Published in final edited form as:
J Clin Epidemiol. 2009 December ; 62(12): 1233–1241. doi:10.1016/j.jclinepi.2008.12.006.

Instrumental variables II: instrumental variable application—in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance

Jeremy A. Rassen[a,b,c,*], **M. Alan Brookhart**[a,b], **Robert J. Glynn**[a,b,d], **Murray A. Mittleman**[b,c,e], and **Sebastian Schneeweiss**[a,b,c]

[a]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham & Women's Hospital, Boston, MA, USA

[b]Harvard Medical School, Boston, MA, USA

[c]Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

[d]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

[e]Cardiovascular Epidemiology Research Group, Beth Israel Deaconess Medical Center, Boston, MA, USA

Abstract

Objective—An instrumental variable (IV) is an unconfounded proxy for a study exposure that can be used to estimate a causal effect in the presence of unmeasured confounding. To provide reliably consistent estimates of effect, IVs should be both valid and reasonably strong. Physician prescribing preference (PPP) is an IV that uses variation in doctors' prescribing to predict drug treatment. As reduction in covariate imbalance may suggest increased IV validity, we sought to examine the covariate balance and instrument strength in 25 formulations of the PPP IV in two cohort studies.

Study Design and Setting—We applied the PPP IV to assess antipsychotic medication (APM) use and subsequent death among two cohorts of elderly patients. We varied the measurement of PPP, plus performed cohort restriction and stratification. We modeled risk differences with two-stage least square regression. First-stage partial r^2 values characterized the strength of the instrument. The Mahalanobis distance summarized balance across multiple covariates.

Results—Partial r^2 ranged from 0.028 to 0.099. PPP generally alleviated imbalances in nonpsychiatry-related patient characteristics, and the overall imbalance was reduced by an average of 36% (±40%) over the two cohorts.

Conclusion—In our study setting, most of the 25 formulations of the PPP IV were strong IVs and resulted in a strong reduction of imbalance in many variations. The association between strength and imbalance was mixed.

Keywords

Pharmacoepidemiology; Antipsychotic agents; Instrumental variable; Mahalanobis distance; Partial *R*-squared; Confounding factor (epidemiology); Physician prescribing preference

* Corresponding author. Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham & Women's Hospital, 1620, Tremont Street, Suite 3030, Boston, MA 02120, USA., jrassen@post.harvard.edu (J.A. Rassen).

1. Introduction

Instrumental variable (IV) analysis joins other techniques [1–4] that attempt to mitigate the bias introduced by measured and unmeasured confounding present in nonexperimental data [5–10]. IVs are of particular interest in pharmacoepidemiology studies, as such studies struggle with potential for bias from confounding by indication and other unmeasured risk factors, particularly in administrative databases [11].

For unbiased IV estimation, the instrument must be valid [12]. A valid instrument is a variable in the observed data that predicts choice of treatment but is not related to the study outcome, except through the effect of treatment. It must also meet several other criteria [13,14]. Although IV validity is not explicitly testable, stratifying the patient population by a valid dichotomous IV should result in more observed balance among the measured covariates than if those same patients had instead been stratified by their actual treatment. If changing study design or IV definition yields even further covariate balance, the increase may correspond to an increase in the validity of the IV.

A strong instrument is one that is a good predictor of actual treatment, with its predictive effect independent of other measured variables. It is important for an IV to be relatively strong: IV estimation involves scaling up an estimate derived by substituting the IV for actual treatment in an outcome model by a factor inversely proportional to IV strength; hence, any residual confounding in that estimate will be amplified if the instrument is weak. Unlike validity, IV strength is a measurable quantity that can be assessed, reported, and compared [15–18]. In nonrandomized research, it is possible that an instrument can be too strong. A variable that is strongly correlated with a confounded exposure cannot plausibly fulfill the requirements for a valid IV: it will likely be associated with the study outcome via the same unmeasured confounders paths that led to the need for IV analysis in the first place [19].

What's new?

- Physician prescribing preference (PPP) has been used as an instrumental variable in clinical epidemiology.
- This article explores variations in the simple definition of PPP by changing the PPP algorithm and considering restriction and stratification schemes.
- The authors evaluate each variation based on the IV strength and reduction in imbalance—two measures derived from basic IV assumptions.
- The article assesses the overall relationship between strength and imbalance.

This article and its companion, “Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships,” together introduce the concept of instrumental variable (IV) analysis and examine some of the key assumptions underlying the technique. Taken together, the articles show how IVs arise in observational data and how IV analysis parallels randomized trial designs, and also examine the key notions of instrument strength and validity. Each of them describes instruments that have been used in clinical epidemiology and gives examples of IV analysis.

In the study presented here, we explore alternative definitions of the physician prescribing preference (PPP) instrument, proposed by Brookhart et al. [7] and related work by other authors [20,21], as well as a series of variations in study design and cohort selection. For each variation, we assess the IV's strength and the reduction in imbalance resulting from the application of the

IV. We compare reductions in imbalance across the variations and assess the overall relationship between strength and imbalance. To accomplish this, we studied two cohorts of elderly patients initiating treatment with antipsychotic medications (APMs), and considered an outcome of mortality within 180 days.

2. Methods

2.1. Physician prescribing preference

Brookhart et al. [7] have proposed that an individual physician's preference for prescribing one drug over another is an IV that predicts which drug a patient will be treated with [22]. They examined physician prescribing patterns and deduced that the variation they observed may be an instrument [23], under the assumption that PPP is unrelated to outcome. They proposed a simple technique for measuring a physician's preference which we term the “base case”.

As in the earlier work, the base case considered the entire cohort; preference at the time of seeing the patient was determined by the treatment a doctor chose for the previous patient who was treated in his or her practice and who also required a new prescription for one of the study drugs [7,24].

