concomitant variable.¹⁶ Measurement of the IV should be prior to the exposure and outcome. In addition, many longitudinal designs can be used in conjunction with an instrument. For example, one can combine two-way fixed effects with an instrument. See Acemoglu and Angrist¹⁷ for one well-known example. In sum, the longitudinal designs championed by Soumerai and Koppel¹ offer little protection from unobserved confounders, while a properly executed IV design may reduce the bias from confounding.

3 | KEYS TO BETTER ANALYSIS OF IV DESIGNS

We conclude with an overview of best practice for the application of IVs. The advice we offer here is a brief summary of information available elsewhere. Interested readers should review more complete guidelines in Baiocchi et al⁴ and Swanson and Hernán.¹⁸ First, analysts should seek to identify instruments that have a haphazard or as-if random element to their assignment. IVs should be accompanied by a bias analysis, so that readers can judge if the IV outcome relationship appears to be less confounded than the exposure-outcome relationship. Second, the stronger the instrument the better. Small and Rosenbaum¹⁹ show that stronger instruments are more resistant

to bias from violations of assumption (b). Matching methods can be used to make instruments stronger.^{20,21}

Next, when possible, researchers should employ falsification tests. We outlined balance testing as one form of falsification testing. However, other forms of falsification tests are possible. More specifically, Yang et al²² show that while assumptions (b) and (c) cannot be tested, one can introduce an additional assumption and test the conjunction of these three assumptions under a single hypothesis. If the hypothesis is rejected, then at least one of the three assumptions must be false. The additional assumption can take a number of forms including: (a) assuming the IV does not affect the exposure in subgroup;^{23,24} (b) assuming the treatment does not affect the outcome in one subgroup;²² and (c) identifying an alternative outcome that is not affected by the treatment but would be affected by potential confounders.²⁵ For example, in Sanwald and Schober,² assume we can identify a subset of patients that would always be treated at a hospital with a catheterization laboratory regardless of how far they lived from this type of hospital. Perhaps these patients exhibit a common comorbidity that requires treatment at this type of hospital. A falsification test would consist of testing whether the instrument has any effect on the outcome in this always-treated subgroup of patients. If the exclusion restriction holds, the IV should have no effect on the outcome within this subgroup. Investigators should also attempt to use other types of falsification tests. Identification of a negative control or placebo outcome is another way to falsify IV assumptions.^{26,27} See Kang et al²⁴ for an example of an IV design of this type, and Davies et al¹¹ for details on negative controls in IV studies. Proposed IVs that fail falsification tests are generally invalid IVs.

An IV provides the investigator with a consistent estimate for compliers—the patient for whom the IV is decisive in their assignment to the treatment.²⁸ As such, effects from IV studies are applicable to marginal patients in a study. This may present additional challenges in terms of understanding whether evidence from an IV study applies to specific set of patients. However, three strategies can help clinicians better understand how well IV effects generalize. First, analysts should present descriptive statistics for compliers to aid in the identification of marginal patients.^{4,29} That is, if the data reveal that the marginal patient is more likely to be septic, have an APACHE II score in a specific range, and be older, this should help clinicians understand whether a specific patient is marginal for treatment based on IV evidence. Alternatively, the investigator may find that the marginal population differs little from the larger patient population. When this is the case, the IV effect may be viewed as more general. See Keele et al⁸ for an example of describing the complier population in a clinical investigation. Second, investigators can use the IV estimate to place bounds on the average treatment effect for the study patient population.³⁰ When the bounds on the average treatment effect clearly show benefits from treatment, concerns about whether a specific patient is marginal are less relevant since treatment should be beneficial on average. Finally, researchers should also attempt to benchmark IV estimates against the results from randomized trials. Benchmarking

against trials serves two purposes. First, it serves to validate the IV assumptions. Second, differences between the IV estimate and trial estimate can also help clinicians understand how generalizable result from an IV study is.

In general, while IV designs depend on untestable assumptions, there are a variety of methods available for probing those assumptions. Evidence based on IVs that pass this litany of tests and diagnostics is, we would argue, an important way to evaluate treatments and protect against bias from confounding by indication. Soumerai and Koppel¹ argue that the IV analysis in Sanwald and Schober² is flawed to the point that IV designs should be ruled out. We would contend that while one could present evidence that the IV in Sanwald and Schober² is invalid, Soumerai and Koppel¹ do not present such evidence.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: We received no financial support, there are no conflicts of interest.

Disclosures: None.

ORCID

Luke Keele <https://orcid.org/0000-0002-3859-2713>

Luke Keele PhD

Dylan Small PhD

University of Pennsylvania, Philadelphia, Pennsylvania

Correspondence

Luke Keele, PhD, University of Pennsylvania, Philadelphia, PA.

