Volume 12 • Number 8 • 2009
VALUE IN HEALTH

Too Much Ado about Instrumental Variable Approach: Is the Cure Worse than the Disease?

Onur Baser, MS, PhD

STATinMED Research, LLC, and Department of Surgery, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Objective: To review the efficacy of instrumental variable (IV) models in addressing a variety of assumption violations to ensure standard ordinary least squares (OLS) estimates are consistent. IV models gained popularity in outcomes research because of their ability to consistently estimate the average causal effects even in the presence of unmeasured confounding. However, in order for this consistent estimation to be achieved, several conditions must hold. In this article, we provide an overview of the IV approach, examine possible tests to check the prerequisite conditions, and illustrate how weak instruments may produce inconsistent and inefficient results.

Methods: We use two IVs and apply Shea's partial *R*-square method, the Anderson canonical correlation, and Cragg–Donald tests to check for weak instruments. Hall–Peixe tests are applied to see if any of these instruments are redundant in the analysis.

Results: A total of 14,952 asthma patients from the MarketScan Commercial Claims and Encounters Database were examined in this study. Patient health care was provided under a variety of fee-for-service, fully capitated, and partially capitated health plans, including preferred provider organizations, point of service plans, indemnity plans, and health maintenance organizations. We used controller–reliever copay ratio and physician practice/prescribing patterns as an instrument. We demonstrated that the former was a weak and redundant instrument producing inconsistent and inefficient estimates of the effect of treatment. The results were worse than the results from standard regression analysis.

Conclusion: Despite the obvious benefit of IV models, the method should not be used blindly. Several strong conditions are required for these models to work, and each of them should be tested. Otherwise, bias and precision of the results will be statistically worse than the results achieved by simply using standard OLS.

Keywords: asthma, instrumental variable, propensity score, regression analysis.

Introduction

Causal inference is challenging in all nonexperimental studies because of the possibility of *overt* and *hidden* bias [1]. When evaluating certain treatment programs, *overt bias* can exist because the treatment and control groups are different in terms of certain observable factors, such as age, gender, and comorbidities. *Hidden bias* may exist as a result of failure to control for unobservable factors, such as doctors' practice/prescription patterns [2]. Practice patterns might be based on physician rules of thumb, experiences and interactions with patients and colleagues as well as comprehensive empirical evidence. Prescription patterns are influenced by nonclinical factors. For example, health plans provide different financial and nonfinancial incentives to doctors or patients to undergo aggressive treatment [3]. It is not feasible to measure all of these characteristics in observational data.

Propensity score matching and regression analysis are two statistical techniques used to remove overt bias. Although regression analysis is widely used in applied economics literature, propensity score methods are increasingly used in medical literature. A systematic literary search by Stürmer et al. found that the annual number of publications using propensity score methods increased from 8 to 71 from 1998 to 2003 [4]. Last year, the number of propensity score methods used was 171.

Baser describes the conditions outlining which method is optimal for controlling for observable bias [2]. Although effectively controlling for observable bias, neither propensity score matching nor regression adjustment addresses problems because of imbalances in unmeasured factors. For this reason, interest in instrumental variable (IV) approach is growing.

Notions of causality in econometrics and their relationship with IVs and other methods are discussed in Heckman [5]. Econometric literature on notions of causality goes back to early work by Ashenfelter [6] and subsequent work by Ashenfelter and Card [7], Heckman and Robb [8], Lalonde [9], Fraker and Maynard [10], Card and Sullivan [11] and Manski [12,13]. The use of IV technique in outcomes research has increased in recent years because even in the presence of hidden bias, such methods may consistently estimate the average causal effects [14]. We are in the beginning stage of this application on outcomes research studies and believe the surge is yet to start [15–19].

However, like many techniques borrowed from one discipline and applied to another, there is a tendency to apply this method blindly. Researchers, unaware of the shortcomings of this technique, may apply it inappropriately. In this article, we draw attention to the problem of using instruments that explain little of the variation in the endogenous explanatory variables (such as treatment choice variables, etc.). These instruments can lead to large inconsistencies in IV estimates. The magnitude of the bias of IV estimates approaches that of ordinary least squares (OLS) estimates as the *R*-square between the instruments and the endogenous variable approaches zero. While these results are known in economics literature, their potential implications for empirical work related to outcomes research have not been fully appreciated.

