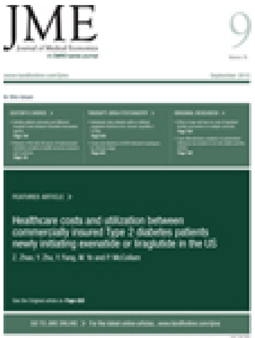
JME

9

Healthcare costs and utilization between commercially insured Type 2 diabetes patients newly initiating exenatide or liraglutide in the US

Taylor & Francis
Taylor & Francis Group

Journal of Medical Economics

ISSN: 1369-6998 (Print) 1941-837X (Online) Journal homepage: http://www.tandfonline.com/loi/ijme20

A selective review of the first 20 years of instrumental variables models in health-services research and medicine

John Cawley

To cite this article: John Cawley (2015) A selective review of the first 20 years of instrumental variables models in health-services research and medicine, Journal of Medical Economics, 18:9, 721-734, DOI: 10.3111/13696998.2015.1043917

To link to this article: http://dx.doi.org/10.3111/13696998.2015.1043917

Accepted author version posted online: 22 Apr 2015.
Published online: 20 May 2015.

Submit your article to this journal

Article views: 149

View related articles

View Crossmark data

1369-6998
doi:10.3111/13696998.2015.1043917

Article 0006.R2/1043917

Original article

A selective review of the first 20 years of instrumental variables models in health-services research and medicine

John Cawley
Department of Policy Analysis and Management, Cornell University, Ithaca, NY, USA, and School of Economics, University of Sydney, Sydney, Australia

Address for correspondence:
John Cawley, Department of Policy Analysis and Management, Cornell University, 2312 MVR Hall,
Ithaca NY 14853, USA.
Tel: 607-255-0952; Fax: 607-255-4071; E-mail: JHC38@cornell.edu

Keywords:
Instrumental variables – Econometrics – Methods – Health

Accepted: 20 April 2015; published online: 20 May 2015
Citation: J Med Econ 2015; 18:721–734

Abstract

Background:
The method of instrumental variables (IV) is useful for estimating causal effects. Intuitively, it exploits exogenous variation in the treatment, sometimes called natural experiments or instruments. This study reviews the literature in health-services research and medical research that applies the method of instrumental variables, documents trends in its use, and offers examples of various types of instruments.

Methods:
A literature search of the PubMed and EconLit research databases for English-language journal articles published after 1990 yielded a total of 522 original research articles. Citations counts for each article were derived from the Web of Science. A selective review was conducted, with articles prioritized based on number of citations, validity and power of the instrument, and type of instrument.

Results:
The average annual number of papers in health services research and medical research that apply the method of instrumental variables rose from 1.2 in 1991–1995 to 41.8 in 2006–2010. Commonly-used instruments (natural experiments) in health and medicine are relative distance to a medical care provider offering the treatment and the medical care provider's historic tendency to administer the treatment. Less common but still noteworthy instruments include randomization of treatment for reasons other than research, randomized encouragement to undertake the treatment, day of week of admission as an instrument for waiting time for surgery, and genes as an instrument for whether the respondent has a heritable condition.

Conclusion:
The use of the method of IV has increased dramatically in the past 20 years, and a wide range of instruments have been used. Applications of the method of IV have in several cases upended conventional wisdom that was based on correlations and led to important insights about health and healthcare. Future research should pursue new applications of existing instruments and search for new instruments that are powerful and valid.

Background

The philosophy of evidence-based medicine is to apply the scientific method to decision-making at every level of the healthcare system, such as a regulator's decision to approve a drug for sale, the physician's decision of whether to prescribe a drug and, if so, which one, and the health insurer's decision of what drugs to include on the formulary. This generally requires knowledge of the causal effects of specific treatments on patient outcomes. It is generally agreed that the most convincing evidence on causal effects comes from well-conducted, properly powered randomized controlled trials (RCT)[1]. However, such evidence is often not available. In some cases, conducting an RCT would be unethical or infeasible. For example, it would be unethical to cause human subjects to become obese in order to estimate the causal effect of obesity on cardiovascular disease, medical care costs, or mortality. Alternatively, there could be high confidence that the treatment is effective, and, thus, it would be unethical to withhold the treatment from a control group[2].

An alternative to RCTs is to estimate models of instrumental variables[3,4]*. The intuition of the approach is to find a natural experiment: something that alters the probability of treatment, but is not otherwise correlated with the outcome. In such cases it may be possible to accurately estimate causal effects using observational data[4–8].

One of the most important early applications of the method of IV to medical research was published by McClellan *et al.*[9] in 1994. This paper recognizes the 20th anniversary of that article with a selective review of the published literature in health-services research and medical research that uses the method of IV.

The purpose of this paper is to provide multiple products that collectively offer an accessible introduction to the method of instrumental variables, its usefulness, and the impact it has had in health-services research and medical research. Specifically, the contribution of this paper is to: (1) provide a brief overview of the method of instrumental variables; (2) measure the increase in its use in health-services research and medicine over the past 20 years; (3) categorize the types of instruments that are commonly used in health-services research and medicine; (4) provide up-to-date examples of high-quality papers that use each category of instrument; and (5) explain how the use of IV changed the way that we think about topics in health-services research and medicine.

Intuition of the method of instrumental variables (IV)

This section provides the basic intuition to the model of IV; readers seeking additional information are referred to detailed guides such as Wooldridge[6], Angrist and Pischke[4], and Martens *et al.*[7].

In many cases researchers want to estimate a causal effect of a treatment X on an outcome Y. A challenge is that in most cases the treatment was not randomly assigned, and some of the factors that affect the choice of treatment (such as education or income or health) may also have a direct effect on the outcome. In such a case, a simple regression of Y on X will yield a biased estimate of the causal effect; specifically, it will suffer bias due to omitted variables.

A randomized controlled trial may not be possible, but the causal effect can still be estimated if one can find a natural experiment Z that affects the probability of assignment to the treatment. (This paper uses the terms natural experiment and instrument interchangeably.)

The simplest form of the model of IV is two-stage least squares. This consists of two sequentially-estimated ordinary least squares regressions.

(1) In the first stage, the endogenous treatment (X) is regressed on the instrument or natural experiment (Z). (For the sake of convenience we assume the following: the treatment and the outcome are continuous variables, there are no other control variables in the model, and the errors are normally distributed.)

$$X = \alpha_1 + \delta Z + \varepsilon_1$$

(2) In the second stage, the outcome (Y) is regressed on the predicted value of the treatment ($\hat{X}$) from the first stage.

$$Y = \alpha_2 + \beta\hat{X} + \varepsilon_2$$

The IV estimator of the effect of X on Y is equal to:

$$\hat{\beta}_{IV} = \frac{\hat{\sigma}_{Z,Y}}{\hat{\sigma}_{Z,X}} \tag{1}$$

The numerator, $\hat{\sigma}_{Z,Y}$, is the sample covariance of Z and Y and equals the effect of the instrument on the outcome, and the numerator $\hat{\sigma}_{Z,X}$ is the sample covariance of Z and X and equals the effect of the instrument on the treatment.

There are three important criteria for instruments (Z).

(1) *Power*: The instrument Z should explain a large proportion of the variance in X. At the extreme, if the instrument Z is uncorrelated with the treatment X, then the denominator in equation (1) equals zero and the IV estimator is undefined.

