HSR Health Services Research

DOI: 10.1111/1475-6773.12355
METHODS CORNER

Falsification Testing of Instrumental Variables Methods for Comparative Effectiveness Research

Steven D. Pizer

Objectives. To demonstrate how falsification tests can be used to evaluate instrumental variables methods applicable to 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. Comparative effectiveness research, instrumental variables, falsification testing, big data, causal inference

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 it 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 of the explanatory variables by treatment and control group included in most

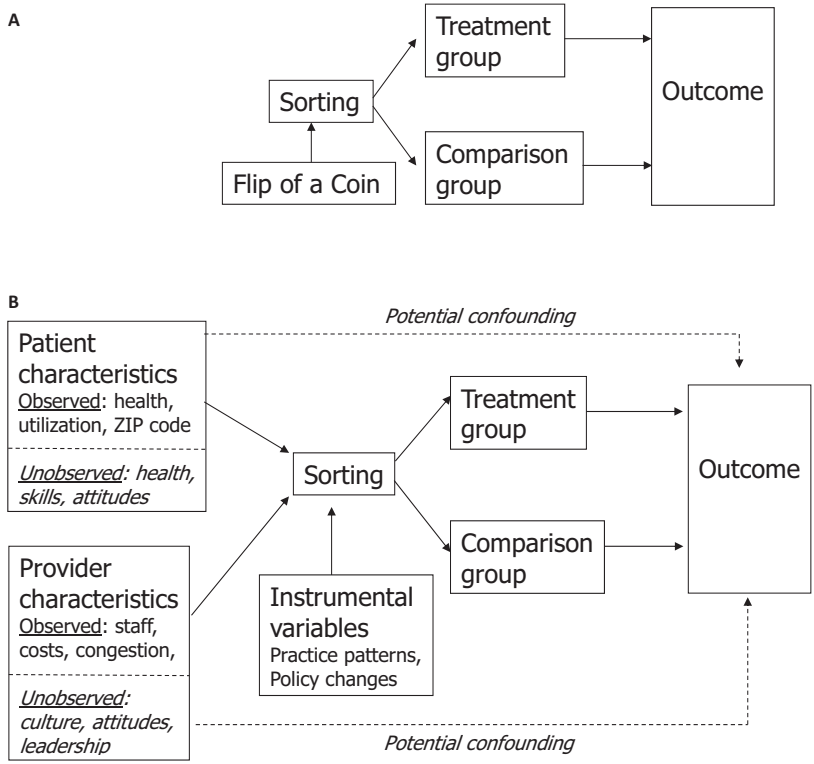
reporting of RCT results. What purpose does this serve? Our expectation is that the sample means will not be significantly different between groups because group assignment was intended to be random. Random assignment is the "identifying assumption" of RCTs because randomization permits us to infer causal 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 RCT's 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) (Krumholz and Selby 2012) 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 percent in 2010 to 38.2 percent in 2012 (Hsiao et al. 2013a). 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 (Selby et al. 2012; Krumholz 2014). RCTs 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 up for many years.

Address correspondence to Steven D. Pizer, Ph.D., 150 South Huntington Avenue (Mail Stop 152H), Boston, MA 02130; e-mails: steven.pizer@va.gov; s.pizer@neu.edu.

Figure 1: (A) Causal Inference in a Randomized Controlled Trial. (B) Causal Inference in an Observational Study



Source: Adapted from Pizer (2009).

Despite these advantages, a key challenge facing observational CER is evaluating the validity of causal inference made with observational data (National Research Council 2013; PCORI Methodology Committee 2013; Velentgas et al. 2013). This is a topic that has generated controversy for hundreds of years (Dowd 2011). Recently, however, 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 (Pizer 2009).[1]

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 (Prentice et al. 2014). Sample STATA code to implement these steps is provided in the Appendix SA1.

FUNDAMENTAL CONCEPTS

What Is IV?

To understand how instrumental variables methods work, it is helpful to start by returning again to why causal inference is valid in an RCT. As illustrated in Figure 1A, participants in an RCT are assigned randomly 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 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 (Pizer 2009).

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 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 (Pizer 2009).

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 (Iezzoni 1997). Propensity score matching or weighting 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 (D'Agostino 1998; Rubin 2007; Garrido et al. 2014). Members of the treatment group are typically 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 methods 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 potentially can 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 on part of the population. 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 (Pizer 2009). This 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.[2] 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.