The discussion in this article does not provide detailed or rigorous treatment of the theory that premises the IV approach. In recent years, several books on IV methods, with various levels of sophistication, have been published. Wooldridge's book is an excellent source for researchers with an elementary level of

Address correspondence to: Onur Baser, President, STATinMED Research, 218 N. Fourth Avenue, Suite 205, Ann Arbor, MI 48104, USA. E-mail: obaser@statinmed.com

10.1111/j.1524-4733.2009.00567.x

© 2009, International Society for Pharmacoeconomics and Outcomes Research (ISPOR) 1098-3015/09/1201 1201–1209

statistical knowledge [20]. His more advanced book provides detailed information on IV for readers with an advanced mathematical background [21]. The work of Bowden and Turkington is geared toward mathematical outcomes researchers [22]. Curious readers are encouraged to consult these books for a more detailed analysis.

Overview of IV Estimation

Suppose we want to estimate the effect of treatment (T) on outcome (Y), i.e., estimate β_1 in:

$$Y = \beta_0 + T\beta_1 + e$$

For simplicity, we assume dichotomous treatment variable (T), homogeneous treatment effect, linear regression, and no covariates. e is unobservable. Least squares estimate of the equation yields the following estimator:

$$\beta_1^{OLS} = \bar{Y}_{T=1} - \bar{Y}_{T=0},$$

which is the difference in mean outcomes. In order to reach a consistent estimator, the key assumption is that treatment (T) is not correlated with the unobserved determinants of the outcome (e).

$$E(T'\varepsilon) = 0 \Rightarrow T'\left(Y - T'\hat{\beta}_{OLS}\right) = 0 \Rightarrow \hat{\beta}_{OLS} = (T'T)^{-1}T'Y$$

The OLS assumption is unlikely to hold, because treatment is related to omitted factors influencing outcome. For example, patients who are more severely ill in ways known to their physicians but not to the analyst might not get the treatment, or vice versa.

In order to obtain a consistent estimator of β_0 and β_1 when treatment and omitted factors are related, we need additional information.

The information comes by way of a new variable—an IV—(Z) that satisfies the following properties:

1. $Cov(Z,e) = 0$ Z should have no partial effect on the outcomes variable and should not be correlated with other factors that affect the outcomes variable.
2. $Cov(Z,T) \neq 0$ Z must be related, either positively or negatively, to the treatment indicator.

If these two conditions are satisfied, then the IV estimator is:

$$E(Z'e) = 0 \Rightarrow Z'\left(Y - T\hat{\beta}_{IV}\right) = 0 \Rightarrow \hat{\beta}_{IV} = (Z'T)^{-1}Z'Y$$

Note that all IV results apply asymptotically. Small sample estimation properties of IV are more complex and, as discussed in the next section, not generally understood. Variants on this approach include two-stage least squares [21,23,24], limited information maximum likelihood estimator [21,25], general method of moments [21,26], and sample selection corrections ("Heckit") [21,27].

The coin toss in the context of randomized controlled trials (RCTs) is a perfect example of IV. The coin toss does not effect the outcome of interest directly (assumption (i)) but it determines treatment assignment (assumption (ii)). Following are some examples of IVs that have been used in applied research:

1. Geography (distance, rivers, small area variation) [28–32].
2. Legal/political institutions (laws, election dynamics) [31,33,34].
3. Administrative rules (wage/staffing rules, reimbursement rules, eligibility rules) [35–37].
4. Naturally occurring randomization (blood type of recipients, draft, birth timing, lottery, roommate assignment) [36,38–40].

Why Not Always Use IV?

The immediate question that arises is if the IV method is superior to risk adjustment methods such as propensity score matching or multivariate regression in the sense that these methods both cover observable and unobservable factors, why not always use the IV method?

First, it is hard to find variables that meet the definition of valid instruments. Conceptually, most variables that have an effect on treatment variables may also have a direct effect on the outcomes variable.

Second, the standard errors of IV estimates are likely to be larger than those of OLS estimates, creating publication bias. It is important to understand that publication bias may exist even without the authors of individual studies being aware of it. The potential problem simply arises because of the desire to report significant findings. Although lack of treatment is useful to report, evidence against the null hypothesis—that is, favorable to the finding of a treatment effect—is more likely to be reported. Because IV estimates have to be larger in order to be significant, published results tend to create publication bias [41].

Third, the desirable properties of the IV estimator hold in large sample sizes. With simulation, Grootendorst showed that in small sample sizes, the estimates can be highly inaccurate [17]. Also, refer to Kinal for related issues with small properties, where the IV estimator may have no expected value [42].