(2) *Validity*: The instrument (Z) should be correlated with the outcome (Y) only through the treatment (X). Put another way, the instrument (Z) should be uncorrelated with the error term in the second stage (ε_2). Put yet another way, all of $\hat{\sigma}_{Z,Y}$ should be due to the effect of Z on the treatment (X), not to a direct effect of Z on Y or a correlation of Z with omitted variables associated with Y. This is known as the identifying assumption. Equation (1) indicates that the bias to the IV estimator $\hat{\beta}_{IV}$ that results from a violation of assumption #2 is greater the weaker the instrument. In other words, the weaker the instrument then the smaller is the denominator $\hat{\sigma}_{Z,X}$ and, thus, any bias in the numerator has a larger effect on $\hat{\beta}_{IV}$[10].

*This paper focuses on the model of instrumental variables, but there are other methods for estimating causal effects, for example: difference-in-differences models (see, e.g., Angrist and Pischke[4], section 5.2), propensity score matching (see, e.g., Morgan and Winship[8], chapter 4[8]), and regression discontinuity designs (see, e.g., Angrist and Pischke[4], chapter 6).

(3) *Monotonicity*: The effect of the instrument (Z) on the probability of treatment (X) must be either weakly positive (i.e., zero or positive) for all individuals in the sample or weakly negative (i.e., zero or negative) for all individuals in the sample. In other words, the instrument should not make some subjects more likely, but other subjects less likely, to get the treatment.

It is relatively easy to test the first criterion for an instrument (power). One can conduct an F-test for the statistical significance of Z in the first-stage regression. The standard benchmark is that the instrument(s) alone should have an F statistic of at least 10 in the first stage regression[11].

It is much more difficult to test the second criterion of validity. If a researcher has more instruments than endogenous regressors (treatments), then the model is said to be over-identified, and one can conduct what is called an over-identification test. In a model with a single endogenous regressor (treatment), this could take the form of testing whether the instrumental variables coefficients are equal when estimated using each instrument individually; the assumption is that if each instrument is valid then the resulting IV coefficients should not be significantly different[4]. However, there are several limitations to over-identification tests[4]: (1) they assume that at least one instrument is valid; (2) IV estimates are often imprecise, so failure to reject their equality may not be informative; and (3) if the different instruments cause variation in the treatment for different sub-groups, and those sub-groups experience heterogeneous treatment effects, then an over-identification test may indicate that the IV estimates are significantly different, even if all of the instruments are valid.

Although it is far from definitive, one simple approach is to test whether the subjects' observable characteristics are correlated with the instrument[9]. Of course, the concern remains that the subjects' *unobservable* characteristics may be correlated with the instrument, but that is untestable, so the intuition is that if observed characteristics are uncorrelated with the instrument then perhaps unobserved characteristics are also uncorrelated with it. Another best practice is that authors should provide a theory of treatment choice and outcomes that explains why the instrument affects treatment choice and yet has no direct effect on outcomes and is uncorrelated with confounders.

In recent years, authors have also investigated the validity of their instrument by conducting falsification tests (also called placebo tests). The approach involves finding a context in which the normally-hypothesized relationship should not be true, and then estimating the usual model and testing for the hypothesized relationship. If the model suggests that the relationship is true in a context in which it cannot possibly be true, then the model is likely flawed. This approach is not specific to IV—it is appropriate for any empirical model. A classic example is Dranove and Wehner[12]; the authors were suspicious of the usual IV method used to test whether physicians can induce demand for their services, so they applied the usual model to a context in which demand inducement is impossible: childbirths. Their application of the usual model indicated (illogically) that obstetricians can induce patients to have additional children, which the authors interpret as evidence that the model was flawed and was measuring not demand inducement, but patients crossing borders for care. Again, one cannot directly test for instrument validity, but results from falsification tests that are consistent with validity are suggestive evidence in favor of the IV.

So far, we have assumed that the treatment effect is identical for all subjects. However, in many cases there may be heterogeneity in the treatment effects; for example, treatment effects may differ by sex, race, age, or socioeconomic status. This is relevant for IV because a given instrument may only create variation in the treatment for certain sub-populations, and, thus, the IV model will measure the treatment effect only for those sub-populations (called the Local Average Treatment Effect or LATE) instead of the average treatment effect for the entire population.

Angrist and Pischke[4] classify subjects into three categories: (1) compliers, whose decision to take up the treatment is influenced by the instrument; (2) always-takers, who take up the treatment irrespective of the instrument; and (3) never-takers, who will not take up the treatment irrespective of the instrument. In IV models, the LATE measures the effect of the treatment on compliers; it cannot provide information on the effect of the treatment on always-takers or never-takers because they are not influenced by the instrument.

This has several implications. First, the LATE measured by the method of IV may differ from the average treatment effect for the entire population, which may be what the researcher wishes to know. Second, instruments that affect different sub-populations may yield different estimates of the treatment effect (which, as mentioned above, is relevant for tests of over-identification). Third, IV can be a complement to, rather than a substitute for, RCTs. IV allows researchers to estimate the treatment effect for specific sub-populations of interest that may be difficult to study using RCTs; for example, if few of the sub-population are among the subjects of the RCT.

Methods

In order to examine the use of IV models in health-services research and medical research, a literature search was conducted in April 2013 of the PubMed and EconLit research databases for articles that satisfied the following criteria: English-language journal articles published after 1990 that satisfied the search terms ('instrumental variable' or 'instrumental variables') in any field and ('health' or 'medical' or 'clinical' or 'patient' or 'patients') in any field.

A total of 1241 hits were returned, 640 from PubMed, and 601 from EconLit. We excluded duplicates, dissertations, working papers, books, and book reviews ($n = 388$), which left 853 papers of interest.

Citations to each were extracted from the Web of Science through the Cited Reference Search function. Given the large number of relevant papers, an exhaustive comprehensive review was infeasible; instead, a selective review was conducted, with exemplar articles chosen on the basis of number of citations, type of instrument (in order to offer a wide variety of examples of the method), and fidelity to best practices for the method of IV (see, e.g., Angrist and Pischke[4]).

In order to gather suggestive evidence on the change in the use of the method of IV over time relative to other methods, PubMed was searched to determine the number of articles that contained each of the following search terms in any field: 'instrumental variables', 'ordinary least squares', 'logit regression', 'regression discontinuity', and 'randomized controlled trial'. The second and third terms were chosen because they are common methods, and the final two terms were chosen because they represent alternate methods of measuring causal effects. Two periods were compared: 1991–1995 (chosen because it is when papers using the method of instrumental variables in health-services research and medical research were first published) and 2006–2010. The search results undoubtedly contain false positives (papers that use the term but do not implement it) but the relative trends provide information about the change over time in the relevance of IV in health-services research and medical research.

Results

The 853 papers that satisfied the search criteria are categorized in Table 1. From 1990–2013, 522 journal articles were published in health-services research and medical research that used the method of IV. In addition, 81 methodological articles were published that explain the method, as well as 14 reviews of literature using the method. The 236 articles that satisfied the search terms but were deemed not relevant mentioned the method in passing, but did not implement it.