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, 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.

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

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

These equations can be estimated simultaneously or sequentially, but naively estimating the outcome equation (2) without accounting for the treatment 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. An IV model could solve this problem by isolating for analysis the component of treatment variation that is due to the instrument and eliminating the component that is due to individual characteristics like self-care skill.[3]

Should IV Be Used for CER?

The use of IV methods in health research has been growing rapidly. Garabedian et al. (2014) performed a systematic search for comparative effectiveness studies relying on IV. 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 (Garabedian et al. 2014). They also found that geographic, facility-level, or provider-level practice pattern differences were used as the IV in fully 46 percent of these studies. 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.

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 et al. (2014) 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." In contrast, Glymour, Tchetgen, and Robins (2012) 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." Although these authors seem to be pointing in different directions, they agree that IV methods can go badly wrong. To assess the potential of IV for CER, it is clearly necessary to understand the potential pitfalls.

How Can IV Go Wrong?

There are two principal threats to the internal validity of IV estimates. First, if the instrument is not strongly enough related to sorting into treatment, IV estimates will be highly imprecise and can be biased, a problem known as "weak instruments" (Bound, Jaeger, and Baker 1995). Second, IV estimates can be biased or misleading because the exclusion restriction is invalid. The exclusion restriction is violated if 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 subpopulation, the IV estimates may not be generalizable to a larger population. In other words, the IV estimate measures a local average treatment effect (LATE) (Imbens and Angrist 1994; Harris and Remler 1998).[4] This issue is analogous to the external validity problem faced by RCTs (Imbens and Angrist 1994).

How Can IV Be Tested?

Instrument strength is directly observable in the treatment equation, so testing is straightforward (Stock and Yogo 2005). In contrast, the exclusion restriction is impossible to prove empirically and is often left to theoretical argument and subject matter expertise (Grootendorst 2007; Rassen et al. 2009; Swanson and Hernán 2013; Garabedian et al. 2014). Naturally, this reduces confidence in IV methods (Swanson and Hernán 2013). Falsification tests can be particularly useful for IV because 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.

The idea of falsification testing dates back at least to Popper (1959), but it has been the subject of more attention recently in health outcomes research because of the increasing opportunities for observational studies discussed above (Glymour, Tchetgen, and Robins 2012; Prasad and Jena 2013; Garabedian et al. 2014). Although falsification tests in general can take many forms, there are two particularly useful strategies for testing the exclusion restriction in IV CER studies: (1) investigating an alternative outcome 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 (2) investigating an alternative population that again ought not to be affected by the treatment but would be affected by potential confounders.

For example, consider a study comparing stroke outcomes among patients receiving alternative anticoagulation therapies for atrial fibrillation (depicted in panel A of Figure 2). Garabedian et al. (2014) argued 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. For example, 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 (panel B of Figure 2). 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.

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

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

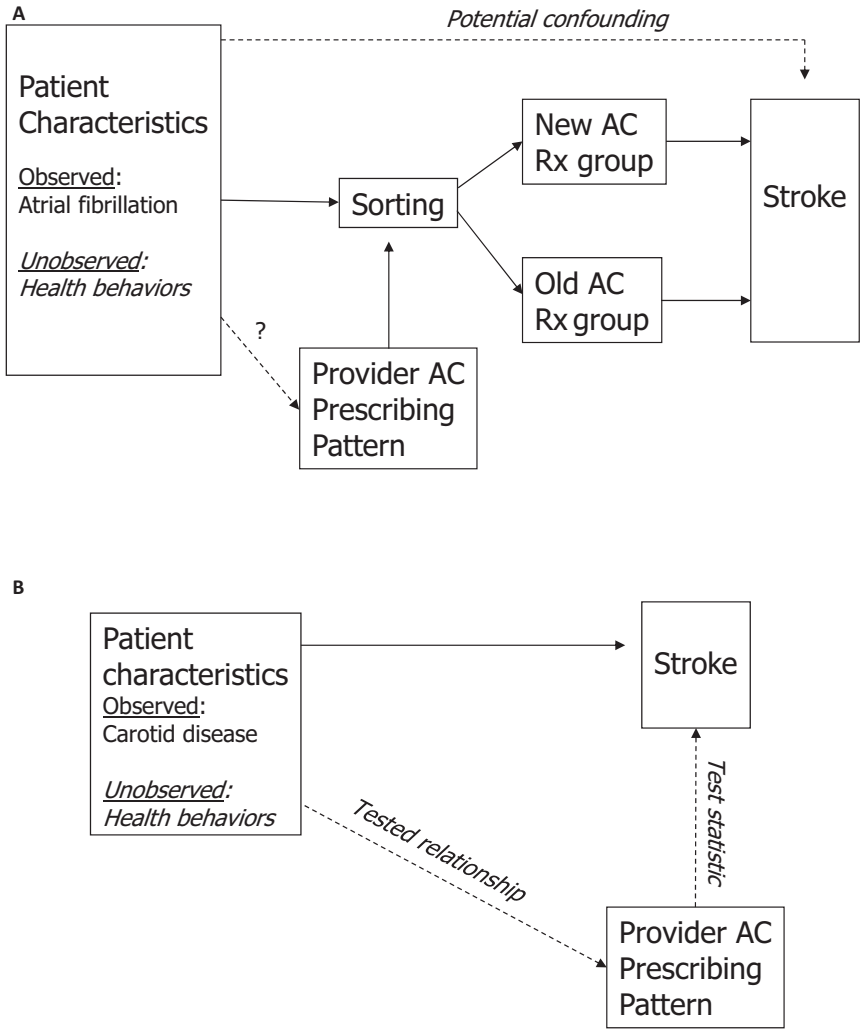
application, so prespecification is valuable to avoid selective reporting (Prasad and Jena 2013).

