Published in final edited form as:
Pharmacoepidemiol Drug Saf. 2010 June ; 19(6): 537–554. doi:10.1002/pds.1908.

Instrumental variable methods in comparative safety and effectiveness research†

M. Alan Brookhart, PhD[1,2,*], **Jeremy A. Rassen, ScD**[1], and **Sebastian Schneeweiss, MD, ScD**[1]

[1]Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital & Harvard Medical School, Boston, MA, USA

[2]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Summary

Instrumental variable (IV) methods have been proposed as a potential approach to the common problem of uncontrolled confounding in comparative studies of medical interventions, but IV methods are unfamiliar to many researchers. The goal of this article is to provide a non-technical, practical introduction to IV methods for comparative safety and effectiveness research. We outline the principles and basic assumptions necessary for valid IV estimation, discuss how to interpret the results of an IV study, provide a review of instruments that have been used in comparative effectiveness research, and suggest some minimal reporting standards for an IV analysis. Finally, we offer our perspective of the role of IV estimation vis-à-vis more traditional approaches based on statistical modeling of the exposure or outcome. We anticipate that IV methods will be often underpowered for drug safety studies of very rare outcomes, but may be potentially useful in studies of intended effects where uncontrolled confounding may be substantial.

Keywords

instrumental variables; confounding factors (epidemiology)

Introduction

Non-randomized studies are necessary to assess the safety and effectiveness of medical interventions as they are used in routine practice. One of the principal problems of such studies is confounding—systematic differences between a group of patients exposed to the intervention *versus* the chosen comparator group. To the extent that differences between the groups can be measured, standard statistical approaches, such as multivariable outcome models and propensity score methods, can be used to remove the confounding effects of these variables. These approaches rely on statistical modeling of either the exposure or outcome, require that all confounding factors are accurately measured, and that the statistical models are correctly specified. Unfortunately, the data available for such studies are frequently missing detailed information on the clinical indications and prognostic variables that guide treatment choices; furthermore, there is often insufficient subject-matter knowledge available to guide the

†The authors declare no conflict of interest.

* Correspondence to: Dr M. A. Brookhart, Department of Epidemiology, University of North Carolina at Chapel Hill, McGavran-Greenberg, CB#7435, Chapel Hill, NC, 27599-7435, USA. abrookhart@unc.edu.

specification of the necessary statistical models. Thus, some residual bias due to uncontrolled confounding is likely to be present in most comparative effectiveness research.

Instrumental variable (IV) analysis is one approach to address the problem of uncontrolled confounding. In the past 10 years, there have been several reviews of IV methods in the statistical and biomedical literature that are presented at varying levels of technical detail with different substantive emphases.[1–6] The present article aims to complement this work by presenting a relatively non-technical introduction to IV methods for applied researchers conducting comparative safety and effectiveness studies. We outline the principles and basic assumptions necessary for valid IV estimation, discuss how to empirically explore the validity of IVs, review a range of IVs that have been used in comparative effectiveness research, and suggest some approaches for reporting results from an IV analysis. Finally, we offer our perspective on the role of IV estimation as compared to more traditional approaches. In the appendix, we discuss some additional issues of IV estimation and illustrate the use of Stata software to conduct a simple IV analysis.

Instrumental Variable Assumptions, Informally

IV analysis begins with the identification of an IV, a factor that is assumed to be related to treatment, but neither directly related to the study outcome nor indirectly related *via* pathways through unmeasured variables. As such, an IV can be thought of as an observed variable that generates (or is associated with) variation in the exposure akin to randomized assignment.

Although the exact requirements of an IV depend on the particular analytic framework that one adopts, typically the following three assumptions are sufficient: (1) an IV should affect treatment or be associated with treatment by sharing a common cause; (2) an IV should be a factor that is as good as randomly assigned, so that it is unrelated to patient characteristics; and (3) an IV should be related to the outcome only through its association with treatment. Thus, an instrument should have no direct or indirect effect on the outcome (e.g., through a direct effect or an association with other medical interventions that could influence the outcome).

Example 1: The Placebo-Controlled Randomized Trial with Non-Compliance

IV methods are often found in association with either a real experiment that involves a randomized intervention or a ‘natural experiment’ that creates an allocation of exposure similar to that of a randomized experiment.[7]

Indeed, the most familiar application of IV methods in medicine is in the analysis of a placebo-controlled randomized controlled trial (RCT) with non-compliance. If the causes of non-compliance are independent risk factors for the outcome, then the association between the actual treatment taken by the patient and the outcome will be confounded. Here, IV methods can be helpful to estimate or place bounds on the effect of treatment.

In an RCT with non-compliance, the treatment arm assignment serves as an IV. In this scenario, the IV assumptions are straightforward and non-controversial. First, unless there is massive non-compliance, the treatment arm assignment will be a strong predictor of the actual exposure. Second, by nature of its random assignment, the IV will be theoretically unrelated to patient characteristics. Finally, if the participants and investigators are blind to the assigned treatment, then the assignment will have no independent effect on the outcome.

In many applications, the instrument may not be the result of an investigator-controlled intervention and, therefore, may be associated with the outcome through pathways that do not go through treatment; in those cases, the assumptions must be evaluated more carefully.

Example 2: The Hospital Formulary

Suppose that a new thrombolytic therapy is introduced for the treatment of acute myocardial infarction (AMI). The new medication is thought to be more effective than existing therapies, but it is expensive and evidence suggests that it may have adverse effect in some patients. The new drug is added to the formulary in some hospitals; however, because of cost and concerns about safety, it is not available in others.

If the side effects are uncommon, it may not be feasible to study the safety of this drug using a conventional pharmacoepidemiologic cohort or case-control study. A study could be done using hospital billing records or insurance claims data, but these data would provide little ability to control for confounding by the severity of the underlying coronary artery disease. If the new medication is used preferentially in patients with the poorest prognosis, confounding by unmeasured indication may cause the new medication to be spuriously associated with poor outcomes.

An IV approach based on the hospital formulary presents one possible approach to studying the safety and effectiveness of the new medication using administrative data. A drug's availability on the formulary is clearly related to whether a patient receives the new medication (Assumption 1). Assumption 2 states that the instrument should be effectively randomized to patients. Although patients are not randomly assigned to hospitals, it may be reasonable to assume that formulary status is effectively randomized to patients in that patients go to hospitals without knowledge of the hospitals' formularies. Assumption 3 states that the instrument should affect the outcome only *via* treatment. This assumption would hold if formulary status were not associated with other practices that might affect the outcomes under study, such as a hospital's overall quality of care.

Reduced Form or Intention-to-Treat Estimators

Given a plausible IV, one often reports the association between the instrument and outcomes. In the IV literature, this is termed the *reduced form* estimate. In the RCT example, this is also known as an intention-to-treat (ITT) estimator, as it is a measure of association between the treatment arm assignment (i.e., the treatment *intention*) and the outcome. For example, one might report the difference in mean outcomes between treatment arms (a risk difference for dichotomous outcomes)

$$\beta_{\mathrm{ITT}} = E[\,Y\,|Z=1\,] - E[\,Y|\,Z=0] \tag{1}$$

where Y is the outcome and Z the IV In the RCT example, $Z = 1$ if the patient is assigned to receive active treatment. In the hospital formulary example, if the new medication is available at the admitting hospital, then $Z = 1$. E[Y|Z = z] is the average value of Y for all subjects with $Z = z$.

