Published in final edited form as:

Med Care. 2019 February ; 57(2): 167–171. doi:10.1097/MLR.0000000000001040.

Falsification Tests for Instrumental Variable Designs With an Application to Tendency to Operate

Luke Keele, PhD*, Qingyuan Zhao, PhD†, Rachel R. Kelz, MD, MSc‡, and Dylan Small, PhD†

*Departments of Surgery, Center for Surgery and Health Economics, Hospital of the University of Pennsylvania, Philadelphia, PA.

†Departments of Statistics, University of Pennsylvania, Center for Surgery and Health Economics, Hospital of the University of Pennsylvania, Philadelphia, PA.

‡Department of Surgery, Center for Surgery and Health Economics, Hospital of the University of Pennsylvania, Philadelphia, PA.

Abstract

Background: Instrumental variable (IV) methods are becoming an increasingly important tool in health services research as they can provide consistent estimates of causal effects in the presence of unobserved confounding. However, investigators must provide justifications that the IV is independent with any unmeasured confounder and its effect on the outcome occurs only through receipt of the exposure. These assumptions, while plausible in some contexts, cannot be verified from the data.

Methods: Falsification tests can be applied to provide evidence for the key IV assumptions. A falsification test cannot prove the assumptions hold, but can provide decisive evidence when the assumption fails. We provide a general overview of falsification tests for IV designs. We highlight a falsification test that utilizes a subpopulation of the data where an overwhelming proportion of units are treated or untreated. If the IV assumptions hold, we should find the intention-to-treat effect is zero within these subpopulations.

Results: We demonstrate the usage of falsification tests for IV designs using an IV known as tendency to operate from health services research. We show that the falsification test provides no evidence against the IV assumptions in this application.

Keywords

surgery; instrumental variables; falsification tests

Reprints: Luke Keele, PhD, McCourt School of Public Policy, Georgetown University, 37th & O Street, NW, Washington, DC 20057. luke.keele@gmail.com.

Qingyuan Zhao, Ph.D., Department of Statistics, University of Pennsylvania, 400 Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104, qyzhao@wharton.upenn.edu

Rachel R. Kelz, Center for Surgery and Health Economics, Department of Surgery, Hospital of the University of Pennsylvania, 3400 Spruce St, Philadelphia, PA 19104, Rachel.Kelz@uphs.upenn.edu

Dylan Small, Professor, Department of Statistics, University of Pennsylvania, 400 Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104, dsmall@wharton.upenn.edu

The authors declare no conflict of interest.

The method of instrumental variables (IV), long used in economics, is now becoming more widely applied in biomedical applications. An IV is an encouragement for receipt of treatment, which may occur in randomized experiments and observational studies. IV methods are particularly valuable because, if certain assumptions hold, valid causal inferences can be drawn even in the presence of unobserved confounding.^{1,2} IVs can be especially useful in analyzing medical claims data due to the inability to measure risk severity, a prevalent confounder in epidemiological studies.

One commonly used instrument is a physician's treatment preferences.³ For example, in a study on the effect of emergency surgery on in-hospital mortality, Keele et al⁴ found that surgeons vary substantially in their preference for operative care or their tendency to operate (TTO). In their study, they measured TTO as the percentage of emergency cases where a surgeon chose to operate for a set of predefined emergency medical conditions. They found that TTO varied—at the surgeon level—from <5% to over 95%, with an average TTO of 59%. As the causal effect of surgery is potentially confounded by patient risk, Keele et al⁴ proposed to use TTO as an IV for whether patients received surgery.

For a variable, such as TTO, to be an IV, 3 key conditions must hold: (1) the IV must be associated with the treatment of interest; (2) there must be no unmeasured confounding between the IV and the outcome; and (3) the IV can only affect the outcome through the treatment—this is known as the exclusion restriction. Critically, IV assumptions (2) and (3) cannot be directly verified with from the data. Instead, investigators must use their best judgement to reason about the plausibility of these assumptions. Although a formal confirmatory test of these 2 assumptions is impossible, these 2 assumptions often have testable implications such that falsification tests are possible.^{5–7} That is, a falsification test can at least cast doubt on these assumptions.