Conducting and Testing a Real IV Analysis

To make the above conceptual discussion more concrete, consider a recent analysis conducted by Prentice et al. (2014). 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 (Alexander et al. 2008; Bognar et al. 2013). SU use increases the risk for hypoglycemia, and concerns about their potential association with cardiovascular disease have been present since the 1970s (Groop 1992). 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 an initial treatment of diabetes (Roumie et al. 2012; O'Riordan 2013). TZDs have also been associated with adverse events, including cardiovascular outcomes (MI and CHF), osteoporosis, and bladder cancer (Nissen and Wolski 2007; Bennett et al. 2011; Hsiao et al. 2013b). 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 (AHRQ 2001, 2013) 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 (Arnold et al. 2006). 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 (AHRQ 2001, 2013). AMI definitions were based on Petersen et al. (1999) and Kyota et al. (2004), and stroke definitions were based on Reker et al. (2002). Due to the overall scarcity of the stroke and AMI outcomes in the data, models that predicted these outcomes separately

Figure 2: (A) Study of Alternative Anticoagulation Therapies for Patients with Atrial Fibrillation. (B) Falsification Test Using Population with Carotid Disease



AC, anticoagulant. Provider characteristics omitted for simplicity.

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, censor-

ing 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[5] (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.

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 (US GAO 2010; VHA 2014). There is significant physician-prescribing practice variation (Gellad et al. 2010, 2012) and VHA patients are assigned to their primary care physicians by variable and often arbitrary methods (Doyle, Ewer, and Wagner 2010; Prentice et al. 2015). 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 2 years later (Prentice et al. 2014). 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 (Prentice et al. 2014). 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 percent 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 versus 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 HbA1c >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.[6]

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 instrument confounders, it is a good idea to include patient characteristics as control variables as well because they will improve the precision of estimates (if they are related to the outcome).

Prentice and colleagues specified three process quality measures to control for potentially confounding provider and facility characteristics: percent of HbA_{1c} labs >9 percent (ACCORD 2008; Turner, Holman, and Cull 1998), percent of blood pressure readings >140/90 mmHg (The State of Health Care Quality 2011), and percent of LDL cholesterol labs >100 mg/dL (NCQA 2011). 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.[7]

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

Table 1: Selected Sample Means or Percentages for Patients Starting Sulfonylureas (SU) or Thiazolidinediones (TZD) as Second Agent and Patients Assigned to Above- and Below-Median SU-Prescribing Providers

	Individual Treatment		*Provider SU Prescribing*	
Demographics	*Start SU (n = 73,726)*	*Start TZD (n = 7,210)*	*Top 50% SU* (n = 40,483)*	*Bottom 50% SU* (n = 40,453)*
Age (years), mean	69.1†	70.1	69.2	69.2
Male	98	98	98	98
White	88	89	90	87
Diabetes management				
$HbA_{1c} \geq 9$	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
Cardiovascular comorbidities				
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
Provider process quality variables				
Provider % $HbA_{1c} > 9$ 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
Outcomes				
ACSC hospitalization	18	13	18	17
All-cause mortality	10	7	10	9
Stroke or AMI	5	4	5	5

Notes. *These two columns show descriptive statistics of patients assigned to providers who prescribe SU below and above the sample median.