2.2. Variations in study design and physician prescribing preference formulation

The use of the previous patient's treatment to estimate preference has the advantage of quickly registering any changes in preference, but two issues arise: first, the previous patient's treatment may not reflect the doctor's true preference, and second, the simple IV as specified may not possess the required strength and validity. To examine these issues, we designed variations on the base case that were meant to exercise the definition of the PPP measure and to create contrasts in strength and validity. We modified (1) preference assignment algorithm, (2) source population, and (3) stratification criteria (Table 1). In all instances, we chose single, dichotomous IVs for interpretability and comparability.

To consider alternative formulas for measuring the doctor's preference, we first altered the preference assignment algorithm. We expanded the time window to calculate preference from more than just the last new prescription filled. We used the previous two, three, and four new prescriptions, and set different targets for prescribing consistency; as an example, in the case of four prescriptions, we considered that “any of the four,” “half of the four,” and “all of the four” were conventional rather than atypical APMs. We hypothesized that expanding the window would increase balance in treatment groups by creating a better, more stable estimate of true underlying preference and therefore better quasi-randomization of patients to the two predicted treatment groups. On the other hand, we thought that this also would likely decrease the IV strength by weakening the correlation between the IV and the treatment, especially at the higher targets of prescribing consistency. Because the need for more data per physician increases as the window expands, we performed all preference assignment variations in Table 1's group R1, a cohort of patients seen by doctors with very high-volume prescribing.

A concern about instruments based on physician preference is that varying physician quality and patients' “shopping” for doctors based on the treatment they expect to receive may introduce confounding of the IV and negatively affect the validity of the instrument [13,19]. To address these concerns, the second category of variations considered cohort restriction schemes in which we limited the patient population by combinations of measured doctor-level confounders (primary care, specialty, year of graduation) and patient-level confounders (age, age relative to the average in the doctor's practice). By restricting, we hoped to isolate subpopulations in which the IV assumptions may have been more consistently held, thereby increasing validity and balance. We hypothesized that the restricted cohorts would show higher

IV strength, because the combination of greater patient homogeneity and greater number of marginal patients would increase the predictive power of the IV within the subgroups.

Finally, because the preference algorithms estimate preference at any given time from physicians' behavior with prior patients, we created stratification schemes that rearranged the data such that the previous patients shared major characteristics (such as age or gender) with the current patient. By this rearrangement, we hoped that the treatments given to prior patients would reflect not just overall preference, but preference within a particular subgroup of patient. Unlike the restriction schemes, stratification always considered the entire cohort. We hypothesized that the stratification method would contribute to higher instrument strength by means of greater prescribing consistency among like patients, and that stratification would not affect residual imbalance.

We expected the estimates of effect on the outcome to be incomparable across these different variations because of the different patient populations and doctor characteristics. We did believe our empirical measures of strength and imbalance, as well as the standard errors of the effect estimates, would be comparable across the variations.

2.3. Example study: antipsychotic medication initiation and risk of short-term mortality

We performed an example study of initiation of APM therapy and the associated risk of short-term mortality. APMs are categorized into two groups: conventional (older) and atypical (newer) agents [25]. They are widely used off-label to control behavioral disturbances in demented elderly patients. Previous studies have found increased rates of death among users of atypical antipsychotic agents as compared with placebo [26]. Nonrandomized studies have indicated that both types of APMs increase risk of death in the elderly, with the atypical drugs showing lesser risk than the conventional ones [8,27].

2.4. Study participants

Our study population, fully described in earlier work [8,27], was comprised of two cohorts of patients aged 65 years and older who initiated APM treatment. The first cohort was drawn from Pennsylvania (PA)'s Pharmaceutical Assistance Contract for the Elderly (PACE), a drug assistance program for the state's low-income seniors, between 1994 and 2003. The second cohort was drawn from all British Columbia (BC) residents aged 65 years or more between 1996 and 2004. Patients with existing cancer diagnoses were excluded.

2.5. Drug exposures, study outcomes, and measured patient characteristics

We defined our exposed group to be initiators of conventional APM treatment and compared them with a referent group of initiators of atypical APM therapy [8,27]. Outcome was defined as death within 180 days from drug initiation. We defined the baseline characteristics of the patients based on the 6 months before each subject's index date and included coexisting illnesses and use of health care services [28–30]. All dates were measured to the level of day; events occurring on the same day were ordered randomly. Because of limitations of the claims data, we were not able to measure several potentially important covariates—frailty, cognitive impairment, and ability to perform activities of daily living—factors which we hoped to adjust for using IV methods.

2.6. Statistical models

Two-stage least squares (2SLS) models were used to estimate risk differences [7,9]. All IV models were run in Stata Version 9 [31] using the ivreg2 module [32]. Reported standard errors are robust and account for clustering within physician practices using the sandwich estimator [33,34].

2.7. Tests of instrument strength

To test for strength, we examined the partial *F* test from the first-stage regression, which predicts treatment as a function of instrument and covariates. The partial *F* test has the null hypothesis that the coefficient for effect of instrument in the first-stage regression model is zero [15]. In the economics literature, an *F* statistic greater than 10 indicates that the instrument is not weak [18,35].

We also computed the partial r^2, the square of the partial correlation between the instrument and the treatment, conditional on other covariates in the model [15]. The partial r^2 can be interpreted as the proportion of the variance explained by the addition of the IV to the model. Large partial r^2 values indicate that the instrument contributes substantially to the prediction of treatment.

2.8. Assessment of residual imbalance

We measured change in imbalance for measured covariates, comparing the population as stratified by the treatment versus stratified by the IV. We assessed the change for each covariate; negative numbers indicated a reduction in imbalance. We also computed a summary measure: percentage change in the Mahalanobis distance [36,37]. In the case of a single dichotomous confounding variable, the Mahalanobis distance reflects the standardized difference in mean prevalence between treatment groups. When additional variables are considered simultaneously, the Mahalanobis distance extends logically and also corrects for observed covariance among the measured characteristics so as to avoid “double-counting” correlated variables.

3. Results

Characteristics of the 36,541 BC initiators of APMs and 20,087 PA initiators are presented in Table 2. There were 4,113 deaths from any cause (11% of cohort) in BC and 2,935 deaths (15%) in PA.