In a placebo-controlled RCT with non-compliance, the reduced form estimate will be a biased estimate of the effect of treatment received, but if treatment effects are the same among the compliers and non-compliers, the bias will be toward the null and will, therefore, represent a conservative bias in a superiority trial. The degree of bias will depend on the amount of non-compliance, with more non-compliance creating greater bias. An ITT estimate that is non-zero is evidence that treatment is affecting the outcome in some patients.

In Example 2, one could look for evidence of beneficial or harmful drug effects by examining the association between hospital formulary status and various outcomes among patients with

AMI. An association between formulary status and reduced mortality risk would suggest that the new medication is reducing mortality risk among AMI patients whose treatment status depends on the hospital's formulary—i.e., those patients who would receive the new medication if it were available. Similarly, one could assess safety in these same patients by exploring the association between formulary status and the risk of adverse outcomes.

The Wald Estimator

The reduced form or ITT estimator can yield evidence about the presence of a treatment effect, but it does not provide an estimate of the effect of treatment. IV approaches can provide such an estimate under certain assumptions.

In the case of a dichotomous instrument and exposure, the classic IV estimator, also called the Wald estimator, is given by

$$\beta_{\mathrm{IV}} = \frac{E[\,Y\,|Z=1] - E[\,Y|\,Z=0]}{E[\,X\,|Z=1] - E[\,X|\,Z=0]} \tag{2}$$

where Y is the outcome, X is the treatment, Z is the instrument, and β is a measure of the effect of X on Y. The numerator of this estimator is the ITT estimate—i.e., the effect of the instrument on the outcome measured as a risk difference. The denominator is the difference in treatment rates between levels of the instruments (e.g., treatment arms of the RCT), and is a measure of compliance. In the case where the instrument perfectly predicts the treatment (e.g., perfect compliance in the RCT example), then $E[X|Z = 1] - E[X|Z = 0] = 1$ and the IV estimator will be identical to the ITT estimator. As the non-compliance increases, the denominator shrinks and the IV estimator increases relative to the ITT estimator. In the case of non-differential compliance, this removes the bias to the null caused by non-compliance.

This simple estimator can be motivated from a variety of different theoretical frameworks. Here, we focus on the connection to linear structural equations models.

Linear structural equation models/two-stage least-squares

The standard approach to IV estimation in economics is based on linear structural equation modeling where the equations specify causal links between variables rather than associations.[8] In the setting of epidemiology or medicine, one would specify a model for the treatment assignment process that depends on the instrument and potential confounding variables. This is similar to a propensity score except that it includes an IV, a factor that should not be included in a propensity score model.[9] Finally, one specifies a model for the outcome that includes the exposure and the additional covariates that are included in Y The resulting system of equations is given by

$$X = \alpha_0 + \alpha_1 Z + \alpha_2 C + \varepsilon_1 \tag{3}$$

$$Y = \Phi + \beta X + \xi C + \varepsilon_2 \tag{4}$$

where C denotes a vector of covariates, α_2 is a vector of parameters, and ε_1 and ε_2 are errors that represent both random error and the effects of unmeasured variables. An ordinary least-squares (OLS) estimate of the treatment effect β requires that exposure is uncorrelated with

ε_2. The more strongly X is correlated with ε_2, the greater the bias will be in a standard least-squares estimate of β. In the setting of the structural equation models, correlation between X and ε_2 is caused by correlation between ε_1 and ε_2. This would occur if there were an unmeasured variable that was a component of both ε_1 and ε_2—e.g., an unrecorded indication or measure of illness severity.

IV approaches can be used to estimate β consistently even if ε_1 and ε_2 are correlated. In the setting of the structural equation models, an IV estimate of β involves the simultaneous solution of the two equations. For linear structural equation models, this can be accomplished by two-stage least-squares (2SLS) estimation. In the first stage, one estimates the parameters in (2) by least squares. In the second stage, one estimates the parameters in (3) by least squares after replacing the confounded exposure with its predicted value from the first stage. If no covariates are included in the first- and second-stage models, then 2SLS estimation of β yields the Wald estimator (1). A 2SLS estimator will be consistent if Z is uncorrelated with ε_2 given C. The additional covariates C that are optionally included in the model should be potential confounders for the instrument-outcome relation. In other words, one would include a covariate if it were expected that the covariate would be related to the outcome and potentially correlated with the IV. In Example 2, if one is worried about correlation of the instrument and the outcome because of hospital quality, C could include measures of quality.

Weak Instruments

An instrument is weak if it is not a strong predictor of the exposure. Weak instruments present several problems. First, standard IV estimators possess a finite sample bias which is inversely proportional to the F statistic that tests whether the included instruments make a significant contribution to the first-stage model.[10,11] Therefore, adding weak instruments to the first stage can increase bias. Second, weak instruments magnify any residual bias resulting from a confounded instrument. Small violations of IV assumptions can lead to tremendous inconsistency in the IV estimator when the IV is weak.[12] Finally, weak instruments also yield highly variable estimates of exposure effects and so may result in studies underpowered to detect small effects, even in very large studies.

Heterogeneous Treatment Effects and the Interpretation of the Estimator

Most medical interventions do not affect all patients in a population the same way. In the structural models above, the effect of treatment is assumed to be constant. When treatment effects are heterogeneous—e.g., sicker patients may benefit more from treatment than healthier patients—the linear structural equations may no longer be a good model for the data. In such settings, various statistical approaches are available that can be used to place bounds on the average effect of treatment in the population.[13–18] However, unless the instrument is very strong, these bounds do not provide much information about treatment effects.

Imbens and Angrist and Angrist *et al.* showed that under a 'monotonicity'* assumption the Wald estimator (1) yields the average effect of treatment among the 'compliers' even if treatment effects are heterogeneous. The compliers (often termed 'marginal patients') are those whose treatment status can be affected by the IV.[19,20] In the RCT example, the compliers would be the patients who would always take their assigned treatment—they would take placebo if assigned placebo or would take the active therapy if assigned it.

*The monotonicity assumption necessary for this interpretation requires that the instrument affects treatment deterministically in one direction, i.e., if the value of the instrument increases for an individual patient, the patient may be induced to start treatment, but never to opt out of treatment (or *vice versa*). In the RCT example, this requires that there are no 'defiers'—patients who would always do the opposite of what they are assigned. The monotonicity assumption is reasonable in most RCT settings, but it may be questionable in other examples in medicine and epidemiology.

In the hospital formulary example, the marginal patients would be those treated with the thrombolytic therapy if admitted to a hospital where the new drug was available, but who would be untreated at other hospitals. If the treatment tends to be given to patients with a poor prognosis, then standard IV methods would estimate the effect of treatment in high-risk patients.

In some cases, the concept of a 'marginal patient' may be problematic as individual patients may be marginal to varying degrees. For example, there may be patients who would be treated with the new medication at some but not all hospitals where it is available. For such cases, the standard IV estimator can also be interpreted as a weighted average of subgroup-specific treatment effects where the weights are related to the strength of the instrument within each sub-group.[21] If the instrument is strong within a particular subgroup that subgroup effect is weighted up. If an instrument is not predictive of treatment within a subgroup, the treatment effect in that subgroup has a zero weight and is not reflected in the IV estimator. Finally, if the effect of the IV on treatment is reversed (e.g., the IV predicts not receiving treatment in a subgroup), then the subgroup effect is negatively weighted. This could have the peculiar effect of making a medication appear to prevent a side effect that it causes. Brookhart and Schneeweiss discuss how this could arise in the case of contraindications and misuse of medical procedures.[21]

For detailed theoretical discussions of treatment effect heterogeneity, see Imbens and Angrist,[19] Angrist *et al.*,[20] Wooldridge,[22] Heckman *et al.*,[23] Hernan and Robins,[5] and Basu *et al.*[24]

Examples of Instrumental Variables in Healthcare Research

The use of IV methods in comparative effectiveness research is limited by the availability of valid and strong instruments. To help researchers identify potential IVs, we review various instruments that have been applied in comparative effectiveness research. For each example, we discuss potential threats to the validity and consider issues related to interpretation of the estimator.