The trend in the use of IV is documented in Table 2. From 1991–1995, an average of 1.2 articles per year in health-services research and medicine used the method of IV. This rose to 9.4 per year in 1996–2000, 14.6 per year in 2001–2005, and 41.8 per year in 2006–2010. Most recently, from 2011–2013, 80.3 articles per year use the method of IV. The annual number of articles published in this literature using the method of IV is plotted in Figure 1; it portrays an exponential increase over time. Unambiguously, the use of IV has become widespread and common in health-services research and medical research.

Table 1. Categories of papers applying the method of instrumental variables to topics in health and medicine.

Type of article	Number of articles
Original research articles	522
Methodological pieces	81
Reviews	14
Not relevant	236
Total	853

Note: This table lists the categories of papers found in the PubMed and EconLit research databases that satisfied the following search criteria: English-language journal articles published after 1990 that satisfied the search terms ('instrumental variable' or 'instrumental variables') in any field and ('health' or 'medical' or 'clinical' or 'patient' or 'patients') in any field. A total of 1241 hits were returned, 640 from PubMed and 601 from EconLit. We excluded duplicates, dissertations, working papers, books, and book reviews ($n = 388$), which left 853 papers of interest. The 'not relevant' category includes articles that satisfied the search terms but only mentioned the method of IV in passing but did not implement it, or were not truly on the topic of health or medicine. See the Methods section for additional details of the literature search.

Table 2. Number of original research articles published that use the method of IV in health-services research or medical research.

Years	Total articles published	Average per year
1991–1995	6	1.2
1996–2000	47	9.4
2001–2005	73	14.6
2006–2010	209	41.8
2011–2013 (partial year)	187	80.3
Total 1991–2013 (partial year)	522	23.4

Note: See the Methods section for information on the details of the literature search.

It is clear that the use of IV has increased in absolute terms, but one might question whether it has increased in relative terms. That is, has its use increased relative to the number of papers published in the field, and has its use increased more than the use of alternate methods? To answer these questions definitively would require analyzing thousands of articles individually to determine the methods they use, which is beyond the scope of this study. To provide preliminary information on these questions, we searched PubMed to determine how the use of alternate models changed over the same period.

Between 1991–1995 and 2006–2010, the number of new articles indexed by PubMed increased by 81%. During that same time, the number of articles that satisfied

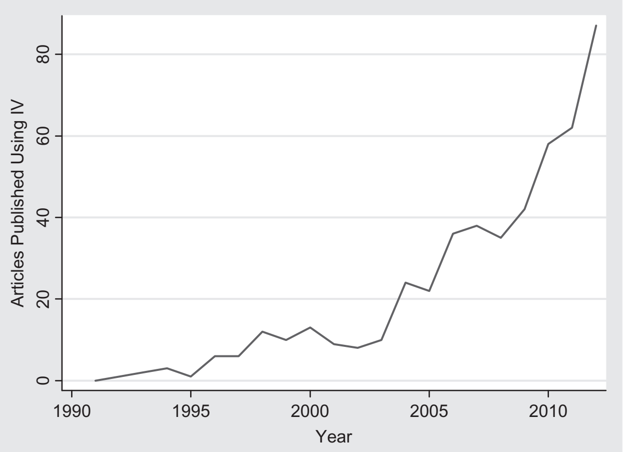

Figure 1. Number of articles published in health-services research and medical research using the method of instrumental variables, 1991–2012. See the Methods section and/or the notes to Table 1 for information on the details of the literature search.

the search term 'instrumental variables' increased 1770%. This is greater than the increase in the numbers of articles using the terms 'logit regression' (557% increase), 'ordinary least squares' (396% increase), 'regression discontinuity' (275% increase), or 'randomized controlled trial' (118% increase). Thus, it seems that the method of IV has become increasing relevant in health-services research and medical research, not only in absolute terms, but also in relative terms.*

The use of instrumental variables has increased more rapidly in health-services research and medicine than in Economics as a whole.† A search of the Econlit database indicates that between 1991–1995 and 2006–2010 the number of new journal articles and working papers indexed by EconLit increased 95.6%; during that same time, the number of journal articles and working papers in EconLit that satisfied the search term 'instrumental variables' increased 654%, which is roughly 40% of the increase seen in the PubMed database of medical journal articles. The slower growth in economics as a whole may be due to the method already being more established in that discipline in 1991–1995.

In the remainder of this section I discuss categories of instruments used in this literature. As shown in Table 3, these categories include: distance to healthcare provider, historic tendency of healthcare provider to administer the treatment, timing of hospital admission, genes, and the heritable condition of a biological relative. For each type of instrument, Table 3 lists an exemplar paper, describing its endogenous treatment of interest, the specific instrument used, the outcome of interest, the finding, and the number of cumulative citations to the paper as well as the number of citations per year since publication. Additional examples of papers using that type of instrument are listed in the final column.

Before discussing the common categories of instruments, I first describe two categories of instruments that are not particularly common but are useful for illustrating how the method of IV, when executed correctly, is analogous to an RCT.

Randomization for purposes other than research

This type of instrument measures random assignment to the treatment, but in contrast to an RCT this random assignment was conducted by someone other than the researcher and for purposes other than research. Thus, although the random assignment was conducted deliberately for a specific purpose, it represents a 'natural' experiment to the researcher.

A classic example in this genre is Angrist *et al.*[13], which seeks to estimate the causal effect of being a military veteran during the Vietnam War era on later-life (civilian) mortality. The challenge is that veteran status is endogenous; the military rejects recruits who are in poor health and this selection would bias any naïve estimate of the effect of military service on later-life mortality.

Angrist *et al.*[13] takes advantage of the fact that, from 1970–1973, priority for the military draft in the US was determined by one's birthday, all of which were ordered based on a lottery (specifically, capsules containing birth dates were drawn from plastic bins). Thus, for men in certain birth-year cohorts during the Vietnam War, one's birthday became a significant predictor of military service. The instrument (draft priority, determined by birth date and the lottery) is presumably uncorrelated with other outcomes that might affect mortality. This random variation in military service was done deliberately by the government to fairly distribute the burdens of military service, but from the perspective of the researchers the draft lottery represents a natural experiment.

The instrument is powerful; the authors estimate that a low draft number (which prioritized one to be drafted) increased the probability of military service by an average of 15.9%. The IV estimates indicate that military service raised the risk of civilian mortality at ages 23–33 years by 0.56 percentage points, or roughly 25%.

This paper was highly influential for several reasons: it was an early (1996) application of the method of IV, a large portion of the paper is methodologically instructive, and the natural experiment it exploits was well-known

*This is not to say that use of IV has increased more than use of every other method of estimating causal effects. A search of PubMed indicates that articles satisfying the search term 'propensity score matching' rose from zero in 1991–1995 to over 350 in 2006–2010.

†The method of instrumental variables is probably most widely used in economics, but is also used in statistics[13], epidemiology[7], sociology[8], and political science[14].

Table 3. Examples of applications of the method of instrumental variables in health and medicine.