When stratified by treatment, many variables showed relative balance, but some showed differences of over 5% (in PA, those included gender, dementia, and mood disorders), especially among measured psychiatric conditions. Table 3 alters the stratification to be by the various IVs rather than by treatment. It shows, for a series of potential confounding variables in each of the cohort restriction variations, the difference in prevalence in the predicted treatment groups after stratifying by the IV. On variables unrelated to psychiatric conditions, balance was broadly achieved (difference tended toward 0%), with the exception of hypertension in the PA cohort.

As a summary measure of the imbalance figures for each covariate, the rightmost columns of the PA and BC sections of Table 3 show the percentage change in Mahalanobis distance between treatment and IV stratification. The Mahalanobis distance was reduced in most cases, indicating improved covariate balance, though several variations, especially among the preference algorithm schemes, showed a greater imbalance. The stratification schemes generally showed good improvement in balance.

As an example of change in balance of a single covariate, the PA cohort was 15.1% male in the atypical APM group and 20.1% male in the conventional APM group (Table 2), for a difference between the groups of 5%. When stratified by IV used in the base case, the difference was reduced to 1.8% (Table 3), that is, stratifying by the base case IV resulted in 1.8% more males in the atypical group than in the conventional group. Overall, the application of the instrument caused an overall reduction of imbalance of 62.7% as compared with stratification by the treatment.

For reference, Table 4 shows results of unadjusted, age and sex adjusted, and fully adjusted ordinary least squares (OLS) models, as well as a two-stage least squares (2SLS) IV analysis. All analyses showed an increased risk of death among those treated with conventional APMs, though some confidence intervals included the null value of zero. We presented figures for both the base case and for the subcohort restricted to patients attended by primary care physicians; we assumed that "doctor shopping" would be minimized when patients were seeing their usual primary care doctor.

Table 5 presents measures of instrument strength for all variations of PPP. Partial r^2 values were generally high as compared with selected values from the economics literature, with values ranging from 0.028 to 0.099. The partial r^2 values observed in the base case were among the highest. The partial r^2 values were similar across the various cohort definitions and were not stronger for the restriction schemes. *P*-values for the *F* statistic were universally less than 0.05.

We compared the partial r^2 with several other measures in our data. Partial r^2 did not vary strongly with study size (BC Spearman $r = 0.068$; PA $r = 0.171$). With regard to change in imbalance (Fig. 1a), the correlation was modest in BC and weak in PA (BC $r = 0.482$; PA $r = -0.049$). With regard to standard errors (Fig. 1b), we observed a consistent decrease in standard error as the IV strengthened (BC $r = -0.496$; PA $r = -0.677$).

In Table 5, one can further observe a decrease in study size as a result of cohort restriction schemes. These decreases correlated with a simultaneous increase in standard error of the IV point estimate ($r^2 = 0.71$).

4. Discussion

The relatively infrequent use of IVs in epidemiology may be the result of a perceived lack of strong instruments or concerns about IV validity. In our two example studies evaluating the effectiveness of medicines in routine care, we found that PPP in almost any of its definitions or study formulations would be considered a strong instrument as compared with typical examples in the economics literature. The results also show a broad reduction in imbalance of measured covariates across our restriction and stratification variants. The reduced imbalance in measured covariates and the IV's strength lend credence to the notion that PPP may be an effective instrument for the selected drug comparison. We also noted that the association between instrument strength and imbalance in measured covariates was a mixed one; the Spearman correlation in BC was fairly high, whereas that of PA was close to zero.

Validity of an IV is an untestable property because it involves quantifying the strength of the association between the instrument and the outcome, potentially mediated through unmeasured paths. As in other approaches to controlling confounding, IV validity can be explored through subject matter expertise or empirical assessment of relationships likely to be correlated with unmeasured factors [19]. Inspection of the reduction in imbalance of measured factors achieved by applying the instrument may also be informative. In our data, application of the IV generally reduced imbalance in measured covariates, but significant imbalance remained among the measured psychiatric conditions. These conditions were each correlated with each other, perhaps because of misclassification of specific psychiatric conditions [38]. Because of these strong correlations, we used the Mahalanobis distance to assess overall balance.

The reduction in Mahalanobis distance in many of the variations, along with previous work [7,8,19,24], suggests that PPP was at least a reasonably valid instrument in this setting. The fact that some imbalance remained, especially in psychiatric conditions, suggests some "nonrandom" assignment of patient to practice, such as a clustering of a particular patient type within practice [19]. (For a violation of the IV assumptions to occur, the selection of patients

to practice would also have to be associated with the outcome of death.) Overall, an observed decrease in Mahalanobis distance may be suggestive of increased validity but is not necessarily indicative; it is possible to imagine a circumstance where the Mahalanobis distance is dramatically reduced but IV validity is not affected. It is also possible that using an IV—even one that yields strong treatment group balance—can lead to greater bias than would occur in a non-IV setting. To avoid this, any numeric evaluation of validity also requires due consideration of potential violations of the IV assumptions based on subject matter expertise and other knowledge [14,39].

The fairly consistent decrease in partial r^2 when additional past prescriptions were added to the preference estimation algorithm suggests that considering the additional prescriptions decreases the proportion of the variance in treatment explained by the instrument and weakens the predictive power of the dichotomous IV. Using a continuous rather than dichotomous IV may have mitigated this effect. Even though the IV was weaker, the additional previous prescriptions may have also yielded a better estimate of the physician's true preference because they estimated preference over a longer period of time and over more patients. This suggests that the somewhat lower partial r^2 values when adding previous prescriptions may be a better estimate of the PPP IV's true strength than the higher value observed in the base case.

At the same time, almost all of the cases in which we saw increases in overall imbalance came from requiring that a doctor be totally consistent in his or her prescribing over the window considered (Table 3, rows P4 through P6). However, the physician may be consistent not because of his or her preference but because he or she is seeing similar patients who may have self-selected to his or her practice (“doctor shopped”), or as a result of other forms of atypical case mix. In these cases, the element of randomness in the “assignment” of patients to doctor may have been reduced or lost.