†For ease of presentation, percentages are presented unless otherwise noted.

Source. Excerpted from Prentice et al. (2014, table 2).

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 submitted for funding consideration and human subjects protection review. An ideal falsification sample would not be exposed to the study treatment, but it 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 (Prentice et al. 2014). First, they selected all individuals who received a new prescription of MET and followed them for 1 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. As 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 (Terza, Basu, and Rathouz 2008; Pizer 2009). The first-stage treatment equation (1) is estimated by logistic or probit regression, and the first-stage residual is calculated as uhat = Treatment − fhat (patient characteristics, provider characteristics, IV), where fhat(.) is the estimated function f(.) and gives the predicted probability of treatment. The second-stage outcome equation (4) is estimated next, after including the estimated residual, uhat, as a covariate along with the original treatment variable. This additional variable controls for possible correlation between unobservable factors affecting treatment (u) and unobservable factors affecting the outcome (v).[8]

$$\text{Outcome} = \text{g}(\text{patient characteristics, provider characteristics, Treatment, uhat}) + v \quad (4)$$

The first-stage residual term, uhat, 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 (Efron 1970).

In the diabetes study, Prentice et al. (2014) used a probit model to estimate their treatment equation and Cox models including the first-stage residual to estimate their outcome equations. 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. As 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

Table 2: Selected First-Stage Probit Results: Receiving Sulfonylureas (SU) Compared to Thiazolidinediones (TZD) (n = 80,936)

	Coefficient	$P < \|t\|$	*95% Confidence Interval*
Instrument			
Provider prescribing history	2.215	0.000	2.098–2.332

Notes. Model also includes baseline demographics, Elixhauser comorbidities, Young severity index, HbA_{1c}, BMI, microalbumin, serum creatinine, provider quality controls, Veterans Affairs Medical Center fixed effects, and year effects that are not shown.
Source. Excerpted from Prentice et al. (2014, table 3).

Table 3: Second-Stage Cox Proportional Hazard Models: Effect of Sulfonylureas (SU) on Mortality, Ambulatory Care–Sensitive Condition (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

Notes. Models include baseline demographics, Elixhauser comorbidities, Young severity index, HbA_{1c}, BMI, microalbumin, serum creatinine, provider quality controls, year fixed effects, and Veterans Affairs Medical Center random effects.
Source. Excerpted from Prentice et al. (2014, table 4).

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) 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

Table 4: Falsification Test: Effect of Sulfonylureas (SU) Prescribing Rate on Mortality, Ambulatory Care–Sensitive Condition (ACSC) Hospitalization, and Cardiovascular Outcomes

	Hazard Ratio	*P < |t|*	*95% Confidence Interval*
MET only sample (n = 76,860)			
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
MET and insulin sample (n = 4,015)			
Mortality	1.30	0.427	0.68–2.52
ACSC hospitalization*	0.81	0.425	0.47–1.37

Notes. Models include baseline demographics, Elixhauser comorbidities, Young severity index, HbA_{1c}, BMI, microalbumin, serum creatinine, provider quality controls, year fixed effects, and Veterans Affairs Medical Center random effects.
*The stroke and heart attack model did not converge in the MET and insulin sample due to small sample sizes.
Source. Excerpted from Prentice et al. (2014, table 5).

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 (Bartel, Chan, and Kim 2014).[9]

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.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: I am grateful to Austin Frakt and Julia Prentice for comments on early drafts of this paper. I thank the Health Services Research & Development Service of the U.S. Department of Veterans Affairs for financial support (Grant no. IIR 10-136). This work does not reflect the official positions of the U.S. Department of Veterans Affairs or Northeastern University. I have no other disclosures to make.

Disclosures: None.

Disclaimers: None.