Type of instrument	Example paper							Other example papers
	Reference	Endogenous treatment	Instrument (natural experiment) for treatment	Outcome	Finding	Citations	Citations/ year	
Explicit randomization for reasons other than research	13	Military service during the Vietnam War era	Indicator for whether draft lottery number made individual eligible for the draft	Mortality	Military service raises mortality risk by ~25%	958	53.2	15
Randomized encouragement	16	Influenza vaccination	Physicians randomly assigned to receive computerized reminders when patient with scheduled appointment is eligible for flu shot	Flu-related hospitalization	Cannot reject null hypothesis that influenza vaccination does not reduce risk of hospitalization for respiratory illness	20	2.5	46
Distance to provider	9	Intensive treatment (e.g., cardiac catheterization) of those with acute myocardial infarction (AMI)	Difference between patient's distance to nearest hospital capable of intensive AMI treatment and distance to nearest hospital	Short-term and long-term mortality	Intensive treatment of AMI increases probability of surviving 4 years by no more than 5 percentage points	385	19.3	17, 18, 19, 20
Provider's historic tendency to provide the treatment	22	Use of a selective COX-2 inhibitor instead of a non-selective non-steroidal anti-inflammatory drug (NSAID)	Whether prescribing physician's previous new NSAID prescription was for a COX-2 inhibitor	Gastrointestinal complications within 120 days of exposure	Use of a COX-2 reduces probability of GI complications by 1.3 percentage points	98	12.3	23, 24, 25, 26
Day of week of admission	29	Waiting time for hip fracture surgery	Day of week of admission	Post-surgery length of stay and inpatient mortality	Wait time for surgery is not a significant predictor of post-surgery length of stay or inpatient mortality	26	1.9	30, 31, 32
Genes (aka Mendelian randomization)	38	Body mass index (BMI)	Genotypes for FTO and MC4R	Hypertension	Each 10% increase in BMI caused increases of 3.85 mm Hg in systolic, and 1.79 mm Hg in diastolic, blood pressure	37	7.4	39, 47
Heritable condition of a biological relative	41	Obesity or BMI	BMI of a biological child	Medical care expenditures	Obesity raises annual medical care costs by $2741 per year (in 2005 dollars)	30	15.0	40, 42, 43

Note: Citation counts from the Web of Science. See the Methods section for details.

and easily understood as providing non-experimental random variation in the treatment. As a result, the paper has been cited 958 times according to the Web of Science, an average of more than 53 citations per year since publication.

Another article that exploits random assignment that was conducted for purposes other than research is Doyle *et al.*[15]. They use data from a Department of Veterans Affairs hospital in one urban area, which, for the purposes of fairly distributing caseloads, randomly assigned patients to one of two teams of physician residents, one of which was from a top-ranked medical school and one of which was from a much lower-ranked medical school. Specifically, patients whose social security number ended in an odd number were assigned to one program, and those whose social security number ended in an even number were assigned to the other program. This assignment ensures that the instrument is both powerful and valid—the ending digit of a social security number is completely deterministic of assignment to a physician team but is otherwise uncorrelated with patient outcomes. All patients received their care at the same VA hospital and from the same nursing staff; the only difference was the presumed quality of the physician residents (proxied for by the ranking of the medical school with which they were affiliated).

Doyle *et al.*[15] found that patients randomly assigned to residents from the higher-ranked program had 10% lower costs than those randomly assigned to residents from the lower-ranked program; this was driven by doctors from the higher-ranked program ordering fewer diagnostic tests. There were no differences in patient health outcomes such as mortality.

Randomized encouragement

Another category of instrument that is a clear analogy to an RCT is that of randomized encouragement. As opposed to an RCT that would assign a person to a treatment, a person can be randomized to receive merely an encouragement to voluntarily undertake the treatment. This instrument's power depends on the extent to which the encouragement increases uptake.

For example, Zhou and Li[16] seek to measure the effectiveness of flu shots, but an RCT that withheld the vaccine from individuals in the control group would be unethical. As a result, the authors report that there exist no published studies of the impact of flu vaccination on pulmonary morbidity. To generate such an estimate, they examine data in which physicians randomly received a computerized reminder that a patient with a scheduled appointment was eligible for a flu shot; this randomization serves as their instrumental variable. The encouragement does seem to affect uptake of the vaccine; 21.46% of patients whose physicians received the encouragement were vaccinated, compared to 13.64% of patients whose physicians did not receive the encouragement. The paper does not discuss the power of the instrument further (e.g., it does not provide the *F* statistic for the IVs in the first stage). The results of their IV models indicate that they cannot reject the null hypothesis of no effect of influenza vaccination on the probability of hospitalization for respiratory illness. However, it is important to keep in mind that the IV measures the Local Average Treatment Effect among compliers. In the context of this paper, the compliers are those who would not have gotten a flu shot in the absence of the reminder, but who did get the flu shot because of the reminder. Such people may benefit less from a flu shot than the 'always-takers' — those who would be given a flu shot whether or not there was a reminder; thus, this paper may under-state the benefits of flu shots for the population as a whole. The study does not examine whether influenza vaccination reduces the probability of contracting influenza.

Distance to provider

McClellan *et al.*[9] is an important landmark paper that pioneered the use of the method of IV in health-services research.* The purpose of that paper was to measure the effect on mortality of more intensive treatments for acute myocardial infarction (AMI), such as cardiac catheterization and revascularization. The challenge once again is selection—more intensive treatment is given to severely ill patients, but not the *most* severely ill patients (who might not survive the procedure), making it difficult to accurately estimate the effect of the treatment on mortality. RCTs that assigned patients with AMI to more intensive and less intensive treatments were seen as costly, difficult, and unethical, so the authors took advantage of naturally-occurring variation in treatment that resulted from heart attack patients living closer to one hospital than another. Patients suffering from an AMI will tend to be taken to the nearest hospital for treatment. If that nearest hospital is capable of the more intensive treatments (catheterization and revascularization) then the patient is more likely to get the more intensive treatment. The authors define their instrument as the distance from the patient's home zip code to the nearest hospital that treats AMI intensively (based on recent history), minus the distance to the nearest hospital; this measures the relative distance to a hospital that treats AMIs intensively.

*At the time of their writing (1994), the authors could accurately say that 'The method, instrumental variables (IV) estimation, is well known in econometrics but generally has not been applied to estimate relationships between medical treatments and health outcomes' (McClellan *et al.*[9], pp. 859–60). Table 2 and Figure 1 in this paper confirm that statement.

Living closer to a catheterization hospital does not guarantee that an AMI patient is admitted to a catheterization hospital, but it does raise the probability from 5.0% to 34.4%. As a result, the instrument is powerful in predicting the treatment; those who live closer to a catheterization hospital were almost twice as likely to undergo catheterization within 7 days of admission (20.7% vs 11.0%).

The authors conclude that failure to adjust for selection can lead to substantial biases. Previous models that simply controlled for observable characteristics had suggested a large mortality benefit to invasive AMI treatment, but the authors' IV models (estimated using a sample of elderly patients) indicate that such procedures cause at best a modest improvement in mortality. Moreover, the authors are suspicious of even the modest improvement in mortality because it is visible at day one of admission, which is before the procedures are performed (they usually take place 7–30 days after the AMI). The authors hypothesize that the hospitals that are more likely to use invasive procedures to treat AMI are also better at other aspects of acute care, which explains the early mortality benefit.