Email: luke.keele@gmail.com

REFERENCES

1. Soumerai SB, Koppel R. The reliability of instrumental variables in health care effectiveness research: less is more. *Health Serv Res.* 2017;52:9-15.
2. Sanwald A, Schober T. Follow your heart: survival chances and costs after heart attacks—an instrumental variable approach. *Health Serv Res.* 2017;52:16-34.
3. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc.* 1996;91:444-455.
4. Baiochi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Stat Med.* 2014;33:2297-2340.
5. Davies NM, Smith GD, Windmeijer F, Martin RM. Cox-2 selective nonsteroidal anti-inflammatory drugs and risk of gastrointestinal tract complications and myocardial infarction: an instrumental variable analysis. *Epidemiology.* 2013;24:352-362.
6. Imbens GW, Rosenbaum P. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *J R Stat Soc Ser A Stat Soc.* 2005;168:109-126.
7. Garabedian LF, Chu P, Toh S, Zaslavsky AM, Soumerai SB. Potential bias of instrumental variable analyses for observational comparative effectiveness research. *Ann Intern Med.* 2014;161:131-138.
8. Keele LJ, Sharoky CE, Sellers MM, Wirtalla CJ, Kelz RR. An instrumental variables design for the effect of emergency general surgery. *Epidemiol Methods.* 2018;7.

9. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?: analysis using instrumental variables. *JAMA*. 1994;272:859-866.
10. Jackson JW, Swanson SA. Toward a clearer portrayal of confounding bias in instrumental variable applications. *Epidemiology*. 2015;26:498-504.
11. Davies NM, Thomas KH, Taylor AE, et al. How to compare instrumental variable and conventional regression analyses using negative controls and bias plots. *Int J Epidemiol*. 2017;46:2067-2077.
12. Zhao Q, Small DS. Graphical diagnosis of confounding bias in instrumental variables analysis. *Epidemiology*. 2018;29:29-31.
13. Soumerai SB, Starr D, Majumdar SR. How do you know which health care effectiveness research you can trust? A guide to study design for the perplexed. *Prevent Chronic Dis*. 2015;12:150-187.
14. O'Neill S, Kreif N, Grieve R, Sutton M, Sekhon JS. Estimating causal effects: considering three alternatives to difference-in-differences estimation. *Health Serv Outcomes Res Method*. 2016;16:1-21.
15. Morgan SL, Winship C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd edn. New York, NY: Cambridge University Press; 2014.
16. Rosenbaum PR. The consequences of adjusting for a concomitant variable that has been affected by the treatment. *J R Stat Soc Ser A Stat Soc*. 1984;147:656-666.
17. Acemoglu D, Angrist J. How large are human-capital externalities? Evidence from compulsory schooling laws. *NBER Macroecon Annual*. 2000;15:9-59.
18. Swanson SA, Hernán MA. Commentary: how to report instrumental variable analyses (suggestions welcome). *Epidemiology*. 2013;24:370-374.
19. Small D, Rosenbaum PR. War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *J Am Stat Assoc*. 2008;103:924-933.
20. Baiocchi M, Small DS, Lorch S, Rosenbaum PR. Building a stronger instrument in an observational study of perinatal care for premature infants. *J Am Stat Assoc*. 2010;105:1285-1296.
21. Baiocchi M, Small DS, Yang L, Polsky D, Groeneveld PW. Near/far matching: a study design approach to instrumental variables. *Health Serv Outcomes Res Method*. 2012;12:237-253.
22. Yang F, Zubizarreta J, Small DS, Lorch S, Rosenbaum P. Dissonant conclusions when testing the validity of an instrumental variable. *Am Stat*. 2014;68:253-263.
23. Glymour MM, Tchetgen Tchetgen EJ, Robins JM. Credible mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *Am J Epidemiol*. 2012;175:332-339.
24. Kang H, Kreuels B, Adjei O, Krumkamp R, May J, Small DS. The causal effect of malaria on stunting: a mendelian randomization and matching approach. *Int J Epidemiol*. 2013;42:1390-1398.
25. Pizer SD. Falsification testing of instrumental variables methods for comparative effectiveness research. *Health Serv Res*. 2016;51:790-811.
26. Lipsitch M, Tchetgen ET, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010;21:383-388.
27. Rosenbaum PR. *Observational Studies*, 2nd edn. New York, NY: Springer; 2002.
28. Harris KM, Remler DK. Who is the marginal patient? understanding instrumental variables estimates of treatment effects. *Health Serv Res*. 1998;33:1337-1360.
29. Angrist JD, Pischke JS. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press; 2009.
30. Small D, Tan Z, Ramsahi R, Lorch S, Brookhart A. Instrumental variable estimation with a stochastic monotonicity assumption. *Stat Sci*. 2017;32:561-579.

SUPPORTING INFORMATION

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