

# **HCFE Working Paper**

**HCFE WP# 2015-01**

**Falsification Testing of Instrumental Variables Methods for  
Comparative Effectiveness Research**

Steven D. Pizer, PhD

Associate Professor of Health Economics, Northeastern University

This work was funded by the Department of Veterans Affairs, Veterans Health Administration, Health Services Research and Development Services. This report presents the findings and conclusions of the authors and does not necessarily represent VA or HSR&D.



Health Care Financing & Economics

Boston VA Healthcare System, 150 South Huntington Avenue, Mail Stop 152H, Boston, MA 02130

• Phone: (857) 364-6058 • Fax: (857) 364-4511 • Web: [www.hcfe.research.va.gov](http://www.hcfe.research.va.gov)

## **Falsification Testing of Instrumental Variables Methods for Comparative Effectiveness Research**

Steven D. Pizer, PhD, Associate Professor of Health Economics, Northeastern University

### **Abstract**

**Objectives:** To demonstrate how falsification tests can be used to evaluate instrumental variables methods that can be used to investigate a wide variety of comparative effectiveness research questions.

**Study Design:** Brief conceptual review of instrumental variables and falsification testing principles and techniques accompanied by an empirical application. Sample Stata code related to the empirical application is provided in the appendix.

**Empirical Application:** Comparative long-term risks of sulfonylureas and thiazolidinediones for management of type 2 diabetes. Outcomes include mortality and hospitalization for an ambulatory care sensitive condition. Prescribing pattern variations are used as instrumental variables.

**Conclusions:** Falsification testing is an easily computed and powerful way to evaluate the validity of the key assumption underlying instrumental variables analysis. If falsification tests are used, instrumental variables techniques can help answer a multitude of important clinical questions.

**Key Words:** falsification testing; instrumental variables; comparative effectiveness research; practice pattern variation.

### **Introduction**

Falsification testing is an old idea that has great potential as a method for evaluating the internal validity of comparative effectiveness research (CER) studies. Though rarely identified as such, falsification tests are familiar to most researchers, as they are a routine, almost automatic component of reporting of randomized controlled trial (RCT) results. Falsification testing of observational studies requires more planning in advance, but is not much more difficult to perform than for RCTs. Given the growing importance of observational studies and instrumental variables methods in CER, falsification testing can play a vital role in improving the reliability and impact of this research.

To understand falsification testing, consider the table of sample means by treatment and control group included in most reporting of RCT results. What purpose does this serve? Our expectation is that the sample means will not be significantly different between groups because group assignment was intended to be random. Random assignment is the “identifying assumption” of RCTs because randomization permits us to infer causal effects of treatment. If the sample means differ by group, the identifying assumption has been falsified and we have reason to doubt the internal validity of the trial. That is, a table of means by group is a falsification test of an RCTs central assumption.

As I will show, similar falsification tests can be implemented for observational studies, which are becoming an increasingly important source of clinical evidence. Wider adoption of electronic medical records and substantial new investments (\$3 billion in research and infrastructure between 2013 and 2019) by the Patient-Centered Outcomes Research Institute (PCORI) [1] are increasing capacity to conduct observational, comparative effectiveness and patient-centered outcomes research. A recent analysis of responses to the National Ambulatory Medical Care Survey showed that the percentage of all physicians who had adopted a basic electronic medical record increased from 25.8% in 2010 to 38.2% in 2012 [2]. These rapid changes in technology and research resources raise the prospect of large observational studies based on clinical data with vastly richer detail than what has been available in the past from administrative or claims-based records.

This emerging “big data” environment holds promise to extend the reach of clinical and health services research to include the study of rare events, heterogeneous treatment effects, long-term outcomes, and other topics that are difficult or impossible to study with RCTs [3,4]. Such trials typically involve numbers of subjects in the hundreds, limiting comparisons to a few treatment options and making patient subgroup comparisons difficult or impossible. In addition, external validity is constrained

by recruitment that frequently excludes the most complex or severely ill patients as well as treatment that is conducted in academic medical centers with research staff supplementing clinical staff. In contrast, observational studies can efficiently exploit electronic medical records and administrative databases containing information on tens or hundreds of thousands of patients of all types, treated in a wide variety of clinical settings and followed for many years.

Despite these advantages, a key challenge facing observational CER is evaluating the validity of causal inference [5,6,7]. Fortunately, important technical strides have been made in design and analytic methods to increase the internal validity of observational studies despite a lack of purposeful, explicit randomization. Depending on the source and strength of treatment variation in observational studies, different statistical methods may be appropriate. For example, if the study is small enough that it is practical to collect data on every potentially confounding variable, propensity score methods can ensure balance of observed variables between treatment and comparison groups, revealing the causal effect of treatment. On the other hand, if the study is too large for practical collection of important variables that might be unavailable in clinical or administrative data, risk-adjusted or propensity score estimates are likely to be biased and quasi-experimental methods like instrumental variables (IV) probably will be more appropriate [8].<sup>1</sup>

This article reviews the fundamental concepts underlying IV estimation and falsification testing, and then demonstrates the steps involved using a specific example comparing the long-term risks associated with alternative oral medications used to manage type 2 diabetes [9]. Sample STATA code to implement these steps is provided in the Appendix.

## **Fundamental Concepts**

### ***What Are Instrumental Variables and Why Use IV for CER?***

The use of IV methods in health research has been growing rapidly. Garabedian and colleagues performed a systematic search for comparative effectiveness studies relying on IV [10]. They found 187 studies published between 1992 and 2011, with the frequency of publication increasing rapidly—from fewer than two per year before 1998 to 34 in 2011 alone [10].

