Reading and conducting instrumental variable studies: a guide, glossary and checklist for clinicians

Venexia Walker research fellow,[1,2] Eleanor Sanderson Lecturer in Medical Statistics[1,2], Michael G Levin cardiologist[3,4], Scott Damraurer William Maul Measey Associate Professor of Surgery[3,4], Timothy Feeney Research editor[6], and Neil M Davies professor of Medical Statistics[7,8,9]

[1] Medical Research Council Integrative Epidemiology Unit at the University of Bristol, BS8 2BN, United Kingdom. Email:

[2] Population Health Sciences, Bristol Medical School, University of Bristol, Barley House, Oakfield Grove, Bristol, BS8 2BN, United Kingdom.

[3] Division of Cardiovascular Medicine, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia PA.

[4] Corporal Michael J. Crescenz VA Medical Center, Philadelphia, PA

[6] Department of Epidemiology, University of North Carolina Chapel Hill, Chapel Hill NC, USA

[7] Division of Psychiatry, University College London, Maple House, 149 Tottenham Court Rd, London W1T 7NF.

[8] Department of Statistical Science, University College London, London WC1E 6BT, UK

[9] K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Norwegian University of Science and Technology, Norway.

Email:neil.m.davies@ucl.ac.uk, Twitter: @nm_davies, https://orcid.org/0000-0002-2460-0508

Twitter: @venexia, @ECSanderson, @MGLevin, @damrauer, @Tfeend

Keywords: Causal inference, methods primer, instrumental variables, epidemiology

Standfirst

Instrumental variable analysis uses naturally occurring variation to estimate the causal effects of treatments, interventions and risk factors on outcomes in the population from observational data. Under specific assumptions, instrumental variable methods can provide unbiased estimates of causal effects. We explain these assumptions and the information and tests typically reported in instrumental variable studies, which can assess the credibility of the findings of instrumental variable studies.

Summary points

- Instrumental variable analysis is a research method that uses naturally occurring variation (i.e., variation not controlled by the researcher), such as policy decisions, clinical preferences, distance, or time to provide evidence about the causal relationships between interventions and outcomes from observational data.
- Instrumental variables can provide credible evidence about the causal effects even if other observational techniques suffer from residual confounding, reverse causation or other forms of bias.
- We explain and illustrate how to use and estimate instrumental variables studies using commonly available packages.
- In common with all empirical research methods, instrumental variable analysis depends on assumptions readers and reviewers must assess.
- Multiple sources of evidence, using a range of assumptions, can help inform clinical decisions.
- We provide a critical appraisal checklist to help assess and interpret instrumental variable studies.

1. Introduction

In clinical practice, establishing causal relationships is crucial for informed decision-making in patient care. Instrumental variable (IV) analysis is increasingly used to provide evidence about causal effects in clinical research (see **Box 1** for glossary). Instrumental variables are variables that are associated with the intervention but not the outcome (other than through the intervention). They can be used to overcome measured and unmeasured confounding of intervention-outcome associations and provide unbiased estimates of the causal effects of an intervention on an outcome using observational data (**Figure 1**). Instrumental variables are defined by three assumptions (see **Box 2**).

Box 2: The instrumental variable assumptions

Three key assumptions define instrument variables [1]:

1. **Relevance:** [IV1] the instrument must be associated with the intervention.
2. **Independence:** [IV2] the instrument and the outcome have no uncontrolled common causes.
3. **The exclusion restriction:** [IV3] the instrument must only affect the outcome through the intervention.

Instrumental variable analysis has a long history (see **Supplementary Box 1**), with applications in many fields, including healthcare and economics; it has increased in popularity due to the availability of larger datasets and the recognition of the need to obtain reliable estimates when key covariates are not measured, and using different analytical assumptions[2,3]. Researchers increasingly use instrumental variables analyses to inform a wide range of clinical questions. For example, institutional variation in testing or treatment practices have been used as instrumental variables to estimate the effects of perioperatively testing for coronary heart disease on postoperative mortality rates[4], the relative safety of robotic versus laparoscopic surgery for cholecystectomy[5], and the length of storage of red blood cells and patient survival[6]. Physicians' preferences for treatments have been used to investigate the effects of COX-2 versus non-selective NSAIDs on gastric complications[7,8] and the effects of conventional versus atypical antipsychotic medication on elderly patients' mortality[9]. Allocation to treatment in randomized controlled trials with non-compliance is

an instrumental variable previously used to investigate the effects of flexible duty-hour conditions for surgeons on patient outcomes and surgeons' training and well-being[10] and the effects of reducing amyloid levels on cognition[11]. Distance from or time to admission to a particular type of hospital has been used as an instrument for receiving a specific treatment[12,13]. One of the most commonly used applications of instrumental variables is Mendelian randomization – the use of genetic variants as instrumental variables. Mendelian randomization has been covered in detail in previous papers, and will not be discussed here. Nevertheless, the core principles of instrumental variable analysis still apply to that method [14,15].

We provide a practical guide for researchers for reading, interpreting and conducting instrumental variable studies using non-genetic observational data. We discuss: first, why a study should use instruments; second, key concepts and assumptions; third, how to assess the validity of instrumental variable assumptions; and fourth, how to interpret results.

2. Clinical and public health implications

Researchers increasingly use large datasets of electronic medical records, registries, or administrative claims data to provide evidence about the relationships between treatments and patient outcomes. A significant limitation of these datasets is that while the large sample size allows for very precise results, they frequently have inadequate measures of critical confounders. Confounders are variables that affect the likelihood of receiving the intervention and also affect the outcome (for example, prior neuropsychiatric diagnoses and likelihood of being prescribed varenicline rather than nicotine replacement therapy for smoking cessation). Patients do not receive most treatments randomly, and key confounders, such as morbidity and other indications for treatment, are often challenging or impossible to measure with sufficient accuracy from diagnosis/billing codes or are unmeasured/unmeasurable; thus, matching treated individuals with sufficiently comparable controls is challenging and often impossible. As a result, observational analysis of large-scale databases may provide unreliable evidence about different treatment options' comparative effectiveness and safety. This issue is challenging for clinicians and patients, as they need reliable evidence of the causal effects of different treatment options to make well-informed

decisions. Instrumental variables can provide an alternative source of evidence about the effects of different treatments and, while less precise than other approaches, may be less affected by individual-level biases such as confounding by indication, where the indications for treatment also affect the likelihood of an outcome.

3. Why use an instrument?

Most observational methods like multivariable regression or propensity score analysis assume that it is possible to measure a sufficient set of confounders to account for all differences in the outcome between individuals given the intervention and control, except those caused by the intervention[16,17]. However, the correct set of confounders is not always known, and even if they have been identified, measuring and accounting for baseline differences is extremely difficult, which can result in multivariable-adjusted and propensity score analyses having serious biases and providing misleading results. For example, COX-2 inhibitors were developed to cause fewer gastrointestinal complications than traditional NSAIDs and marketed to patients and physicians. As a result, patients prescribed these medications typically were at higher risk of gastrointestinal complications at baseline. Thus, in observational datasets, patients prescribed COX-2s tended to have higher rates of gastrointestinal complications than patients prescribed NSAIDs, a difference that was not fully attenuated after adjustment for measured confounders. This is because the pre-existing differences in the risk of gastrointestinal complications are very challenging to measure sufficiently, especially in electronic medical records, resulting in residual confounding by indication. Alternatively, consider patients prescribed nicotine replacement therapy for smoking cessation differ from those prescribed pharmaceuticals like varenicline. They tend to be sicker, older and have worse mental health[18]. However, these differences are often not recorded in electronic medical records or other datasets. For example, patients may discuss smoking cessation with their GP when they have pre-clinical symptoms of heart disease; these symptoms may not be perfectly recorded in medical records or even mentioned to the physician.