There are two threats to validity for this instrument. The first is that patients who live closer to hospitals that tend to provide intensive treatments may be different than those who live further from such hospitals, and these differences may be correlated with outcomes. To investigate this, the authors divide the sample by the value of the instrument to inspect whether those with high and low values of the instrument differ on unobserved characteristics. Their Table 4[9] lists the mean characteristics of those who lived relatively far and relatively close to a hospital that intensively treated AMI and found that those who lived relatively far were more likely to be rural and less likely to be black (which the authors can control for in their IV models) but are very similar to each other in terms of co-morbid disease characteristics such as cancer, diabetes, and renal disease.

The second threat to validity for this instrument is that hospitals that were capable of the more intensive treatments may have also had other advantages—they may have been more modern or efficient or aggressive about treatment generally in ways that benefitted patient outcomes, aside from whether catheterization or revascularization took place. In essence, that is what the authors acknowledge when they hypothesize that their results showing a mortality benefit from day one of admission are not due to the intensive treatment that has not yet occurred, but instead reflect that hospitals that treat AMI invasively are also better at acute care in other unobserved ways.

The McClellan *et al.*[9] paper also illustrates why it matters that the method of IV measures a Local Average Treatment Effect (LATE). The IV method measures the effect of the treatment for compliers—in this paper, the compliers are heart attack patients who would not have received intensive treatment if they had lived relatively far from a hospital that tends to treat heart attacks intensively, but who did receive treatment because they lived relatively close to such a hospital. The benefit of intensive treatment for compliers may be less than what it would be for the 'always-takers' — heart attack patients who are always intensively treated, irrespective of their distance from such hospitals. For this reason, this paper may understate the average treatment effect of such surgeries. This explains why intensive treatment for heart attacks continues, despite this paper's finding of minimal benefits among compliers.

The McClellan *et al.*[9] paper is seen as a landmark application of IV to questions in medicine and health services. It has been cited 385 times, or roughly 19 times per year since publication.

Distance to the provider has been one of the more popular categories of instruments. Others have used it to study the impact of high-level vs low-level Neonatal Intensive Care Units[17], for-profit vs not-for-profit dialysis centers[18], level I vs level II trauma centers[19], and high-quality vs low-quality hospitals[20].

Provider's historic tendency to administer the treatment

Another common IV approach is to exploit historic differences in the ways that providers (either hospitals or individual physicians) treat similar cases; these are sometimes called 'preference-based' IVs[21]. The logic is that providers may have different decision rules or preferences when deciding between alternative treatments. For example, Brookhart *et al.*[22] wish to measure the impact on the probability of gastrointestinal (GI) toxicity of a patient being prescribed one type of painkiller rather than another; specifically, a COX-2 inhibitor instead of a non-selective non-steroidal anti-inflammatory drug (NSAID). This would seem to be a research question that could be addressed with an RCT, which are routine in testing the efficacy of pharmaceuticals, but the authors note that RCTs are generally under-powered to detect uncommon adverse events; a strength of using opportunistic data is that it may provide the statistical power necessary to detect even uncommon side-effects.

The empirical challenge is that physicians try to prescribe the most appropriate pharmacotherapy for each patient, which includes trying to avoid adverse side-effects. This is good for the patient, but this causes selection bias in a naïve regression of outcomes on the drug prescribed. The authors seek to address this through the IV method, using as an instrument the prescribing physician's historic tendency to prescribe a COX-2 instead of a non-selective NSAID. Specifically, the authors use the decision the same physician made regarding their immediately preceding patient who needed either a COX-2 or a non-selective NSAID. As a robustness check, they also use an alternative instrument that is the historical proportion of a physician's new NSAID prescriptions that were for a COX-2 inhibitor.

This instrument is powerful; the probability of being prescribed a COX-2 was 55% if the same physician's last NSAID prescription was for a non-selective NSAID, but that rises to 77% if the physician's last NSAID prescription was for a COX-2.

Conventional non-IV analyses yielded no evidence that COX-2s were protective against GI toxicity, but the IV models indicate that prescribing a COX-2 instead of a non-selective NSAID lowered the probability of GI toxicity within 60 days, 120 days, and 180 days by 1.02, 1.31, and 1.21 events per hundred patients, respectively. A nice advantage of this research question is that the conventional analyses and IV results can be compared to evidence from RCTs, which confirms that the IV method produces more accurate results than the non-IV methods. The advantages of the IV method relative to the existing RCTs are greater statistical power and the ability to examine different populations, such as an older and more frail sample of Medicare beneficiaries, among whom the authors find a greater protective effect. Thus, IV is not only a substitute for RCTs, it can be a complement to them.

The potential challenges to validity are that patients may sort to physicians based on factors such as income and severity of illness, and thus the current patient may have similar unobserved characteristics as the previous patient and thus the instrument (the provider's previous decision) may be correlated with unobserved correlates of the subject's health. Brookhart *et al.*[22] examine the observable characteristics of patients whose physicians tend to prescribe COX-2, and find that these tend to be higher-risk patients. Another risk is that providers who prefer the treatment to the alternative may also do other unobserved things differently that affect patient outcomes. For example, Brookhart *et al.*[22] acknowledge that physicians who tend to prescribe COX-2 may also influence the outcome in other ways, such as prescribing proton pump inhibitors as a precaution, and this additional GI protection may be incorrectly attributed to the COX-2.

Once again, it is important to keep in mind that the IV measures the Local Average Treatment Effect (LATE) among compliers. In the context of Brookhart *et al.*[22], the compliers are those who would not have been prescribed a COX-2 by certain physicians, but would be prescribed them by others. The treatment effect among this group may not generalize to other groups in the population, such as the 'always-takers' who would be prescribed a COX-2 by any physician.

'Preference-based' IVs have been used widely. Some utilize preferences of individual physicians[22], while others use facility-level data[23,24]. At the extreme, some have used regional variation in procedures as an instrument[25,26], but this may be less convincing than using the history of a single provider; geographic variation in treatment may be due to differences in unobserved correlates of demand, such as disparities in income, education, cancer clusters, or other aspects of unobserved health. Applications of this category of instrument differ in whether they use the provider's immediately previous decision or average decision over a longer period of time.

Preference-based IVs may violate the monotonicity condition. In the case of Brookhart *et al.*[22], different physicians may have different decision rules for matching patients to drugs, and seeing one physician rather than another may raise the probability of being prescribed a COX-2 for some patients, but lower that probability for others. A similar issue has been raised about the use of random assignment to disability examiners as an instrument for the probability of being approved for Social Security Disability Insurance[27]; examiners may look more favorably on certain applicants and less favorably on others, violating the monotonicity requirement of IV[28].

The previous category of instruments, differential distance to a provider, also involves the provider's historic tendency to provide the treatment; for example, in McClellan *et al.*[9], whether the nearest hospital is classified as capable of intensive treatment of AMI is determined by the hospital's history of providing such treatments. However, using distance to provider as an instrument has an added advantage of utilizing variation in the matching of patients to providers. Using the provider's historic treatment patterns alone as an instrument, without any correction for matching of patients to providers, risks selection bias due to the endogenous matching of patients to providers.