The increasing popularity of IV among comparative effectiveness researchers is leading to intensifying debate in the literature about the strengths and weaknesses of the approach, with different authors reaching seemingly conflicting conclusions. For example, Garabedian and colleagues conclude, “Although no observational method can completely eliminate confounding, we recommend against treating instrumental variable analysis as a solution to the inherent biases in observational CER studies” [10]. In contrast, Glymour and colleagues conclude, “Given that it will often be nearly free to conduct IV analyses with secondary data, they may prove extremely valuable in many research areas . . . [however if IV] is uncritically adopted into the epidemiologic toolbox, without aggressive evaluations of the validity of the design in each case, it may generate a host of false or misleading findings” [11].

To understand how instrumental variables methods work, it is helpful to start by returning again to why causal inference is valid in a randomized clinical trial. As illustrated in Figure 1A, participants in an RCT are randomly assigned between treatment and control groups. Because this sorting is accomplished by a mechanism (flip of a coin) that is uncorrelated with any patient or provider characteristics, we expect the mean values of all these variables (whether observed or not) to be the same in both groups. Furthermore, because the coin flip has no direct effect on the outcome, any mean difference observed at the end of the trial must be due to treatment itself [8].

Causal inference in observational studies is more complex, as illustrated in Figure 1B. Sorting into treatment and comparison groups is not determined by one, random factor; instead, numerous

---

<sup>1</sup> Quasi-experimental methods include interrupted time series, regression discontinuity, instrumental variables, and many other designs. This article focuses on IV, although the principles apply to all of these designs.

patient and provider characteristics, both observed and unobserved, can play a role. Many of these variables may also directly affect the outcome, resulting in potential confounding (illustrated by the dotted lines in the Figure). For example, sicker patients are more likely to choose more aggressive treatments, leading unadjusted comparisons to suggest that aggressive treatments are associated with poor outcomes [8].

The standard method of reducing this confounding is to try to control for individual characteristics that might affect outcome risk, using a regression model to statistically adjust for between group differences in risk factors [12]. Propensity score matching is a variant on this approach, whereby propensity scores are calculated using a long list of variables (including interactions and transformations) that might be related to the outcome [13,14,15]. Members of the treatment group are matched by propensity score with members of the comparison group through a process ensuring that observable characteristics are balanced between groups.

Unfortunately, neither standard risk adjustment nor propensity score matching can ensure that *unobserved* patient and provider characteristics will be balanced or adjusted for in the analysis. In Figure 1B, one such unobservable confounder is level of self-care skill. Patients with more skills may seek more aggressive treatment, having more confidence that they will be able to manage any additional complexity that may be involved. Because such patients are likely to have better outcomes than those with less well developed skills, failing to adjust for unobserved skill differences could lead to an erroneous finding of a beneficial treatment effect.

An IV approach can potentially solve this problem. Imagine a situation where the flip of a coin does not exclusively determine assignment to treatment like it does in an RCT, but it has a strong influence. An IV model statistically isolates the component of variation in treatment that can be traced back to the coin flip and then examines differences in outcomes that are due to that component alone, separated from observed and unobserved potential confounders [8]. It is like finding a little RCT inside a lot of observational data.

Of course, coin flips like this are rarely found in real data, so the researcher must find another variable (an instrument) that has the experimental properties of the coin flip: it must be strongly related to sorting into treatment, and it must not be related to the outcome, except through its effect on treatment.<sup>2</sup> The first property (instrument strength) is illustrated in Figure 1B by the solid arrow connecting the IV to sorting. The second property, known as the exclusion restriction, is illustrated in the Figure by the lack of any arrow connecting the IV directly to the outcome. In CER, a promising and frequently used IV is geographic, facility-level, or provider-level practice pattern differences. In Garabedian's review, fully 46% of identified IV studies featured this type of instrument. Practice pattern instruments can be easily constructed and applied to an enormous variety of CER questions, so it is vital to be able to evaluate the validity of this approach.

Instrument strength is straightforward to test [17]. If the instrument is not strongly enough related to sorting into treatment, IV estimates will be highly imprecise and can be biased [18]. The exclusion restriction is more difficult to test and is often left to theoretical argument and subject matter expertise [10,16,19,20]. Naturally, this reduces confidence in IV methods [10]. The contribution of falsification tests is that they help evaluate the validity of the exclusion restriction, thereby identifying cases where the instrument is confounded and strengthening confidence in cases where no evidence of confounding is revealed.

IV models in CER are implemented and tested by translating the diagram in Figure 1B into two equations for estimation. The first explains variation in treatment as a function of patient characteristics,

---

<sup>2</sup> Some authors make a distinction among non-treatment pathways through which a potential instrument might be associated with the outcome. The instrument might have a direct effect, or it might be partly caused by another variable that also affects the outcome (e.g., [16]). For our purposes, we do not need to make this distinction.

provider characteristics, instrumental variables, and unobserved factors (denoted by  $u$ ). The second explains variation in outcomes as a function of patient characteristics, provider characteristics, receipt of treatment, and unobserved factors (denoted by  $v$ ), some of which might be the same as in the first equation.

$$(1) \text{ Treatment} = f(\text{patient characteristics, provider characteristics, IV}) + u$$

$$(2) \text{ Outcome} = g(\text{patient characteristics, provider characteristics, Treatment}) + v$$

These equations can be estimated simultaneously or sequentially, but naively estimating the outcome equation (Equation 2) without accounting for the treatment equation (Equation 1) will lead to bias if there are unobservable factors that influence both treatment and outcomes. For example, if the unobserved confounder is the patient's self-care skill as mentioned above, naïve estimation of Equation (2) will falsely attribute some of the effect of self-care skill to the treatment. A practice pattern based IV model could solve this problem by isolating for analysis the component of treatment variation that is due to practice patterns and eliminating the component that is due to individual characteristics like self-care skill.