Instrumental variable analysis offers an approach to addressing these problems. It relies on a distinct set of assumptions from other methods, which do not require measuring or even knowing all the potential confounders of the intervention and outcome.

4. What is an instrumental variable?

Instrumental variables are defined by following three assumptions. First, the instrument associates with the intervention of interest (relevance), second, it shares no common cause with the outcome (independence) and third, it only affects the outcome through the intervention (exclusion restriction). Note that instruments only need to be associated with the likelihood of receiving the intervention; they do not necessarily need to cause it[1]. Instrumental variable analyses exploit naturally occurring variation (the instrument) to estimate the impact of the intervention on an outcome. This variation can be due to clinical or policy decisions which are not related to the key unmeasured confounders of the relationships of interest. **Box 2** defines these assumptions, and **Figure 2** uses a directed acyclic graph to represent these relationships. Assessing the plausibility of these assumptions is critical to determining whether a proposed instrumental variable is valid, and we discuss this in detail in the following section.

These assumptions can be defined unconditionally, or more frequently, as conditional on other important covariates in a dataset; for example, physicians' prescribing preferences are usually conditioned on a patient's age. If these assumptions are violated, for example, by residual confounding of the instrument-outcome association, then the results of an instrumental variable analysis can be more biased than other approaches, such as multivariable adjustment and propensity score. Thus, a key challenge for authors and readers of instrumental variable studies is determining whether the assumptions are plausible for the research question.

5. Types of instruments

Numerous natural experiments have been proposed and assessed as potential instruments. These commonly include physician preference (for example, a clinician's preference to

prescribe one treatment versus another for a given diagnosis), access to intervention (for example, distance to a hospital with specific speciality staff or equipment), or randomisation (for example, in the context of an RCT with non-compliance); examples of these instruments are given below. Other sources of variation, including calendar time, have also been used and are covered elsewhere [19–21].

Physician Preference

Clinicians have preferences for many clinical decisions, such as testing, treatments or diagnoses. These pre-existing preferences may be independent of the subsequent patients they see. For example, a physician may prefer prescribing nicotine replacement therapy over pharmaceutical treatments such as varenicline[18]. Studies generally cannot measure physicians' preferences for one treatment or another, so they measure preferences in other ways. For instance, physicians' prescribing preferences may be captured by looking at previous prescriptions for the treatments under consideration or, more rarely, surveys used to elicit preferences. Physicians' prescriptions to their previous patients are often associated with the prescriptions they issue to their future patients. If this occurs in a way that is unrelated to the patient-level confounders of their current patients, the independence assumption may hold. Physicians' previously demonstrated preferences are consistently associated with their prescriptions to their current patients[7,8]. A potential weakness of physicians' prescribing preferences as an instrument is that the physicians' preferences may not be specific to the treatment of interest and may be associated with broader differences in care.

Access

Examples of access instruments include distance to hospitals [12], travel time to the hospital as a proxy for quicker treatment[22], the raising of the school leaving age as a proxy for education [23], and date of treatment as a proxy for choice of treatment[24]. Here, for the instrumental variable assumptions to hold, access must associate with the likelihood of receiving the intervention but not directly affect the outcome or share any unmeasured

confounders with the outcome. A potential weakness of studies using access-based instruments is that geographic location and distance to healthcare facilities are often highly non-random and related to important unmeasured confounders such as socioeconomic position.

Random assignment in the presence of non-compliance

Treatment assignment in a randomised controlled trial with non-compliance or an encouragement design can be an instrumental variable[25–27]. By design, random assignment should balance confounders between individuals assigned to the intervention and those not. Conventional analyses of randomised trials report the intention to treat (ITT) estimate, which is the difference in outcomes between participants allocated to the intervention and participants assigned to the control. However, if some trial participants do not comply with their treatment allocation, the ITT will underestimate the effects of taking intervention as it will also reflect the effects of compliance. Instrumental variable analysis can be used to estimate the effects of taking the intervention, which can be estimated by assuming that the treatment assignment affects the likelihood of receiving the intervention in the same direction for everyone (i.e. the instrument has a monotonic effect if it increases the likelihood of the exposure for some individuals it does not decrease it for others). Under the monotonicity assumption, the instrumental variable estimate will reflect the *complier* or local average treatment effects (LATE) (see **Box 4** for definitions). This parameter is the effect of the intervention on participants whose treatment status was affected by the instrument. A limitation of random assignment is that assignment may alter behaviour in other ways, leading to violations of the exclusion restriction (e.g., if people assigned to control in an unblinded trial seek treatment via other means). Examples of using allocation to treatment as an instrument include a cluster randomized trial of vitamin A supplementation with non-compliance [25]. Treatment allocation can be used to estimate the effects of an underlying continuous risk factor, for example, the effects of reducing amyloid levels on cognition, rather than the effect of being allocated to amyloid-lowering medication[11]. If the risk factor is continuous, then it is more challenging to interpret under monotonicity, and studies may make other assumptions (e.g. assuming a constant effect of the risk factor).

The instrumental variable assumptions need to be assessed and considered for each application, and just because the assumptions are plausible for one treatment or population does not mean that they will be valid in another.

6. How can the core instrumental variable assumptions be assessed?

Directed acyclic graphs (DAG) provide a convenient and transparent way to depict and explain the assumptions required for an applied instrumental variable analysis[28–30]. Researchers can adapt the structure used in **Figure 2** for specific research questions. Studies can then use empirical data to assess whether the three core instrumental variable assumptions hold.

The first instrumental variable assumption, relevance, states that the instrument must be strongly associated with likelihood of taking the intervention. The strength of the instrument-intervention association is easily testable. For example, in the study of smoking cessation medications, we found that physicians who had previously prescribed varenicline were 24 percentage points (95% CI: 23 to 25) more likely to prescribe varenicline to their subsequent patients than physicians who had previously prescribed nicotine replacement therapy. However, a difference in treatment rates across instrument values is insufficient to measure instrument strength because it does not reflect the sample size. In a small study of a few hundred patients, even a very large difference in treatment rates across the instrument's value will provide very little information about the effects of treatment. This is why many instrumental variable studies report *the partial F-statistic* of the regression of the intervention on the instrument, which reflects both the strength of the association and the total sample size. The partial F-statistic in an instrumental variable analysis is analogous to the sample size in a randomized controlled trial. Most instrumental variable estimation packages in Stata and R, such as ivreg2 or AER respectively[31,32], will report this F-statistic by default. A value above ten is typically considered ‘strong’ and unlikely to lead to weak instrument bias[33]. However, an F-statistic above ten does not guarantee that an

instrumental variable study will have sufficient statistical power to detect an effect size of interest.

The remaining assumptions are untestable, so they cannot be proven to hold, but they are falsifiable[34,35]. An assumption is falsifiable if it is possible to use empirical data to disprove it. The independence assumption can be falsified by testing the instrument-baseline covariates association using covariate balance and bias component plots[19,20], or randomization tests[36]. If the instrumental variable assumptions hold, there should be no detectable associations between the instrument and alternative pathways or other baseline covariates that predict the outcome[37]. The exclusion restriction is falsifiable by demonstrating that other variables are affected by the instrument, which also affect the outcome. For example, consider a study of ACE inhibitors for cardiovascular disease. If physicians who are more likely to prescribe ACE inhibitors are also more likely to prescribe statins, which also affect cardiovascular disease, the exclusion restriction assumption would be violated. An alternative way to falsify the independence and exclusion restriction assumptions is to investigate whether the instrument predicts the outcome in subgroups of the population for which the instrument does not affect the likelihood of receiving the intervention. Suppose there is evidence that the instrument affects the outcome, even in subgroups where the instrument does not affect the likelihood of receiving the intervention. In that case, the instrumental variable assumptions are unlikely to be plausible. For example, if other non-hypertensive patients (e.g. children) who were treated by physicians who preferred ACE inhibitors also had better outcomes. Falsification tests are useful indicators of how plausible the assumptions are likely to be however, failure to falsify an assumption does not mean the assumption can be assumed to be satisfied. For example, if the instrument was associated with an unmeasured confounder of the intervention and the outcome, this would not be evident in a covariate plot that only included measured covariates.