We had hypothesized that a stronger instrument would be associated with somewhat greater imbalance: as instrument strength increases, the IV starts to resemble more closely the treatment variable. If this resemblance becomes too strong, then the IV may be confounded by the same factors that confound treatment, and stratification by the strong IV should reduce imbalance less than stratified by a weaker IV that is less correlated with the treatment's confounders. In our data, by Spearman's rank-based measure of correlation between strength and balance, this played out in BC ($r = 0.482$) but not in PA ($r = -0.049$). Using Pearson's measure based on an assumed linear relationship, there was moderate correlation in both populations (BC $r = 0.270$; PA $r = -0.249$). The divergent findings suggest no clear answer to whether there was a trade-off between imbalance and strength.

The IV methods measure the effect in the marginal patient rather than the effect in the entire cohort [12,40,41]. By varying the cohort definitions, we may have also affected who the marginal patient would be, and therefore, any measures of effect drawn from these variations may not be comparable. We did not present second-stage-effect estimates for all variations, as the choice of the “right” estimate would be very much a decision of study design and subject matter expertise, and should not be driven by the results that appear most reasonable based on previous knowledge.

This study examined a range of implementations of the PPP instrument in two pharmacoepidemiologic studies on APM treatment. In these limited examples, the application of the PPP instrument did generally reduce imbalances, but created imbalances in some cases of the very stringent IV definitions. Imbalances in measured covariates can be controlled for in the analysis, but the remaining imbalances suggest that the unmeasured covariates may be imbalanced as well, and may therefore lead to bias in a traditional outcome model.

In summary, we have demonstrated a number of variants of the PPP instrument and shown how empirically assessing the strength of an IV and its reduction in imbalance of covariates may inform the use of PPP in practical settings relevant to pharmacoepidemiology using claims data.

Acknowledgments

Funding: Dr. Schneeweiss received support from the National Institute on Aging (R01-AG021950), National Institute of Mental Health (U01-MH078708), and the Agency for Healthcare Research and Quality (2-RO1-HS10881), Department of Health and Human Services, Rockville, MD. He is Principal Investigator of the Brigham & Women's Hospital DEcIDE Research Center on Comparative Effectiveness Research funded by the Agency for Healthcare Research and Quality.

References

1. Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, Solomon DH. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: the example of COX2 inhibitors and myocardial infarction. Epidemiology 2005;16:17–24. [PubMed: 15613941]
2. Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. Pharmacoepidemiol Drug Saf 2005;14:465–76. [PubMed: 15651087]
3. Sturmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol 2005;162:279–89. [PubMed: 15987725]
4. Davey Smith G, Ebrahim S. “Mendelian randomization”: can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 2003;32:1–22. [PubMed: 12689998]
5. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. JAMA 1994;272:859–66. [PubMed: 8078163]
6. Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. Annu Rev Public Health 1998;19:17–34. [PubMed: 9611610]
7. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology 2006;17:268–75. [PubMed: 16617275]
8. Wang PS, Schneeweiss S, Avorn J, Fischer MA, Mogun H, Solomon DH, Brookhart MA, et al. Risk of death in elderly users of conventional vs. atypical antipsychotic medications. N Engl J Med 2005;353 (22):2335–41. [PubMed: 16319382]
9. Angrist JD, Imbens G, Rubin DB. Identifcation of causal effects using instrumental variables. JASA 1996;94:444–55.
10. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007;297:278–85. [PubMed: 17227979]
11. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol 2005;58:323–37. [PubMed: 15862718]
12. Greene, WH. Econometric analysis. 5th. Upper Saddle River, NJ: Prentice Hall; 2003.
13. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology 2006;17:360–72. [PubMed: 16755261]
14. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. Int J Biost 2007;3 article 14.
15. 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 Stat Assoc 1995;90:443–50.
16. Murray MP. Avoiding invalid instruments and coping with weak instruments. J Econ Perspect 2006;20:111–32.