Fourth, the interpretation of IV is difficult, especially when the treatment effect is heterogeneous. Randomized clinical trials estimate the treatment effect in well-defined populations. Therefore, there are always issues of external validity. Analogous issues arise with IV estimates. They estimate treatment effect for the "marginal" patients whose treatment is affected by the instruments. Therefore, they often do not estimate treatment effect in the general population [43].

The last sets of problems are related to weak instruments, which are the focus of this article.

Weak Instruments

There is a very important difference between the two requirements for an IV. Because assumption (i) is a covariance between the IVs and the unobservable error u, it can never be directly checked or even tested. Rather, we must maintain this assumption by appealing to clinical behavior (in the presence of multiple instruments, indirect tests can be conducted. See Wooldridge for details [21]).

By contrast, assumption (ii) that IV is correlated with treatment choice can be tested, given a random sample from the population. If this correlation is *weak*, this may lead to large inconsistencies in IV estimates with the bias in the same direction as that of OLS estimates [44]. Because IV estimates also have larger standard errors than those of OLS estimates, as pointed out by Bound et al., ". . . the cure can be worse than the disease" [45].

Testing for Weak Instruments

Staiger and Stock formalized the definition of weak instruments and most researchers appear to have concluded (incorrectly) from that work that if F-statistics on coefficients of exogenous

variables on the endogenous treatment indicator is greater than 10, one need to worry no further about weak instruments [43].

Another statistic commonly used, as recommended by Bound et al., is the R^2 of the regression with instruments partialed out [45]. However, Shea showed, in general, the distribution of this *F*-statistic is nonstandard [46]. Also, for models with multiple endogenous variables, these indicators may not be sufficiently informative [46].

To grasp the pitfalls facing empirical researchers here, consider the following simple example. We have a model with two endogenous covariates (treatment and insurance choice) and two instruments (distance to nearest specialized hospital and small area variation). Distance to nearest specialized hospital increases the likelihood of being admitted in a specialized hospital. Therefore, patients near a specialized hospital are more likely to be treated by specialized medical staff, in a special care unit, and with other dimensions of higher intensity. Distance to nearest specialized hospital might also affect the insurance choice. The differences between each managed care plan lie mainly in the degree of compensation one receives for medical treatment outside the managed care network. Patients who live close to a specialized hospital are more likely to choose a low premium insurance plan.

Small area variations in hospital surgical volumes might affect the treatment quality. Because volume is positively correlated with surgical quality, patients who live in high volume areas might get better treatment choices than patients who live in low volume areas. Therefore, both distance to nearest specialized hospital and small area variation in hospital volumes are valid instruments because they have a direct effect on treatment or insurance choice but indirectly related to the outcome.

Suppose the distance to the nearest specialized hospital is highly correlated with treatment and insurance choice, but the small area variation is just a noise. In this case, because we have one instrument for two endogenous variables, this model is underidentified. Bound et al.'s *F*-statistics [45] and partial *R*-squared measures from regression with instruments will not reveal this weakness. Indeed, the *F*-statistics are statistically significant and without investigation, but we may not realize the model cannot be estimated in this form. The statistics proposed by Bound et al. [45] diagnose instruments relevant only in the presence of one endogenous covariate.

When multiple endogenous variables are used, other statistics can be used. Shea provided such statistics [46]. Shea's *R*-squares take the intercorrelations among instruments into account. As a rule of thumb, if an estimated equation yields a large value of the standard partial *R*-squares and a small value of the Shea measure, we should conclude that the instruments lack sufficient relevance to explain all the endogenous regressors.

A more general approach to weak instruments was proposed by Anderson [47] and discussed in Hall and Peixe [48]. Anderson's approach considers the canonical correlations of the excluded and included instruments. This test shows whether some instruments are redundant. Stock and Yogo go into more details and provide useful rules of thumb regarding the weakness of instruments based on a statistic from Cragg and Donald [49,50].

Data Sources and Construction of Variables

We illustrate the implications of a weak IV using MarketScan data to examine the effect of controller medication on health-care expenditures for asthma patients. Briefly, the MarketScan Commercial Claims and Encounters Database contains detailed descriptions of inpatient, outpatient, medical, and outpatient prescription drug services for approximately 13 million persons in 2005 who were covered by corporate-sponsored health-care plans.