Distance to specialty care provider

In a classic paper, McClellan *et al.* proposed an IV that was an indicator of whether the nearest hospital managed AMI admissions with cardiac catheterizations.[25] Distance has been used as an IV in a variety of other applications, including a study of the effect of dialysis center profit status on survival[26] and two studies of the effect of treatment in specialized trauma centers.[27,28]

Using distance as an IV depends on an assumption that distance is associated with receiving care. This is certainly reasonable in McClellan *et al.*'s example as patients who are experiencing an AMI are likely to be taken to the nearest hospital because of the urgency of the condition. Distance must also not be related to patient characteristics. An IV estimator based on distance would tend to reflect the effect of treatment in patients whose treatment status depends more on distance. In McClellan *et al.*'s example, this would be patients who would be catheterized if taken to hospitals that performed catheterizations.

Preference-based IVs

A variety of papers have applied IVs, defined at the level of the geographic region,[29–32] hospital,[33–35] dialysis center,[36] or individual physician.[37–40] We have termed such IVs 'preference-based instruments' since they are derived from the assumption that different providers or groups of providers have different preferences or treatment algorithms dictating how medications or medical procedures are used.[21] These approaches exploit naturally occurring variation in medical practice patterns to estimate treatment effects. Variation in

formularies, hospital capacity, and prescription drug benefit plan structure can also lead to region-, hospital-, or physician-level differences in medical practice.[41]

In order for preference to be a valid instrument, it must predict treatment—so there must be variations in treatment preference across the providers or aggregations of providers being studied. It also must be independent of the characteristics of the patients, so there cannot be differences in case-mix across levels of the instrument. Preference must only affect the outcome through its influence on the treatment under study. Thus, providers who preferentially use one treatment must not also preferentially use other treatments that may affect the outcome.

A previous paper considered how treatment effect heterogeneity could bias a preference-based IV estimator relative to the average effect of treatment in the population.[21] It discussed how commonly available subject matter knowledge, such as whether medications or medical procedures tend to be overused or underused, could be used to help anticipate the direction of the bias relative to the average effect of treatment in a population. For example, in effectiveness studies, underuse of a therapy results in the IV up-weighting the treatment effects in high-risk patients and, thus, could cause the IV estimator to be biased high for the beneficial effects at the population level (to the extent that treatment is more beneficial in high-risk patients). Conversely, overuse up-weights treatment effects low-risk patients and, thus, may cause an IV estimator to understate the beneficial effects at the population level (to the extent that treatment is less beneficial in low-risk patients). Hennessey *et al.* and Rassen *et al.* explore the practical issue of how to estimate preference given a time series of treatment decisions.[42,43]

Day of the week of hospital admission as an instrument for waiting time for surgery

Ho *et al.* sought to determine the effect of waiting time to surgery on length of stay and inpatient mortality among patients admitted to hospital with a hip fracture.[44] They used the day of the week of the index hospital admission as an instrument for wait time for surgery under the assumption that many surgeons operate only on weekdays and, therefore, patients admitted on the weekend may have to wait longer for surgical treatment. The IV could be confounded if patients admitted on the weekend were different from those admitted on the weekday. The instrument could be independently related to the outcome if other aspects of hospital care that could affect the outcome were different over the weekend.

Treatment effect heterogeneity would alter the interpretation of the IV estimator if surgeons were more willing to come in on the weekend to operate on severely injured patients. In this case, the IV estimator would tend to yield the effect of treatment in patients who are in better health.

Randomized encouragement studies

IVs can also be created when patients or providers are randomized to receive programs encouraging the use of a particular treatment. These studies are often called ‘randomized encouragement designs’.[45] The randomization inherent in such studies guarantees that there is no systematic association between patient and provider characteristics and the IV. Treatment effect heterogeneity may be an issue, as there is evidence that there may exist ‘physician defiers’ who would always do the opposite of what is encouraged.[46]

Drug co-payment amount

Studying the effect of adherence to a preventive therapy is challenging because it is likely to be confounded by other health-related behaviors.[47,48] Cole *et al.* used drug co-payment amount as an instrument to study the effect of β-blocker adherence on clinical outcomes and health care expenditures after a hospitalization for heart failure.[41] They assumed that co-payment would affect β-blocker adherence, but would be otherwise unrelated to the outcomes being

studied. The results from this study would generalize to those patients whose adherence is likely to be affected by co-payment change.

Calendar time

IV can also arise from secular trends in medication use. Variation in medication use across time could result from changes in guidelines, changes in formularies or reimbursement policies, changes in physician preference (e.g., due to marketing activities by drug makers), release of new effectiveness or safety information (e.g., resulting from new studies, 'Dear Doctor' letters, FDA 'black box' advisories), or the arrival of new agents to the market.

Many studies have explored associations between secular trends in medication use and outcomes, although these are often not done in a formal IV framework.[49,50] Calendar time has been used in a formal IV analysis in a study of β-blocker therapy on outcomes among patients hospitalized with heart failure,[51] in a study of antivirals in HIV patients,[52] and in a study of hormone replacement therapy on myocardial infarction risk.[53]

An IV based on calendar time can be confounded by the things that change in time such as the characteristics of patients who enter the cohort, other medical practices (such as the changing use of medications), or medical coding systems. Since these are likely in many examples, IVs based on calendar time are most reasonable in situations where a dramatic change in practice occurs over a relatively short period of time.

Reporting Results

An IV analysis depends on many assumptions and can result in biased and imprecise estimates if these assumptions do not hold. Here, we suggest a few reporting standards that may help readers properly interpret and evaluate the validity of an IV analysis.

Justify need for and role of IV in the study

IV methods are inefficient and should not be used as a primary analysis unless unmeasured confounding is thought to be strong. Researchers should discuss why substantial unmeasured confounding is expected. In many cases, this will be an appeal to confounding by unmeasured indications or disease severity. If unmeasured confounding is possible, but not expected to be severe, IV may be more appropriately used as a secondary analysis.

Describe theoretical basis for the choice of IV

A good IV should have a theoretical motivation, i.e., why it is expected to influence treatment, but be otherwise unrelated to patient characteristics and outcome. This could be an appeal to a real or natural experiment.[7] For example, in the hospital formulary example, one could assert that patients chose hospitals without knowledge of their formulary and, therefore, formulary status may be effectively randomly assigned.