NOTES

1. Quasi-experimental methods include instrumental variables, interrupted time series, regression discontinuity, sample selection, and many other designs. This article focuses on IV, although the principles apply to all of these designs.
2. Some authors make a distinction among nontreatment pathways through which a potential instrument might be associated with the outcome (thus violating the exclusion restriction). The instrument might have a direct effect, or it might be partly caused by another variable that also affects the outcome (e.g., Swanson and Hernán 2013). For this article I do not need to make this distinction.
3. Other quasi-experimental designs can be thought of as IV models. For example, natural experiments, interrupted time series, and regression discontinuity designs use policy changes or discontinuities as instruments. These are strong designs because the reasons for variation in the instruments are well understood and there are good conceptual reasons to believe the exclusion restriction is valid.
4. Technically, generalizability may be limited because of the combination of heterogeneous treatment effects across individuals and instruments that have influence on treatment decisions in a limited part of the population. Bias can also be introduced if the size of the treatment effect is correlated with the instrument.
5. Falsification tests can also be useful when program location is used as an instrument. See Edwards et al. (2014) for a recent example.
6. Another way of thinking about this test is that the demonstrated balance in patient characteristics is important because these variables serve as proxies for unobservable potential confounders.
7. It is not necessary that the IV balance these factors if they are controlled, but showing balance can increase confidence in the assumption that there are no unobserved and uncontrolled provider or facility characteristics that are strongly correlated with treatment and outcome.
8. The estimated coefficient on the residual term is a test statistic for the presence of unobserved confounders in the outcome equation if estimated alone (Davidson and MacKinnon 1989).

9. This example illustrates how an instrument can be valid only for part of a population, which is separate, but related to the fact that an instrument can have varying degrees of influence on treatment (LATE). Consequently, different instruments can have distinct patterns of influence and validity across the population.

REFERENCES

Action to Control Cardiovascular Risk in Diabetes Study (ACCORD) Group. 2008. "Effects of Intensive Glucose Lowering in Type 2 Diabetes." *New England Journal of Medicine* 358: 2545–59.

AHRQ. 2001. *AHRQ Quality Indicators–Guide to Prevention Quality Indicators: Hospital Admission for Ambulatory Care Sensitive Conditions.* AHRQ Pub. No. 02-R0203. Rockville, MD: Agency for Healthcare Research and Quality.

Agency for Health Care Research and Quality [AHRQ]. 2013. "Prevention Quality Indicators Technical Specifications–Version 4.5" [accessed on August 1, 2013]. Available at http://www.qualityindicators.ahrq.gov/Modules/PQI_TechSpec.aspx

Alexander, G. C., N. L. Sehgal, R. M. Moloney, and R. S. Stafford. 2008. "National Trends in Treatment of Type 2 Diabetes Mellitus, 1994-2007." *Archives of Internal Medicine* 168: 2088–94.

Arnold, N., M. W. Sohn, C. Maynard, and D. M. Hynes. 2006. *VIReC Technical Report 2: VA Mortality Data Merge Project.* Hines, IL: VA Information Resource Center.

Bartel, A. P., C. W. Chan, and S. H. Kim. 2014. *Should Hospitals Keep Their Patients Longer? The Role of Inpatient and Outpatient Care in Reducing Readmissions.* NBER Working Paper# 20499. Cambridge, MA: National Bureau of Economic Research.

Bennett, W. L., N. M. Maruthur, S. Singh, J. B. Segal, L. M. Wilson, R. Chatterjee, S. S. Marinopoulos, M. A. Puhan, P. Ranasinghe, L. Block, W. K. Nicholson, S. Hutfless, E. B. Bass, and S. Bolen. 2011. "Comparative Effectiveness and Safety of Medications for Type 2 Diabetes: An Update Including New Drugs and 2-Drug Combinations." *Annals of Internal Medicine* 154: 602–13.

Bognar, K., K. F. Bell, D. Lakdawalla, A. Shrestha, J. T. Snider, N. Thomas, and D. Goldman. 2013. "Clinical Outcomes Associated with Rates of Sulfonylurea Use among Physicians." *American Journal of Managed Care* 19: 16–221.

Bound, J., D. A. Jaeger, and R. M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of American Statistical Association* 90: 443–50.

D'Agostino, R. B. 1998. "Tutorial in Biostatistics Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group." *Statistics in Medicine* 17: 2265–81.

Davidson, R., and J. G. MacKinnon. 1989. "Testing for Consistency Using Artificial Regressions." *Econometric Theory* 5 (3): 363–84.

Dowd, B. E. 2011. "Separated at Birth: Statisticians, Social Scientists, and Causality in Health Services Research." *Health Services Research* 46 (2): 397–420.