Details of the patient selection criteria are provided in Crown et al. [51] and summarized as follows:

1. Patients with evidence of asthma were selected from the intersection of the medical claims and encounter records, enrollment files, and pharmaceutical data files.
2. Individuals meeting at least one of the following criteria were deemed to show evidence of asthma:
   - At least two outpatient claims with primary or secondary diagnoses of asthma.
   - At least one emergency room (ER) claim with primary diagnosis of asthma, and a drug transaction for an asthma medication 90 days before or 7 days after the ER claim.
   - At least one inpatient claim with a primary diagnosis of asthma.
   - A secondary diagnosis of asthma and a primary diagnosis of respiratory infection in an outpatient or inpatient claim.
   - At least one drug transaction for an anti-inflammatory agent, oral antileukotrienes, long-acting bronchodilator, or inhaled or oral short-acting beta-agonists.
3. Patients with a diagnosis of chronic obstructive pulmonary disease and having one or more diagnoses or procedure codes indicating pregnancy or delivery, or who were not continuously enrolled for 24 months, were excluded from our study group.

The sociodemographic characteristics include age of the household, percentage of the patients who were female, and geographic region (northeast, north-central, south, west, and "other" region). Charlson comorbidity index scores are generated to capture the level and burden of comorbidity. Point-of-service plans and other plan types, including health maintenance organizations and preferred provider organizations, are included. The analytic file contains patients with fee-for-service (FFS) health plans and those with partially or fully capitated plans. Data on costs are not available for the capitated plans however. Therefore, the value of patients' service utilization under the capitated plans is priced and imputed using average payments from the MarketScan FFS inpatient and outpatient services by region, year, and procedure.

The outcomes variable is total health-care costs. The MarketScan database contains information on all payments processed with regard to reimbursement for particular services, including secondary payers and patient out-of-pocket costs. For services in which these MarketScan employers are a secondary payer (i.e., the patient has other primary insurance), the amounts paid by other insurers is also documented and included in the cost. In cases where services delivered are completely covered by another primary insurance, these claims are not included in the database. Data on costs were not available for the capitated plans. Therefore, the value of patients' service utilization under the capitated plan was priced and imputed using average payments from the MarketScan FFS inpatient and outpatient services by region, year, and procedure.

The endogenous variable is treatment choice (= 1 if controller, = 0 if reliever). Asthma drugs can be reliever medications (used to relieve symptoms in an acute asthmatic exacerbation or asthma attack) or primarily controller medications (used to control pulmonary inflammation and prevent an attack).

The IVs are controller/reliever copay ratio and physician/practice prescribing pattern. Copayments for outpatient pharma-

ceuticals are calculated by first stratifying all prescription drug claims by year, then by plan within year. We then calculate the average out-of-pocket patient copayments for asthma drugs by therapeutic class for each plan to calculate the ratio of mean controller payments to mean reliever copayments. These plan-level ratios are attached to each patient's record for a given plan. Our second IV involves calculation of the proportion of patients obtaining controller medication for each physician/provider tax identification number. In many cases, this tax identification number includes a multiphysician medical practice, but in some cases is unique to one physician.

Results

The objective of this study was to estimate the cost of illness for asthma patients treated with controller and reliever medications using the IV approach.

Table 1 reports the demographic characteristics of the sample, stratified by treatment choice. Patients using controller medication had a mean age of 40 years (compared with 30 years for patients using reliever medication) and were more likely to be female. The racial distribution in counties was similar between the two groups. Patients treated with controller medication were more likely to receive their health-care coverage as an employee compared with those with reliever medication. Significant differences in mean income between the two groups were evident from county-level US census data compared with the claims data. Patients treated with controller medication had higher numbers of major diagnostic categories, higher Charlson comorbidity scores, and higher rates of asthma specific comorbidities. The descriptive table shows that the ratio of mean controller copayments to mean reliever copayments was lower for the treated controller group relative to the reliever group. Physicians were more likely to prescribe controller medication to the patient group treated with controller medication. The unadjusted total health-care costs were significantly higher for patients treated with controller medication relative to the ones who were treated with reliever medication.