### ***How Can IV Go Wrong?***

In addition to the problem of weak instruments, IV estimates can be biased or misleading because the exclusion restriction is invalid or because the IV estimates are not generalizable to the population of interest. If the exclusion restriction is invalid, the IV is correlated with the outcome through some pathway other than treatment. For example, if practice patterns for the treatment in question are related to diffusion of new knowledge, receipt of the treatment may be correlated with receipt of other services that are sensitive to new knowledge and also have effects on the outcome. In this case, the IV estimate would falsely attribute some of the beneficial effects of other treatment improvements to the treatment under study.

IV estimates can be misleading even if the instrument is strong and the exclusion restriction is valid. This can occur because IV estimates measure outcome differences that can be attributed to treatment variations caused by the instrument. If the instrument only affects a small sub-population, the IV estimates may not be generalizable to a larger population. In other words, the IV estimate measures a local average treatment effect (LATE) [21,22]. This issue is analogous to the external validity problem faced by RCTs [23].

### ***How Does a Falsification Test Help?***

The idea of falsification testing dates back at least to Popper [24], but has been the subject of more attention recently in health outcomes research because of the increasing opportunities for observational studies discussed above [11,16,25]. In IV CER studies, a falsification test of the exclusion restriction will typically involve identifying a dependent variable or a population that ought not to be affected by the treatment under study, but would be affected by potential confounders that might be correlated with the proposed IV and the outcome.

For example, consider a study comparing stroke outcomes among patients receiving alternative anticoagulation therapies for atrial fibrillation. Garabedian and colleagues (2014)[10] argue that practice pattern IV studies are often vulnerable to bias because they fail to control for one or more of the following patient characteristics: race, education, income, age, insurance status, health status, and health behaviors. If health behaviors are correlated with anticoagulant prescribing patterns and the outcomes under study, this could indeed be a problem. However, patients without atrial fibrillation but who have carotid artery disease are also at elevated risk for stroke and should not be treated with anticoagulants. If anticoagulant prescribing patterns are unrelated to stroke outcomes for carotid disease patients, then it is less likely that confounding health behaviors are correlated with anticoagulant prescribing patterns. Instead of using an alternative population (those with carotid disease), another option would be to choose an alternative outcome that should not be affected by the treatment but would be affected by health behaviors (e.g., incident lung cancer).

More formally, an ideal falsification test for the exclusion restriction would estimate an alternative specification for Equation (2) that excludes treatment but includes the practice pattern IV.

$$(3) \text{ Outcome} = g(\text{patient characteristics, provider characteristics, IV}) + v$$

This equation is estimated for an alternative population or an alternative outcome, selected to be as close as possible to the outcomes and populations of interest without being subject to the treatment under study. If the IV has no significant estimated effect on the outcome in Equation (3) then the exclusion restriction is not rejected. Note that multiple tests are possible for the same application, so prespecification is valuable to avoid selective reporting [25].

### **Conducting and Testing a Real IV Analysis**

To make the above conceptual discussion more concrete, consider a recent analysis conducted by Prentice and colleagues [9]. The investigators set out to compare the effects on long-term outcomes of two classes of oral medications used as second-line agents to control type 2 diabetes: sulfonylureas (SU), like glyburide and glipizide, and thiazolidinediones (TZD), like rosiglitazone and pioglitazone. SUs are well-established, inexpensive and often used as first and second line agents in diabetes treatment [26,27]. SU use increases the risk for hypoglycemia and concerns about their potential association with cardiovascular disease have been present since the 1970s [28]. Several recent studies have reported an increased risk of cardiovascular disease and death among patients who started on an SU compared to metformin (MET) as initial treatment of diabetes [29,30]. TZDs have also been associated with adverse events, including cardiovascular outcomes (MI and CHF), osteoporosis and bladder cancer [31-33]. To compare the effectiveness and risks of these two medication classes, Prentice and colleagues applied a practice pattern IV technique to a large administrative database combining data elements from the Veterans Health Administration (VHA) and Medicare.

The outcomes chosen for study were readily computable from the administrative data and included all-cause mortality, hospital admission (VHA or Medicare) for any of 13 ambulatory care sensitive conditions (ACSC) as defined by the Agency for Healthcare Research and Quality [34,35] and AMI or stroke. The VA Vital Status File which determines the date of death from VA, Medicare, and Social Security Administration data was used to determine all-cause mortality [36]. ACSC hospitalizations are hypothesized to be preventable with high quality outpatient care and include several diabetes and cardiovascular complications such as uncontrolled diabetes, short and long-term complications of diabetes, or congestive heart failure [34,35]. AMI definitions were based on Petersen et al. (1999) and Kyota et al. (2004) and stroke definitions were based on Reker et al. (2002) [37-39]. Due to the overall scarcity of the stroke and AMI outcomes in the data, models that predicted these outcomes separately were unstable. Consequently, AMI and stroke were combined into one outcome. The modeled outcome was the amount of time between the initiation date of SU or TZD and the earliest date of any of the three outcomes, censoring on the date an individual started a third drug or the end of the study period.

#### ***Step One: Choose and Specify IV***

When considering a quasi-experimental design, it is vital to identify a source of variation in treatment that is arbitrary or random with respect to potentially confounding variables. This source of variation could be a policy change or boundary (as in interrupted time series or regression discontinuity) or it could be practice variation or program location<sup>3</sup> (as in many IV studies). In Figure 1B the source of arbitrary or random variation in treatment that is only related to the outcome through its effect on treatment is labeled the instrumental variable. The choice and specification of the IV should be determined through consideration of institutional factors and the causal diagram in Figure 1B.

---

<sup>3</sup> Falsification tests can also be useful when program location is used as an instrument. See Edwards et al (2014[46]) for a recent example.