Timing of admission

Day of admission can be useful as an instrument for waiting time prior to surgery or for whether a given treatment is ever received. The first paper to use this category of instrument, Ho *et al.*[29], estimated the effect of waiting time for hip surgery on post-surgery length of stay and mortality. Their instrument is the day of week of admission to the hospital. The authors assume that surgeons prefer to operate on weekdays rather than weekends and, therefore, patients admitted on Friday may wait longer for surgery than those admitted on Monday. They found in their data for the US and Canada that average wait times for surgery do differ significantly by the day of week of hospital admission; for example, in Massachusetts the percentage of patients delayed 3 or more days before surgery was 27% for those admitted on a Wednesday, 25% for those admitted on a Thursday, and 32% for those admitted on a Friday.

However, in other geographic areas they examine (Manitoba, Quebec, and California) it is not true that those admitted near the weekend end up waiting longer for surgery. The paper does not provide detailed information about the power of the instrument.

In contrast to the descriptive statistics that suggested that longer waiting time for hip surgery was associated with higher mortality, the IV models failed to reject the null hypothesis of no effect.

A risk to validity for this type of instrument is that day of week of admission may affect post-surgery outcomes for reasons other than surgery wait time. For example, Ho *et al.*[29] acknowledge that day of week of admission could affect patient outcomes through lower levels of nursing staff on the weekend. However, they argue that if this was the case then the IV models should show some effect of day of week of admission on post-surgery outcomes when in fact they cannot reject the null hypothesis of no effect.

One concern is that patients who are knowledgeable about this feature of the healthcare system, and whose conditions are less severe, could wait and time their admission to minimize their inpatient wait for surgery; this could bias estimates, making longer wait times more strongly associated with worse post-surgery outcomes. For this reason, day of week is a more suitable instrument for acute conditions such as heart attacks or broken hips because individuals with such conditions are unlikely to wait and choose the optimal day to be admitted to the hospital. In other words, the more urgent or acute the condition the more exogenous may be the day of week of admission to the hospital.

Bhattacharya *et al.*[30] apply this category of instrument in their investigation of whether use of the Swan-Ganz catheter (which helps physicians monitor heart functioning) affects mortality. Previous research had found the controversial result that patients in the intensive care unit (ICU) who receive Swan-Ganz catheterization were significantly more likely to die within 180 days. Bhattacharya *et al.*[30] re-analyze the data from an earlier study, but estimate IV models that take advantage of the fact that patients admitted on weekends are less likely to be catheterized on the day of admission; the *F* statistic on the instrument in the first stage exceeds the minimum threshold of 10 (it is 14.53). As a test of validity, they check whether patient health status at admission varies by the day of week of admission and find no statistically significant differences. Their models indicate that the Swan-Ganz treatment either decreases mortality or has no detectable effect on mortality while the patient is in the ICU, although the treatment may increase mortality after the patient leaves the ICU.

This category of instrument has also been applied to estimate the impact of alternative treatments for patients with kidney stones (medical expulsive therapy vs early endoscopic removal) on medical expenditures and re-admission[31] and the effect of very early rehabilitation for stroke victims[32].

Genetic endowment

A more recent IV approach is to instrument for a heritable condition using an individual's genotype. In epidemiology, this is called Mendelian randomization[33,34]; in essence, it takes advantage of the very natural experiment of the randomization of genetic endowment. The identifying assumption is that one's genotype predicts the heritable characteristic, but is otherwise uncorrelated with the outcome.

This strategy faces several challenges. First, individual genes tend to be weak instruments. An ideal instrument would be a gene that acts as a switch, turning on or off one characteristic while affecting no other characteristics that could bias the IV estimate. However, Conley[35] is skeptical. He notes that humans have only 21,000 genes to code for a vast number of characteristics and, thus, genes tend not to act as isolated switches for single characteristics; instead, each gene acts in concert with many others to affect a large number of characteristics. Thus, any single gene or small number of genes may not be very powerful as instruments, and they also may not be valid because each could affect characteristics other than the endogenous one of interest[36,37].

A second threat to validity is genetic linkage. People do not inherit individual genes, but instead strips of DNA from their parents. As a result, the inheritance of one gene may be correlated with the inheritance of other nearby genes that are not controlled for in the model but that are correlated with the outcome.

Timpson *et al.*[38] seek to measure the effect of body mass index (BMI) on blood pressure and the risk of hypertension. RCTs of weight reduction had suggested that weight loss reduces blood pressure, but these estimates were suspect because certain aspects of the weight reduction treatments, such as restricted dietary intake or increased exercise, could have reduced blood pressure independently of BMI. In order to estimate the causal effect of BMI on blood pressure, the authors instrument for BMI using rs9939609 (FTO) and rs17782313 (MC4R) genotypes, which were linked to higher weight in multiple studies. The IV models imply that each 10% increase in BMI raises systolic blood pressure by 3.85 mm Hg and diastolic blood pressure by 1.79 mm Hg.

The genetic instruments in Timpson *et al.*[38] are weak; together the two genetic loci account for only 0.6–0.7% of the variation in BMI in European populations. Investigating validity, the authors find no robust associations between their genetic instruments and confounding factors such as income and education.

Voight *et al.*[39] exploit Mendelian randomization to measure the impact of high levels of high density lipoproteins (HDL) or 'good' cholesterol on the risk of AMI. The authors instrument for HDL cholesterol levels using a single nucleotide polymorphism (SNP) in the endothelial lipase gene (LIPG Asn396Ser). Multiple prospective cohort studies had documented that carriers of the LIPG 396Ser allele had higher HDL, but similar risk factors for AMI compared to non-carriers. Although observational studies in epidemiology had found that an increase in HDL cholesterol was associated with a significantly reduced risk of AMI, the authors' IV models indicate a small and statistically insignificant reduction in risk.

The power of this single allele as an instrument is not discussed much by the authors, but only 2.6% of the authors' sample are carriers of the allele. In search of a stronger instrument, the authors create a score that reflects whether subjects have HDL-raising alleles at 14 SNPs and find similar results. Regarding validity, they note that it is impossible to exclude the possibility that the genes in question have multiple functions (which is called pleiotropy), but find that the allele in question is not associated with a wide variety of other risk factors for AMI, such as LDL cholesterol, triglycerides, BMI, blood pressure, or diabetes.

Heritable conditions of biological relatives

A related approach is to instrument for a heritable condition or characteristic of the subject using the same condition or characteristic of a biological relative. This can be a powerful instrument in situations in which the condition in question has a large heritable component. In particular, it is generally more powerful than using a single SNP as an instrument, but it does not necessarily avoid the threats to validity posed by pleiotropy or genetic linkage.

The first application of this method was by Cawley[40], who sought to measure the impact of BMI on employment disability. Previous studies had documented that heavier individuals were more likely to be disabled, but there was no evidence of the causal effect of obesity on disability. To measure that causal effect, Cawley[40] instrumented for the BMI of the subject using the BMI of a biological child, controlling for the child's age and gender. This relies on a large literature that finds that there is a large heritable component of BMI. The instrument had an F statistic slightly above the minimum threshold of $F=10$; it explained roughly 4% of the variation in BMI and 10% of the variation in weight in pounds, controlling for other observed characteristics. The validity of the instrument depends on the child's weight not being correlated with parental disability, except through parental weight. Cawley[40] cites research which finds no consistent pattern between childhood obesity and socioeconomic status, and no measurable effect of common household environment on body weight. The author also follows the example of McClellan *et al.*[9] and compares the observable characteristics of respondents based on whether they have a low or high value of the instrument (BMI of their biological child, controlling for age) and finds that the mean characteristics are similar, consistent with the identifying assumption. The IV models yield no evidence that body weight causes employment disability; instead, the observed correlation may be due to disability causing weight gain or the influence of omitted variables.