Because the Hausman test showed the treatment choice is endogenous, the IV method has been applied [52]. ($P < 0.000$) We have two possible candidates for instrument: controller/reliever copay ratio and physician/practice prescribing pattern. The first key assumption for IV is that it does not independently affect the outcome, so it is not associated with measured and unmeasured health status. Table 2 shows a different division of the sample from Table 1, namely division according to quintile of IVs. The first assumption, that copay ratios and prescribing patterns affect health-care costs only through its effect on the

Table 1 Summary of asthma patient characteristics

	Controller medication (n = 3,903)		Reliever medication (n = 11,049)			
Variables	Mean	STD	Mean	STD	*P*-values	STD difference
Explanatory Variables						
Age	39.74	16.52	30.50	18.07	0.0000	53.38
(%) Female	0.63	0.48	0.57	0.49	0.0000	11.87
Race						
White	0.84	0.14	0.85	0.13	0.2737	5.49
Black	0.09	0.11	0.09	0.10	0.3819	4.41
Other	0.02	0.02	0.02	0.02	0.7958	3.47
Geographic Regions						
Northeast	0.02	0.15	0.02	0.13	0.1442	2.66
North-Central	0.81	0.39	0.85	0.36	0.0000	10.86
South	0.02	0.15	0.02	0.13	0.1442	2.66
West	0.04	0.20	0.04	0.19	0.4918	1.27
Year of Patient Identification						
1996	0.40	0.49	0.39	0.49	0.3035	1.91
1997	0.26	0.44	0.28	0.45	0.0142	4.59
1998	0.31	0.46	0.31	0.46	0.5362	1.15
1999	0.02	0.15	0.02	0.13	0.0209	4.15
Member Type						
Employee	0.49	0.50	0.35	0.48	0.0000	29.54
Spouse	0.27	0.45	0.20	0.40	0.0000	18.72
Dependents						
4–11 Years	0.08	0.27	0.19	0.39	0.0000	32.31
12–18 Years	0.11	0.32	0.21	0.41	0.0000	25.59
Others	0.04	0.20	0.06	0.24	0.0000	9.99
County Mean Household Income	$25,829	$6,686	$24,997	$6,141	0.0000	12.95
Number of Major Diagnosis Categories	6.80	2.39	6.06	2.15	0.0000	32.58
Charlson Comorbidity Index	0.92	1.18	0.49	0.92	0.0000	40.73
Asthma-Specific Comorbidities						
Allergic Rhinitis	0.28	0.45	0.18	0.39	0.0000	23.94
Migraine	0.07	0.26	0.05	0.22	0.0000	8.43
Depression	0.10	0.30	0.10	0.30	0.3604	1.69
Gastrointestinal Disorder	0.28	0.45	0.21	0.41	0.0000	16.44
Sinusitis	0.27	0.44	0.23	0.42	0.0000	7.81
Anxiety	0.03	0.16	0.02	0.15	0.4556	1.38
Instrumental Variables						
Controller/Reliever Copayment	1.32	0.30	1.29	0.26	0.0000	14.08
Tax Provider ID Controller %	0.61	0.05	0.60	0.04	0.3144	20.73
Tax Provider ID Reliever %	0.39	0.05	0.40	0.04	0.3144	20.73
Outcomes Variable						
Total Cost	$4,321	$7,011	$2,792	$6,151	0.0000	23.18

ID, identification number; STD, standard deviation.

Table 2 Descriptive statistics on asthma patients by quintiles of instrumental variables