The VHA is the largest integrated health care system in the United States serving over 8.3 million patients each year and spending nearly 4 billion dollars on prescriptions in 2009 [40,41]. There is significant physician-prescribing practice variation [42,43] and VHA patients are assigned to their primary care physicians by variable and often arbitrary methods [44,45]. Consequently, provider-level prescribing variation is unlikely to be related to the observable or unobservable patient characteristics shown in Figure 1B and identified by Garabedian and colleagues. This is a promising start for a potential instrument.

Prentice and colleagues defined treatment as initiating either SU or TZD as a second hypoglycemic agent after experience with metformin, noting that most patients who initiated one or the other remained on it two years later [Prentice et al]. They defined their instrument as the proportion of second line agent prescriptions (SU or TZD) written for SU by each provider (for all of their patients) during the year prior to the patient's initiation date for their second line agent [9]. Providers and patients were paired based on that initiation date to minimize confounding that could occur if patients later switched providers. If a provider had < 10 patient-level second line agent prescriptions during the prior year (70% of the time), the rate at the community based outpatient clinic (CBOC) or VHA medical center (VAMC) where the provider practiced was used.

To check whether this instrument was random with respect to patient characteristics, Prentice and colleagues performed a simple falsification test by comparing sample means between SU and TZD initiators (columns 1 and 2 of Table 1), and then between those paired with high vs. low SU prescribers (columns 3 and 4 of Table 1). Although there were some notable differences by initiation group—for example SU initiators were more likely to have baseline HbA<sub>1c</sub> > 9—these differences were no longer evident when patients were grouped by provider prescribing pattern (Table 1), indicating that the first falsification test did not reject the proposed instrument.

#### ***Step Two: Choose and Specify Control Variables***

Once an IV has been chosen, consider other potential confounders that might be correlated with the IV as well as the outcome. In the practice pattern example, patient characteristics are not expected to be correlated with the IV for institutional reasons and Table 1 demonstrates that this appears to be true in the data. In contrast, as shown in Figure 1B, provider and facility characteristics like the quality of care delivered might be correlated with practice patterns and might also affect the outcome. If possible, this danger can be mitigated by including provider and facility quality measures as control variables in the outcome equation. Although they are less likely to be confounders, it is a good idea to include patient characteristics as control variables as well because they will improve the precision of estimates.

Prentice and colleagues specified three process quality measures to control for potentially confounding provider and facility characteristics: percent of HbA<sub>1c</sub> labs > 9% [47,48], percent of blood pressure readings >140/90 mm Hg [49], and percent of LDL cholesterol labs > 100 mg/dL [49]. These variables were computed at the same provider, CBOC or VAMC level and time periods as the IV prescribing rate. Sample means for these variables are shown in Table 1, which also demonstrates that the IV appears to balance these factors as well.

#### ***Step Three: Choose Falsification Sample and Outcomes***

Once the IV and control variables have been specified it is tempting to proceed with the study, but a little more advance planning is essential to support falsification testing. If the instrument is valid it should affect the outcome only through treatment. Therefore, it should have no effect on outcomes that are not in the treatment pathway. Such outcomes could be the result of unrelated disease processes affecting the study population or they could be study outcomes experienced by those not subject to the study treatment. In either case, investigators will usually have to specify the necessary data when the study protocol is approved. An ideal falsification sample would not be exposed to the study treatment, but would be exposed to all of the potential confounders that might be correlated with the instrument and the outcome, like provider- or facility-level quality of care.

In the diabetes study, Prentice and colleagues specified two populations for falsification testing that were closely related to the study population but not subject to treatment by SU or TZD [9]. First, they selected all individuals who received a new prescription of MET and followed them for one year. They assumed these patients were being treated with MET as their first line agent and their disease had not progressed to the point of needing a second line agent in that time period. Consequently, the SU prescribing rate should not affect the outcomes for these individuals. They used provider SU prescribing rates to predict all-cause mortality, ACSC hospitalization, and stroke or AMI controlling for all the demographics, comorbidities and process quality variables. Since no individuals in this population were on SU, no treatment equation was estimated and the falsification test was performed by including the instrument in an alternative specification of the outcome equation.

Using the same analyses, the second falsification test used a sample of individuals who initiated insulin after MET and took no other diabetes drugs during the study period. Again, the conceptual model indicated that SU prescribing rates should not affect the outcomes for these individuals if there were no important instrument-outcome confounders. An appealing feature of this pair of falsification tests is that the falsification populations bracket the study population in terms of disease severity, with MET only patients the least severe and insulin patients the most severe. If the falsification tests support the exclusion restriction, it is difficult to imagine why it would fail only among those with moderate disease.

#### ***Step Four: Estimate IV Model***

Linear IV models can be estimated easily in most statistical packages, but health outcomes of interest are often more appropriately estimated by nonlinear methods like logistic regression or survival models. Nonlinear IV models can also be estimated, but methods often involve specialized programming, making implementation more difficult. Two-stage residual inclusion is a widely applicable and easily implemented approach that does not involve specialized programming beyond the use of standard commands in a statistical package like STATA [8,50]. The first stage treatment equation (Equation 1) is estimated by logistic or probit regression, and the first stage residual is calculated as  $\hat{u} = \text{Treatment} - \hat{f}(\text{patient characteristics, provider characteristics, IV})$ , where  $\hat{f}(\cdot)$  is the estimated function  $f(\cdot)$  and gives the predicted probability of treatment. The second stage outcome equation (Equation 4) is estimated next, after including the estimated residual,  $\hat{u}$ , as a covariate. This additional variable controls for possible correlation between unobservable factors affecting treatment ( $u$ ) and unobservable factors affecting the outcome ( $v$ ).

(4) Outcome =  $g(\text{patient characteristics, provider characteristics, Treatment, } \hat{u}) + v$   
The first-stage residual term,  $\hat{u}$ , is an estimated quantity, but statistical software will not automatically account for the increased uncertainty that implies, so standard errors for estimates from Equation (4) must be recalculated by bootstrapping [51].