In this paper, we present a practical guide to the application of falsification tests for 2 key IV assumptions. While falsification tests come in different forms, we focus on a falsification test that exploits the fact that certain subgroups in the data are almost uniformly exposed to the treatment. We focus on this type of falsification test, as we think it is the most widely applicable. We review both the logic of and statistical methods for this test. We then apply the test to data on the TTO IV to further show how to apply these tests in clinical investigations. We find that the falsification test does not cast doubt on the IV assumptions for TTO—an original result that further validates TTO as an instrument. Next, we outline notation and review the IV assumptions.

METHODS

Notations and IV Assumptions

Let D be a binary treatment, Y is the outcome, Z is the instrument, and $X = (X_1, \dots, X_k)$ is a set of baseline covariates, where $X_1 = S$ is a binary indicator that stratifies the sample into 2 subgroups. In the TTO application, D is whether a patient receives operative care, Y is in-hospital mortality, Z is a physicians' TTO, and X contains measures of patient and surgeon characteristics.

We assume we have observed n independent and identically distributed samples of (D, Y, Z, X) . Furthermore, for $z = 0, 1$ and $d = 0, 1$, let $Y(z, d)$ be the potential outcome of Y when an intervention sets $Z = z$ and $D = d$, and let $D(z)$ be the potential outcome of D when Z is set to z . The following set of equations define the relationship between the potential outcomes, $Y(z, d)$ and $D(z)$, and the observed values (Y, D, Z) :

$$\begin{aligned} D &= D(Z) = ZD(1) + (1 - Z)D(0), \\ Y &= Y(Z, D) = Y(Z, D(Z)) = ZY(1, D(1)) + (1 - Z)Y(0, D(0)). \end{aligned}$$

This notation implicitly assumes the stable unit treatment value assumption.⁸ In the context of TTO, stable unit treatment value assumption primarily rules out any interference among the patients, that is one patients' mortality is unaffected by other patients' exposure to operative care.

Next, the 3 "core" IV assumptions are as follows²:

- (1) Relevance: $E[D|X, Z = 1] \neq E[D|X, Z = 0]$.
- (2) Effective random assignment: Z is independent of $D(z)$ and $Y(z, d)$ given X for all $z = 0, 1$ and $d = 0, 1$. This means that Z is independent of unmeasured confounders conditional on covariates X .
- (3) Exclusion restriction: $Y(z, d) = Y(z', d)$ for any z, z' , and d . This means that the potential outcome $Y(z, d)$ does not depend on z , that is, the effect of the instrument on the outcome entirely goes through the exposure D . This assumption allows us to write $Y(0, d) = Y(1, d) = Y(d)$.

For TTO to be a valid instrument, it must satisfy these 3 assumptions. We briefly review the plausibility of each of these assumptions for TTO. See Keele et al⁴ for a detailed discussion of these assumptions in the context of TTO. First, patients who receive care from high TTO surgeons must be more likely to have an operation. Second, conditional on the baseline covariates, TTO must be effectively randomly assigned to patients. That is, the assignment of surgeons to patients should behave in an as if random manner. Recent work has proposed a series of diagnostic tests for this assumption.^{9–11} Finally, any effect TTO has on mortality must be solely due to having an operation. One chief concern is that TTO is proxy for surgical skill. This might manifest itself as confounding between the TTO and the outcome or if TTO is a measure of surgical skill, TTO may have a direct effect on the outcome. Besides these core assumptions, point identification of the causal effect requires additional assumptions such as a deterministic,¹ stochastic¹² monotonicity assumption, or a no-interaction assumption.^{13,14}

Among the 3 core assumptions, assumption (1) is immediately testable using observational data, but assumptions (2) and (3) are unverifiable because they involve counterfactuals that are not observable. Next, we focus on using falsification tests to provide evidence for or against IV assumptions (2) and (3).

