Journal of Clinical Epidemiology 149 (2022) 203–205

Journal of Clinical Epidemiology

KEY CONCEPTS IN CLINICAL EPIDEMIOLOGY

A Gentle Introduction to Instrumental Variables

Tarjei Widding-Havneraas[a,b,*], Henrik Daae Zachrisson[c]

[a]Department of Clinical Medicine, University of Bergen, Bergen, Norway
[b]Centre for Research and Education in Forensic Psychiatry, Haukeland University Hospital, Bergen, Norway
[c]Department of Special Needs Education, University of Oslo, Oslo, Norway

Accepted 30 June 2022; Published online 7 July 2022

Abstract

Instrumental variables (IV) is a central strategy for identifying causal effects in absence of randomized experiments. Clinicians and epidemiologists may find the intuition of IV easy to grasp by comparison to randomized experiments. Randomization is an ideal IV because treatment is assigned randomly, and hence unaffected by everything else. IV methods in nonexperimental settings mimic a randomized experiment by using a source of "as good as" random variation in treatment instead. The main challenge with IV designs is to find IVs that are as good as randomization. Discovering potential IVs require substantive knowledge and an understanding of design principles. Moreover, IV methods recover causal effects for a subset of the population who take treatment when induced by the IV. Sometimes these estimates are informative, other times their relevance is questionable. We provide an introduction to IV methods in clinical epidemiology. First, we introduce the main principles and assumptions. Second, we present practical examples based on Mendelian randomization and provider preference and refer to other common IVs in health. Third, practical steps in IV analysis are presented. Fourth, the promise and perils of IV methods are discussed. Finally, we suggest further readings.

1. Mimicking a randomized experiment

Clinicians and epidemiologists strive to improve health through interventions. Decisions on which treatments or policies to pursue require causal knowledge [1]. Causal effects demands comparable treatment and control groups (i.e., exchangeability) [2]. While comparable groups are expected by design in randomized controlled trials (RCT), observational studies are often challenged by confounding bias. As most observational methods can only adjust for measured confounders, ruling out unmeasured confounding is often unrealistic. A common issue in clinical epidemiology is confounding by indication where treatment decisions are based on potentially unmeasured patient characteristics such as disease severity [2]. Figure 1A presents a data-generating process, a directed acyclic graph (DAG) [2], for this scenario. Here the interest lies in the effect of a treatment, D, on an outcome, Y, but any causal interpretation is precluded by unmeasured patient characteristics, U. Instrumental variables (IV) are appealing as these methods can provide causal effects from observational data even with unmeasured confounding [4].

Conceptually, an IV can be compared to randomization in RCTs. IV methods, like RCTs, depend on random variation in treatment for comparable groups. But in contrast to RCTs with investigator-led randomization, IV methods instead exploit a source of as good as random variation in treatment and are thus considered quasi-experimental designs [4]. The main challenge is to find a credible IV which is as good as random in allocating people to treatment.

In Figure 1B, Z is an IV to the effect of D on Y. IV relies on three main conditions. A valid IV, Z, must (i) predict treatment status ("relevance"), (ii) only affect Y through D ("exclusion"), and (iii) be as good as randomly assigned ("independence"). These conditions are met as there's a causal path $Z \rightarrow D$ and no open paths between Z and Y except through D.

The randomization indicator in a double-blind placebo-controlled trial is a valid IV that meets conditions (i)-(iii) by design. Randomization increases the probability of receiving treatment among people assigned to treatment, while exclusion and independence is expected by double-blindness and random assignment of Z [2]. With observational data, however, researchers must combine

Declaration of interest: The authors declare that they have no conflict of interest.

* Corresponding author. Department of Clinical Medicine, University of Bergen, Jonas Lies Vei 87, 5021, Bergen, Norway, Tel.: +4792451487; fax: +4755973321.

E-mail address: tarjei.widding-havneras@uib.no (T. Widding-Havneraas).

https://doi.org/10.1016/j.jclinepi.2022.06.022

creativeness, knowledge of the field, and design principles to find potential IVs.

A fourth condition (iv) of identical or homogenous treatment effects is required to estimate the average treatment effect (ATE). In health settings, however, effects are often heterogenous (i.e., vary by people) which require an alternative fourth condition of monotonicity (i.e., *Z* only affects *D* in one direction). IV methods only exploit variation in treatment induced by the IV. Hence, under conditions (i)-(iii) and monotonicity, IV methods retrieve the local average treatment effect (LATE), which is the average treatment effect for people whose treatment was determined by the IV (''compliers''). Compliers are expected to consist of comparable treatment and control groups and thus link IV to the RCTs we mimic [5].

