Instrumental variables: The power of wishful thinking vs the confounded reality of comparative effectiveness research

1 | INTRODUCTION

Keele and Small responded to our article on instrumental variables (IVs) published in Health Services Research in February 2017.^{1,2} Here, we address their efforts to defend IVs and we present additional evidence of the unreliability of IVs in comparative effectiveness research (CER). We appreciate that some economists, statisticians, and other IV adherents are emboldened by their faith in the power of weak cross-sectional associations to accurately reflect the world. But health outcomes research requires confronting the interrelatedness of social and medical factors—almost always a confounded reality with unmeasured and, indeed, unknown variables.³ That is, most IV studies assume life is far less confounded than it is.

Rassen et al⁴ define an instrumental variable (IV) as “an unconfounded proxy for a study exposure that can be used to estimate a causal effect in the presence of unmeasured confounding.” In other words, it assumes that a common variable (eg, differences between regions in rates of a medical treatment [eg, prostatectomy]) can assume randomization in estimating the effect on mortality—without having to worry about unknown confounding. In this wonderfully convenient but almost always unrealistic theory, IVs can randomize exposure to medical interventions. For example, IV proponents used distance from patients' homes to the hospital as an instrumental variable (IV) that presumably “randomizes” early MI treatments to estimate effects on mortality rates.^{2,5-7} In these examples, however, the IV is likely associated with other critical variables (eg, urban/rural status, socioeconomic characteristics, health status) that impact health outcomes. The IV is confounded and therefore produces a biased effect estimate.

Even after more than a quarter century of IV research in health care, most instrumental variables still violate the assumptions that the IV-Outcome relationship is not confounded.^{2,3} Studies that use IV also ignore or accept inadequate measures of countless confounders (eg, health status, clinical risk factors, geography, procedure volume, access to care, and SES—part of a longer list we discuss below).³

2 | RESPONSES TO SEVERAL OF KEELE AND SMALL'S STATEMENTS ABOUT OUR ARTICLE

1. Keele et al state that we “generally argue against [the use of IVs].” We agree that we are deeply skeptical of most uses of

IV, but we do not reject all uses. For instance, we believe that some lotteries (eg, the draft board lotteries in the 1960s)^{8,9} and other randomized experiments usually serve as valid IVs. The authors, however, equate our and Cochrane's rejection of most IVs as a weak research design^{10,11} with “throwing the baby out with the bathwater.” The IV mentioned above, distance to the hospital, or other IV favorites, such as regional variations in health care, or variation in care by hospitals or physicians, are untrustworthy for the obvious reason that they do not “randomize” and are biased by a multitude of likely confounders.³

2. We disagree with Keele and Small's statement that our previous article failed to provide evidence that the IV in Sanwald and Schober is invalid.^{2,7} In fact, the IV they use (distance to medical facility again) could only distinguish hospitals with or without cardiac catheterization laboratories. But it ignored other medical interventions for acute MI plus other unmeasured differences in care quality that affect mortality.¹² Absent knowledge of causes of mortality the IV provides no useful information.⁵ It also suffers from many other confounders of distance IVs, a small list of which are reproduced in Table 1 below. For example, the landmark McClellan et al's³ catheterization laboratories study showed that the small “mortality effects” of cardiac catheterization occurred before exposure to the procedure, an illustration of the reverse causality seen in many cross-sectional IV studies. The “randomization” assumption was violated. The “effect” was due to many confounders of distance, such as rurality, SES, health, and access to other treatments, like medications.

This misuse of IVs (generally, weak cross-sectional associations) has consequences beyond academic debates. Inappropriate research methods offer no or false guidance for policy makers and can even cause wasteful resource allocation and patient harms.¹³

3 | COMMENTS ON KEELE AND SMALL'S DEFENSE OF IVS

1. The authors offer the general IV justification that IVs are a “haphazard element” that “randomize” or “as-if” randomize treatment (eg, distance from the hospital randomizes catheterization for acute MIs). Oddly for a defense, they admit this

assumption “cannot be tested with data”, but can be examined via “falsification tests” to see whether analyses of a few possible confounding variables change their conclusion. Unfortunately, IVs are confounded by numerous known or unmeasured variables.^{3,14} Many powerful confounders are usually ignored by proponents of IV and are rarely included in the study data.