A Falsification Test for IV Designs

In general, assumptions (2) and (3) cannot be proven using data, but they can sometimes be tested indirectly via falsification. Angrist and Krueger¹⁵ refer to such tests as instances of “refutability.” This type of falsification test arises from the fact that causal theories may do more than predict the presence of a causal effect; causal theories may also predict an absence of causal effects.^{16,17} In general, falsification tests do not prove an assumption holds, but they can provide decisive evidence that an assumption is likely invalid. Falsification tests for IVs cannot prove that assumptions (2) and (3) hold, but failing a falsification test shows that these assumptions are implausible and the proposed instrument should generally not be used for clinical investigations.

Next, we review the small literature on falsification tests for IV designs. Yang et al⁶ first proposed a logical basis for indirect testing of the key IV assumptions. The general idea is that, although assumptions (2) and (3) cannot be tested directly, one can introduce an additional assumption and test the conjunction of these 3 assumptions. If the conjunction is rejected, then we know at least 1 of the 3 assumptions must be false. Depending on the problem and the data structure, the additional assumption can be: (1) The IV does not affect the treatment exposure in 1 subgroup^{5,18}; (2) the treatment does not affect the outcome in 1 subgroup⁶; (3) there is an alternative outcome that is not affected by the treatment but would be affected by potential confounders.⁷

Next, we describe a falsification test that encompasses the first 2 approaches. We outline the logic of this test and present how to apply it with statistical models. We think this form of falsification test is most broadly applicable in clinical studies. In general, for this type of test, we require a subpopulation (indicated by $S = 1$) such that the subjects are almost certainly treated or conversely that none of the subjects are treated. In the TTO application, these would be subgroups where all or none of the patients received surgery. Formally, we describe this condition as:

- (4) Entirely treated/untreated subpopulation: $P(D = 1 | S = 1) = 1$ or 0.

Under this condition, it is straightforward to verify that:

Proposition 1 Under assumptions (2), (3), and (4), we have $Y \perp Z | X_{-1}, S = 1$ where $X_{-1} = (X_2, \dots, X_k)$ are all the covariates besides $S = X_1$.

Proof. Without loss of generality, we assume the subpopulation is entirely treated: $P(D = 1 | S = 1) = 1$. By using the 3 assumptions, the conditional density (probability mass) function of Y given $X_{-1}, S = 1$, and $Z = 1$ can be written as

$$\begin{aligned} P(Y | S = 1, X_{-1}, Z = 1) &= P(Y(1, D) | S = 1, X_{-1}, Z = 1) \\ &\stackrel{(4)}{=} P(Y(1, 1) | S = 1, X_{-1}, Z = 1) \\ &\stackrel{(2)}{=} P(Y(1, 1) | S = 1, X_{-1}) \\ &\stackrel{(3)}{=} P(Y(1) | S = 1, X_{-1}). \end{aligned}$$

By the same reasoning, we can show that $P(Y|S=1, X_{-1}, Z=0)$ is equal to the same quantity. Thus, Y and Z are conditionally independent given X_{-1} and $S=1$.

In general, this implies that we can empirically test the hypothesis $Y \perp Z|X_{-1}, S=1$ which should be true if assumptions (2), (3), and (4) hold. For example, when Y is binary, we can fit the following logistic regression model in the $S=1$ subgroup:

$$P(Y=1|S=1, Z, X) = \text{expit}(\alpha Z + \beta^T X). \quad (1)$$

Then we test whether the intention-to-treat (ITT) effect $\hat{\alpha}$ equal to 0 using standard methods. As such, the falsification test does not require any special statistical methods. For any subpopulation where all units are exposed to treatment, the investigator simply estimates the ITT effect of Z on Y controlling for X_{-1} . If we are able to reject the null hypothesis that $\alpha = 0$, we have evidence that assumptions (2) and (3) may not hold. That is, when $\hat{\alpha}$ is nonzero, we have evidence that the instrument has a direct effect on the outcome in subgroup where all patients are treated. Moreover, the confidence interval for α is indicative of the direction and magnitude of the bias of an IV analysis using the subpopulation $S=0$. Large estimates of α indicate larger amounts of bias due to violating assumptions (2) and (3).