A further way to assess the plausibility of assumptions is to investigate if there are any differences (heterogeneity) in the effect sizes implied by different instruments. This requires more than one instrument (which, when there are more instruments than interventions, is

technically known as being "*over-identified*"). If there is more than one instrument affecting the likelihood of receiving the intervention, e.g. physicians' preferences and distance to the healthcare facility, the heterogeneity in the effects of the intervention implied by each instrument may indicate violations of the instrumental variable assumptions. Bonet's instrumental variable inequality tests can also falsify binary interventions' exclusion restriction and independence assumptions[38].

7. How to generate instrumental variable estimates

Instrumental variables can test whether an intervention affects an outcome and estimate the magnitude of that effect. The simplest estimator is the instrument-outcome association (*reduced-form*, **Box 1**), which can be estimated using regression methods, e.g. linear or logistic regression methods. Importantly, this estimator does not estimate the magnitude of the effect of the *intervention* on the outcome. However, under the instrumental variable assumptions, it is a valid test of the null hypothesis that the intervention does not affect the outcome. An advantage of this test is that it is simple, requires the fewest and weakest assumptions, and can indicate the direction of effect. A disadvantage of this test is that it does not provide a scale for the effect of the intervention on the outcome, limiting the interpretation of the results. Ideally, we want to know the average effect of the intervention (also known as the average treatment effect or ATE), not just the effects of the instrument. For example, researchers and readers may be more interested in the effect of *prescribing varenicline or nicotine replacement therapy* (the intervention) on their current patient than the effect of *physicians' previous prescriptions* for smoking cessation treatment (the instrument) on smoking cessation rates (the outcome).

Several instrumental variable estimators can estimate the ATE parameter; we cover some of the most used below. It should be noted, however, that these methods were largely developed to estimate ATE for normally distributed instruments, exposures, and outcomes with linear relationships. While, in practice, these methods are widely used for binary outcomes and/or non-linear relationships (sometimes with the same name, sometimes under a different name), the interpretation can be difficult and more advanced methods might be required. We note several instances of this below.

If only one instrument is available, then the average effect of the intervention on an outcome can be estimated using instrumental variable estimators, such as the “Wald estimator”. The Wald estimator is the ratio of the instrument-outcome association divided by the instrument-intervention association. This estimator rescales the instrument-outcome association to the intervention scale and indicates the effect of a unit change in the intervention on the outcome. For example, if patients prescribed smoking cessation treatments by physicians who previously prescribed varenicline were 1 percentage point more likely to cease smoking (the instrument-outcome association) and 10 percentage points more likely to be prescribed varenicline (the instrument-intervention association), then the Wald estimate would be -0.01/0.1=-0.1. This would imply that prescribing varenicline increases the absolute probability of stopping smoking by 10 percentage points.

When a study has one or more instruments available, for example, if a study used the physicians' preferences and distance to healthcare facility as instruments, then the effects of the intervention on the outcome can be estimated using a two-stage least squares (2SLS) estimator. This estimator comprises two regressions or “stages”. The “first stage” is a regression of the intervention on the instruments, which can predict the intervention value based on the instrument values. The “second stage” is a regression of the predicted intervention status on the outcome. The estimated coefficient on the predicted value is the instrumental variable estimate of the effect of the intervention on the outcome. The interested reader can work through a simulated example and the formulae in the **Supplementary Materials**. This analysis can be conducted via two separate regressions, as described above. It is usually essential that both stages of instrumental variable analysis contain the same covariates[39]. However, this will not account for the estimation error from the first stage and is likely to give incorrect standard errors and confidence intervals. Typically, most analyses use a package like ivreg2 in Stata or AER package in R[31,32]. These packages compute the instrumental variable estimates in a single step and integrate all the estimation errors from both stages.

Different types of outcomes require different instrumental variable estimators, which rely on logic similar to the 2SLS estimator described above. Commonly used estimators include:

1. *Continuous outcome:* Mean differences, for example, the effects of smoking cessation treatment on body mass index using physicians' prescribing preferences[18], can be estimated using additive structural mean models[40].
2. *Binary outcomes*: Causal risk differences, odds-ratio and risk ratios, for example, estimating the effects of coronary bypass surgery on mortality[12], can be estimated using additive, logistic and multiplicative structural mean models and control function approaches[41–43].
3. *Survival outcomes*: There are methods for using instrumental variables with survival outcomes, which use a similar approach to two-stage-least squares, or the control function approach[44], and have been developed to allow for covariate and outcome-dependent censoring[45]. For example, estimating the effects of screening frequency on colorectal cancer diagnoses using international differences in screening policies[46].
4. *Quantile instrumental variable regression*: Non-linear effects of the intervention can be estimated using instrumental variable quantile regression[47–49]. These can estimate non-linear dose-response relationships. For example, investigating whether the effects of a unit increase in body mass index on healthcare costs differ for those underweight versus those who are overweight[50].

Methods for instrumental variable estimation is an area of active methodological development, spanning statistics, econometrics and computer science. See, for example, estimators combining instrumental variable analysis and matching[51] and estimators using machine learning[52–54].

8. How should instrumental variable estimates be interpreted?

The interpretation of instrumental variable estimates depends on a further fourth *"point-identifying assumption",* which can be used to interpret the instrumental variable estimates as an average treatment effect. Without this assumption, the three core instrumental

variable assumptions are only sufficient to identify the "bounds" of a causal effect[38,55,56]. However, instrumental variable "bounds" are typically very wide, so interpretation of most instrumental variable studies requires a further fourth "point identifying" assumption. The interpretation of instrumental variable estimates, including those produced by all of the above methods, depends on the point-identifying assumption made (see **Box 4: The fourth instrumental variable assumption**).

The interpretation of the estimates from instrumental variable estimation depends on assumptions about the relationships between the instrument, intervention, and outcome. Several assumptions can be made, each leading to a different interpretation. Four assumptions and the resulting interpretation are outlined below, using an exemplar study of the effect of statins on the risk of myocardial infarction (MI) in which statin prescriptions was instrumented using prescriber preference:

1. *Constant treatment effect*: The intervention has the same effect on everyone. This allows the estimate to be interpreted as the average effect of the intervention for the entire population. For example, this could hold if receiving a prescription for statins has the same effect on the risk MI for everyone. However, this assumption is often implausible, especially for binary outcomes, as it could only hold if the treatment entirely cured or caused the outcome.
2. *No effect modification*: The intervention has variable effects on the outcome across individuals, but the instrument does not affect the effect of the intervention on the outcome. This allows the estimate to be interpreted as the average effect of the intervention on individuals who received it. For example, this could hold if physicians' preference for statins or not does not affect the effects of being prescribed a statin on risk of MI. This could be interpreted as the average effect of prescribing statins on the risk of MI for those prescribed statins.
3. *No simultaneous heterogeneity*: The intervention has variable effects on the outcome across individuals, and the instrument has variable effects on the likelihood of receiving the intervention, but this variation in effects is independent. This allows the estimate to be interpreted as the average effects of the intervention across the population. For example, this could hold if physicians who prefer prescribing statins

are more likely to prescribe statins to men than women, but these differences in the effects of physicians' preferences are not related to any differences in the effects of statins. This could be interpreted as the average effect of prescribing statins on the risk of MI in the population.