For example, in the study mentioned above where investigators examined the effects of distance to hospitals with catheterization laboratories on acute MI mortality,⁵ their Medicare datasets do not capture use of life-saving drug treatment in hospitals—like aspirin and reperfusion therapy.¹⁵ But hospital medication use covaries with the IV, distance to the hospital, rurality, and mortality.¹⁶ Thus, the absence of data on life-saving hospital medications violates a basic assumption of IV use and offers a strong alternative pathway to mortality.

2. For another example of critical but unmeasured variables, consider research on rural health outcomes that routinely highlight the hundreds of associations of health-related (potentially confounding) variables with the IV, distance to the hospital, and mortality.¹⁷⁻¹⁹ They are not in the IV dataset in over half of distance IV/mortality studies,³ nor can they be wished away by a limited subgroup analysis of a few variables. Thus, if a rural resident (ie, “distant” from a hospital) is poor, less educated, cannot get an ambulance quickly, and is dying of a heart attack in transit), IV users assume the death may be treated as a consequence of lack of access to revascularization in the hospital rather than poor SES, poor health, limited rural ER access, or the deficit in many alternative medical procedures they do not examine (see Figure 1).³
3. Keele and Strong also discuss the comparative strength of IVs vs “other feasible designs.” But, the leading research design texts show that only a few quasi-experimental designs protect against threats to internal validity, such as history and selection.^{13,20,21} Cross-sectional, correlational IV studies do not control for any of these biases.

4 | TWO STUDIES ILLUSTRATING THE BIASES CAUSED BY CONFOUNDED IVS

It is instructive to expand on the example of the landmark JAMA study by McClellan et al.⁵ It is the most cited, extant IV study, finding a small 5 percent reduction in mortality among acute MI patients with closer relative distance to hospitals with cardiac catheterization laboratories. It was followed by a large increase in similar IV studies in the following 25 years.³ Ironically, the study provides solid evidence for the invalidity of such methods, even before many researchers attempted to replicate the unreliable findings:

- The more distant patients already experienced the **total** reported 5 percent greater mortality than the closer patients in the first **day** of hospitalization—largely before any treatment or mortality benefits occur. This small inversely causal difference was due to IV-Outcome confounding and nonequivalence.
- The more distant patients were sicker, not randomly equivalent to closer patients. For example, distant MI patients receiving fewer cardiac procedures were almost 10 times more likely to have a rural residence,⁵ a predictor of MI mortality.³ Also, more distant patients were about half as likely to be admitted to a high-volume hospital, a major predictor of better outcomes. These and other variables were associated with both mortality and distance to the hospital, a violation of IV assumptions.
- We also found that rural residents (the majority of “distant patients”) are less likely to receive life-saving medications for acute MI than closer patients.¹⁵

5 | IV STUDY SUGGESTING THE MORTALITY CONSEQUENCES OF ADVANCED AMBULANCES

- Perhaps the most dramatic example of heroic assumptions is seen in Sanghavi et al's^{10,14,22,23} use of an equally biased IV—comparing (cross-sectionally) the rate of use of advanced vs basic ambulances (the latter with less life-saving equipment and trained personnel) by county—on survival among Medicare patients transported to the hospital. Moreover, many IV-Outcome confounders were poorly measured in the Medicare claims, such as illness severity: often the real cause of increased mortality, not the dispatching of advanced ambulances.
- Emergency medical clinicians often triage advanced life support ambulances to save critically ill patients.¹⁴ The figure below, from a large study in Washington State, indicates only a few of the patients' illnesses at the time they were transported by advanced vs basic ambulances. Those transported in advanced ambulances with trained personnel were already more likely to die before they reached the hospital: an example of a correlational IV study of reversed cause and effect (Figure 2).

Despite these severe limitations, the article's authors proclaimed the national policy implications of their study, even calculating that many millions could be saved if the United States abandoned more expensive advanced ambulances. Such overconfidence in interpretation of a flawed correlational study demonstrates yet another IV study, that if believed by policy makers, could have unintended adverse health consequences by defunding life-saving interventions.