In the diabetes study, Prentice and colleagues used a probit model to estimate their treatment equation and Cox models including the first-stage residual to estimate their outcome equations [9]. The strength of their practice pattern IV is demonstrated by the size and precision of its estimated effect in the treatment equation (Table 2). The IV estimates of treatment effects are expressed as hazard ratios in Table 3, indicating that SU prescribing significantly increased the risk of mortality and ACSC hospitalization relative to TZD prescribing, but did not have a significant effect on the risk of stroke or heart attack. Since SUs are widely used and considered safe while TZDs are used less frequently and typically considered more risky, these are surprising and potentially important results.

#### ***Step Five: Compute Falsification Test***

The falsification tests specified above can be computed by estimating Equation (3) with either the falsification sample and the study outcomes or with the study sample and the falsification outcomes. The exclusion restriction is rejected if the IV in Equation (3) has a statistically significant effect on the outcome. No bootstrapping is necessary because none of the covariates in Equation (3) are estimated.

Although presenting multiple falsification tests is better than only one, it is not possible to prove conclusively that there is no confounding. As with other aspects of analytic design, specification of falsification tests at the proposal stage of a project helps to allay concerns that investigators might be presenting only the results that support their design.

In the diabetes study, Prentice and colleagues found no significant effects of the IV on any of the outcomes in either falsification sample (Table 4). These results are consistent with validity of the IV and improve confidence in the IV estimates, but it is always possible that a different test specification or a larger sample could detect a problem.

It is also possible that an instrument that is not rejected for one population will be rejected for another, closely related population. Bartel, Chan and Kim (2014)[52] use day of the week admitted to the hospital as an instrument for length of stay when measuring the effect of length of stay on rehospitalization and other outcomes for patients with heart failure. For institutional or personal preference reasons, patients admitted on Monday or Tuesday tend to have shorter lengths of stay than those admitted on Thursday or Friday (who are more likely to stay over the weekend). Bartel, Chan and Kim tabulate patient characteristics by their instrument to try to falsify the assumption that admission day is uncorrelated with observed and unobserved health status and they find that the instrument is not rejected for patients with the most severe disease, but it is rejected for less severe cases. This makes sense because severe cases might have to respond to symptoms immediately, making admission day effectively random, but less severe cases might choose their admission day with a desired length of stay in mind. The investigators appropriately proceed to use the instrument only for the population supported by the falsification test [52].

## **Conclusion**

Falsification testing is a fundamental scientific tool that is particularly useful when considering an instrumental variables approach to an observational study. With proper advance planning, falsification tests can be easily applied to potential instruments, with the results either rejecting the instruments or increasing confidence in them. Causal inference from an instrumental variables observational study will never be as strong as it could be from a well-executed randomized clinical trial, but, if testing supports the strength and validity of the instruments, these studies can shed light on a multitude of important clinical questions that would otherwise be too confounded to investigate with other observational study designs.

## **References**

1. Krumholz HM, Selby JV. Seeing through the eyes of patients: the Patient-Centered Outcomes Research Institute Funding Announcements. *Ann Intern Med* 2012 Sep 18;157(6):446-7.
2. Hsiao CJ, Jha AK, King J, et al. Office-based physicians are responding to incentives and assistance by adopting and using electronic health records. *Health Aff (Millwood)* 2013 Aug;32(8):1470-7.
3. Selby JV, Fleurence R, Lauer M, et al. Reviewing hypothetical migraine studies using funding criteria from the Patient-Centered Outcomes Research Institute. *Health Aff* 2012;31(10):2193-2199.
4. Krumholz HM. Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. *Health Aff* 2014;33(7):1163-70.

5. National Research Council. 2013. *Frontiers in Massive Data Analysis*. Washington DC: The National Academies Press.
6. PCORI (Patient-Centered Outcomes Research Institute) Methodology Committee. 2013. *The PCORI Methodology Report*. Available at: <http://www.pcori.org/research-we-support/research-methodology-standards>. Accessed on 12-27-2013.
7. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, eds. *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. AHRQ Publication No. 12(13)-EHC099. Rockville, MD: Agency for Healthcare Research and Quality; January 2013.
8. Pizer SD. An intuitive review of methods for observational studies of comparative effectiveness. *Health Serv Outcomes Res Method* 2009;9:54–68.
9. Prentice JC, Conlin PR, Gellad W, et al. Capitalizing on Prescribing Pattern Variation to Compare Medications for Type 2 Diabetes. *Value in Health*. In Press.
10. Garabedian LF, Chu P, Toh S, et al. Potential bias of instrumental variable analyses for observational comparative effectiveness research. *Ann Intern Med* 2014;161:131-138.
11. Glymour MM, Tchetgen EJ, Robins JM. Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions. *Am J Epidemiol* 2012;175(4):332–9.
12. Iezzoni LI. Risk adjustment for measuring healthcare outcomes. *Health Administration Pr*, 1997.
13. Rubin, DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statist Med* 2007;26:20-36.
14. D'Agostino RB. Tutorial in biostatistics propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statist Med* 1998;17:2265-81.
15. Garrido MM, Kelley AS, Paris J, et al. Methods for Constructing and Assessing Propensity Scores. *Health Serv Res* 2014 Oct;49(5):1701-20.
16. Swanson SA, Hernán MA. How to Report Instrumental Variable Analyses (Suggestions Welcome). *Epidemiology* May 2013;24(3).
17. Stock JH, Yogo M. Testing for Weak Instruments in Linear IV Regression. In: *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenber*. Chapter 5 pages 80-108. Cambridge University Press; 2005.
18. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J Am Statistical Assoc* 1995; 90 443-50.
19. Rassen JA, Brookhart MA, Glynn RJ, et al. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J Clin Epidemiol* 2009 December;62(12):1226–32.