4. *Monotonicity*: The instrument affects the likelihood of receiving the intervention in the same way for everyone, i.e. increases or has no effect on, and never decreases the likelihood of receiving the intervention, or vice versa. This allows the estimate to be interpreted as the effect of the intervention on the outcome among patients who received the intervention or not due to the instrument. For example, this could hold if patients who were prescribed a statin by physicians who generally preferred not to prescribe statins, would also have been prescribed statin by physicians who preferred to prescribe statins (and vice versa). This could be interpreted as the effect of prescribing statins on the risk of MI among patients who would have received a different treatment (statins or not) if they had attended physicians with different preferences. This is known as a Local Average Treatment Effect (LATE) or Complier Average Causal Effect (CACE). A weakness of this assumption is that it is impossible to know which individuals are in this group.[57,58].

9. Data for instrumental variable studies

Instrumental variable studies typically require measures of the instrument, the intervention, and the outcome for individual-level data analysis using the same sample of people. This straightforward approach allows the most flexibility to test and evaluate the instrumental variable assumptions. However, integrating additional external datasets can improve the power and precision of instrumental variable analyses using an approach known as two-sample instrumental variable analysis[59]. This approach estimates the instrument-intervention association in one sample and the instrument-outcome association in another, from which the Wald estimator can be calculated. For example, a study could estimate the effects of policy reform on educational attainment using census data from the entire population but estimate the effects on health outcomes in a cohort study sub-sampled from the same underlying population[60]. A significant advantage of two-sample instrumental

variable analysis is that it does not need measures of the intervention or the outcome in all samples; this can significantly increase power, particularly when the outcome is rare or difficult to measure.

5. Summary

Instrumental variable analysis can provide reliable evidence about the causal effects of an intervention, even if the intervention-outcome association is affected by unmeasured confounding. Key to conducting and reading instrumental variable studies is assessing the plausibility of the three core instrumental variable assumptions. Does the instrument strongly associate with the intervention? Is there a rationale for why the instrument-outcome association is less likely to suffer from confounding than the intervention-outcome association? Is there evidence that measured covariates are less strongly associated? Are there alternative pathways that could mediate the effects of the instrument?

Instrumental variable analysis can provide a valuable complement to other forms of observational analysis. It depends on distinct assumptions to other approaches, and combined with other sources of evidence can strengthen inferences. The increasing size of data available for clinical research means there is a growing opportunity to use these methods to improve patient care.

Box 1: Glossary of terms used in instrumental variable studies

Concepts

Natural experiment: A source of variation in the likelihood of receiving an intervention in the real world that can be used to investigate the causal impact of an intervention.

Instrumental variable: A specific variable in a dataset that 1) is associated with an intervention, 2) only affects the outcome via its effect on the intervention, and 3) has no common cause with the outcome.

Fourth *point identifying* assumption: The assumption used to estimate the mean effect of the intervention on the outcome, without which it is only possible to estimate bounds for the effect of the intervention on the outcome.

Local average treatment effect/Complier Average Causal Effect: The effect of an intervention on individuals whose intervention status is affected by the instrument.

Counterfactual values: the outcomes that the patients would have had, had they been allocated to intervention or control.

Statistical methods

Reduced form: The instrument-outcome association, which, under the instrumental variable assumptions, is a valid test of the null hypothesis that the intervention affects the outcome.

Wald estimator: The ratio of the instrument-outcome association and the instrument-intervention association[61].

Two-stage least squares: An instrumental variable estimator. The first stage estimates the instrument(s)-intervention association(s) and uses these associations to predict the intervention values[40]. The second stage uses the predicted interventions in a regression to estimate the effect of the intervention on the outcome.

Box 3: A critical appraisal checklist for evaluating instrumental variable studies

Readers of instrumental variable studies could consider the following questions:

Core instrumental variable assumptions

Is there evidence that the instruments are associated with the intervention of interest? Does the study report a partial F-statistic?

Are the instruments associated with measured potential confounders of the intervention-outcome relationship?

Are there likely to be confounders of the instrument-outcome relationship that do not confound the intervention-outcome relationship?

Is the proposed instrument likely to affect the outcome via mechanisms other than the intervention of interest?

Do the authors use negative control outcomes to investigate the plausibility of the instrumental variable assumptions?

Fourth instrumental variable assumption

Do the authors report the fourth instrumental variable assumption?

Do the authors describe their estimand, and how it relates to clinical practice?

Methods reporting

All studies

Does the study clearly state the instrumental variable estimator used in the analysis?

For two-stage least squares, are the same covariates included in both stages of the analysis?

Data presentation

Do the authors present the instrument-outcome association, an instrumental variable estimate, or both?

If they provide an instrumental variable estimate, do they compare it with the multivariable-adjusted estimate?

Was the definition of the instrument pre-specified, or was the definition of the instrument chosen based on the data under analysis?

Do the authors provide the code they used to allow researchers to reproduce their findings?

Interpretation

If the instrumental variable estimate is similar to the multivariable-adjusted estimate and provides evidence consistent with a causal effect, could it be due to weak instrument bias in a single study or confounding of the instrument-outcome association?

If the instrumental variable estimate differs from the observational estimate and provides little evidence of a causal effect, could this be due to weak instrument bias or confounding?

Are the 95% confidence intervals of the estimate sufficiently precise to test for differences with the multivariable-adjusted estimate and detect a clinically meaningful difference?

Clinical implications

Do the results triangulate with other forms of evidence?

If a randomised clinical trial is not feasible or unlikely to be conducted in the short term, and there is existing evidence from multiple instrumental variable studies, and other robust study designs converge on consistent results, this information may help guide patient care; for example, informing clinical guidelines or regulatory decisions.

Box 4: The fourth *"point identifying assumption"* assumption.

The three core instrumental variable assumptions are only sufficient to estimate the "bounds" of a causal effect, which are the largest and smallest values consistent with the observed data. However, instrumental variable "bounds" are typically very wide, so most instrumental variable studies require a further fourth "point identifying" assumption. Options for the fourth assumption include the constant treatment effect [IV4h], no effect modification (IV4n), no simultaneous heterogeneity (NOSH) [IV4nosh], and monotonicity [IV4m] [40,62,63].

1. *The constant treatment effect assumption* requires that the effect of the intervention on the outcome is the same for everyone. For example, suppose the intervention of interest was an anti-hypertensive medication such as ACE inhibitors. In that case, ACE inhibitors should give the same reduction in systolic blood pressure for everyone, regardless of any other characteristics.
2. *The no-effect modification assumption* requires that the intervention has the same effect on the outcome irrespective of the instrument's value. For example, if the effects of ACE inhibitors are the same irrespective of physicians' preference.
3. *The no simultaneous heterogeneity (NOSH) assumption* requires that any heterogeneity in the effects of the instrument on the intervention is independent of heterogeneity in the effects of the intervention on the outcome. This assumption would hold if the variation in the effect of physician preferences on prescribing were not related to the treatment's expected efficacy- i.e., the instrument implicitly samples a representative sample of causal effects from the population.
4. *The monotonicity assumption* requires that the effect of the instrument on the likelihood of receiving the intervention is always in the same direction, e.g. the instrument only increases or decreases the likelihood of receiving the intervention. For example, a patient who attends a physician who prefers ACE inhibitors to other anti-hypertensives will be more likely to receive an ACE inhibitor than a patient who attends a physician who prefers another anti-hypertensive.