Doyle, J. J., S. M. Ewer, and T. H. Wagner. 2010. "Returns to Physician Human Capital: Evidence from Patients Randomized to Physician Teams." *Journal of Health Economics* 29: 866–82.

Edwards, S. T., J. C. Prentice, S. R. Simon, and S. D. Pizer. 2014. "Home-Based Primary Care and Risk of Preventable Hospitalization in Older Veterans with Diabetes." *Journal of the American Medical Association Internal Medicine* 174 (11): 1796–803.

Efron, B. 1970. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7 (1): 1–26.

Garabedian, L. F., P. Chu, S. Toh, A. M. Zaslavsky, and S. B. Soumerai. 2014. "Potential Bias of Instrumental Variable Analyses for Observational Comparative Effectiveness Research." *Annals of Internal Medicine* 161: 131–8.

Garrido, M. M., A. S. Kelley, J. Paris, K. Roza, D. E. Meier, R. S. Morrison, and M. D. Aldridgeet. 2014. "Methods for Constructing and Assessing Propensity Scores." *Health Services Research* 49 (5): 1701–20.

Gellad, W. F., C. B. Good, J. C. Lowe, and J. M. Donohue. 2010. "Variation in Prescription Use and Spending for Lipid-Lowering and Diabetes Medications in the VA Healthcare System." *American Journal of Managed Care* 16: 741–50.

Gellad, W. F., M. Mor, X. Zhao, J. Donohue, and C. Good. 2012. "Variation in Use of High-Cost Diabetes Medications in the VA Healthcare System." *Archives of Internal Medicine* 172: 1608–11.

Glymour, M. M., E. J. Tchetgen, and J. M. Robins. 2012. "Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions." *American Journal of Epidemiology* 175 (4): 332–9.

Groop, L. C. 1992. "Sulfonylureas in NIDDM." *Diabetes Care* 15: 737–54.

Grootendorst, P. 2007. "A Review of Instrumental Variables Estimation of Treatment Effects in the Applied Health Sciences." *Health Services Outcomes Research Methodology* 7: 159–79.

Harris, K. M., and D. K. Remler. 1998. "Who is the Marginal Patient? Understanding Instrumental Variables Estimates of Treatment Effect." *Health Services Research* 33 (5 Pt 1): 1337–60.

Hsiao, C. J., A. K. Jha, J. King, V. Patel, M. F. Furukawa, and F. Mostashari. 2013a. "Office-Based Physicians Are Responding to Incentives and Assistance by Adopting and Using Electronic Health Records." *Health Affairs (Millwood)* 32 (8): 1470–7.

Hsiao, F. Y., P. H. Hsieh, W. F. Huang, Y. W. Tsai, and C. S. Gau. 2013b. "Risk of Bladder Cancer in Diabetic Patients Treated with Rosiglitazone or Pioglitazone: A Nested Case-Control Study." *Drug Safety* 36: 643–9.

Iezzoni, L. I. 1997. *Risk Adjustment for Measuring Healthcare Outcomes.* Chicago, IL: Health Administration Press.

Imbens, G. W., and J. D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Krumholz, H. M. 2014. "Big Data and New Knowledge in Medicine: The Thinking, Training, and Tools Needed For a Learning Health System." *Health Affairs* 33 (7): 1163–70.

Krumholz, H. M., and J. V. Selby. 2012. "Seeing through the Eyes of Patients: The Patient-Centered Outcomes Research Institute Funding Announcements." *Annals of Internal Medicine* 157 (6): 446–7.

Kyota, Y., S. Schneeweiss, R. J. Glynn, C. C. Cannuscio, J. Avorn, and D. H. Solomon. 2004. "Accuracy of Medicare Claims-Based Diagnosis of Acute Myocardial Infarction: Estimating Positive-Preidct Value on the Basis of Review of Hospital Records." *American Heart Journal* 148: 99–104.

National Research Council. 2013. *Frontiers in Massive Data Analysis.* Washington, DC: The National Academies Press.

NCQA. 2011. *The State of Health Care Quality 2011.* Washington, DC: National Committee for Quality Assurance.

Nissen, S. E., and K. Wolski. 2007. "Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes." *New England Journal of Medicine* 356: 2457–71.