Report strength of instrument and results from first-stage model

An IV should be strongly related to treatment. The authors should report the Wald statistic or, if using a multivariable-adjusted IV estimator, the entire first-stage model. This permits the reader to assess IV strength and simultaneously to understand the treatment assignment mechanism and the confounding potential of various observed covariates. The first-stage F statistic and the partial r^2 attributable to the inclusion of the IVs can be reported as a means of assessing whether the instruments are sufficiently strong.[10] As discussed in the appendix, the finite sample bias in 2SLS estimator is proportional to the inverse of the F statistic. As a rule of thumb, F statistics less than 10 are thought to be potentially problematic when using 2SLS

estimator and multiple instruments.[11] The partial r^2 can be interpreted as the proportion of the variance explained by the addition of the IV to the model.[54]

Report distribution of patient risk factors across levels of the IV and exposure

Ideally, an IV should be unrelated to the characteristics of the patient as it would be if it were randomly assigned. To evaluate this assumption, one should report the means and frequencies of the observed variables across levels of the instrument. It may also be helpful to report standardized differences, i.e., the difference in means divided by the standard error of the difference. Reporting *p*-values is probably not helpful as they are more a reflection of sample size than covariate balance. It can also be helpful to report the means and frequencies of patient variables across levels of the treatment. This allows the reader to assess the potential for confounding in the IV relative to the confounding in the exposure. In the appendix, we provide an example of such a table (Table A1).

Explore concomitant treatments

The instrument should only influence the outcome through its influence on the treatment under study. For many clinical problems, there may be a variety of other interventions that could be used alongside the intervention under study. For example, statins (cholesterol lowering medications) may be prescribed to patients after AMI along with ACE inhibitors, β-blockers, and aspirin. If an instrument for statin prescribing also affects the prescribing of the other medications, the IV approach may be biased for the effect of statin exposure. Exploring the association between the instrument and concomitant treatments can help determine whether the resulting effect is attributable solely to the intervention being studied or is likely to be affected by co-interventions.

Evaluate sensitivity of IV estimator to modeling assumptions

If using a multivariable IV approach, we suggest reporting an unadjusted IV estimate and exploring the sensitivity of the results to the inclusion/exclusion of covariates, particularly if there is not a strong theoretical reason to believe that they confound the instrument-outcome association. Results that are highly sensitive to the included covariates may reveal a validity problem with the IV—e.g., a possible association between the instrument and the unmeasured covariates.

Discuss issues related to interpretation of the estimator

Because IV approaches do not always yield the average effect of treatment in a population, researchers should consider the patients to whom the treatment effect generalizes. Typically, these would be patients whose treatment status depends strongly on the instrument. As described earlier, the Wald estimator can be interpreted as a weighted average of subgroup-specific treatment effects in which the weights are related to the strength of the IV within the subgroups. When clinically important subgroups can be created using observed data, we suggest reporting the strength of the IV across these different subgroups. If the IV is considerably stronger or weaker in a particular subgroup, then the treatment effect in that subgroup will be weighted up or down, respectively. Such an analysis can help to identify the 'marginal patients'. In the appendix, we provide an example of such a table (Table A2).

Conclusions

IVs methods represent a potential approach to control confounding in studies of the safety and effectiveness of medical interventions. In this paper, we have provided an overview of the assumptions and methods necessary to conduct a simple IV analysis. Also, we have suggested some approaches for reporting the results of such an analysis.

It is important to realize that even if a valid IV is available, IV methods will not always be helpful. If the instrument is weak, an IV study will be underpowered to detect anything less than a very strong effect, even with large samples. Furthermore, if only a small amount of unmeasured confounding is expected, there may be no reason to use an IV—the larger variance of the IV estimator may result in a much larger mean-squared error relative to the only slightly biased conventional estimator. Here, one might consider using the IV in a secondary or sensitivity analysis. Because of their inefficiency, IV methods may be poorly suited for studies of very rare safety outcomes, as they are likely to be underpowered even with a reasonably strong instrument and a large study.

KEY POINTS

- Instrumental variable methods can reduce confounding bias in comparative effectiveness research.
- These methods are closely connected to natural experiments and depend on identifying an occurrence that leads to a random or pseudorandom assignment of exposure to some patients.
- We review the assumptions necessary for valid IV estimation and provide guidance on reporting an IV analysis.

In our view, IV methods have the most potential for studies of intended effects. Here, substantial uncontrolled confounding is likely due to confounding by indication or confounding by disease severity.[55] For such problems, conventional approaches may often be substantially biased and, thus, IV methods may deserve status as a primary analysis—provided a valid IV is available.

The primary barrier to the use of IV methods is the need to have a plausible IV. Unfortunately, such variables have been difficult to find in epidemiology and medicine. Further work in this area may help to identify new instruments or reveal ways to improve existing instruments, thereby expanding the potential applications of these methods.

Acknowledgments

This project was funded under contract no. 290-2005-0016-I-TO3-WA1 from the Agency for Healthcare Research and Quality (AHRQ), US Department of Health and Human Services (DHHS) as part of the Developing Evidence to Inform Decisions about Effectiveness (DECIDE) program. The authors of this report are responsible for its content. Statements in the report should not be construed as endorsement by AHRQ or DHHS. Dr Brookhart was additionally supported by a career development award from the National Institute on Aging (K25AG027400). The content of this paper does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health. Dr Rassen is supported by a career development award from AHRQ (1 K01 HS018088).

Appendix

Additional issues

Finite sample bias in small samples

IV estimators based on 2SLS possess substantial finite sample bias. The bias results from some over-fitting in the first-stage regression that leads to a correlation between the true error term (that includes the unmeasured confounders) and the predicted value of treatment.[10] The bias is toward the OLS estimate and decreases with sample size but increases with the number of instruments included in the first-stage model. To see this, imagine including a very large number of random numbers as instruments in the first stage (an experiment that was done in Bound *et al.*10). Although the random numbers are theoretically unrelated to the error term,

they will have some chance associations with the unmeasured variables, thus, some confounding will get transferred to the predicted value of treatment. Various approaches have been proposed to deal with this bias, such as limited information maximum likelihood (LIML) as well as approaches based on jackknife and split-sample methods.[56,57] Finite sample bias is likely to be more of a problem in the presence of many weak instruments and can persist even in very large samples.

Dichotomous outcomes and relative measures of effect

The simple Wald estimator and the linear structural equation models can be used with dichotomous outcomes. The linear structural models require the use of appropriate software to conduct inference,[58] correctly specified models, and the predicted values of exposure in the 0–1 range.[5]

However, in medicine and epidemiology interest often focuses on ratio measures such as relative risks or rates. IV approaches based on the Wald estimator or linear structural equation models yield estimates of an absolute measure of effect (e.g., a risk difference). A variety of IV approaches can be used to estimate relative measures of effect, and each imposes somewhat different assumptions. Sommer and Zeger propose an IV estimator of a risk ratio appropriate for an RCTwith a specific kind of non-compliance.[59] Cuziak *et al.* propose a more general IV estimator of the risk ratio.[60] Greenland offers some discussion on these approaches.[2]

In the general framework of the structural nested mean model of Robins,[61] Hernan and Robins provide an estimator of the causal relative risk.