Assessing the point-identifying assumptions

These point-identifying assumptions are untestable but falsifiable. The constant treatment effect assumption is falsifiable by checking for differences in instrument strength across covariates. For binary interventions with causal binary proposed instruments and binary outcomes, monotonicity inequalities can falsify the monotonicity assumption[64]. Cumulative distribution graphs for continuous interventions can assess this assumption[40]. If the proposed instrument is a preference, assessing the plausibility of the monotonicity assumption is possible by conducting a preference survey[34]. These surveys suggest that a strict definition of monotonicity is unlikely to be plausible, as there is substantial heterogeneity in clinical treatment decisions. However, Small and colleagues (2017) proposed more plausible assumption: stochastic monotonicity, which requires monotonicity to hold on average[65]. Finally, suppose the monotonicity assumption holds, and the instrument causally affects the interventions. In that case, the counterfactual values (the outcomes that the patients would have had, had they been allocated to intervention or control) among always-takers, compliers, and never-takers can falsify the constant treatment effect assumption.

Interpreting instrumental variable estimates

Instrumental variable estimates can be interpreted as the ATE under any of the constant treatment effects, the no-effect modification or the No Simultaneous Heterogeneity (NOSH) assumptions. The constant treatment effect assumption identifies the ATE by assuming the intervention has the same effect on everyone. This assumption is most commonly used to identify the effects on continuous outcomes. However, this assumption can be implausible. For example, an intervention could only have a constant effect on a binary outcome if it entirely determined the outcome or did not affect it. In the example of statin use, it is implausible to assume that statins have the same effect on every individual in the population. The no-effect modification assumption identifies the intervention's effect on those exposed by assuming that the effect of the intervention is independent of the instrument's value. For example, in a randomised controlled trial (RCT) with an encouragement design where the intervention is an encouragement to take a treatment, allocation to the intervention or control arm does not change the effect of the treatment.

This assumption can identify interventions' effects on binary outcomes regarding causal risk and odds ratios. Finally, the NOSH assumption requires that heterogeneity in the effects of the instrument on the likelihood of receiving the intervention must be independent of heterogeneity in the effect of the intervention on the outcome to be interpreted as the ATE[63].

Instrumental variable estimates can be interpreted as reflecting a LATE using the monotonicity assumption. The monotonicity assumption identifies the effects of the intervention on those whose intervention status was affected by the instrument. This assumption is typically, but not exclusively, applied to binary instruments and interventions[42]. Individuals who are always exposed (e.g. those with very strong indications) or who never receive the intervention (e.g. those with very strong contraindications) will not be affected by the instrument. There are two remaining groups of individuals: those for whom the instrument increases (or decreases) their likelihood of receiving the intervention, known as compliers, and those who do the opposite to their allocation, known as defiers. The monotonicity assumption assumes that there are no defiers in the sample. For example, physicians' prescribing preferences could have a monotonic effect if patients prescribed nicotine replacement therapy who attended a physician who previously prescribed varenicline would also have been prescribed nicotine replacement therapy by a physician who previously prescribed nicotine replacement therapy (and vice versa).

Figure 1: Similarities and differences between instrumental variable analysis and randomized controlled trials.

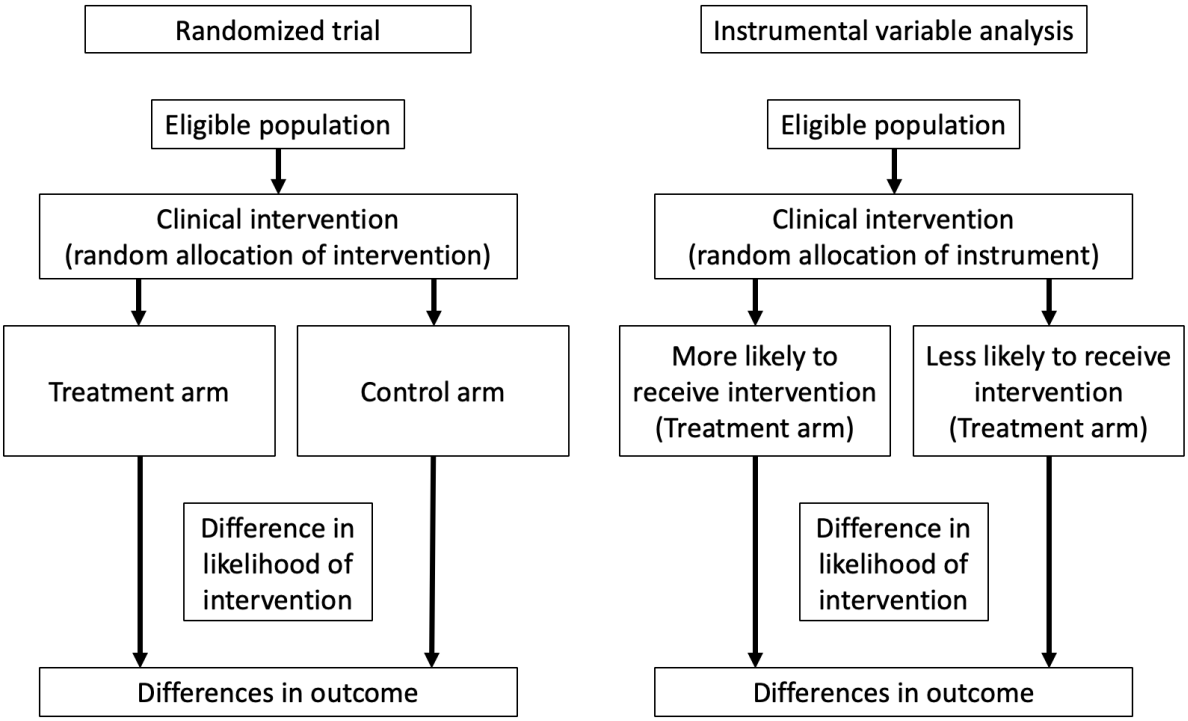
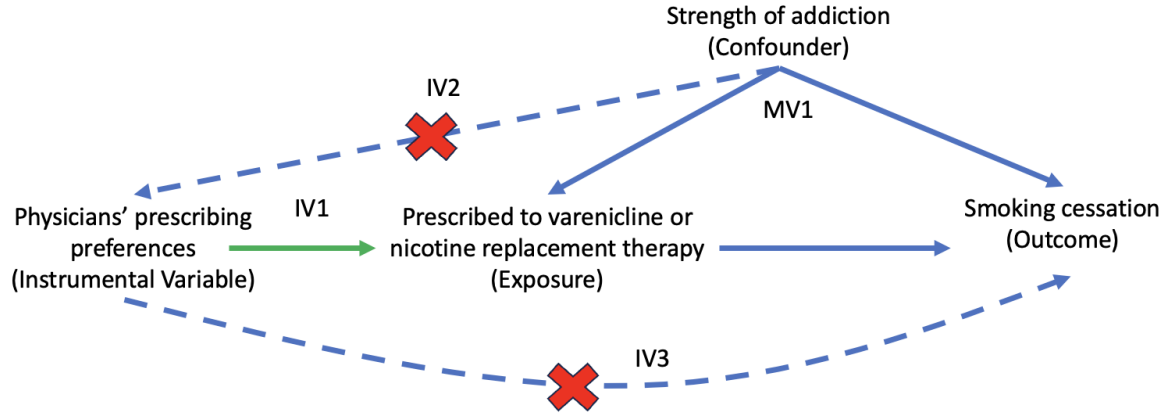


Figure 2: Assumptions of multivariable-adjusted and instrumental variable studies



Notes: Physicians' prescribing preferences are typically unmeasured; thus, typically, studies use prescriptions issued to the physicians' previous patients to proxy for their preferences. Multivariable adjustment assumes that a sufficient set of confounders can be measured to control for all open paths between the intervention and the outcome. In contrast, instrumental variable analysis assumes that there is an instrument that associates with the intervention (relevance, IV1), has no uncontrolled common cause with the outcome (independence, IV2), and only affects the outcome via its association with the intervention (exclusion restriction, IV3).