Next, we review 3 key points on the implementation and interpretation of this falsification test. First, in many applications, including TTO, it may be unlikely that there are subgroups where every unit is treated. In the original TTO study, patients were included if they had 1 of 51 specific acute medical conditions. For none of these 51 subconditions do all patients receive surgery, but for 5 conditions, 90% of patients receive surgery. In such instances, we can apply an approximate falsification test within these subgroups. However, the results should be interpreted with greater caution as the small proportion of untreated units also contribute to the ITT effect.

Second, notice that we only need assumptions (2) and (3) to hold in the $S=1$ subpopulation in the proof of proposition. This means that the falsification test does not give direct evidence about whether assumptions (2) and (3) are true in the $S=0$ subpopulation. Still, it is often reasonable to expect that the 2 subpopulations share similar unmeasured confounders or direct causal pathways from the IV to the outcome, so the falsification test still provides indirect evidence for or against the IV assumptions.

Third, while the test is based on the statistical significance of α , the magnitude of α is also important because it is indicative of the bias of an IV analysis due to violations of assumptions (2) and (3). In fact, we might reject the hypothesis that $\alpha = 0$, but the point estimate for α is very small. This observation would suggest that the IV assumptions are violated but the bias in the IV analysis is small and generally insignificant.¹¹ In general, investigators should examine the magnitude of α and interpret smaller estimates as better evidence for the validity of the IV assumptions.

RESULTS: AN APPLICATION TO TTO

Next, we apply the falsification test to the TTO application. We use the same data from Keele et al.⁴ The data are based on all-payer hospital discharge data from New York, Florida, and Pennsylvania in 2012–2013 linked to the American Medical Association (AMA) Physician Masterfile to join data on patient claims with surgeons. The data include patient sociodemographic and clinical characteristics including a measure of patient frailty, an indicator for severe sepsis, and 31 indicators for comorbidities based on Elixhauser indices.¹⁹ The study population was restricted to patients admitted for inpatient emergency care, urgently, or through the emergency department with a diagnosis of an acute general surgical condition. All subjects were categorized by the nature of the surgical conditions including indicators for 51 specific conditions. Surgeons were excluded if they could not be identified in the AMA Masterfile, did not meet criteria for general surgery training, did not attend an allopathic program, did not train within the United States, or did not perform at least 5 operations for 1 of the 51 specific acute surgical conditions per year within the 2-year study timeframe. Using these inclusion criteria, the cohort included 4094 surgeons and 605,498 patients with a diagnosis for a condition where either nonoperative or operative management may be considered.

One motivation for performing a falsification test in this setting is that TTO may be correlated with surgical skill. That is, more skilled surgeons may choose operative care more often, and thus the exclusion restriction—assumption (3)—might be violated. Although Keele et al.⁴ found that TTO is only weakly correlated with measures of surgical skill, there may be other unmeasured aspects of surgical skill that cause TTO to have some direct effect on mortality. Here, we investigate this possibility using the falsification test outlined above. In 5 of the 51 conditions included in the study, 90% of patients receive surgery. Within each of these strata, we implemented the falsification test using a linear probability model, where X includes baseline covariates as well as hospital-fixed effects. As noted by Keele et al.,⁴ the exclusion restriction is unlikely to hold across hospital comparisons, thus we believe adding hospital-fixed effects is crucial for the validity of TTO as an IV. We used a linear probability model due to the large number of fixed effects. Logistic regression may be inconsistent with many fixed effects due to the incidental parameter problem.^{20,21} Finally, we include TTO in the model as a proportion. If we conduct the test Using TTO split into a binary variable, the results are similar.