Cawley and Meyerhoefer[41] estimate the effect of BMI and obesity on medical care expenditures. They instrument for respondent BMI using the BMI of a biological child, controlling for child age and gender. The instrument is powerful, with F statistics across models ranging from 31 to 281, and averaging 144. Their IV models indicate that the medical care costs of obesity are much higher than previously thought. Whereas non-IV models indicate that obesity is associated with $656 higher annual medical care costs, the IV models indicate that obesity raises medical care costs by $2741 per year (in 2005 US dollars). To examine the validity of the instrument, the authors conduct several falsification tests, the results of which are consistent with the identifying assumption. The IV model finds a stronger impact of obesity on medical expenditures for diabetes (clearly linked to obesity) than on medical expenditures for other conditions, and the model does not find an impact of obesity on medical care costs for conditions that are unrelated to obesity (e.g., epilepsy, brain damage, and central nervous system disorders). Moreover, biologically unrelated children (e.g., stepchildren) are not significant predictors of respondent weight. Still, the authors acknowledge the potential threats to validity posed by pleiotropy and genetic linkage. Cawley *et al.*[42] use a similar approach with more and updated data.

Another example of this category of instrument is Smith *et al.*[43], which estimates the impact of BMI on mortality by instrumenting for respondent BMI with offspring BMI, and concludes that BMI may raise all-cause mortality and mortality from cardiovascular disease substantially more than previously appreciated.

Additional observations

The selective review yielded additional observations. First, standards for IV have risen over time. Early papers tended to provide only vague information about power, whereas more recent papers are more likely to report F statistics from the first stage and acknowledge the minimum standard of power of $F \geq 10$. Unfortunately, even some recent papers are relatively vague about instrument power.

Second, the standards for validity seem to have risen as well. Falsification tests of instruments have become more common. Some instruments that were used decades ago would be unlikely to pass peer review today; see, e.g., French and Popovici[44] and Rashad and Kaestner[45] for discussions of specific instruments used in the earlier literature that are now considered weak and invalid.

More recent studies are also more likely to discuss the marginal population affected by the instrument, and how the LATE measured by IV likely relates to the average treatment effect for the population.

Conclusion

Many studies in public health estimate only correlations. Descriptive studies are valuable, because it is important to establish basic facts and patterns in the data, but correlations can be misleading estimates of the true causal effects. Remler and van Ryzin[3] offer the example of autism. Because the prevalence of autism rose during a period when vaccinations were rising, and because the onset of autism symptoms occurs around the same time as some childhood vaccinations, some people mistakenly concluded that vaccines caused autism, which led some parents to refuse to vaccinate their children, which in turn led to the resurgence of previously suppressed childhood illnesses.

Estimating causal effects as opposed to (or in addition to) correlations can be extremely useful for clinical decision-making and formulating policy, because causal effects provide information about the effects of specific treatments on health, and the ability of interventions to influence important outcomes.

Instrumental variables is one useful method of measuring causal effects, and should be considered alongside and in conjunction with other methods of identification, such as randomized controlled trials, propensity score matching, difference-in-differences models, and regression discontinuity designs[3,4,8].

One risk is that the method of IV can be used incorrectly, such as with a weak or invalid instrument. In addition, one must be cautious when generalizing from IV studies; the Local Area Treatment Effect (LATE) measured in one study may be quite different than average treatment effect for the population as a whole.

This paper conducts a selective review of the first 20 years of applications of the method of instrumental variables in health-services research and medical research. It finds a dramatic increase in the use of the method, with at least 80 papers that use the method published annually in recent years. Based on searches of PubMed, the method of IV has become increasingly relevant in health-services research and medical research in relative as well as absolute terms. The selective review indicates that types of natural experiments that are exploited using the method of IV include: distance to provider, the provider's historic tendency to administer the treatment, day of week of admission, and genetic endowment.

The method of IV can be a substitute for RCTs when RCTs would be unethical or infeasible (as in the case of intensive treatment of AMI examined by McClellan *et al.*[9]) or can be a complement to RCTs when there are advantages of greater statistical power or the ability to examine new sub-populations (as in the case of pharmacological treatment examined in Brookhart *et al.*[22]).

Results from studies using the method of IV have challenged conventional wisdom that was based on correlations rather than causal effects. In some cases, IV indicated that treatments that were previously believed to be efficacious in fact had little or no effect. McClellan *et al.*[9] show that, contrary to popular wisdom at the time, more intensive treatment of heart attacks may have had no beneficial impact on mortality at all. Doyle *et al.*[15] show that, surprisingly, the quality of physician training has no detectable impact on patients' health outcomes such as mortality. Ho *et al.*[29] find that waiting time for hip fracture surgery does not affect post-surgery length of stay in the hospital or mortality. Voight *et al.*[39] conclude that raising HDL cholesterol does not in fact reduce the risk of heart attack. Zhou and Li[16] find the influenza vaccination had no impact on the probability of hospitalization for respiratory illness.

In other cases, IV indicated that treatments had much bigger effects than earlier, non-IV studies had suggested. Brookhart *et al.*[22] find that prescribing a COX-2 instead of a non-selective NSAID lowered the probability of GI toxicity by more than 1 percentage point. Bhattacharya *et al.*[30] find that use of the Swan-Ganz catheter does not increase the risk of mortality but may in fact decrease it. Cawley and Meyerhoefer[41] found that obesity raises medical care costs more than previously appreciated.

An important caveat for all of these studies is that the treatment effects were measured for compliers—those whose treatment decisions were influenced by the instrument. The LATE measured for the compliers may be quite different from the average treatment effect for the population as a whole.

Limitations of this paper include that the review was selective. A comprehensive review of all 522 papers was beyond the scope of this study. Still, this review is able to identify major categories of IVs and offer illustrative examples. The citation counts used in this study are from the Web of Science and do not include citations of these articles by unpublished working papers, books, or government reports; as a result, they under-state the full impact of the IV papers. This review was based on the databases PubMed and EconLit; there may have been additional applications of the method of IV to topics in health and medicine that were published in journals other than those in medicine, public health, and economics.*

The method of IV, when properly conducted with powerful and valid instruments, can estimate causal effects when randomized controlled trials are infeasible or unethical, or as a complement to RCTs in order to estimate causal effects for specific sub-groups. Readers should pay careful attention to the extent to which instruments satisfy the requirements for power, validity, and monotonicity, and consider how the LATE that was estimated may differ from the average treatment effect for the population. Applications of the method of IV have in several cases upended conventional wisdom that was based on correlations and led to important insights about health and healthcare. Future research will no doubt energetically pursue new applications of existing instruments and search for novel instruments that are powerful and valid.

Transparency

Declaration of funding

Funding for this paper was provided by Novo Nordisk, Inc., and the Cornell University Institute on Health Economics, Health Behaviors, and Disparities.