6 | POWERFUL CONFOUNDERS OF THE MOST COMMON IV-MORTALITY STUDIES IN THE ONLY SYSTEMATIC REVIEW THAT OBSERVED NON-STUDY DATA

Methodological leaders in CER and IV recognized the crucial importance of closely examining other data sources and common sense/experience to identify IV-Outcome confounders, such as concomitant medical treatments.²⁴

- To our knowledge, Garabedian et al³ published the only study that searched the worldwide literature to identify unmeasured confounders of the most common four IVs and mortality (including distance to the hospital). They discovered hundreds of confounders (Table 1). Because the individual confounders are too numerous to display, here we only list the categories in Table 1. Every IV study examined (n = 65) had at least one confounder.³ Of course, the number of true confounders is grossly underestimated because studies of many IV-Outcome confounders have not been published.

TABLE 1 Instrumental variables and categories of associated confounders (with example and variables)

	Instrumental variable category			
	Distance to facility instrument	Regional variation instrument	Facility variation instrument	Physician variation instrument
Confounder category violating IV assumptions (associated with IV and mortality)				
Geographic location	Urban/rural, U.S. region, and absolute distance (for relative difference)	Urban/rural, U.S. region	Urban/rural	
Patient	Race, education, income, age, insurance status, health status/comorbid conditions, and health behaviors	Race, education, income, age, insurance status, health status/comorbid conditions, and health behaviors	Race, education, income, age, insurance status, health status/comorbid conditions, and health behaviors	Race, education, income, age, sex, insurance status, health status/comorbid conditions, and health behaviors
Provider supply		Number of hospital beds or nursing homes		
Technology adoption and utilization		Invasive cardiac procedures, radical prostatectomies, prescribing behavior, and practice patterns		
Treatment	Receipt of other treatments, time to treatment, and transfer status	Receipt of other treatments, time to treatment, and transfer status	Receipt of other treatments	Receipt of other treatments
Facility	Procedure volume, facility volume, clinical services offered, departments, teaching status, profit status, trauma designation, delivery system type, and practice type	Procedure volume, facility volume, clinical services offered, departments, teaching status, profit status, trauma designation, delivery system type, and practice type	Procedure volume, facility volume, clinical services offered, departments, teaching status, profit status, trauma designation, delivery system type, and practice type	
Physician				Use other treatments
Health system				Reimbursement policies, regional variation, and facility variation confounders, that is, geographic location, provider supply, technology adoption and utilization, and facility characteristics

Note: There is an overwhelming bias toward undercounting confounders. Many are unknown, many are unmeasured or unmeasurable, and all are potential threats to the validity of results by those using IVs.

Adapted from: Garabedian et al.³ Printed with permission from Annals of Internal Medicine.

FIGURE 1 The arrows to IV-Outcome confounders show that confounders, such as SES, poor health, and rural residence, may cause hospital mortality, possibly more than the hypothesized access to catheterization laboratories. [Arrows indicate associations]. The IV is biased if the IV Outcome Related Through Unadjusted 3rd Variable. Adapted from: Garabedian et al.³ Printed with permission from Annals of Internal Medicine

FIGURE 2 Several serious conditions of patients transported in advance life support vs basic life support ambulance
Source: Prekker et al.²⁷

7 | USE OF WEAK STUDY DESIGNS FOR CONTROLLING THREATS TO THE VALIDITY OF IVS

In the majority of cases, observational IV studies (eg, distance IVs that prompted these debates) only observe outcomes after treatment. They cannot measure pre- and post- changes in outcomes and

are at the bottom of the hierarchy of nonexperimental (or quasi-experimental) designs (Table 2). This is why the most respected international body reviewing medical evidence (the Cochrane Collaboration) still does not include most IV designs in its systematic reviews of health interventions, based on the hierarchy of strong and weak study designs and the relative inability of cross-sectional designs (most IV studies) to control for most biases.²¹

TABLE 2 Hierarchy of strong and weak designs, based on capacity to control for biases (Cochrane Effective Practice and Organisation of Care (EPoC) 2017, Soumerai SB, Starr D, Majumdar SR 2015, and Naci H, Soumerai SB, 2016)^{11,13,28}