	IV-1: controller/reliever ratio					IV-2: controller/reliever ratio				
Variables	Q1	Q2	Q3	Q4	Q5	Q1	Q2	Q3	Q4	Q5
Explanatory Variables										
Age	31.62	31.81	30.52	30.93	40.22	31.20	31.49	26.59	31.55	41.61
(%) Female	0.59	0.61	0.56	0.56	0.62	0.57	0.59	0.52	0.59	0.62
Race										
White	0.88	0.87	0.86	0.79	0.78	0.87	0.86	0.75	0.85	0.79
Black	0.08	0.07	0.08	0.08	0.13	0.08	0.08	0.11	0.07	0.12
Other	0.02	0.02	0.02	0.03	0.03	0.02	0.02	0.04	0.02	0.03
Geographic Regions										
Northeast	0.00	0.00	0.02	0.02	0.06	0.00	0.00	0.21	0.00	0.04
North-Central	1.00	0.95	0.92	0.61	0.54	0.99	0.91	0.28	0.86	0.59
South	0.00	0.00	0.02	0.02	0.06	0.00	0.00	0.21	0.00	0.04
West	0.00	0.02	0.02	0.17	0.07	0.00	0.00	0.17	0.07	0.09
Year of Patient Identification										
1996	0.08	0.00	0.87	0.39	0.37	0.93	0.09	0.02	0.00	0.29
1997	0.85	0.00	0.04	0.04	0.38	0.07	0.86	0.87	0.00	0.17
1998	0.07	0.93	0.09	0.56	0.24	0.00	0.06	0.10	1.00	0.44
1999	0.00	0.07	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.10
Member Type										
Employee	0.40	0.40	0.37	0.36	0.40	0.35	0.40	0.33	0.40	0.43
Spouse	0.19	0.20	0.18	0.24	0.31	0.20	0.19	0.18	0.20	0.31
Dependents										
4–11 Years	0.16	0.14	0.17	0.21	0.14	0.16	0.17	0.27	0.16	0.12
12–18 Years	0.19	0.19	0.22	0.16	0.12	0.22	0.18	0.19	0.17	0.12
Others	0.07	0.07	0.06	0.03	0.03	0.07	0.06	0.02	0.06	0.02
County Mean Household Income	$24,691	$26,040	$24,584	$27,610	$24,897	$24,127	$24,640	$26,216	$26,631	$26,069
Number of Major Diagnosis Categories	6.27	6.31	6.30	5.72	6.34	6.30	6.27	5.64	6.18	6.38
Charlson Comorbidity Index	0.54	0.57	0.57	0.55	0.80	0.57	0.56	0.56	0.56	0.80
Asthma-Specific Comorbidities										
Allergic Rhinitis	0.20	0.21	0.21	0.26	0.19	0.21	0.21	0.21	0.22	0.19
Migraine	0.06	0.06	0.05	0.06	0.05	0.06	0.06	0.05	0.06	0.05
Depression	0.11	0.11	0.10	0.10	0.07	0.11	0.11	0.08	0.11	0.07
Gastrointestinal Disorder	0.22	0.23	0.22	0.21	0.28	0.22	0.22	0.23	0.23	0.28
Sinusitis	0.24	0.24	0.25	0.23	0.23	0.25	0.24	0.24	0.24	0.22
Anxiety	0.02	0.03	0.03	0.02	0.02	0.03	0.02	0.03	0.02	0.02
Treatment										
Controller Medication	0.22	0.24	0.25	0.26	0.34	0.25	0.23	0.26	0.25	0.34
Outcomes Variable										
Total Cost	$3,043	$3,223	$2,805	$2,806	$4,134	$2,940	$3,064	$3,150	$3,110	$3,973

IV, instrumental variable.

likelihood of treatment choice, cannot be tested directly. We can, however, indirectly test this assumption and see how realistic it is a priori. This assumption would be satisfied if a person's copay ratio or physician practice/prescribing pattern was not associated with the clinical severity of the asthma, the primary unobserved variable that will determine the treatment. If this is true, IVs should also be independent of observed variables such as age, gender, and comorbidities, associated with health status and hence the likelihood of treatment choice. The data in Table 2 show that observable factors are independent across the quintiles.

We also tested whether these instruments satisfied the second assumption: are they highly correlated with treatment?

By looking at the quintiles in Table 2, we can see the correlation between treatment choice and the IVs across the quintiles. Among the physicians who are most likely to give prescriptions for controllers, the number of patients who are getting controller medication is almost two times higher. However, the likelihood of being prescribed controller medications was similar across the quintiles of copay ratios.

First, we used only controller/reliever copay ratio as an IV (see Table 3). Shea's partial R-square was very small for this equation. The Cragg–Donald statistic failed to reject its null hypothesis of underidentification. The Anderson canonical correlation failed to reject its null hypothesis at the 10% level, suggesting that the instrument may be inadequate to identify the equation. Second, we used only physician/practice prescribing patterns as an IV. Shea's partial R-square was 0.58 for this equation. The Cragg–Donald statistic rejected its null hypothesis of underidentification. The Anderson canonical correlation rejected its null hypothesis, suggesting that the instrument was adequate to identify the equation. We also attempted both instruments at the same time, but the Hall–Peixe test showed that ratio of copays as an instrument was redundant. The Hausman–Taylor

Table 3 Testing the strength of the instruments

Instruments	Shea's R^2	Partial R^2	F-stat	P-value	Cragg–Donald (P-value)	Anderson's test (P-value)
IV-1	0.0023	0.0023	1.78	0.1563	2.79 (0.145)	2.57 (0.165)
IV-2	0.58	0.58	18.65	0	28.54 (0.000)	28.57 (0.000)

IV, instrumental variable.