IV probit models are often used in economics for dichotomous outcomes and have been used in health research (e.g., Pracht and Tepas [28] and Bhattacharya *et al.* 62). These models use probit link functions to constrain probabilities of exposure and treatment in the range of 0–1 and can be fit using the *ivprobit* function in Stata (Stata Corp., College Station, Texas), although this procedure may not be appropriate for dichotomous exposures. Although parameters in probit models are not readily interpretable to epidemiologists, they can be converted to approximate odds ratios by multiplying by 1.6 and marginalized to estimate various parameters of interest. Parameters in generalized linear models can also be estimated using IVs by constructing moment-based estimators that are based on the assumption that the instrument should be orthogonal to the regression error.[51,58,63]

Grouped-treatment approaches

Several recent papers have implemented ‘grouped treatment’ approaches that implicitly use a two-stage estimation approach, but are not based on linear models. For example, Johnston used average treatment rate within a hospital as a predictor in a logistic regression model of individual outcomes that also included patient-level covariates.[33] Schmoor *et al.* used the average treatment rate within a clinic in a Cox proportional hazards model of patient-level mortality.[35] To our knowledge, these approaches are not motivated by a theoretical model and, thus, may not yield parameters that are causally interpretable. However, such approaches, like reduced form IV estimates, can be used to assess the presence and direction of an average treatment effect among those whose treatment is influenced by the instrument. In a simulation study, two-stage logit models have been found to yield parameters that, in the presence of strong residual confounding, are closer to the true parameter than logistic regression.[64] Also, in several practical data analysis examples, two-stage logit approaches have been found to yield estimates that are substantively indistinguishable from estimates yielded from theoretically-motivated IV methods.[65]

Continuously valued treatments

When the treatment or exposure under study takes on continuous values, for example, if one is studying the effect of the dose of a medication, then one needs to adopt appropriate statistical approaches for these data. For example, one could assume a structural equation model that assumes a linear effect of dose. However, this is clearly a strong modeling assumption. Angrist and Imbens discuss a more general approach to IV estimation of exposures with variable intensities (such as doses).[66]

It is natural to consider dichotomizing a continuous exposure in order to take advantage of IV methods for dichotomous exposures. However, this can lead to bias if the IV is associated with the actual dose within a dose category.[66] To see how this could be, suppose that one is using a clinic as the basis of an IV to study the dosing of a medication that is dichotomized into a 'high' and 'low' dose. It could be that certain clinics that aggressively use a medication tend to have higher 'high' doses than clinics that use a medication more conservatively. Thus, an IV estimate of the effect of a high dose may be exaggerated. This could be empirically explored by examining whether the IV predicts dose level within a dose category.

Multiple instruments

Many applications of IV methods in economics involve settings in which many instruments are available. As mentioned earlier, a large number of weak instruments can increase finite sample bias in an IV estimator. Therefore, when using many instruments, the researcher should pay close attention to the *F* statistic to ensure that the instruments are making an important contribution to the first-stage model. Also, in theory, each instrument can identify a slightly different treatment effect by weighting each stratum-specific estimate differently. If different IV approaches yield substantively different results, the researcher should consider whether this is due to sampling variability, invalidity of one of the instruments, or differences in treatment effects identified by each instrument.

Testing for the need of IV methods

There is a large literature in economics on testing for whether an IV is necessary. These tests assume that one has a valid IV and the test then attempts to determine whether there is a sufficiently large difference between the conventional estimate and the IV estimate to conclude that the conventional estimate is biased. If the conventional appears to be unbiased, then it would be preferred given its smaller variance. Tests of this sort have been proposed by Durbin,[67] Hausman,[68] and Wu.[69] One limitation of these tests is that they assume homogeneous treatment effects. Therefore, if such a test rejects the hypothesis that the IV is unnecessary, one cannot be sure whether it is because of treatment effect heterogeneity or confounding. Despite this ambiguity, these tests may be useful in certain situations and can be implemented in *Stata* add-on modules (see Appendix).

Longitudinal studies

Our discussion so far has assumed that exposure is determined at baseline and is invariant during the follow-up period. However, many exposures, particularly those involved with medications, are time varying. Conventional IV methods do not have a facility for handling time-varying exposures.[5] Robins has proposed general IV approaches based on nested structural models that can use both time-varying confounders and instruments to estimate the effects of time-varying exposures.[61] These methods have been relatively understudied and represent an important area for future research and application.

IV as a secondary analysis

Because IV methods rely on assumptions that are entirely different from conventional methods, IV analysis may be useful as a confirmatory or secondary analysis. The use of IV methods as a sensitivity analysis has been suggested by Greenland[2] and has been applied in various substantive projects.[30,39,40,70,71] To the extent that such approaches disagree, one must carefully evaluate the assumptions underlying each approach and consider the possibility that treatment effect heterogeneity may be leading to divergent results.[72]

Suggested tables for IV analysis

Table A1

Balance of patient characteristics across treatment groups and levels of the instrument

Patient characteristics	Treatment status			IV status		
	Drug A	Drug B	Std. Dif.	Predicted A	Predicted B	Std. Dif.
	Mean (SD) or Freq (%)			Mean (SD) or Freq (%)		
C_1						
C_2						
C_3						
…						
C_k						

Table A2

Instrument strength overall and within subgroups

Instrument definition and subgroups	RD (Wald denominator)	95%CI	*F* statistic	Partial r^2
Subgroup 1				
Subgroup 1				
…				
Subgroup k				
Overall				

IV estimation using Stata

The following section illustrates a 2SLS IVanalysis using Stata (StataCorp. 2009. Stata Statistical Software: Release 9. College Station, TX: StataCorp LP). We use the built-in ivreg command and the extension ivreg2. ivreg2 is available for download from the Statistical Software Components archive *via* the Stata command ssc install ivreg2.

ivreg and ivreg2 will yield equal point estimates and standard errors, but ivreg2 offers much more in terms of diagnostic output and analysis options.

Below, let outcome be the outcome, exp be the exposure, iv be the instrument, age be a continuous age variable, sex be an indicator for male sex (1 = male, 0 = female), and c1, c2, and c3 be three dichotomous confounders.

We begin with simple crude and adjusted models using OLS estimation

```
. reg outcome exp
      Source |       SS       df       MS              Number of obs = 7 8731
-------------+------------------------------           F(1, 78729) = 165.66
       Model |  3.17645234     1  3.17645234            Prob >F = 0.0000
    Residual |  1509.58325 78729  .019174424            R-squared = 0.0021
-------------+------------------------------           Adj R-squared = 0.0021
       Total |  1512.7597  78730  .019214527            Root MSE = .13847

------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |  -.0158182   .001229   -12.87   0.000    -.018227   -.0134094
       _cons |   .0227949   .0005525    41.26   0.000    .0217121    .0238778
------------------------------------------------------------------------------

. reg outcome exp age sex c1 c2 c3
      Source |       SS       df       MS              Number of obs = 78730
-------------+------------------------------           F(6, 78723) = 218.16
       Model |  24.7418045     6  4.12363408            Prob >F = 0.0000
    Residual |  1488.01751 78723  .018901941            R-squared = 0.0164
-------------+------------------------------           Adj R-squared = 0.0163
       Total |  1512.75932 78729  .019214766            Root MSE = .13748

------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |  -.0206018   .0012311   -16.73   0.000    -.0230147   -.0181889
         age |   .0012435   .0000408    30.49   0.000    .0011635    .0013234
         sex |  -.0047138   .0010838    -4.35   0.000    -.0068381   -.0025896
          c1 |  -.0126105   .0012686    -9.94   0.000    -.015097    -.010124
          c2 |    .003701   .0022383     1.65   0.098    -.000686     .008088
          c3 |  -.0030277   .0011144    -2.72   0.007    -.0052119   -.0008434
       _cons |  -.0455348   .0029075   -15.66   0.000    -.0512335   -.0398362
------------------------------------------------------------------------------
```

The bold lines show a crude risk difference of −1.58 per 100 and an adjusted risk difference of −2.06 per 100. Turning to IVs, we run a simple ivreg of exposure, instrument, and outcome

```
. ivreg outcome (exp = iv)
Instrumental variables (2SLS) regression
      Source |       SS       df       MS              Number of obs = 63053
-------------+------------------------------           F(1, 63051) = 12.32
       Model |  1.9715631     1  1.9715631             Prob > F = 0.0004
    Residual |  1241.48864 63051  .019690229            R-squared = 0.0016
-------------+------------------------------           Adj R-squared = 0.0016
       Total |  1243.4602  63052  .019721186            Root MSE = .14032
```

outcome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exp	**-.0122035**	**.0034773**	**-3.51**	**0.000**	**-.019019**	**-.005388**
_cons	.0220232	.0007775	28.32	0.000	.0204993	.0235472

Instrumented: exp

Instruments: iv

The bold line shows the desired point estimates: an absolute risk difference of −1.2 per 100 people, with a 95% confidence interval of −0.5 to −1.9 per 100. Adding the IV has moved the point estimate toward the null, but increased the standard error by a factor of three.