We implemented the test for 2 different outcome measures: mortality and prolonged length of stay (PLOS). PLOS is measured using an indicator variable which equals one when the hospital and operation-specific length of stay is greater than the 75th percentile. For one of the conditions—Meckel's diverticulum—there are not enough data for the test using the mortality outcome.

The results of the falsification test in the 5 subgroups can be found in Table 1. For the mortality outcome, we pass the falsification test for each of the 4 conditions as all the P -values are well above standard thresholds of statistical significance. Moreover, in the 3 subgroups where > 95% are treated, the point estimates of risk difference are all very small. For example, the point estimate for the ITT effect in the appendix subgroup is quite small

($\hat{\alpha}$). The variation in the test results for the PLOS outcome is useful for understanding how to interpret falsification tests. For PLOS, 2 of the P -values are below 0.05, which would strictly speaking indicate a failure of the test. However, in both cases the estimated point estimates are very small. For these 2 subgroups, the estimated regression coefficients are 0.3 and 0.04—smaller in magnitude than the estimated coefficients in some of the other tests (such as bowel perforation). Thus, while we do not formally pass the test in these cases, any bias in the IV analysis should be inconsequential, as the estimated coefficients are quite small which indicates a very weak direct effect of the IV on the outcomes.

In conclusion, we find little indication of an ITT effect of TTO on either outcome in these 5 subgroups. While not proof, this provides indirect evidence that the IV assumptions might indeed hold or if violated any bias is inconsequential. That is, we find that TTO has little effect on the outcomes in subgroups where a high percentage of patients are treated.

DISCUSSION

In this paper, we outlined one form of falsification test for key IV assumptions. The test is conducted by estimating an ITT effect within subgroups in which all units are either treated or untreated and requires no special software to implement. The presence of such subgroups may not be present in all applications, but is a fairly common feature of many datasets.

In this paper, we focused on 1 form of falsification test for IV designs. Negative control methods are another type of falsification test that can be applied to IV designs.¹⁰ Here, investigators test whether the IV is associated with an outcome that is likely to be affected by the same confounders as the outcome of interest, but that outcome is unlikely to be directly affected by the exposure. Alternatively, one can seek to find a negative control population of patients that were not exposed to the treatment of interest. In general, falsification tests are a useful technique for making the case for a specific variable as an instrument. When possible falsification tests should be included in any IV analysis.

Finally, we applied the test to TTO and found additional evidence that TTO seems to contain little information about surgical skill. That is, we found that among the set of patients where > 90% receive surgery, we find no evidence of a direct effect of TTO on the 2 outcomes. We interpret this as evidence that we have not violated the exclusion restriction, as evidence of an association in this subpopulation would clearly indicate a direct effect of TTO on the outcomes. Although these results do not prove the IV assumptions hold, they do provide evidence in favor of TTO as a valid IV.

Acknowledgments

R.R.K. is funded by a grant from the National Institute on Aging, R01AG049757-01A1. The dataset used for this study was purchased with a grant from the Society of American Gastrointestinal and Endoscopic Surgeons.

The content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Although the AMA Physician Masterfile data are the source of the raw physician data, the tables and tabulations were prepared by the authors and do not reflect the work of the AMA. The Pennsylvania Health Cost Containment Council (PHC4) is an independent state agency responsible for addressing the problems of escalating health costs, ensuring the quality of health care, and increasing access to health care for all citizens. While PHC4 has provided data for this study, PHC4 specifically disclaims responsibility for any analyses, interpretations or conclusions.