Strong design: often trustworthy effects	
Multiple RCTs	The "gold standard" of evidence, incorporating systematic review of all RCTs of an intervention (eg, random assignment of smoking cessation treatment)
Single RCT	A single, strong randomized experiment, but sometimes not generalizable
Interrupted time series with control series	Baseline trends often allow visible effects and control for biases. This design has two controls: baseline trend and control group to measure sudden discontinuities in trend soon after an intervention
Intermediate designs: sometimes trustworthy effects	
Single interrupted time series	Controls for trends, but no comparison group (see above)
Before and after with comparison group	Pre- and postchange using single observations. Comparability of baseline trend often unknown
Weak designs: rarely trustworthy effects (no controls for common biases. excluded from literature syntheses)	
Uncontrolled before and after (pre-post)	Simple observations before and after intervention, no baseline trend or control group
Cross-sectional designs	Simple correlation, no baseline, and no measure of change. Includes most IVs

Source: Soumerai et al.²⁹

8 | CONCLUSION

The evidence presented here should give scholars pause before relying on IVs in CER research. The complexity of health, health care, and social factors almost ensures the presence of major confounding factors that seldom, if ever can be addressed with incomplete datasets, unfounded assumptions and complex but unrealistic statistics. While IVs are extraordinarily convenient, they can also be extraordinarily misleading.

Scholars and policy makers are eager to find efficiencies in medicine. Health care consumes almost one-fifth (18 percent) of the gross national product.²⁵ But we cannot rely on research methods that often mislead us because of unmeasured and ignored variables. Increasingly, we should insist on strong experimental and quasi-experimental research designs (such as RCTs, controlled interrupted time-series designs, and systematic reviews) in pilot tests of expensive policies. Private and taxpayer investments should be based on solid evidence of safety and efficacy.²⁶ We should not base policy on weak research designs that may lead to perverse effects, such as unsustainable costs and policies that fail to improve medical care.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: Parts of this article were presented at the International Society for Pharmacoepidemiology

(ISPE) 10th Asian Conference on Pharmacoepidemiology on October 29, 2017, in Brisbane, Australia. The title of the presentation was: "Instrumental Variable Analysis and Interrupted Time Series Analysis in Health Policy Research: You Can't Fix by Adjustment What You Bungled by Design."

ORCID

Stephen B. Soumerai <https://orcid.org/0000-0001-6811-2496>

Stephen B. Soumerai ScD¹

Ross Koppel PhD^{2,3}

¹Harvard Medical School Department of Population Medicine and Harvard Pilgrim Health Care, Boston, Massachusetts

²Department of Biomedical Informatics, University of Pennsylvania, Philadelphia, Pennsylvania

³Department of Biomedical Informatics, University at Buffalo (SUNY), Buffalo, New York

Correspondence

Stephen B. Soumerai, ScD, Harvard Medical School Department of Population Medicine and Harvard Pilgrim Health Care, Boston, MA.

Email: ssoumerai@hms.harvard.edu

REFERENCES

- Keele L, Small D. Instrumental variables: don't throw the baby out with the bathwater. *Health Serv Res.* 2018. <https://doi.org/10.1111/1475-6773.13130>
- Soumerai SB, Koppel R. The reliability of instrumental variables in health care effectiveness research: less is more. *Health Serv Res.* 2017;52(1):9-15.
- 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(2):131.
- Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J Clin Epidemiol.* 2009;62(12):1226-1232.
- 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(11):859-866.
- Chen J, Krumholz HM, Wang Y, et al. Differences in patient survival after acute myocardial infarction by hospital capability of performing percutaneous coronary intervention: implications for regionalization. *Arch Intern Med.* 2010;170(5):433-439.
- Sanwald A, Schober T. Follow your heart: survival chances and costs after heart attacks—an instrumental variable approach. *Health Serv Res.* 2017;52(1):16-34.
- Selective Service System. The Vietnam Lotteries. The Vietnam Lotteries. <https://www.sss.gov/About/History-And-Records/lotter1>. Published 2018. Accessed October 11, 2018.
- Angrist JD. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. *Am Econ Rev.* 1990;80(3):313-336.
- Soumerai SB. Outcomes of basic versus advanced life support for out-of-hospital medical emergencies. (Letter to the editor). *Ann Intern Med.* 2016;165(1):68-69.
- Cochrane Effective Practice and Organisation of Care (EPoC). What study designs can be considered for inclusion in an EPoC review and what should that be called? <https://epoc.cochrane.org>