17. Staiger D, Stock JH. Instrumental variable regression with weak instruments. Econometrica 1997;65:557–86.
18. Stock, JH.; Yogo, M. Testing for weak instruments in linear IV regression. In: Andrews, DWK.; Stock, JH., editors. Identification and inference for econometric models: essays in honor of Thomas Rothenberg . Cambridge; New York: Cambridge University Press; 2005. p. 80-108.
19. Brookhart MA, Rassen J, Wang PS, Dormuth CA, Mogun H, Schneeweiss S. Evaluating the validity of an instrumental variable study of neuroleptics: can between-physician differences in prescribing patterns be used to estimate treatment effects? Med Care 2007;45:S116–22. [PubMed: 17909369]
20. Wen SW, Kramer MS. Uses of ecologic studies in the assessment of intended treatment effects. J Clin Epidemiol 1999;52:7–12. [PubMed: 9973068]
21. Johnston SC. Combining ecological and individual variables to reduce confounding by indication: case study–subarachnoid hemorrhage treatment. J Clin Epidemiol 2000;53:1236–41. [PubMed: 11146270]
22. Schneeweiss S, Glynn RJ, Avorn J, Solomon DH. A Medicare database review found that physician preferences increasingly outweighed patient characteristics as determinants of first-time prescriptions for COX-2 inhibitors. J Clin Epidemiol 2005;58:98–102. [PubMed: 15649677]
23. Solomon DH, Schneeweiss S, Glynn RJ, Levin R, Avorn J. Determinants of selective cyclooxygenase-2 inhibitor prescribing: are patient or physician characteristics more important? Am J Med 2003;115:715–20. [PubMed: 14693324]
24. Schneeweiss S, Solomon DH, Wang PS, Rassen JA, Brookhart MA. Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: an instrumental variable analysis. Arthritis Rheum 2006;54:3390–8. [PubMed: 17075817]
25. Salzman, C. Clinical geriatric psychopharmacology. 4th. Philadelphia: Lippincott Williams and Wilkins; 2005.
26. Ray WA, Meredith S, Thapa PB, Meador KG, Hall K, Murray KT. Antipsychotics and the risk of sudden cardiac death. Arch Gen Psychiatry 2001;58:1161–7. [PubMed: 11735845]
27. Schneeweiss S, Setoguchi S, Brookhart MA, Dormuth CA, Wang PS. Mortality in users of conventional and atypical antipsychotic medications in British Columbia seniors. Can Med Assoc J 2007;126(5):627–32. [PubMed: 17325327]
28. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992;45:613–9. [PubMed: 1607900]
29. Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J Clin Epidemiol 1993;46:1075–9. 81–90. discussion 81–90. [PubMed: 8410092]
30. Schneeweiss S, Maclure M. Use of comorbidity scores for control of confounding in studies using administrative databases. Int J Epidemiol 2000;29:891–8. [PubMed: 11034974]
31. Stata Version 9 . College Station TX: StataCorp LP;
32. Baum, CF.; Schaffer, ME.; Stillman, S. Statistical software components. Boston College Department of Economics; 2006. ivreg2: Stata module to extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression.
33. Huber, PJ. The behavior of maximum likelihood estimates under non-standard conditions In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability . Berkeley, CA: University of California Press; 1967. p. 221-33.
34. White H. Maximum likelihood estimation of misspecified models. Econometrica 1982;50:1–26.
35. Cragg JG, Donald SG. Testing identifiability and specification in instrumental variable models. Econometric Theory 1993;9:222–40.
36. Mahalanobis PC. On the generalized distance in statistics. Proc Natl Inst Sci India 1936;12:49–55.
37. Stata procedure for calculating the Mahalanobis distance. 2007 [January 28, 2008]. Available at http://personalpages.manchester.ac.uk/staff/mark.lunt.
38. Inouye SK, Foreman MD, Mion LC, Katz KH, Cooney LM Jr. Nurses' recognition of delirium and its symptoms: comparison of nurse and researcher ratings. Arch Intern Med 2001;161:2467–73. [PubMed: 11700159]

39. Wooldridge, JM. Introductory econometrics: a modern approach. 3rd. Mason, OH: Thomson/South-Western; 2006.

40. Greenland S. An introduction to instrumental variables for epidemiologists. Int J Epidemiol 2000;29:722–9. [PubMed: 10922351]

41. Harris KM, Remler DK. Who is the marginal patient? Understanding instrumental variables estimates of treatment effects Health Serv Res 1998;33:1337–60.

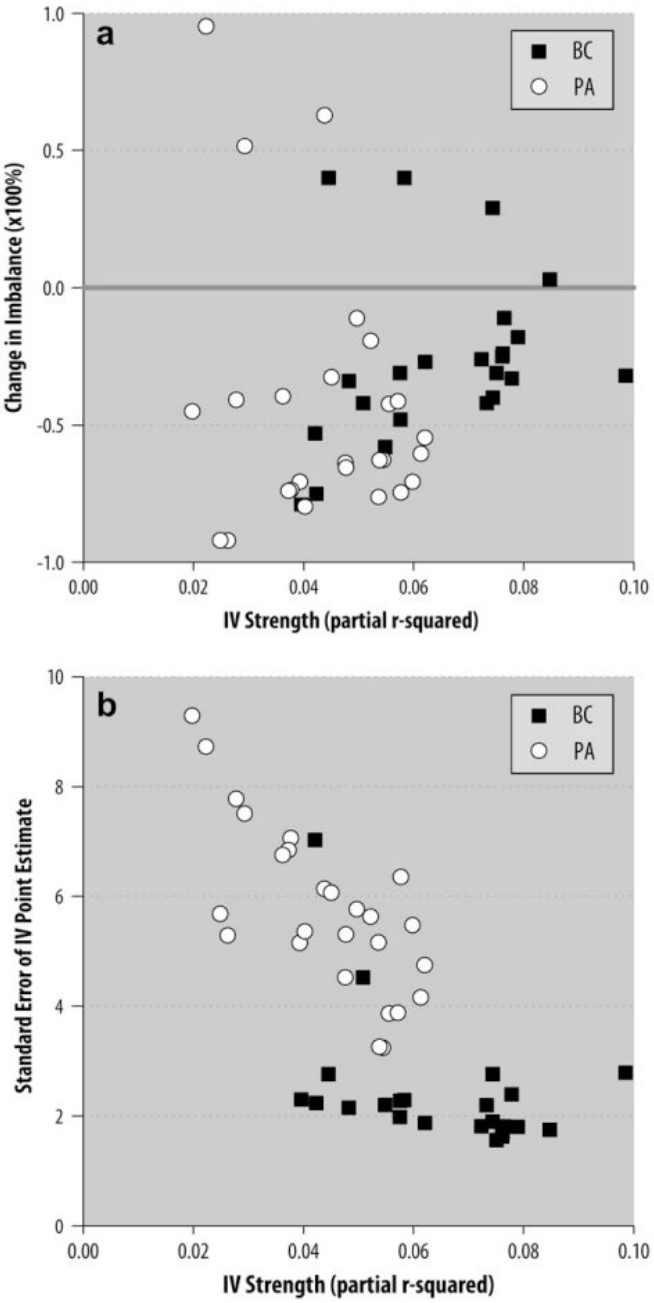

Fig. 1.

(a, b) For each variation (square or circle), the strength of the instrument (horizontal axis) is compared with change in imbalance and standard error (vertical axes). More negative changes in imbalance are better.