References

1 Hernán MA, Robins J. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*. 2006;17:360–72.

2 Jena AB, Worsham C. *Random acts of medicine: the hidden forces that sway doctors, impact patients, and shape our health*. New York: Doubleday 2023.

3 Khullar D, Jena AB. "Natural Experiments" in Health Care Research. *JAMA Health Forum*. 2021;2:e210290.

4 Cheng XS, Liu S, Han J, *et al.* Association of Pretransplant Coronary Heart Disease Testing With Early Kidney Transplant Outcomes. *JAMA Intern Med*. 2023;183:134.

5 Kalata S, Thumma JR, Norton EC, *et al.* Comparative Safety of Robotic-Assisted vs Laparoscopic Cholecystectomy. *JAMA Surg*. Published Online First: 20 September 2023. doi: 10.1001/jamasurg.2023.4389

6 Halmin M, Rostgaard K, Lee BK, *et al.* Length of Storage of Red Blood Cells and Patient Survival After Blood Transfusion: A Binational Cohort Study. *Ann Intern Med*. 2017;166:248.

7 Brookhart MA, Wang PS, Solomon DH, *et al.* Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology*. 2006;17:268–75.

8 Davies NM, Davey Smith G, Windmeijer F, *et al.* COX-2 Selective Nonsteroidal Anti-inflammatory Drugs and Risk of Gastrointestinal Tract Complications and Myocardial Infarction: An Instrumental Variable Analysis. *Epidemiology*. 2013;24:352–62.

9 Wang PS, Schneeweiss S, Avorn J, *et al.* Risk of death in elderly users of conventional vs. atypical antipsychotic medications. *N Engl J Med*. 2005;353:2335–41.

10 Bilimoria KY, Chung JW, Hedges LV, *et al.* National Cluster-Randomized Trial of Duty-Hour Flexibility in Surgical Training. *N Engl J Med*. 2016;374:713–27.

11 Ackley SF, Zimmerman SC, Brenowitz WD, *et al.* Effect of reductions in amyloid levels on cognitive change in randomized trials: instrumental variable meta-analysis. *BMJ*. 2021;n156.

12 McClellan M, McNeil B, Newhouse J. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *Journal of American Medical Association*. 1994;272:859–66.

13 Svedahl ER, Pape K, Austad B, *et al.* Impact of altering referral threshold from out-of-hours primary care to hospital on patient safety and further health service use: a cohort study. *BMJ Qual Saf*. 2023;32:330–40.

14 Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ*. 2018;k601.

15 Sanderson E, Glymour MM, Holmes MV, *et al.* Mendelian randomization. *Nat Rev Methods Primers*. 2022;2:6.

16 VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol*. 2019;34:211–9.

17 Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*. 1983;70:41.

18 Thomas KH, Martin RM, Davies NM, *et al.* Smoking cessation treatment and risk of depression, suicide, and self harm in the Clinical Practice Research Datalink: prospective cohort study. *BMJ*. 2013;347:f5704–f5704.

19 Davies NM, Davey Smith G, Windmeijer F, *et al.* Issues in the reporting and conduct of instrumental variable studies: a systematic review. *Epidemiology*. 2013;24:363–9.

20 Garabedian LF, Chu P, Toh S, *et al.* Potential Bias of Instrumental Variable Analyses for Observational Comparative Effectiveness Research. *Ann Intern Med*. 2014;161:131.

21 Widding-Havneraas T, Chaulagain A, Lyhmann I, *et al.* Preference-based instrumental variables in health research rely on important and underreported assumptions: a systematic review. *Journal of Clinical Epidemiology*. 2021;139:269–78.

22 Guo Z, Cheng J, Lorch SA, *et al.* Using an instrumental variable to test for unmeasured confounding. *Statist Med*. 2014;33:3528–46.

23 Davies NM, Dickson M, Davey Smith G, *et al.* The causal effects of education on health outcomes in the UK Biobank. *Nat Hum Behav*. 2018;2:117–25.

24 Gokhale M, Buse JB, DeFilippo Mack C, *et al.* Calendar time as an instrumental variable in assessing the risk of heart failure with antihyperglycemic drugs. *Pharmacoepidemiol Drug Saf*. 2018;27:857–66.

25 Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*. 2000;29:722–9.

26 Hirano K, Imbens GW, Rubin DB, *et al.* Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*. 2000;1:69–88.

27 Li S-M, Ran A-R, Kang M-T, *et al.* Effect of Text Messaging Parents of School-Aged Children on Outdoor Time to Control Myopia: A Randomized Clinical Trial. *JAMA Pediatr*. 2022;176:1077.

28 Pearl J. *Causality: models, reasoning, and inference*. Cambridge, U.K. ; New York: Cambridge University Press 2000.

29 Feeney T, Hartwig FP, Davies N. How to use directed acyclic graphs (DAGs), a guide for clinician researchers.

30 Steiner PM, Kim Y, Hall CE, *et al.* Graphical Models for Quasi-experimental Designs. *Sociological Methods & Research*. 2017;46:155–88.

31 Kleiber C, Zeileis A. *Applied Econometrics with R*. New York: Springer 2008.

32 Baum CF, Schaffer ME, Stillman S. IVREG2: Stata module for extended instrumental variables/2SLS and GMM estimation. 2013. http://EconPapers.repec.org/RePEc:boc:bocode:s425401

33 Stock J, Yogo M. Testing for weak instruments in linear IV regression. *National Bureau of Economic Research Technical Working Paper Series*. 2002;No. 284.

34 Labrecque J, Swanson SA. Understanding the Assumptions Underlying Instrumental Variable Analyses: a Brief Review of Falsification Strategies and Related Tools. *Curr Epidemiol Rep*. 2018;5:214–20.

35 Keele L, Zhao Q, Kelz RR, *et al.* Falsification Tests for Instrumental Variable Designs With an Application to Tendency to Operate. *Med Care*. 2019;57:167–71.

36 Branson Z, Keele L. Evaluating a Key Instrumental Variable Assumption Using Randomization Tests. *American Journal of Epidemiology*. 2020;189:1412–20.

37 Lipsitch M, Tchetgen Tchetgen E, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology*. 2010;21:383–8.

38 Balke A, Pearl J. Bounds on Treatment Effects from Studies with Imperfect Compliance. *Journal of the American Statistical Association*. 1997;92:1171–6.

39 Wooldridge J. *Econometric analysis of cross section and panel data*. Cambridge, Massachusetts: The MIT press 2002.

40 Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *J Am Stat Assoc*. 1995;90:431–42.

41 Clarke PS, Windmeijer F. Identification of causal effects on binary outcomes using structural mean models. *Biostatistics*. 2010;11:756–70.

42 Clarke PS, Windmeijer F. Instrumental Variable Estimators for Binary Outcomes. *Journal of the American Statistical Association*. 2012;107:1638–52.

43 Newey WK. Nonparametric Instrumental Variables Estimation. *American Economic Review*. 2013;103:550–6.

44 Tchetgen Tchetgen EJ, Walter S, Vansteelandt S, *et al.* Instrumental variable estimation in a survival context. *Epidemiology*. 2015;26:402–10.