- org/sites/epoc.cochrane.org/files/public/uploads/Resources-for-authors2017/what_study_designs_should_be_included_in_an_epoc_review.pdf. Published 2017. Accessed October 11, 2018.
12. Ibanez B, James S, Agewall S, et al. 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation. The Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). *Eur Heart J*. 2018;39(2):119-177.
 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. *Prev Chronic Dis*. 2015;12:E101.
 14. Soumerai SB, Koppel R. Don't Let Weak Research Influence Policies with Life and Death Consequences | THCB. 2017. <http://thehealthcareblog.com/blog/2017/06/14/dont-let-weak-research-influence-policies-with-life-and-death-consequences/>. Accessed September 25, 2018.
 15. Willison DJ, Soumerai SB, Palmer RH. Association of physician and hospital volume with use of aspirin and reperfusion therapy in acute myocardial infarction. *Med Care*. 2000;38(11):1092.
 16. Taylor HA, Hughes GD, Garrison RJ. Cardiovascular disease among women residing in rural America: epidemiology, explanations, and challenges. *Am J Public Health*. 2002;92(4):548-551.
 17. Kulshreshtha A, Goyal A, Dabhadkar K, Veledar E, Vaccarino V. Urban-rural differences in coronary heart disease mortality in the United States: 1999-2009. *Public Health Rep*. 2014;129(1):19-29.
 18. Abrams TE, Vaughan-Sarrazin M, Fan VS, Kaboli PJ. Geographic isolation and the risk for chronic obstructive pulmonary disease-related mortality: a cohort study. *Ann Intern Med*. 2011;155(2):80.
 19. Khan JA, Casper M, George M, et al. Geographic and sociodemographic disparities in drive times to joint commission-certified primary stroke centers in North Carolina, South Carolina, and Georgia. *Prev Chronic Dis*. 2011;8(4):A-79.
 20. Shadish W, Cook T, Campbell D. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth Cengage Learning; 2002.
 21. Cochrane Effective Practice and Organisation of Care (EPOC). EPOC resources for review authors. EPOC resources for review authors. <https://epoc.cochrane.org/resources/epoc-resources-review-authors>. Published 2017. Accessed September 24, 2018.
 22. Sanghavi P, Jena AB, Newhouse JP, Zaslavsky AM. Outcomes of basic versus advanced life support for out-of-hospital medical emergencies. *Ann Intern Med*. 2015;163(9):681-690.
 23. Soumerai S, Koppel R. How bad science can lead to bad science journalism — and bad policy. *The Washington Post*. https://www.washingtonpost.com/posteverything/wp/2017/06/07/how-bad-science-can-lead-to-bad-science-journalism-and-bad-policy/?utm_term=.a61f018a89e9. Published 2017. Accessed September 24, 2018.
 24. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf*. 2010;19(6):537-554.
 25. Centers for Medicare & Medicaid Services. National Health Accounts Historical. <https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html>. Published January 8, 2018. Accessed September 25, 2018.
 26. Wharam JF, Daniels N. Toward evidence-based policy making and standardized assessment of health policy reform. *JAMA*. 2007;298(6):676-679.
 27. Prekker ME, Feemster LC, Hough CL, et al. The epidemiology and outcome of prehospital respiratory distress. *Acad Emerg Med*. 2014;21(5):543-550.
 28. Naci H, Soumerai SB. History bias, study design, and the unfulfilled promise of pay-for-performance policies in health care. *Prev Chronic Dis*. 2016;13:E82.
 29. Soumerai SB, Ceccarelli R, Koppel R. False dichotomies and health policy research designs: randomized trials are not always the answer. *J Gen Intern Med*. 2017;32:204-209.

SUPPORTING INFORMATION

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