Since this is an IV analysis, it may not be necessary to include covariates, but we run another simple model with age, sex, and three major covariates adjusted for

. ivreg outcome (exp = iv) age sex c1 c2 c3

Instrumental variables (2SLS) regression

Source	SS	df	MS		
				Number of obs =	63052
				F(6, 63045) =	154.85
Model	20.9205856	6	3.48676426	Prob >F =	0.0000
Residual	1222.53921	63045	.019391533	R-squared =	0.0168
				Adj R-squared =	0.0167
Total	1243.4598	63051	.019721492	Root MSE =	.13925

outcome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
exp	**-.0154683**	**.0034772**	**-4.45**	**0.000**	**-.0222835**	**-.0086531**
age	.0012849	.0000469	27.42	0.000	.0011931	.0013768
sex	-.0053853	.0012283	-4.38	0.000	-.0077928	-.0029779
c1	-.0123998	.0014531	-8.53	0.000	-.0152479	-.0095518
c2	.002082	.0025857	0.81	0.421	-.002986	.00715
c3	-.0041958	.0012639	-3.32	0.001	-.006673	-.0017187
_cons	-.0478727	.0032737	-14.62	0.000	-.0542892	-.0414561

Instrumented: exp

Instruments: age sex c1 c2 c3 iv

In this case, adjusting for covariates made little difference: the point estimate changed from −1.2 per 100 to −1.5 per 100.

However, this simple output offers little in the way of diagnostic information. In particular, we are interested in the first-stage regression fit statistics, in order to know how well the instrument predicted treatment. The ivreg2 command offers more output, and adding the first and ffirst options yields extensive information about the first-stage.

Again, the simple regression with no covariates

```
. ivreg2 outcome (exp = iv), first ffirst
First-stage regressions
-----------------------
First-stage regression of exp:
OLS estimation
--------------
Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only
                                                 Number of obs = 63053
                                                 F(1, 63051) = 15438.69
                                                 Prob > F = 0.0000
Total (centered) SS = 8278.904255                Centered R2 = 0.1967
Total (uncentered) SS = 9803                     Uncentered R2 = 0.3216
Residual SS = 6650.467981                        Root MSE = .3248
------------------------------------------------------------------------------
         exp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iv |   .5487822   .0044167   124.25   0.000     .5401255    .5574389
       _cons |   .1034863   .0013594    76.13   0.000      .100822    .1061507
------------------------------------------------------------------------------
Included instruments: iv
-----------------------
Partial R-squared of excluded instruments: 0.1967
Test of excluded instruments:
  F(1, 63051) = 15438.69
  Prob > F = 0.0000
Summary results for first-stage regressions
-------------------------------------------
Variable /     Shea Partial R2 /    Partial R2 /    F(1, 63051)    P-value
exp /          R2 0.1967 /          0.1967 /        15438.69       0.0000
Underidentification tests
Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)
Ha: matrix has rank=K1 (identified)
Anderson canon. corr. N*CCEV LM statistic  Chi-sq(1)=12402.34  P-val=0.0000
Cragg-Donald N*CDEV Wald statistic         Chi-sq(1)=15439.18  P-val=0.0000
Weak identification test
Ho: equation is weakly identified
Cragg-Donald Wald F-statistic                                  15438.69
See main output for Cragg-Donald weak id test critical values
Weak-instrument-robust inference
Tests of joint significance of endogenous regressors B1 in main equation
Ho: B1=0 and overidentifying restrictions are valid
Anderson-Rubin Wald test                   F(1, 63051)=12.30   P-val=0.0005
Anderson-Rubin Wald test                   Chi-sq(1)=12.30     P-val=0.0005
Stock-Wright LM S statistic                Chi-sq(1)=12.30     P-val=0.0005
Number of observations                                 N = 63053
```

Number of regressors K = 2
Number of instruments L = 2
Number of excluded instruments L1 = 1
IV (2SLS) estimation

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

Number of obs = 63053
F(1, 63051) = 12.32
Prob > F = 0.0004
Total (centered) SS = 1243.4602 Centered R2 = 0.0016
Total (uncentered) SS = 1269 Uncentered R2 = 0.0217
Residual SS = 1241.488637 Root MSE = .1403

outcome	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
exp	**-.0122035**	**.0034772**	**-3.51**	**0.000**	**-.0190187**	**-.0053882**
_cons	.0220232	.0007775	28.33	0.000	.0204993	.0235471

Underidentification test (Anderson canon. corr. LM statistic): 1.2e+04
Chi-sq(1) P-val = 0.0000

Weak identification test (Cragg-Donald Wald F statistic): 1.5e+04
Stock-Yogo weak ID test critical values: 10% maximal IV size 16.38
15% maximal IV size 8.96
20% maximal IV size 6.66
25% maximal IV size 5.53
Source: Stock-Yogo (2005). Reproduced by permission.

Sargan statistic (overidentification test of all instruments): 0.000
(equation exactly identified)

Instrumented: exp
Excluded instruments: iv

In the bold section, one can see that the estimates and confidence intervals are equal in both ivreg and ivreg2. However, looking at the italicized section, one sees additional information about the first-stage regression, in particular the partial r^2 value and the first-stage F statistic, both indicators of instrument strength. (They are labeled ‘partial’ as they examine the instrument independent of other specified covariates, which are nil in this case.) In this case, the partial r^2 value is 0.1967, indicating that the instrument adds significantly to the prediction of the exposure. The first-stage partial F statistic is 15 348 with a negligible p-value, both indicating a very strong instrument.

Note that the output is customized for a common situation in econometrics, where multiple IVs are used simultaneously. As such, many of the reported values are the same, where in the case of multiple IVs, they would not be. The partial r^2 and Shea partial r^2 are examples.

Adding covariates to the model

. ivreg2 outcome (exp = iv) age sex c1 c2 c3, first ffirst

First-stage regressions

First-stage regression of exp:

OLS estimation

Estimates efficient for homoskedasticity only

Statistics consistent for homoskedasticity only

Number of obs = 63052

F(6, 63045) = 2732.49

Prob > F = 0.0000

Total (centered) SS = 8278.880083 Centered R2 = 0.2064

Total (uncentered) SS = 9803 Uncentered R2 = 0.3298

Residual SS = 6570.270997 Root MSE = .3228

exp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0024803	.0001063	23.34	0.000	.002272	.0026885
sex	-.017055	.0028439	-6.00	0.000	-.0226292	-.0114809
c1	-.0025468	.0033684	-0.76	0.450	-.0091488	.0040552
c2	.0089066	.0059953	1.49	0.137	-.0028442	.0206573
c3	.0235374	.0029224	8.05	0.000	.0178095	.0292654
iv	.5450276	.0043934	124.06	0.000	.5364165	.5536387
_cons	-.0547891	.0075906	-7.22	0.000	-.0696666	-.0399116