Declaration of financial/other relationships

JC has disclosed that he has no significant relationships with or financial interests in any commercial companies related to this study or article. JME peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

Acknowledgments

JC thanks three anonymous referees, Mette Hammer and Neil Wintfeld for their helpful comments. JC also thanks Anna Choi, Mariann Lucas, and Jeffrey Swigert for research assistance, and Eric Maroney for editorial assistance.

References

1. U.S. Preventive Services Task Force. Procedure Manual – Section 4: Evidence Report Development. 2014. Washington, DC. http://www.uspreventiveservicestaskforce.org/Page/Name/procedure-manual—section-4. Accessed October 23, 2014
2. Freedman B. Equipoise and the ethics of clinical research. N Engl J Med 1987;317:141-5
3. Remler DK, Van Ryzin GG. Research methods in practice: strategies for description and causation. 2nd edn. New York: Sage Publications, 2015
4. Angrist JD, Pischke J-S. Mostly harmless econometrics: An Empiricist's Companion. Princeton, NJ: Princeton University Press, 2009
5. Greene W. Econometric analyses. 7th edn. New York: Prentice Hall, 2011
6. Woodridge JM. Econometric analysis of cross-sectional and panel data. Cambridge: MIT Press, 2002
7. Martens EP, Pestman WR, deBoer A, et al. Instrumental variables: application and Limitations. Epidemiology 2006;17:260-7
8. Morgan SL, Winship C. Counterfactuals and causal inference: methods and principles for social research. New York: Cambridge University Press, 2007
9. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. JAMA 1994;272:859-66
10. Bound J, Jaeger DA, Baker R. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. J Am Statist Assoc 1995;90:443-450
11. Stock JH, Wright JH, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. J Bus Econ Stat 2002;20:518-29
12. Dranove D, Wehner P. Physician-induced demand for childbirths. J Health Econ 1994;13:61-73
13. Angrist J, Imbens G, Rubin D. Identification of causal effects using instrumental variables. J Am Statist Assoc 1996;91:444-55
14. Sovey AJ, Green DP. Instrumental variables estimation in political science: a reader's guide. Am J Polit Sci 2010;55:188-200
15. Doyle JJ, Ewer SM, Wagner TH. Returns to physician human capital: evidence from patients randomized to physician teams. J Health Econ 2010;29:866-82
16. Zhou X-H, Li SM. ITT analysis of randomized encouragement design studies with missing data. Stat Med 2006;25:2737-61
17. Lorch S, Baiocchi M, Ahlberg C, et al. The differential impact of delivery hospital on the outcomes of premature infants. Pediatrics 2012;130:270-8
18. Brooks JM, Irwin CP, Hunsicker LG, et al. Effect of dialysis center profit-status on patient survival: a comparison of risk adjustment and instrumental variable approaches. Health Serv Res 2006;41:2267-89
19. McConnell KJ, Newgard CD, Mullins RJ, et al. Mortality benefit of transfer to level I versus level II trauma centers for head injured patients. Health Serv Res 2005;40:435-57
20. Gowrisankaran G, Town R. Estimating the quality of care in hospitals using instrumental variables. J Health Econ 1999;18:747-67
21. Brookhart M, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. Int J Biostatist 2007;14:1-23
22. 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
23. Schneeweiss S, Seeger JD, Landon J, et al. Aprotinin during coronary-artery bypass grafting and risk of death. N Engl J Med 2008;358:771-83
24. Schmoor C, Caputo A, Schumacher M. Evidence from nonrandomized studies: a case study on the estimation of causal effects. Am J Epidemiol 2008;167:1120-9
25. Brooks J, Chrischilles E, Scott S, et al. Was breast conserving surgery underutilized for early stage breast cancer? Instrumental variables evidence for stage II patients from Iowa. Health Serv Res 2004;38:1385-402
26. Stukel TA, Fisher ES, Wennberg DE, et al. 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;297:278-85
27. Maestas N, Mullen KJ, Strand A. Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. Am Econ Rev 2013;103:1797-29
28. Chaisemartin C. Tolerating defiance? Local average treatment effects without monotonicity. Working paper. Warwick: University of Warwick, 2014
29. 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;34:1499-518

*Few such articles appear to exist; the Soc Abstracts database was searched for journal articles satisfying the terms 'instrumental variables' and ('health' or 'medicine'), and there were only 43 results for the entire period 1990–2013. Many of those were also included in the EconLit or PubMed databases and, thus, were included in this review.

30. Bhattacharya J, Shaikh AM, Vytlacil E. Treatment effect bounds: an application to Swan-Ganz catheterization. J Econometrics 2012;168:223-43

31. Hollingsworth JM, Norton EC, Kaufman SR, et al. Medical expulsive therapy versus early endoscopic stone removal for acute renal colic: an instrumental variable analysis. J Urol 2013;190:882-7

32. Matsui H, Hashimoto H, Horiguchi H, et al. An exploration of the association between very early rehabilitation and outcome for the patients with acute ischaemic stroke in Japan: a nationwide retrospective cohort survey. BMC Health Serv Res 2010;10:213

33. Palmer TM, Thompson JR, Tobin MD. Meta-analysis of Mendelian randomization studies incorporating all three genotypes. Stat Med 2008; 27:6570-82

34. Wehby GL, Ohsfeldt RL, Murray JC. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. Stat Med 2008;27:2745-9

35. Conley D. The promise and challenges of incorporating genetic data into longitudinal social science surveys and research. Biodemogr Soc Biol 2009;55:238-51

36. Cawley J, Han E, Norton EC. The validity of genes related to neurotransmitters as instrumental variables. Health Econ 2011;20:884-8

37. von Hinke Kessler Scholder S, Davey Smith G, Lawlor DA, Propper C, Windmeijer F. Mendelian randomization: the use of genes in instrumental variable analyses. Health Econ 2011;20:893-6

38. Timpson NJ, Harbord R, Davey Smith G, et al. Does greater adiposity increase blood pressure and hypertension risk?: Mendelian randomization using the FTO/MC4R genotype. Hypertension 2009;54:84-90

39. Voight BF, Peloso GM, Orho-Melander M, et al. Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. Lancet 2012; 380:572-80

40. Cawley J. An instrumental variables approach to measuring the effect of body weight on employment disability. Health Serv Res 2000;35:1159-79

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

42. Cawley J, Meyerhoefer C, Biener A, et al. Forthcoming. Savings in medical expenditures associated with reductions in Body Mass Index among adults with obesity, by Diabetes Status. Pharmacoeconomics DOI: 10.1007/s40273-014-0230-2

43. Smith GD, Jonathan ACS, Abigail F, et al. The association between BMI and mortality using offspring BMI as an indicator of own BMI: large intergenerational mortality study. Br Med J 2009;339:b5043

44. French MT, Popovici I. That instrument is lousy! In search of agreement when using instrumental variables estimation in substance use research. Health Econ 2011;20:127-46

45. Rashad I, Kaestner R. Teenage sex, drugs and alcohol use: problems identifying the cause of risky behaviors. J Health Econ 2004;23:493-503

46. Have TR, Ten MR, Elliott MJ, et al. Causal models for randomized physician encouragement trials in treating primary care depression. J Am Statis Assoc 2004;99:16-25

47. Wehby G, Jugessur A, Murray JC, et al. Genes as instruments for studying risk behavior effects: an application to maternal smoking and orofacial clefts. Health Serv Outcomes Res Methodol 2011;11:54-78