45 Lee Y, Kennedy EH, Mitra N. Doubly robust nonparametric instrumental variable estimators for survival outcomes. *Biostatistics*. 2023;24:518–37.

46 Engel C, Vasen HF, Seppälä T, *et al.* No Difference in Colorectal Cancer Incidence or Stage at Detection by Colonoscopy Among 3 Countries With Different Lynch Syndrome Surveillance Policies. *Gastroenterology*. 2018;155:1400-1409.e2.

47 Chernozhukov V, Hansen C. An IV Model of Quantile Treatment Effects. *Econometrica*. 2005;73:245–61.

48 Victor Chernozhukov, Ivan Fernandez-Val, Sukjin Han, *et al.* CQIV: Stata module to perform censored quantile instrumental variables regression. 2012. https://ideas.repec.org/c/boc/bocode/s457478.html

49 Chernozhukov V, Fernández-Val I, Han S, *et al.* Censored quantile instrumental-variable estimation with Stata. *The Stata Journal*. 2019;19:768–81.

50 Cawley J, Meyerhoefer C. The medical care costs of obesity: an instrumental variables approach. *J Health Econ*. 2012;31:219–30.

51 Kang H, Kreuels B, May J, *et al.* Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting. *Ann Appl Stat*. 2016;10. doi: 10.1214/15-AOAS894

52 Chen Y, Xu L, Gulcehre C, *et al.* On Instrumental Variable Regression for Deep Offline Policy Evaluation. *J Mach Learn Res*. 2022;23.

53 Kreif N, DiazOrdaz K. Machine learning in policy evaluation: new tools for causal inference. Published Online First: 2019. doi: 10.48550/ARXIV.1903.00402

54 Takatsu K, Levis AW, Kennedy E, *et al.* Doubly robust machine learning for an instrumental variable study of surgical care for cholecystitis. Published Online First: 2023. doi: 10.48550/ARXIV.2307.06269

55 Chesher A. Instrumental variable models for discrete outcomes. *Econometrica*. 2010;78:575–601.

56 Swanson SA, Hernán MA, Miller M, *et al.* Partial Identification of the Average Treatment Effect Using Instrumental Variables: Review of Methods for Binary Instruments, Treatments, and Outcomes. *Journal of the American Statistical Association*. 2018;113:933–47.

57 Lundberg I, Johnson R, Stewart BM. What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory. *Am Sociol Rev*. 2021;86:532–65.

58 Swanson SA, Hernán MA. Commentary: How to Report Instrumental Variable Analyses (Suggestions Welcome). *Epidemiology*. 2013;24:370–4.

59 Inoue A, Solon G. Two-Sample Instrumental Variables Estimators. *Review of Economics and Statistics*. 2010;92:557–61.

60 Zhao Q, Wang J, Spiller W, *et al.* Two-Sample Instrumental Variable Analyses Using Heterogeneous Samples. *Statist Sci*. 2019;34. doi: 10.1214/18-STS692

61 Wald A. Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*. 1949;20:595–601.

62 Hernán MA, Robins JM. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC 2020.

63 Hartwig FP, Wang L, Davey Smith G, *et al.* Average Causal Effect Estimation Via Instrumental Variables: the No Simultaneous Heterogeneity Assumption. *Epidemiology*. 2023;34:325–32.

64 Swanson SA, Miller M, Robins JM, *et al.* Definition and Evaluation of the Monotonicity Condition for Preference-based Instruments. *Epidemiology*. Published Online First: 14 March 2015. doi: 10.1097/EDE.0000000000000279

65 Small DS, Tan Z, Ramsahai RR, *et al.* Instrumental Variable Estimation with a Stochastic Monotonicity Assumption. *Statist Sci*. 2017;32. doi: 10.1214/17-STS623

66 Wright P. Letter from Philip Wright to Sewall Wright, 4 March 1926. 1926. https://ase.tufts.edu/economics/documents/wrightPhilipAndSewall.pdf

67 Angrist JD, Krueger A. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*. 1991;106:979–1014.

Acknowledgements

We are grateful from extremely helpful comments from Brian Lee, Luke Keele, Ting Ye, and Robert Platt, and to our reviewers Christopher Worsham and Tarjei Widding-Havneraas.

Funding and disclosures

The Medical Research Council (MR/V002147/1) and Norwegian Research Council (295989) and the UCL Division of Psychiatry support NMD. M.G.L. received support from the Institute for Translational Medicine and Therapeutics of the Perelman School of Medicine at the University of Pennsylvania, the NIH/NHLBI National Research Service Award postdoctoral fellowship (T32HL007843), the Measey Foundation, and Doris Duke Foundation (Award 2023-0224). The

funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Competing interests

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare support from MRC, Norwegian Research Council, NIH/NHLBI, the Measey Foundation and Doris Duke Foundation for the submitted work") no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work. TF has received funding from Pfizer, Takeda, Acadia, and iHeed for unrelated consulting work, and receives funds from The BMJ for editorial work. NMD receives funds from The BMJ for editorial work. SMD receives research support from RenalytixAI and Novo Nordisk, outside the scope of the current research.

Contributions

VW, ES, and NMD conceived the paper and wrote the first draft. All other authors revised the manuscript and provided critical feedback. All authors act as guarantors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Data and availability

Not applicable.

Supplement

Supplementary Box 1: A Brief History of Instrumental Variables

Instrumental variable (IV) analysis has a rich history in econometrics and has expanded to various fields, including medicine and social sciences. This statistical technique emerged as a powerful tool for addressing endogeneity and establishing causal relationships. Sewall Wright and Philip Wright originally proposed instrumental variables in 1926[66]. Here are some key examples:

Angrist and Krueger (1991): In their influential study, Angrist and Krueger used changes in compulsory schooling laws as an instrumental variable to estimate the causal effect of education on labour market outcomes, such as earnings[67]. This work demonstrated the power of instrumental variables in social sciences and economics.

Newhouse and McClellan (1993): Newhouse and McClellan conducted a seminal study using geographic variation in healthcare spending as an instrumental variable to examine the causal impact of healthcare expenditures on patient outcomes[12]. This study shed light on the effectiveness and efficiency of healthcare spending, contributing to health policy discussions.

Brookhart et al. (2006): A significant contribution to pharmacoepidemiology, Brookhart et al. used prescribing preferences of physicians as an instrumental variable to assess the causal effects of medications on patient outcomes[7]. This study highlighted the applicability of instrumental variable analysis in addressing confounding bias in observational studies.

Hernán and Robins (2006): Hernán and Robins provided an exposition of the assumptions required for instrumental variable analysis[1]. They highlighted the approach's potential value in epidemiological and clinical research, particularly in the absence of randomised controlled trials.

Supplementary Box 2: Instrumental variable estimation

The two most common instrumental variable estimators are the Wald estimator and the 'two-stage least squares' (2SLS) estimator. The Wald estimator can only be used with a single instrument. In contrast, two-stage least squares can use multiple instruments.

With a binary instrument (Z) and binary or continuous intervention (X) and outcome (Y) the Wald estimator is:

$$\widehat{\beta_{IV}} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(X|Z=1) - E(X|Z=0)} \quad \text{(i)}$$

This is the difference between the outcome when the instrument takes the value 1 and when the instrument takes the value 0 divided by the difference in the intervention when the instrument takes the value 1 and the value 0.