REFERENCES

1. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996;91:444–455.
2. Baiocchi M, Cheng J, Small DS. Instrumental variable methods for causal inference. *Stat Med* 2014;33:2297–2340. [PubMed: 24599889]
3. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *Int J Biostat* 2007;3:14.
4. Keele LJ, Sharoky CE, Sellers MM, et al. An instrumental variables design for the effect of emergency general surgery. *Epidemiol Methods* 2018 [In press].
5. Glymour MM, Tchetgen Tchetgen EJ, Robins JM. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *Am J Epidemiol* 2012;175:332–339. [PubMed: 22247045]
6. Yang F, Zubizarreta J, Small DS, et al. Dissonant conclusions when testing the validity of an instrumental variable. *Am Stat* 2014;68: 253–263.
7. Pizer SD. Falsification testing of instrumental variables methods for comparative effectiveness research. *Health Serv Res* 2016;51:790–811. [PubMed: 26293167]
8. Rubin DB. Randomization analysis of experimental data: the Fisher randomization test comment. *J Am Stat Assoc* 1980;75:591–593.
9. Jackson JW, Swanson SA. Toward a clearer portrayal of confounding bias in instrumental variable applications. *Epidemiology* 2015;26:498. [PubMed: 25978796]
10. Davies NM, Thomas KH, Taylor AE, et al. How to compare instrumental variable and conventional regression analyses using negative controls and bias plots. *Int J Epidemiol* 2017;46:2067–2077. [PubMed: 28398582]
11. Zhao Q, Small DS. Graphical diagnosis of confounding bias in instrumental variables analysis. *Epidemiology* 2018;29:e29–e31. [PubMed: 29570473]
12. Small D, Tan Z, Ramsahi R, et al. Instrumental variable estimation with a stochastic monotonicity assumption. *Stat Sci* 2017;32:561–579.
13. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologists dream. *Epidemiology* 2006;17:360–372. [PubMed: 16755261]
14. Wang L, Tchetgen Tchetgen E. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *J R Stat Soc Ser B* 2018;80:531–550.
15. Angrist JD, Krueger AB. Empirical strategies in labor economics. In: Ashenfelter O, Card D, eds. *Handbook of Labor Economics* Amsterdam: Elsevier Science Publishers; 1999;3A:1277–1366.
16. Rosenbaum PR. *Observational Studies* Springer Series in Statistics, 2nd ed. New York, NY: Springer-Verlag; 2002.
17. Lipsitch M, Tchetgen ET, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 2010;21:383–388. [PubMed: 20335814]
18. Kang H, Kreuels B, Adjei O, et al. The causal effect of malaria on stunting: a Mendelian randomization and matching approach. *Int J Epidemiol* 2013;42:1390–1398. [PubMed: 23925429]
19. Elixhauser A, Steiner C, Harris DR, et al. Comorbidity measures for use with administrative data. *Med Care* 1998;36:8–27. [PubMed: 9431328]
20. Neyman J, Scott EL. Consistent estimates based on partially consistent observations. *Econometrica* 1948;16:1–32.
21. Lancaster T The incidental parameter problem since 1948. *J Econ* 2000;95:391–413.

TABLE 1.

A Falsification Test the Surgeon's TTO as an Instrumental Variable

	Intestinal Obstruction: Adhesions	Upper GI: Appendix	Upper GI: Meckel's	Intestinal Obstruction: Incarcerated Hernias	Upper GI: Bowel Perforation
% Treated	96.3	95.9	94.8	91.1	90.5
Mortality					
α	0.8	-0.2	-	-1.1	-0.7
95% CI	-1.6 to 3.2	-0.8 to 0.4	-	-3.1 to 0.9	-8.5 to 7.2
<i>P</i>	0.502	0.460	-	0.293	0.862
Sample size	3230	61,478	-	20,173	3924
PLOS					
α	0.3	0.04	-0.3	0.08	-0.2
95% CI	0.14–0.49	0.02–0.06	-1.2 to 0.7	-0.03 to 0.19	-0.51 to 0.07
<i>P</i>	0.000	0.000	0.550	0.149	0.129
Sample size	3313	61,486	346	20,188	3966

Here, we used 5 subgroups of patients who are almost always treated regardless of surgeon's TTO.

CI indicates confidence interval; GI, gastro-intestinal; PLOS, prolonged length of stay; TTO, tendency to operate.