Table 1

Study variations considered

Base case	
Base cohort with no restrictions and physician's previous prescription as instrument	
1. Preference assignment algorithm changes[a]	
Lenient criteria	
P1	At least 1 conventional APM rx within last 2 rx's
P2	At least 1 conventional APM rx within last 3 rx's
P3	At least 1 conventional APM rx within last 4 rx's
Strict criteria	
P4	2 conventional APM rx's within last 2 rx's
P5	3 conventional APM rx's within last 3 rx's
P6	4 conventional APM rx's within last 4 rx's
Moderate criteria	
P7	At least 2 conventional APM rx's within last 3 rx's
P8	At least 2 conventional APM rx's within last 4 rx's
2. Cohort restrictions	
Cohort restriction based on doctor characteristics	
R1	Doctor has a very high-volume practice
R2	Doctor has a high-volume practice
R3	Doctor has a low-volume practice
R4	Doctor sees many older patients
R5	Doctor sees many younger patients
R6	Doctor is a primary care physician
R7	Doctor is a specialist
R8	Doctor graduated before 1980 (PA[b])
R9	Doctor graduated after 1980 (PA[b])
Cohort restriction based on patient characteristics	
R10	Patient above median patient age
R11	Patient below median patient age
R12	Patient in the middle quartiles of age
Cohort restriction based on patient and doctor characteristics	
R13	Patient is older than the median age in the doctor's practice
3. Stratification changes	
S1	Last patient was in the same age category
S2	Last patient was also above/below the median patient age
S3	Last patient was also above/below the median patient age within doctor's practice
S4	Last patient was in the same quartile of propensity score

Abbreviations: rx, prescription; APM, anti-pschotic medication.

[a] All preference assignment algorithm changes were carried out within cohort R1, very high-volume prescribers.

[a]Data available in Pennsylvania only.

Table 2

Characteristics of adults 65 years and older in British Columbia and Pennsylvania stratified by type of APM received

Characteristic	British Columbia		Pennsylvania	
Atypical treatment	**Conventional treatment**	**Atypical treatment**	**Conventional treatment**	
Number of new drug starts	23,785	12,756	12,031	8,056
Age (mean)	80.32	79.89	83.58	83.30
Male (%)	35.1	39.7	15.1	20.1
Cardiac arrhythmia (%)	0.1	0.0	1.4	1.4
Cerebrovascular disease (%)	9.9	10.8	30.2	28.3
Congestive heart failure (%)	6.0	8.4	30.4	31.8
Hypertension (%)	24.1	22.3	64.2	57.2
Diabetes (%)	13.8	15.0	26.3	25.5
Myocardial infarction (%)	2.3	2.7	3.3	3.4
Other ischemic heart disease (%)	2.7	3.8	23.8	28.3
Other cardiovascular disorders (%)	16.6	20.2	57.7	55.4
Dementia (%)	12.6	9.7	19.0	7.8
Delirium (%)	8.4	7.4	15.2	11.7
Mood disorders (%)	25.3	15.6	35.5	21.8
Psychotic disorders (%)	16.7	11.2	24.4	21.7
Other psychiatric disorders (%)	4.5	3.1	7.9	5.7
Nursing home residence in previous 180 days (%)	26.8	31.0	20.2	15.5
Number of drugs used (mean)	7.34	7.36	7.82	6.65

Table 3

Differences in prevalance (imbalances) between treatment groups after stratification by the instrumental variable created in each of the study variations