Included instruments: age sex c1 c2 c3 iv

Partial R-squared of excluded instruments: 0.1962

Test of excluded instruments:

F(1, 63045) = 15389.75

Prob > F = 0.0000

Summary results for first-stage regressions

Variable /	*Shea Partial R2 /*	*Partial R2 /*	*F(1,63045)*	*P-value*
exp /	*0.1962 /*	*0.1962 /*	*15389.75*	*0.0000*

Underidentification tests

Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)

Ha: matrix has rank=K1 (identified)

*Anderson canon. corr. N*CCEV LM statistic* *Chi-sq(1)=12371.49* *P-val=0.0000*

*Cragg-Donald N*CDEV Wald statistic* *Chi-sq(1)=15391.46* *P-val=0.0000*

Weak identification test

Ho: equation is weakly identified

Cragg-Donald Wald F-statistic *15389.75*

See main output for Cragg-Donald weak id test critical values

Weak-instrument-robust inference

Tests of joint significance of endogenous regressors B1 in main equation

Ho: B1=0 and overidentifying restrictions are valid

Anderson-Rubin Wald test	*F(1,63045)=19.74*	*P-val=0.0000*
Anderson-Rubin Wald test	*Chi-sq(1)=19.74*	*P-val=0.0000*
Stock-Wright LM S statistic	*Chi-sq(1)=19.74*	*P-val=0.0000*
Number of observations	*N = 63052*	
Number of regressors	*K = 7*	
Number of instruments	*L = 7*	
Number of excluded instruments	*L1 = 1*	

IV (2SLS) estimation

Estimates efficient for homoskedasticity only

Statistics consistent for homoskedasticity only

Number of obs = 63052

F(6,63045) = 154.85

Prob >F = 0.0000

Total (centered) SS = 1243.459795 Centered R2 = 0.0168

Total (uncentered) SS = 1269 Uncentered R2 = 0.0366

Residual SS = 1222.53921 Root MSE = .1392

outcome	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
exp	**-.0154683**	**.003477**	**-4.45**	**0.000**	**-.022283**	**-.0086536**
age	.0012849	.0000469	27.42	0.000	.0011931	.0013768
sex	-.0053853	.0012282	-4.38	0.000	-.0077927	-.002978
c1	-.0123998	.001453	-8.53	0.000	-.0152477	-.009552
c2	.002082	.0025856	0.81	0.421	-.0029856	.0071497
c3	-.0041958	.0012638	-3.32	0.001	-.0066728	-.0017188
_cons	-.0478727	.0032736	-14.62	0.000	-.0542887	-.0414566

Underidentification test (Anderson canon. corr. LM statistic): 1.2e+04

Chi-sq(1) P-val = 0.0000

Weak identification test (Cragg-Donald Wald F statistic): 1.5e+04

Stock-Yogo weak ID test critical values: 10% maximal IV size 16.38

15% maximal IV size 8.96

20% maximal IV size 6.66

25% maximal IV size 5.53

Source: Stock-Yogo (2005). Reproduced by permission.

Sargan statistic (overidentification test of all instruments): 0.000

(equation exactly identified)

Instrumented: exp

Included instruments: age sex c1 c2 c3

Excluded instruments: iv

Again, the final results are equivalent to that of ivreg. With the addition of covariates, the partial *F* statistic and partial r^2 values remain the same, but the F statistic for the first-stage regression (reported in the first block of output) decreases. Nonetheless, it is the *partial* values that are of interest.

ivreg2 offers many other analytic and diagnostic options, all of which are described in the procedure's documentation. Notably, it offers generalized method of moments (GMM) and LIML estimation, as well as various tests of endogeneity.

References

1. Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. Annu Rev Public Health 1998;19:17–34. [PubMed: 9611610]
2. Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol 2000;29:722–729. [PubMed: 10922351]
3. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. Epidemiology 2006;17:260–267. [PubMed: 16617274]
4. Glymour, MM. Natural experiments and instrumental variable analyses in social epidemiology. In: Oakes, JM.; Kaufman, JS., editors. Methods in Social Epidemiology. John Wiley and Sons; San Francisco, CA: 2006. p. 429-468.
5. Hernan MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology 2006;17:360–372. [PubMed: 16755261]
6. Grootendorst P. A review of instrumental variables estimation of treatment effects in the applied health sciences. Health Serv Outcomes Res Methodol 2007;10:159–179.
7. Angrist JD, Krueger AB. Instrumental variables and the search for identification: from supply and demand to natural experiments. J Econ Perspect 2001;15:69–85.
8. Goldberger AS. Structural equation methods in the social sciences. Econometrica 1972;40:979–1001.
9. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149–1156. [PubMed: 16624967]
10. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variables is weak. J Am Stat Assoc 1995;90:443–450.
11. Staiger D, Stock JH. Instrumental variables regression with weak instruments. Econometrica 1997;65:557–586.
12. Small DS, Rosenbaum P. War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. J Am Stat Assoc 2008;103:924–933.
13. Robins, JM. The Analysis of randomized and nonrandomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest, L.; Freeman, H.; Mulley, A., editors. Health Services Research Methodology: A focus on AIDS. U.S. Public Health Service; Washington, D.C.: 1989. p. 113-159.
14. Manski CF. Nonparametric bounds on causal effects. Am Econ Rev 1990;80:319–323.
15. Balke A, Pearl J. Bounds on treatment effects from studies with imperfect compliance. J Am Stat Assoc 1997;92:1171–1176.
16. MacLehose RF, Kaufman S, Kaufman JS, Poole C. Bounding causal effects under uncontrolled confounding using counterfactuals. Epidemiology 2005;16:548–555. [PubMed: 15951674]
17. Cheng J, Small DS. Bounds on causal effects in three-arm trials with noncompliance. J R Stat Soc B 2006;68:815–836.
18. Kaufman S, Kaufman JS, MacLehose RF. Analytic bounds on causal risk differences in directed acyclic graphs involving three binary variables. J Stat Plan Inference 2009;139:3473–3487. [PubMed: 20161106]
19. Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. Econometrica 1994;62:467–475.

20. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc 1996;81:444–455.
21. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. Int J Biostat 2007;3
22. Wooldridge J. On two-stage least-squares estimation of the average treatment effect in a random coefficient model. Econ Lett 1997;56:129–133.
23. Heckman JJ, Urzua S, Vytlacil EJ. Understanding instrumental variable models with essential heterogeneity. NBER Working Pap 2006;12:574.
24. Basu A, Heckman JJ, Navarro-Lozano S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Econ 2007;16:1133–1157. [PubMed: 17910109]
25. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. JAMA 1994;272:859–866. [PubMed: 8078163]
26. Brooks JM, Irwin CP, Hunsicker LG, Flanigan MJ, Chrischilles EA, Pendergast JF. Effect of dialysis center profit-status on patient survival: a comparison of risk-adjustment and instrumental variable approaches. Health Serv Res 2006;41:2267–2289. [PubMed: 17116120]
27. McConnell KJ, Newgard CD, Mullins RJ, Arthur M, Hedges JR. Mortality benefit of transfer to level I versus level II trauma centers for head-injured patients. Health Serv Res 2005;40:435–457. [PubMed: 15762901]
28. Pracht EE, Tepas JJ 3rd, Langland-Orban B, Simpson L, Pieper P, Flint LM. Do pediatric patients with trauma in Florida have reduced mortality rates when treated in designated trauma centers? J Pediatr Surg 2008;43:212–221. [PubMed: 18206485]
29. Wen SW, Kramer MS. Uses of ecologic studies in the assessment of intended treatment effects. J Clin Epidemiol 1999;52:7–12. [PubMed: 9973068]
30. Earle CC, Tsai JS, Gelber RD, Weinstein MC, Neumann PJ, Weeks JC. Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis. J Clin Oncol 2001;19:1064–1070. [PubMed: 11181670]
31. Brooks JM, Chrischilles EA, Scott SD, Chen-Hardee SS. Was breast conserving surgery underutilized for early stage breast cancer? Instrumental variables evidence for stage II patients from Iowa (erratum appears in Health Serv Res. 2004 Jun;39(3):693). Health Serv Res 2003;38:1385–1402. [PubMed: 14727779]
32. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007;297:278–285. [PubMed: 17227979]
33. Johnston SC. Combining ecological and individual variables to reduce confounding by indication: case study—subarachnoid hemorrhage treatment. J Clin Epidemiol 2000;53:1236–1241. [PubMed: 11146270]
34. Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. N Engl J Med 2008;358:771–783. [PubMed: 18287600]
35. Schmoor C, Caputo A, Schumacher M. Evidence from nonrandomized studies: a case study on the estimation of causal effects. Am J Epidemiol 2008;167(9):1120–1129. [PubMed: 18334500]
36. Bradbury BD, Do TP, Winkelmayer WC, Critchlow CW, Brookhart MA. Greater Epoetin alfa (EPO) doses and short-term mortality risk among hemodialysis patients with hemoglobin levels less than 11g/dL. Pharmacoepidemiol Drug Saf 2009;18:932–940. [PubMed: 19572312]
37. Korn EL, Baumrind S. Clinician preference and the estimation of causal treatment effects. Stat Sci 1998;13:209–235.
38. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology 2006;17:268–275. [PubMed: 16617275]
39. 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;353:2335–2341. [PubMed: 16319382]

40. Schneeweiss S, Solomon DH, Wang PS, Rassen J, Brookhart MA. Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: an instrumental variable analysis. Arthritis Rheum 2006;54:3390–3398. [PubMed: 17075817]
41. Cole JA, Norman H, Weatherby LB, Walker AM. Drug copayment and adherence in chronic heart failure: effect on cost and outcomes. Pharmacotherapy 2006;26:1157–1164. [PubMed: 16863491]
42. Hennessy S, Leonard CE, Palumbo CM, Shi X, Ten Have TR. Instantaneous preference was a stronger instrumental variable than 3- and 6-month prescribing preference for NSAIDs. J Clin Epidemiol 2008;61(12):1285–1288. [PubMed: 18495427]
43. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. J Clin Epidemiol 2009;62(12):1233–1241. [PubMed: 19345561]
44. Ho V, Hamilton BH, Roos LL. Multiple approaches to assessing the effects of delays for hip fracture patients in the United States and Canada. Health Serv Res 2000;34:1499–1518. [PubMed: 10737450]
45. Ten Have, tr; Elliot, MR.; Joffe, MM.; Zanutto, E.; Datto, C. Causal models for randomized physician encouragement trials in treating primary care depression. J Am Stat Assoc 2004;99:16–25.
46. Rosenstein AH, Shulkin D. Changing physician behavior is tool to reduce health care costs. Health Care Strateg Manage 1991;9:14–16. [PubMed: 10113903]
47. White HD. Adherence and outcomes: it's more than taking the pills. Lancet 2005;366:1989–1991. [PubMed: 16338439]
48. Brookhart MA, Patrick AR, Dormuth C, et al. Adherence to lipid-lowering therapy and the use of preventive health services: an investigation of the healthy user effect. Am J Epidemiol 2007;166:348–354. [PubMed: 17504779]
49. Mamdani M, Juurlink DN, Kopp A, Naglie G, Austin PC, Laupacis A. Gastrointestinal bleeding after the introduction of COX 2 inhibitors: ecological study. BMJ (Clin Res) 2004;328:1415–1416.
50. Juurlink DN, Mamdani MM, Lee DS, et al. Rates of hyperkalemia after publication of the randomized aldactone evaluation study. N Engl J Med 2004;351:543–551. [PubMed: 15295047]
51. Johnston KM, Gustafson P, Levy AR, Grootendorst P. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. Stat Med 2008;27:1539–1556. [PubMed: 17847052]
52. Cain LE, Cole SR, Greenland S, et al. Effect of highly active anti-retroviral therapy on incident AIDS using calendar period as an instrumental variable. Am J Epidemiol 2009;169:1124–1132. [PubMed: 19318615]
53. Shetty KD, Vogt WB, Bhattacharya J. Hormone replacement therapy and cardiovascular health in the United States. Med Care 2009;47:600–606. [PubMed: 19319001]
54. Shea J. Instrument relevance in multivariate linear models: a simple measure. Rev Econ Statist 1997;79:348–352.
55. Walker AM. Confounding by indication. Epidemiology 1996;7:335–336. [PubMed: 8793355]
56. Angrist JD, Krueger AB. Split-sample instrumental variables estimates of the return to schooling. J Bus Econ Stat 1995;13
57. Angrist JD, Imbens GW, Krueger AB. Jacknife instrumental variables estimation. J Appl Econometrics 1999;14:5767.
58. Baum CF, Schaffer ME, Stillman S. Instrumental variables and GMM: estimation and testing. Stata J 2003;3:1–31.
59. Sommer A, Zeger SL. On estimating efficacy from clinical trials. Stat Med 1991;10:45–52. [PubMed: 2006355]
60. Cuzick J, Edwards R, Segnan N. Adjusting for non-compliance and contamination in randomized clinical trials. Stat Med 1997;16:1017–1029. [PubMed: 9160496]
61. Robins JM. Correcting for non-compliance in randomized trials using structural nested mean models. Commun Stat Theory Methods 1994;23:2379–2412.
62. Bhattacharya J, Goldman D, McCaffrey D. Estimating probit models with self-selected treatments. Stat Med 2006;25:389–413. [PubMed: 16382420]

63. Foster EM. Instrumental variables for logistic regression: an illustration. Soc Sci Res 1997;26:487–504.
64. Johnston SC, Henneman T, McCulloch CE, van der Laan M. Modeling treatment effects on binary outcomes with grouped-treatment variables and individual covariates. Am J Epidemiol 2002;156:753–760. [PubMed: 12370164]
65. Rassen J, Schneeweiss S, Glynn RJ, Mittleman M, Brookhart MA. Instrumental variables methods for dichotomous outcomes. Am J Epidemiol 2009;169(3):273–284. [PubMed: 19033525]
66. Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. J Am Stat Assoc 1995;90:431–442.
67. Durbin J. Errors in variables. Rev Int Stat Inst 1954;22:23–32.
68. Hausman JA. Specification tests in econometrics. Econometrica 1978;46:1251–1271.
69. Wu DM. Alternative tests of the independance between stochastic regessors and disturbances. Econometrica 1973;41:529–546.
70. Schneeweiss S, Setoguchi S, Brookhart A, Dormuth C, Wang PS. Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. CMAJ 2007;176:627–632. [PubMed: 17325327]
71. Wang PS, Schneeweiss S, Setoguchi S, et al. Ventricular arrhythmias and cerebrovascular events in the elderly using conventional and atypical antipsychotic medications. J Clin Psychopharmacol 2007;27:707–710. [PubMed: 18004143]
72. Brooks JM, Chrischilles EA. Heterogeneity and the interpretation of treatment effect estimates from risk adjustment and instrumental variable methods. Med Care 2007;45:S123–S130. [PubMed: 17909370]