20. Grootendorst P. A review of instrumental variables estimation of treatment effects in the applied health sciences. *Health Serv Outcomes Res Method* 2007;7:159-79.
21. Imbens GW, Angrist JD. Identification and Estimation of Local Average Treatment Effects. *Econometrica* Mar 1994;62(2):467-75.
22. Harris KM, Remler DK. Who is the marginal patient? Understanding instrumental variables estimates of treatment effect. *Health Serv Res* December 1998 part 1;33(5):1337-60.
23. Imbens GW. Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *J Economic Literature* June 2010;48: 399–423.
24. Popper KR. *The Logic of Scientific Discovery*, 1934, English translation 1959. New York: Basic Books, Inc.
25. Prasad V, Jena AB. Prespecified Falsification End Points Can They Validate True Observational Associations? *JAMA* January 16, 2013;309(3):241-2.
26. Bogner K, Bell KF, Lakdawalla D, et al. Clinical outcomes associated with rates of sulfonylurea use among physicians. *Am J Manag Care* 2013;19:16-221.
27. Alexander GC, Sehgal NL, Moloney RM, Stafford RS. National trends in treatment of type 2 diabetes mellitus, 1994-2007. *Arc Intern Med* 2008;168:2088-94.
28. Groop LC. Sulfonylureas in NIDDM. *Diabetes Care* 1992;15:737-54.
29. Roumie CL, Hung AM, Greevy RA, et al. Comparative effectiveness of sulfonylurea and metformin monotherapy on cardiovascular events in Type 2 diabetes mellitus: a cohort study. *Ann Intern Med* 2012;157:601-10.
30. O'Riordan M. Sulfonylurea Use Increases All-Cause Mortality Risk. European Association for the Study of Diabetes (EASD) 2013 Meeting; 2013; Barcelona, Spain; 2013:25
31. Bennett WL, Maruthur NM, Singh S, et al. Comparative effectiveness and safety of medications for type 2 diabetes: An update including new drugs and 2-drug combinations. *Ann Intern Med* 2011;154:602-13.
32. Hsiao FY, Hsieh PH, Huang WF, et al. Risk of bladder cancer in diabetic patients treated with rosiglitazone or pioglitazone: A nested case-control study. *Drug Safety* 2013;36:643-9.
33. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *N Engl J Med* 2007;356:2457-71.
34. AHRQ Quality Indicators-Guide to Prevention Quality Indicators: Hospital Admission for Ambulatory Care Sensitive Conditions. In: Agency for Health Care Research and Quality, editor. Rockville, MD; 2001.

35. Agency for Health Care Research and Quality (AHRQ). 2013. Prevention Quality Indicators Technical Specifications– Version 4.5, May 2013. Available at: [http://www.qualityindicators.ahrq.gov/Modules/PQI\\_TechSpec.aspx](http://www.qualityindicators.ahrq.gov/Modules/PQI_TechSpec.aspx). [Accessed August 1, 2013].
36. Arnold N, Sohn MW, Maynard C, et al. VA-NDI Mortality Merge Project. In VIREC Technical Report 2, edited by VA Information Resource Center, Hines, IL: VA Information Resource Center, 2006.
37. Petersen LA, Wright S, Normand SLT, et al. Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database. *J Gen Intern Med* 1999;14:555-8.
38. Kyota Y, Schneeweiss S, Glynn RJ, et al. Accuracy of medicare claims-based diagnosis of acute myocardial infarction: estimating positive-predict value on the basis of review of hospital records. *Am Heart J* 2004;148:99-104.
39. Reker DM, Rosen AK, Hoenig H, et al. The hazards of stroke case selection using administrative data *Med Care* 2002;40:96-104.
40. Veterans Health Administration. About VHA. Available from: <http://www.va.gov/health/aboutvha.asp/>. [Accessed July 15, 2014].
41. United States General Accounting Office. Drug Review Process is Standardized at the National Level, but Actions Are Needed to Ensure Timely Adjudication of Nonformulary Drug Requests. 2010. GAO 10-776. Available from: <http://www.gao.gov/assets/310/308933.pdf>. [Accessed July 31, 2014].
42. Gellad WF, Good CB, Lowe JC, et al. Variation in prescription use and spending for lipid-lowering and diabetes medications in the VA healthcare system. *Am J Manag Care* 2010;16:741-50.
43. Gellad WF, Mor M, Zhao X, et al. Variation in Use of High-Cost Diabetes Medications in the VA Healthcare System. *Arc Intern Med* 2012;172:1608-11.
44. 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.
45. Prentice JC, Conlin PR, Gellad WF, et al. Long term outcomes of analogue insulin compared to NPH for patients with Type 2 diabetes. *Am J Manag Care*. In press.
46. Edwards ST, Prentice JC, Simon SR, et al. Home-Based Primary Care and Risk of Preventable Hospitalization in Older Veterans with Diabetes. *JAMA Internal Medicine* (In Press).
47. Action to Control Cardiovascular Risk in Diabetes Study Group. Effects of intensive glucose lowering in type 2 diabetes. *N Engl J Med* 2008;358:2545-59.
48. Turner RC, Holman RR, Cull CA. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* 1998;352:837-53.
49. The State of Health Care Quality 2011. Continuous Improvement and the Expansion of Quality Measurement. National Committee for Quality Assurance, editor. Washington DC; 2011.

50. Terza JV, Basu A, Rathouz PJ. Two-stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling. *J Health Economics* 2008;27:531-43.

51. Efron B. Bootstrap methods: another look at the jackknife. *Ann Stat* 1970;7(1):1-26.

52. Bartel AP, Chan CW, Kim SH. Should Hospitals Keep Their Patients Longer? The Role of Inpatient and Outpatient Care in Reducing Readmissions. September 2014. NBER Working Paper# 20499. Cambridge, MA: National Bureau of Economic Research.