	British Columbia Differences										Pennsylvania Differences									
Study variation[b]	**Age**	**Male (%)**	**Htn. (%)**	**Diab. (%)**	**MI (%)**	**Dem. (%)**	**Mood Dis. (%)**	**Nurs. home (%)**	**Num. drugs**	**Summary[d] (%)**	**Age (%)**	**Male (%)**	**Htn. (%)**	**Diab. (%)**	**MI (%)**	**Dem. (%)**	**Mood Dis. (%)**	**Nurs. home (%)**	**Num. drugs**	**Summary (%)**
Prev.[c]	—	*35.1*	*24.1*	*13.8*	*2.3*	*12.6*	*25.3*	*26.8*	—		—	*15.1*	*64.2*	*26.3*	*3.3*	*19.0*	*35.5*	*20.2*	—	
Base case	−0.07	0.0	−2.0	−0.2	−0.1	−2.1	**−6.0**	**3.7**	−0.12	−30.6	0.06	1.8	**3.5**	0.6	0.1	**7.5**	**8.7**	1.5	0.78	−62.7
R1	0.00	0.3	−2.4	0.1	−0.1	−2.0	**−6.6**	**5.0**	−0.07	−18.3	−0.38	−1.6	**−3.2**	1.3	0.6	**−9.2**	**−7.4**	0.6	−0.82	−19.3
R2	−0.03	0.1	−2.4	−0.2	−0.1	−2.1	**−6.1**	**4.4**	−0.11	−24.5	−0.13	−1.3	**−3.3**	−.6	−0.1	**−8.2**	**−8.1**	−1.1	−0.77	−42.4
R3	0.68	−0.7	0.6	−1.1	−1.0	−1.7	**−5.9**	1.3	−0.10	−52.6	0.33	**−3.2**	**−4.2**	−0.3	−0.1	**−6.0**	**−8.8**	−2.7	−0.72	−73.6
R4	−0.04	0.3	−2.6	−0.3	−0.2	−2.2	**−5.9**	**4.7**	−0.14	−23.5	−0.11	−1.3	**−3.3**	−0.6	−0.1	**−8.3**	**−8.3**	−1.1	−0.80	−41.3
R5	1.63	−2.7	1.5	−0.3	0.2	0.1	**−10.0**	2.0	0.11	−42.4	0.39	−2.7	**−4.1**	−0.3	−0.1	**−5.9**	**−8.6**	−2.3	−0.67	−73.9
R6	−0.42	−0.3	−2.6	−0.5	−0.5	−2.2	−2.5	**3.9**	−0.11	−27.4	−0.21	−1.0	**−4.4**	−0.9	−0.6	**−7.0**	**−6.6**	−3.8	−0.88	−63.8
R7	−1.92	3.0	**3.8**	1.6	2.0	−2.8	**−10.6**	**−6.5**	−0.17	−32.4	−0.23	−3.0	−1.7	0.1	0.8	**−8.3**	**−7.1**	−0.4	−0.68	−54.6
R8	—	—	—	—	—	—	—	—	—	—	−0.18	−1.5	−2.3	−1.0	0.0	**−6.5**	**−9.5**	−2.6	−0.76	−60.4
R9	—	—	—	—	—	—	—	—	—	—	0.33	−1.8	**−4.9**	0.1	−0.1	**−7.5**	**−7.7**	0.2	−0.78	−65.5
R10	0.00	0.1	**−3.1**	−0.1	−0.2	−2.2	**−3.6**	**4.2**	−0.10	−33.1	−0.22	−1.7	**−3.4**	−1.5	0.6	**−6.4**	**−7.3**	−1.3	−0.83	−70.6
R11	0.14	−0.4	−0.7	−0.4	0.3	−1.0	**−8.9**	2.1	−0.22	−47.7	−0.03	−1.8	−2.4	0.1	0.2	**−5.6**	**−8.8**	−0.7	−0.64	−70.6
R12	0.02	1.4	−2.0	−0.3	−0.1	−2.7	**−5.6**	**4.6**	−0.04	−42.1	0.07	−2.3	−2.8	−1.1	0.4	**−8.4**	**−8.1**	0.1	−0.65	−76.2
R13	0.09	0.1	−2.7	0.0	0.0	−2.2	**−4.5**	**3.4**	−0.14	−40.2	0.19	−1.1	−2.6	−2.3	0.4	**−6.5**	**−7.6**	0.8	−0.47	−74.6
S1	−0.13	0.9	−1.5	0.1	0.2	−1.6	**−6.1**	**3.9**	−0.15	−57.7	0.01	−1.7	−2.2	−1.7	0.2	**−6.2**	**−7.5**	−0.5	−0.44	−79.8
S2	−0.57	1.0	−0.9	0.6	0.2	−1.4	**−5.2**	1.8	−0.14	−75.3	−0.39	−1.4	−1.5	−0.4	0.1	**−3.6**	**−6.4**	−1.0	−0.31	−92.2
S3	−0.49	1.3	−1.0	0.9	0.2	−1.6	**−5.0**	1.9	−0.12	−78.7	−0.45	−1.5	−0.2	−0.4	0.0	**−3.3**	**−5.6**	−1.1	−0.16	−92.1
S4	−0.07	0.0	−2.0	−0.2	−0.2	−0.1	−2.1	**−6.0**	**3.7**	−0.12	−30.6	0.06	−1.9	**−3.6**	−0.6	−0.1	**−7.4**	**−8.7**	−1.6	−0.78	−62.7
P1 (R1[a])	−0.02	0.9	−2.6		0.1	**−1.6**	**−6.2**	4.9	−0.15	−26.4	−0.24	−1.6	**−3.3**	−0.1	1.3	**−9.2**	**−6.8**	2.3	−0.73	−32.6
P2 (R1[a])	0.10	0.5	−2.9	−0.7	0.1	**−1.4**	**−6.4**	5.0	−0.21	−31.4	0.06	−0.8	**−3.6**	−0.8	1.0	**−8.0**	**−7.0**	2.3	−0.63	−40.9
P3 (R1[a])	0.10	0.7	−2.2	−0.9	0.1	**−1.2**	**−6.5**	5.0	−0.18	−34.3	0.16	−1.2	**−3.5**	−0.4	1.0	**−7.2**	**−7.1**	2.5	−0.58	−45.0
P4 (R1[a])	0.11	−1.0	**−3.1**	0.1	0.3	**−2.3**	**−7.1**	4.8	−0.15	29.2	−0.47	−0.9	−1.8	2.2	0.6	**−9.3**	**−8.5**	0.3	−0.59	62.9
P5 (R1[a])	0.12	−0.9	−1.9	0.3	0.5	**−2.1**	**−6.8**	4.5	−0.13	40.4	−0.21	−0.6	−1.0	0.4	0.8	**−8.6**	**−9.5**	−0.9	−0.53	51.5
P6 (R1[a])	0.22	−1.4	−1.5	0.3	0.9	**−2.3**	**−7.9**	4.6	−0.08	39.7	−0.05	−0.9	1.0	−0.4	0.8	**−10.2**	**−8.7**	−1.2	−0.48	95.3
P7 (R1[a])	0.16	−0.1	**−3.0**	−0.1	0.0	**−2.2**	**−6.9**	4.9	−0.16	3.4	−0.29	−0.6	**−3.0**	−0.5	0.5	**−7.7**	**−7.8**	1.3	−0.43	−11.2

	British Columbia Differences										Pennsylvania Differences									
Study variation[b]	**Age**	**Male (%)**	**Htn. (%)**	**Diab. (%)**	**MI (%)**	**Dem. (%)**	**Mood Dis. (%)**	**Nurs. home (%)**	**Num. drugs**	**Summary[d] (%)**	**Age (%)**	**Male (%)**	**Htn. (%)**	**Diab. (%)**	**MI (%)**	**Dem. (%)**	**Mood Dis. (%)**	**Nurs. home (%)**	**Num. drugs**	**Summary (%)**
P8 (R1[a])	0.17	0.1	**−3.0**	−0.7	0.2	**−2.2**	**−7.1**	4.8	−0.24	−10.7	0.03	0.2	−2.5	−1.7	0.9	**−7.1**	**−7.3**	2.0	−0.38	−39.6

NOTE. Smaller figures indicate greater balance. The italicized row indicates the covariate's prevalence in the base case variation's untreated group. Differences of over 3% are shown in bold.

Abbreviations: Prev., prevalence; Htn., hypertension; Diab., diabetes; MI, myocardial infarction; Dem., dementia; mood Dis., mood disorders; Nurs. home, nursing home.

[a] R1: All variations in the P group were based on the very high-volume prescribers in population R1.

[b] See Table 1 for a description of the variations.

[c] Prevalence of the covariate in the unexposed (atypical) group in the base cohort.

[d] Summary change in imbalance is calculated by subtracting the Mahalanobis distance from the data stratified by treatment from the data stratified by the instrumental variable, and then dividing by the distance as stratified by treatment. Positive numbers represent increased imbalance, whereas negative numbers represent reduced imbalance.