O'Riordan, M. 2013. "Sulfonylurea Use Increases All-Cause Mortality Risk." European Association for the Study of Diabetes (EASD) 2013 Meeting. Barcelona, Spain; 201325

Patient-Centered Outcomes Research Institute (PCORI) Methodology Committee. 2013. "The PCORI Methodology Report" [accessed on December 27, 2013]. Available at http://www.pcori.org/research-we-support/research-methodology-standards

Petersen, L. A., S. Wright, S. L. T. Normand, and J. Daley. 1999. "Positive Predictive Value of the Diagnosis of Acute Myocardial Infarction in an Administrative Database." *Journal of General Internal Medicine* 14: 555–8.

Pizer, S. D. 2009. "An Intuitive Review of Methods for Observational Studies of Comparative Effectiveness." *Health Service Outcomes Research, Methodology* 9: 54–68.

Popper, K. R. 1959. "*The Logic of Scientific Discovery, 1934.*" English translation 1959. New York: Basic Books, Inc.

Prasad, V., and A. B. Jena. 2013. "Prespecified Falsification End Points Can They Validate True Observational Associations?" *Journal of the American Medical Association* 309 (3): 241–2.

Prentice, J. C., P. R. Conlin, W. Gellad, D. Edelman, T. A. Lee, and S. D. Pizer. 2014. "Capitalizing on Prescribing Pattern Variation to Compare Medications for Type 2 Diabetes." *Value in Health* 17 (8): 854–62.

Prentice, J. C., P. R. Conlin, W. F. Gellad, D. Edelman, T. A. Lee, and S. D. Pizer. 2015. "Long Term Outcomes of Analogue Insulin Compared to NPH for Patients with Type 2 Diabetes." *American Journal of Managed Care* 21 (3): e235–44.

Rassen, J. A., M. A. Brookhart, R. J. Glynn, M. A. Mittleman, and S. Schneesweiss. 2009. "Instrumental Variables I: Instrumental Variables Exploit Natural Variation in Nonexperimental Data to Estimate Causal Relationships." *Journal of Clinical Epidemiology* 62 (12): 1226–32.

Reker, D. M., A. K. Rosen, H. Hoenig, D. R. Berlowitz, J. Laughlin, L. Anderson, C. R. Marshall, and M. Rittman. 2002. "The Hazards of Stroke Case Selection Using Administrative Data." *Medical Care* 40: 96–104.

Roumie, C. L., A. M. Hung, R. A. Greevy, C. G. Grijalva, X. Liu, H. J. Murff, T. A. Elasy, and M. R. Griffin. 2012. "Comparative Effectiveness of Sulfonylurea and Metformin Monotherapy on Cardiovascular Events in Type 2 Diabetes Mellitus: A Cohort Study." *Annals of Internal Medicine* 157: 601–10.

Rubin, D. B. 2007. "The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine* 2007 (26): 20–36.

Selby, J. V., R. Fleurence, M. Lauer, and S. Schneewiss. 2012. "Reviewing Hypothetical Migraine Studies Using Funding Criteria from the Patient-Centered Outcomes Research Institute." *Health Affairs* 31 (10): 2193–9.

Stock, J. H., and M. Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression, Chapter 5." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenber*, edited by D. W. K. Andrews, pp. 80–108. New York: Cambridge University Press.

Swanson, S. A., and M. A. Hernán. 2013. "How to Report Instrumental Variable Analyses (Suggestions Welcome)." *Epidemiology* 24 (3): 344–5.

Terza, J. V., A. Basu, and P. J. Rathouz. 2008. "Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling." *Journal of Health Economics* 27 (3): 531–43.

The State of Health Care Quality. 2011. *Continuous Improvement and the Expansion of Quality Measurement.* Washington, DC: National Committee for Quality Assurance (NCQA).

Turner, R. C., R. R. Holman, and C. A. Cull. 1998. "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* 352: 837–53.

United States General Accounting Office [US GAO]. 2010. "Drug Review Process is Standardized at the National Level, but Actions Are Needed to Ensure Timely Adjudication of Nonformulary Drug Requests." GAO 10-776 [accessed on July 31, 2014]. Available at http://www.gao.gov/assets/310/308933.pdf

Velentgas, P., N. A. Dreyer, P. Nourjah, S. R. Smith, and M. M. Torchia (eds.). 2013. *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.

Veterans Health Administration [VHA]. 2014. "About VHA" [accessed on July 15, 2014]. Available at http://www.va.gov/health/aboutvha.asp/

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Annotated STATA Code.