## Appendix: Annotated STATA code

### Data Organization and Variable Definitions

Start by defining a baseline period, an index date, and an outcome period. The index date separates the baseline period from the outcome period and indicates when the patient received either the study treatment or a comparator. Data in the baseline period should be organized to have each observation represent a patient-event (either a prescription, a lab value or an outcome). Data in the outcome period should be organized with each observation representing a patient. The following variables are used in the code below.

**Trxx:** Indicator variable equal to 1 if baseline prescription event is study treatment and zero otherwise.

**Allrx:** Indicator variable equal to 1 if baseline event is a prescription and zero otherwise.

**provider:** Unique provider ID.

**facility:** Unique facility ID.

**year:** Year of index date.

**treatment:** Indicator variable equal to 1 if index treatment is study treatment and zero otherwise.

**outcomedays:** Number of days from index date to first outcome.

**censored:** Number of days from index date to censoring.

### Step One: Choose and Specify IV

In practice pattern IV applications, the investigator often will be computing a provider- or facility-level mean in the baseline period, excluding the particular patient whose record is being processed. Thus, the patient's value or values are excluded from both the numerator and denominator of the rate used to predict his or her treatment. This can be done efficiently by calculating the overall numerator and overall denominator for all records in the baseline data and then subtracting the patient-specific values on each line before combining to form the rate. The rate is then saved as a patient-level variable and added to the outcome data.

```
egen numer1 = sum(Trxx), by(provider)
egen denom1 = sum(Allrx), by(provider)
numer2 = numer1 - Trxx
denom2 = denom1 - Allrx
IVrate = numer2/denom2
```

### Step Two: Choose and Specify Control Variables

Control variables typically include standard demographics, risk-adjustment variables (based on diagnosis codes), and baseline medications and lab values if available. If provider-level process quality variables are included, they can be constructed excluding the individual patient using the same coding technique as Step One.

### Step Three: Choose Falsification Sample and Outcomes

The falsification sample should ideally include the same control variables as the study sample. Falsification outcomes should be as close as possible to study outcomes without being affected by the study treatment.

### Step Four: Estimate IV Model

The IV model is estimated on the outcome period data, using two equations bootstrapped together. The first is the treatment equation, which can be estimated by logistic regression or probit if treatment is binary. Results of this regression are used to calculate the predicted residual, which is

included in the outcome equation. The following treatment equation example includes fixed effects for facilities and years:

```
xi: probit treatment IVrate {control variables} i.facility i.year  
predict Txprob  
Txres = treatment – Txprob
```

The following outcome equation example estimates a Cox proportional hazards model including fixed effects for years and random effects for facilities:

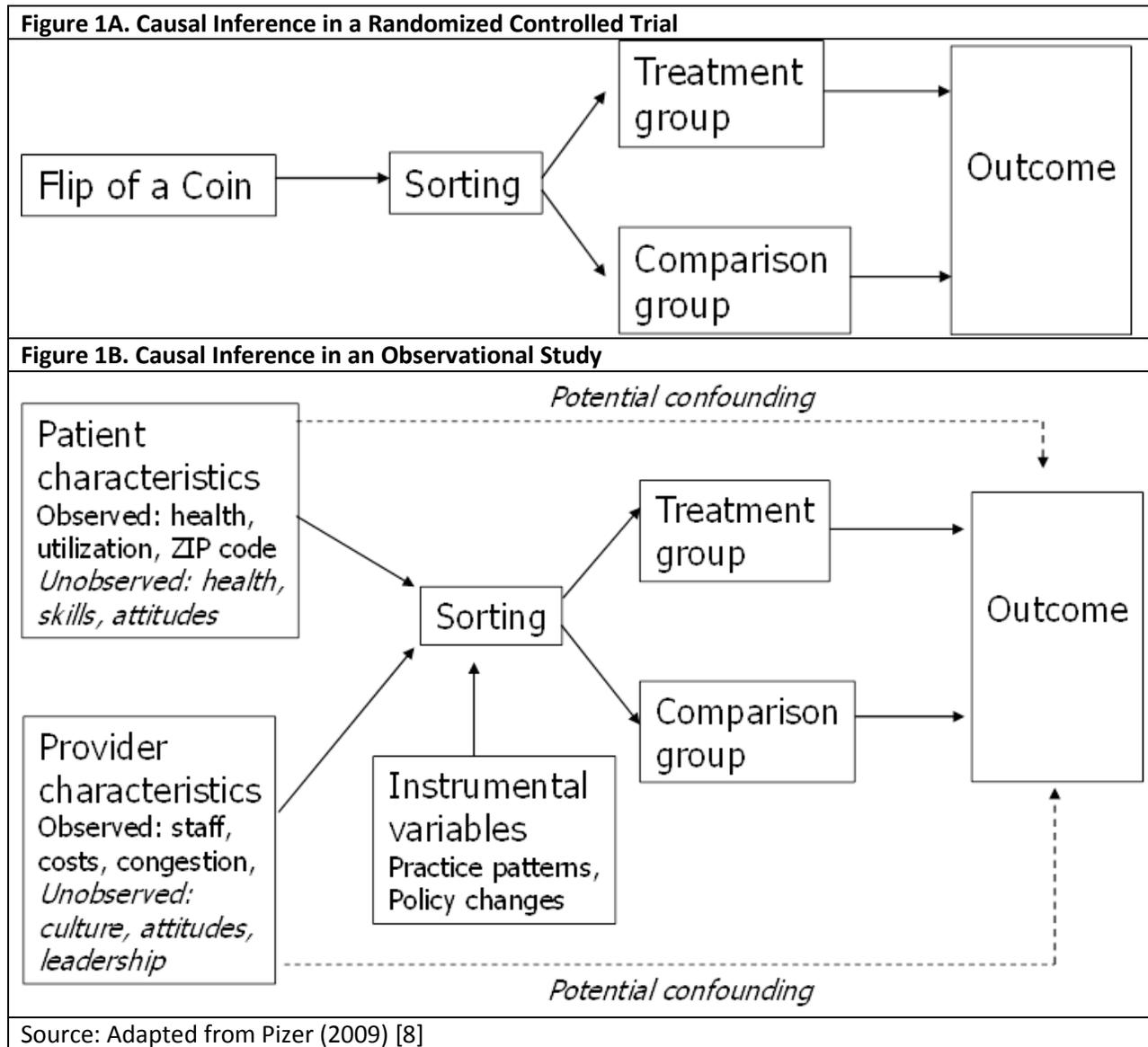
```
stset outcomedays, failure(censored)  
xi: stcox treatment Txres {control variables} i.year, shared(facility)
```