Table 4

Difference in risk of all-cause mortality within 180 days of initiation of conventional versus atypical APM treatment.

Population and variation	Events in conventional APM group	Events in atypical APM group	Unadjusted OLS estimate	Age/sex-adjusted OLS estimate	Fully adjusted OLS estimate[a]	IV analysis estimate
British Columbia						
Base case (unrestricted)	1,806	2,307	4.46 (3.69, 5.23)	4.49 (3.75, 5.22)	3.55 (2.74, 4.37)	4.00 (0.94, 7.06)
Restricted to PCPs (R6)	1,735	2,115	4.24 (3.41, 5.06)	4.48 (3.68, 5.28)	3.59 (2.70, 4.48)	3.11 (−0.57, 6.79)
Pennsylvania						
Base case (unrestricted)	1,307	1,628	2.69 (1.65, 3.73)	2.47 (1.46, 3.49)	3.91 (2.68, 5.13)	7.69 (1.26, 14.12)
Restricted to PCPs (R6)	960	1,129	2.39 (1.07, 3.71)	2.29 (0.98, 3.60)	4.32 (2.71, 5.93)	5.34 (−3.53, 14.21)

NOTE. The values within brackets are 95% confidence intervals. Risk differences are expressed per 100 patients.

Abbreviations: APM, antipsychotic medication; OLS, ordinary least squares; IV, instrumental variable; PCP, primary care physician.

[a] Adjusted for age, sex, race, year of treatment, and history of diabetes, arrhythima, cerebrovascular disease, myocardial infarction, congestive heart failure, hyptertension, other ischemic heart disease, other cardiovascular disorders, dementia, delirium, mood disorders, psychotic disorders, other psychiatric disorders, antidepressant use, nursing home residence, and hospitalization.

Table 5

Characteristics of the first-stage instrumental variable regression model for each study variation

	British Columbia					Pennsylvania				
	First-stage models					First-stage models				
Study variation	*N*	Unadj. IV to treatment OR	Adj. IV to treatment OR[a]	First-stage partial *F* statistic[a]	Partial r^2 value study variation	*N*	Unadj. IV to treatment OR	Adj. IV to treatment OR[a]	First-stage partial *F* statistic[a]	Partial r^2 value[a]
Base case	31,976	6.15	3.80	909	0.075	13,131	6.60	3.29	428	0.054
R1	24,085	6.68	3.98	613	0.079	4,577	8.06	3.19	250	0.052
R2	29,741	6.37	3.86	814	0.076	9,198	7.46	3.30	539	0.055
R3	2,235	3.97	2.66	78	0.042	3,933	5.19	2.81	153	0.038
R4	28,205	6.44	3.88	753	0.076	9,024	7.64	3.36	545	0.057
R5	3,771	4.44	2.93	142	0.051	4,107	5.08	2.79	158	0.037
R6	27,352	5.65	3.37	736	0.062	8,602	6.38	3.10	428	0.048
R7	4,462	8.79	5.27	88	0.099	4,184	7.08	3.62	275	0.062
R8	—	—	—	—	—	6,538	6.76	3.51	425	0.061
R9	—	—	—	—	—	6,148	6.19	3.10	306	0.048
R10	16,774	6.51	4.00	589	0.078	5,432	6.51	3.59	343	0.060
R11	12,369	4.71	3.19	347	0.058	5,256	5.43	2.81	214	0.039
R12	14,914	5.91	3.80	531	0.073	5,362	6.12	3.35	302	0.054
R13	12,585	6.00	3.90	420	0.074	4,149	6.27	3.63	252	0.058
S1	22,242	4.70	3.21	448	0.055	7,092	5.20	2.93	296	0.040
S2	29,143	3.33	2.72	465	0.042	10,688	2.74	2.28	287	0.026
S3	28,563	3.14	2.64	421	0.040	9,608	2.66	2.26	244	0.025
S4	31,976	6.15	3.80	909	0.075	13,131	6.57	3.27	744	0.054
P1 (R1)	24,085	6.73	4.10	582	0.072	4,577	8.62	3.48	214	0.045
P2 (R1)	24,085	6.56	4.01	478	0.057	4,577	7.89	3.15	130	0.028
P3 (R1)	24,085	6.72	4.10	450	0.048	4,577	7.67	3.09	92	0.020
P4 (R1)	24,085	7.37	4.41	607	0.074	4,577	7.06	3.01	208	0.044
P5 (R1)	24,085	7.26	4.56	509	0.058	4,577	6.12	2.75	137	0.029
P6 (R1)	24,085	7.10	4.64	422	0.045	4,577	5.64	2.80	104	0.022
P7 (R1)	24,085	6.46	4.22	624	0.085	4,577	6.19	3.08	237	0.050

	British Columbia					Pennsylvania				
	First-stage models					First-stage models				
Study variation	***N***	**Unadj. IV to treatment OR**	**Adj. IV to treatment OR[a]**	**First-stage partial *F* statistic[a]**	**Partial r^2 value study variation**	***N***	**Unadj. IV to treatment OR**	**Adj. IV to treatment OR[a]**	**First-stage partial *F* statistic[a]**	**Partial r^2 value[a]**
P8 (R1)	24,085	5.83	3.95	557	0.076	4,577	4.99	2.72	171	0.036

See Table 1 for description of variations. The first stage model predicts atypical or conventional APM treatment dependent on the instrumental variable and the measured covariates. Higher F statistics and partial r^2 values indicate stronger ability of the instrument to predict treatment.

R1: All variations in the P group were based on the very high-volume prescribers in population R1.

Abbreviations: OR, odds ratio; Unadj., unadjusted; Adj., adjusted.

[a] Adjusted for age, sex, race, year of treatment, and history of diabetes, arrhythima, cerebrovascular disease, myocardial infarction, congestive heart failure, hyptertension, other ischemic heart diseases, other cardiovascular disorders, dementia, delirium, mood disorders, psychotic disorders, other psychiatric disorders, antidepressant use, nursing home residence, and hospitalization.