Only condition (i) is possible to empirically verify while conditions (ii)-(iv) must be assumed [2]. Applications of IV methods therefore strongly depend on building a case for the validity of the IV design based on substantive knowledge, logic, and empirical justifications [4].

2. Proposed IVs

An increasingly popular type of IV analysis is Mendelian randomization (MR) which uses genetic variation associated with a treatment of interest as an IV. MR applies the randomization of the genome at conception [6]. For example, MR has been used to estimate causal effects of alcohol consumption on cardiovascular health. RCTs are infeasible and conventional observational studies are hampered, e.g., by reverse causation due to lower alcohol consumption in people with poor health and confounding bias induced by other health and social characteristics.

Specifically, the rs671-A allele in aldehyde dehydrogenase 2 (ALDH2) gene, involved in alcoholic metabolism and prevalent in Asian populations, has been used as an IV [7]. Carriers consume less alcohol on average compared to noncarriers, likely due to adverse reactions (e.g., nausea) [7]. Cho et al. [7] use this IV in a Korean sample; the proportion of variance in alcohol use caused by the allele variant (i.e., to which individuals are genetically randomized), is used to estimate the association between alcohol consumption and cardiovascular health. All other causes of alcohol use are by design excluded from the estimated effect. Results show that alcohol consumption increases risk of hypertension and blood pressure. IV methods here contribute to a topic where the evidence is mixed and causal knowledge is crucial.

Several nongenetic IVs are proposed in health research. Provider preference IVs (PP IV) are increasingly used, too [8]. PP IV designs assume that variation in clinicians' treatment preference for similar patients induces random variation in patients' treatment status. PP IVs have been used for treatment effects in medical specialities such as cancer, cardiology, and psychiatry [8]. A considerable literature exist on methodological concerns in MR, PP IV, and other IVs in health [2,6,9]. For more proposed IVs, including distance to provider, day of hospital admission, and calendar time, see, e.g., Brookhart et al. [10] and Glymour and Swanson [9].

3. Practical steps

Researchers should start by considering the estimand of interest and whether conditions (i)-(iv) are likely to hold [11]. Second, data availability is key. IV usually requires large datasets. Several databases include genetic data suitable for MR analyses (see, e.g., overview in Davies et al. [12]). For nongenetic IV designs, large health surveys, cohort studies, or administrative data are well-suited. Third, preregistration of statistical analyses on platforms such as Open Science Framework can improve overall transparancy and credibility. Fourth, with data, relevance can be empirically verified while plausibility of exclusion, independence, and monotonicity can be assessed by combining substantive knowledge and falsification tests [13]. Finally, reporting guidelines makes it easier to evaluate validity and interpret findings for researchers, clinicians, and other readers [14]. Guidelines are developed for MR analyses [14], and Swanson & Hernan [15] and Brookhart et al. [10] are helpful for IV in general.

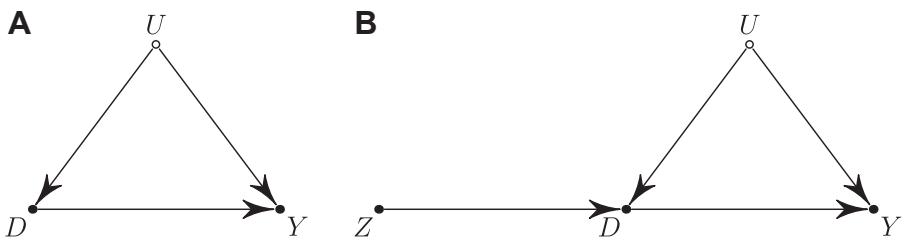

Fig. 1. Directed acyclic graphs for confounding bias and instrumental variables. (A) presents a DAG where the effect of *D* on *Y* is biased by unmeasured confounding, *U*. The total association between *D* and *Y* consists of a causal ($D\rightarrow Y$) and confounding ($D\leftarrow U\rightarrow Y$) component. As *U* is unmeasured, the confounding back-door ($D\leftarrow U\rightarrow Y$) cannot be blocked by conditioning and the effect of *D* on *Y* remains confounded [3]. (B) presents a DAG where *Z* is a valid IV for the effect of *D* on *Y*. Instead of aiming to close the confounding back-door through conditioning, *Z* isolates causal covariation in *D* and *Y* due to *Z* and ignores confounded covariation. By only using this causal covariation, IV analysis identify causal effects of *D* on *Y* for people whose value on *D* is determined by *Z*.