If the instrument is continuous, the Wald estimator is the ratio of the instrument-outcome ($\widehat{\beta_{ZY}}$) and instrument-intervention ($\widehat{\beta_{ZX}}$) associations estimated using linear regression.

$$\widehat{\beta_{IV}} = \frac{\widehat{\beta_{ZY}}}{\widehat{\beta_{ZX}}} \quad \text{(ii)}$$

Where β_{ZY} is obtained from the estimation of the regression:

$$Y = \alpha_Y + \beta_{ZY} Z + u_Y \quad \text{(iii)}$$

Where Y is a vector containing the outcome for each individual, Z is a vector containing the instrument for each individual and u_Y is a random error term. And β_{ZX} is obtained from the estimation of the regression

$$X = \alpha_X + \beta_{ZX}Z + u_X \quad \text{(iv)}$$

Where X is a vector containing the outcome for each individual, Z is a vector containing the instrument for each individual and u_X is a random error term. The two-stage least squares estimator can be estimated by predicting the value of the intervention using the observed values of the instrument. This is obtained by estimating equation (iv) and predicting X, indicated X from the estimated values of $\hat{\alpha}_X$and $\hat{\beta}_{ZX}$.

$$\hat{X} = \hat{\alpha}_X + \hat{\beta}_{ZX} \quad Z \quad \text{(v)}$$

The instrumental variable estimate of the effect of receiving the intervention is then obtained by replacing the intervention with the predicted intervention status in the 'second stage' regression of the intervention on the outcome.

$$Y = \alpha_{Y_{iv}} + \beta_{IV}\hat{X} + u_{Y_{iv}} \quad \text{(vi)}$$

This second stage regression can be estimated using linear regression for continuous outcomes or logistic regression for a binary outcome. In each case, the standard errors must account for the uncertainty in the prediction of intervention X in the second stage. This correction is done automatically in instrumental variable regression packages such as ivreg2 in Stata or ivreg in the AER R package.

Supplementary Box 3: Instrumental variable analysis of COX-2s vs traditional NSAIDs using physicians' prescribing preferences

This example uses a simulated random sample from the population. It simulates a study investigating the effects of two types of anti-inflammatory drugs, traditional NSAIDs (e.g. ibuprofen) vs. COX-2 selective inhibitors (COX-2s, e.g. celecoxib). The dataset contains data from 100,000 patients, and it is a patient-level file, i.e. each patient has a single row. In the dataset, the intervention is indicated by the variable 'prescribed_cox_2'. It equals one if the patient had a COX-2 and zero if they had a traditional NSAID. The outcome of interest is whether the patient subsequently had a gastrointestinal complication (variable 'has_gi_event') equal to one or did not have a complication when the outcome is equal to zero. The dataset is called iv_example.csv. There are 100,000 observations with variables on treatment, physician who prescribed the treatment, age, and sex. We will use the information on the physician who prescribed the treatment to create an instrument and estimate the effects of prescribing COX-2s versus traditional NSAIDs.

1. **Create the instrument:** the physician's previous prescription. The variable visit order indicates the order in which the patients visited their GP. Create a variable equal to one if the physician previously prescribed a COX-2 and equal to zero if they previously prescribed a traditional NSAID.
2. **Test the relevance assumption (IV1)**. Are the physicians' previous prescriptions associated with their subsequent prescriptions?

Supplementary Table 1: Association of instrument and likelihood of treatment

		Confidence intervals			
	Risk difference per 100	**Lower**	**Upper**	**R^2**	**F-statistic**
Prior Rx Cox-2	10.7	10.1	11.4	1.1%	1156

R^2 is the proportion of variability explained in the outcome variable of a regression by the covariates. The R^2 value for the prior prescription instrument is 1.1%. Since the R^2 statistic is small, we know that the resulting IV estimates will be imprecise and have wide confidence intervals.

The F statistics for the (first stage) regression of prescribed COX-2 on prior prescription is 1156. Economists often refer to an instrument with a first-stage F statistic less than 10 as a "weak instrument", i.e. an instrument which will give an IV estimate with a relatively large finite sample bias. However, it is essential not to just select instruments, or search through different definitions of the instrument with F statistics greater than 10 in the dataset under analysis as this can lead to bias via winner's curse.

The linear probability model estimates the association between the instrument and the likelihood of receiving the intervention on the absolute probability scale, i.e. risk differences.

3. **Evaluate the independence assumption**: Investigate the plausibility of the third instrumental variable assumption, independence. Do the instruments associate with the measured confounders?

Supplementary Table 2: Association of instrument, prior prescription and measured confounders

Outcome:		**Confidence intervals**		
	Mean/Risk difference per 100	**Lower**	**Upper**	**p-value**
Age	-6.1	-12.3	0.00	0.05
Female	-0.4	-1.0	0.2	0.22

There was little evidence of association between any of the instruments and age or sex. Of course, we cannot check for associations with unmeasured confounders. The simulated dataset includes unmeasured confounders.

4. **Estimate multivariable-adjusted regression**: Are prescriptions of COX-2s associated with a higher or lower risk of gastrointestinal events? What happens when you adjust for the observed covariates?

Supplementary Table 3: Association of prescriptions of COX-2s and gastrointestinal complications, unadjusted and adjusted for age and sex

Outcome:		**Confidence intervals**		
	Risk difference per 100	**Lower**	**Upper**	**p-value**
Unadjusted	5.1	4.9	5.4	<0.001
Adjusted	0.2	0.0	0.5	0.14

The unadjusted linear regression estimates show that patients prescribed COX-2s are 5.1 (95% CI: 4.9, 5.4) percentage points more likely to have a gastrointestinal adverse event compared to those prescribed traditional NSAIDs. This observational estimate attenuates after adjustment for age and sex to 0.2 (95% CI: 0.0, 0.5). The true effect of COX-2s on the gastrointestinal events was simulated in this dataset as -8.3 per 100 patients treated (i.e. fewer events in those prescribed COX-2s), so we can see that this analysis is still biased by an unmeasured confounder.

4. **Estimate the Wald estimator:** The instrumental variable ratio estimator (Wald type) is the instrument outcome association divided by the instrument-intervention association.

The instrument-outcome association is -0.006. The instrument-exposure association is 0.11. Putting these values into the Wald estimator results in the following estimate = -0.006/0.11=-0.06, or 6.0 fewer events per 100 patients treated. Since this dataset was

simulated under the instrumental variable assumptions, this IV estimate is close to the true effect of -8. Of course, it is not possible to know the "true" effect in real datasets.

6. Estimate two-stage least squares estimator: again using the prior prescription as the instrumental variable. Compare this with the ratio estimate above.

Supplementary Table 4: Two-stage least squares estimates of the effect of COX-2s on gastrointestinal complications, unadjusted and adjusted for age and sex

Outcome:		**Confidence intervals**		
	Risk difference per 100	**Lower**	**Upper**	**p-value**
2SLS	-6.0	-8.7	-3.3	<0.001
2SLS 3 prior Rx	-6.6	-8.3	-4.9	<0.001

We can see that with a single instrument, two-stage least squares gives the same estimate as the ratio estimator. We can increase the precision of the instrumental variable estimator by including multiple instruments, in this case, prior prescriptions.

7. Specification tests. Investigate the endogeneity test and overidentification test using -ivregress- postestimation commands.

We can test whether there is any evidence of differences between the multivariable-adjusted and instrumental variable results using an endogeneity test, which gives a $chi^2(1)=198$, p-value<0.001. This test rejects the null hypothesis that there are no differences between the two estimates. One explanation is if the multivariable-adjusted regression suffers from residual confounding.

Because we have multiple instruments, we can use the over-identification tests to investigate if there is any heterogeneity in the effects implied by each instrument. Here, we have three instruments, and the test gives $chi^2(2)=2.2$, p-value=0.34. This test means we cannot reject the null hypothesis that the effects of COX-2s on gastrointestinal complications implied by each of the three instruments are the same.