#### **Step Five: Compute Falsification Test**

The falsification test is computed using an alternative formulation of the outcome equation. In this example, a falsification sample is used with the study outcome. If the IVrate variable has a statistically significant effect on the outcome, the instrument is rejected.

```
stset outcomedays, failure(censored)  
xi: stcox IVrate {control variables} i.year, shared(facility)
```

Figures and Tables



**Table 1: Selected sample means or percentages for patients starting SU or TZD as second agent and patients assigned to above and below-median SU prescribing providers**

	Individual Treatment		Provider SU prescribing	
	Start SU n=73726	Start TZD n=7210	Top 50% SU <sup>a</sup> n=40483	Bottom 50% SU <sup>a</sup> n=40453
<b>Demographics</b>				
Age (y), mean	69.1 <sup>b</sup>	70.1	69.2	69.2
Male	98	98	98	98
White	88	89	90	87
<b>Diabetes management</b>				
HbA <sub>1c</sub> >=9	9	5	8	8
Retinopathy complications	14	16	14	14
Nephropathy complications	10	12	10	10
Neuropathy complications	19	22	20	19
Cerebrovascular complications	13	14	13	13
Cardiovascular complications (some)	24	28	25	25
Cardiovascular complications (severe)	26	23	25	25
Peripheral vascular complications	14	16	14	14
Metabolic complications	1	1	1	1
<b>Cardiovascular comorbidities</b>				
BMI obese	41	39	41	41
Congestive heart failure	13	12	13	13
Cardiac arrhythmias	21	21	21	21
Valvular disease	10	11	9	10
Hypertension	84	84	84	84
Pulmonary circulatory disorder	1	1	1	1
Chronic pulmonary disease	23	21	24	23
<b>Provider Process Quality Variables</b>				
Provider % HbA <sub>1c</sub> > 9 in baseline period, mean	10	10	10	10
Provider BP % > 140 or >90 in baseline period, mean	41	42	41	41
Provider LDL % > 100 in baseline period, mean	38	40	38	38
<b>Outcomes</b>				
ACSC hospitalization	18	13	18	17
All-cause mortality	10	7	10	9
Stroke or AMI	5	4	5	5

<sup>a</sup> These two columns show descriptive statistics of patients assigned to providers who prescribe SU below and above the sample median.

<sup>b</sup> For ease of presentation, percentages are presented unless otherwise noted.

Excerpted from Prentice et al. (In Press) [9] Table 2.

**Table 2. Selected First-stage Probit Results: Receiving SU Compared to TZD (n=80,936)**

	Coefficient	P< t	95% Confidence Interval	
<b>Instrument</b>				
Provider prescribing history	2.215	0.000	2.098	2.332

Model also includes baseline demographics, Elixhauser comorbidities, Young severity index, HbA<sub>1c</sub>, BMI, microalbumin, serum creatinine, provider quality controls, Veterans Affairs Medical Center fixed effects and year effects that are not shown.

Excerpted from Prentice et al (In Press) [9] Table 3

**Table 3. Second Stage Cox Proportional Hazard Models: Effect of SU on Mortality, ACSC hospitalization and Cardiovascular Outcomes (n=80,936)**

	Hazard Ratio	P< t	95% Confidence Interval	
All-cause mortality	1.50	0.014	1.09	2.09
ACSC hospitalization	1.68	<0.001	1.31	2.15
Stroke or heart attack	1.15	0.457	0.80	1.66

Models include baseline demographics, Elixhauser comorbidities, Young severity index, HbA<sub>1c</sub>, BMI, microalbumin, serum creatinine, provider quality controls, year fixed effects and Veterans Affairs Medical Center random effects.

Excerpted from Prentice et al (In Press) [9] Table 4.

**Table 4. Falsification Test: Effect of SU Prescribing Rate on Mortality, ACSC hospitalization and Cardiovascular Outcomes**

	Hazard Ratio	P< t	95% Confidence Interval	
<b><i>MET only sample (n=76,860)</i></b>				
All-cause mortality	1.30	0.115	0.94	1.79
ACSC hospitalization	1.23	0.149	0.93	1.62
Stroke or heart attack	1.11	0.657	0.70	1.77
<b><i>MET and Insulin sample (n=4,015)</i></b>				
Mortality	1.30	0.427	0.68	2.52
ACSC hospitalization <sup>a</sup>	0.81	0.425	0.47	1.37

Models include baseline demographics, Elixhauser comorbidities, Young severity index, HbA<sub>1c</sub>, BMI, microalbumin, serum creatinine, provider quality controls, year fixed effects and Veterans Affairs Medical Center random effects.

<sup>a</sup>The stroke and heart attack model did not converge in the MET and insulin sample due to small sample sizes.

Excerpted from Prentice et al (In Press) [9] Table 5.