4. Promise and perils

Becoming familiarized with IV methods forces any researcher to explicitly consider risks associated with confounding bias in any nonexperimental study. Because IVs rely on observational data, as opposed to experiments, they may also have stronger external validity. Combining IV methods with existing data bases can give causal estimates for long-term outcomes whereas RCTs require time to pass. Yet IV methods are not a panacea for causal inference with observational data. Credible IVs are rare and the methodological literature vast. The main concern with IV designs is that the unverifiable assumption of no unmeasured confounding between D and Y is replaced with other unprovable assumptions (e.g., no unmeasured confounding between Z and Y) [2,4]. In sum, IV designs is an attractive solution to the key issue of unmeasured confounding which haunts causal inference from observational data. Applications of IV methods, nonetheless, must consider the relevance of estimates and address strong assumptions.

Acknowledgments

Tarjei Widding-Havneraas is supported by funding from the Western Norway Regional Health Authority (912197) and Research Council Norway (288585/IAR). Henrik Daae Zachrisson's contribution was supported by funding from the European Research Council Consolidator Grant ERC-CoG-2018 EQOP [grant number 818425].

References

[1] Westreich D. Epidemiology by Design: A Causal Approach to the Health Sciences. New York, NY: Oxford University Press; 2019.
[2] Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020.
[3] Elwert F. Graphical Causal Models. In: Morgan SL, editor. Handbook of Causal Analysis for Social Research. Dordrecht: Springer; 2013: 245−73.
[4] Huntington-Klein N. The Effect: An Introduction to Research Design and Causality. Chapman and Hall/CRC; 2022.
[5] Angrist JD. Empirical Strategies in Economics: Illuminating the Path from Cause to Effect. Cambridge: National Bureau of Economic Research; 2022.
[6] Pingault JB, O'Reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. Nat Rev Genet 2018;19:566−80.
[7] Cho Y, Shin S, Won S, Relton CL, Davey Smith G, Shin M. Alcohol intake and cardiovascular risk factors: a Mendelian randomisation study. Sci Rep 2015;5:18422.
[8] Widding-Havneraas T, Chaulagain A, Lyhmann I, Zachrisson HD, Elwert F, Markussen S, et al. Preference-based instrumental variables in health research rely on important and underreported assumptions: a systematic review. J Clin Epidemiol 2021;139:269−78.
[9] Glymour MM, Swanson SA. Instrumental Variables and Quasi-Experimental Approaches. In: Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ, editors. Modern Epidemiology. New York, NY: Wolters Kluwer; 2021:677−709.
[10] Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf 2010;19(6):537−54.
[11] Lundberg I, Johnson R, Stewart BM. What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. Am Soc Rev 2021;86(3):532−65.
[12] Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. BMJ 2018;362:k601.
[13] Labrecque J, Swanson SA. Understanding the Assumptions Underlying Instrumental Variable Analyses: a Brief Review of Falsification Strategies and Related Tools. Curr Epidemiol Rep 2018;5(3): 214−20.
[14] Skrivankova VW, Richmond RC, Woolf BA, Yarmolinsky J, Davies NM, Swanson SA, et al. Strengthening the Reporting of Observational Studies in Epidemiology using Mendelian Randomization: The STROBE-MR Statement. JAMA 2021;326:1614−21.
[15] Swanson SA, Hernán MA. Commentary: How to Report Instrumental Variable Analyses (Suggestions Welcome). Epidemiology 2013;24: 370−4.

Further reading

[1] Huntington-Klein N. Instrumental Variables. The Effect: An Introduction to Research Design and Causality. CRC Press Chapman & Hall; 2021:469−504. Accessible introduction including DAGs and estimation.
[2] Glymour MM, Swanson SA. Instrumental Variables and Quasi-Experimental Approaches. In: Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ, editors. Modern Epidemiology. New York, NY: Wolters Kluwer; 2021:677−709. State-of-the-art epidemiology-oriented introduction with overview of candidate IVs and methodological considerations.
[3] Hernán MA, Robins JM. Instrumental variable estimation. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC; 2020: 193−206. Rigorous standard reference for IV in epidemiology.