Table 4 First stage estimation using probit regression

	With IV-1		With IV-2		With IV-1 and IV-2	
Variables	Coefficient	STD	Coefficient	STD	Coefficient	STD
Explanatory Variables						
Age	0.01	0.00	0.01	0.00	0.0128	0.00
% Female	0.02	0.02	0.02	0.02	0.0205	0.02
Race						
White	0.04	0.22	0.09	0.21	0.0663	0.22
Black	–0.17	0.25	–0.13	0.24	–0.1547	0.25
Other	–0.57	0.74	–0.53	0.74	–0.5517	0.74
Geographic Regions						
Northeast	–0.04	0.09	–0.04	0.09	–0.0377	0.09
North-Central	–0.29	0.05	–0.28	0.05	–0.2753	0.05
West	–0.20	0.08	–0.21	0.08	–0.2033	0.08
Year of Patient Identification						
1996	0.09	0.03	0.06	0.03	0.0645	0.03
1998	–0.02	0.03	0.00	0.03	–0.0047	0.03
1999	0.16	0.09	0.20	0.09	0.1898	0.09
Member Type						
Employee	0.11	0.07	0.14	0.07	0.1367	0.07
Spouse	0.12	0.07	0.14	0.08	0.1426	0.08
Dependents						
12–18 Years	0.05	0.05	0.06	0.05	0.0591	0.05
Others	0.10	0.06	0.11	0.06	0.1100	0.06
County Mean Household Income per $10,000	0.12	0.02	0.12	0.02	0.12	0.02
Number of Major Diagnosis Categories	0.03	0.01	0.03	0.01	0.0255	0.01
Charlson Comorbidity Index	0.16	0.01	0.16	0.01	0.1622	0.01
Asthma-Specific Comorbidities						
Allergic Rhinitis	0.33	0.03	0.33	0.03	0.3320	0.03
Migraine	0.05	0.05	0.05	0.05	0.0507	0.05
Depression	–0.08	0.04	–0.07	0.04	–0.0744	0.04
Gastrointestinal Disorder	0.00	0.03	0.00	0.03	–0.0017	0.03
Sinusitis	0.02	0.03	0.02	0.03	0.0180	0.03
Anxiety	–0.06	0.07	–0.07	0.07	–0.0654	0.07
IVs						
Controller/Reliever Copayment	0.08	0.06	N/A	N/A	0.0286	0.07
Tax Provider ID controller %	N/A	N/A	0.72	0.33	0.6340	0.38

ID, identification number; IV, instrumental variable; STD, standard deviation.

test showed the rejection of controller–reliever copay ratio as an adequate instrument [53].

Our estimation method has two stages. We first estimated the likelihood of prescribing controller medication as a function of the exogenous covariates and IVs via conventional probit analysis. The results are presented in Table 4. For the second stage outcome (health-care expenditures) regression, following the Park test as described by Manning and Mullahy, we chose generalized linear models (GLMs) with log links and gamma family [54]. We regressed the total expenditure on treatment choice, exogenous covariates, and the first-stage probit residual as explanatory variables [55]. The results are presented in Table 5.

In the outcomes tables (Table 6), we compared raw outcomes between the patient groups that used controller medication only and reliever medication only, then we used the standard regression technique to adjust the raw outcome differences for observable differences in demographic and comorbid diseases characteristics between these two groups. These estimates were then compared with three IV estimators: one with weak IV, one with strong IV, and one with using both of them as an IV (one being a redundant IV).

The predicted cost differences were similar between standard regression and IV regression with weak instruments ($260 vs. $270). However, standard errors of the IV estimator increased almost 10-fold. Therefore, the differences were insignificant according to IV estimation with a weak instrument. However, using the right IV, with a strong relationship between treatment variables, there was a significant relationship between the health-care cost of the controller-only user group and reliever-only user group ($894, $P = 0.000$). Because the coefficient on treatment choice in standard GLM regression is positive, downward bias can explain the positive relationship between the unobserved severity level and reliever medication use.

Discussion

A widely recognized problem in observational research is that because of unobservable differences between individuals, it is unclear to what extent differences in outcomes reflect differences in treatment choices, even if we follow standard risk adjustment models, such as regression analysis or propensity score matching.

IV approach is a novel method to control for both observed and unobserved differences between individuals. However, this method is based on two strong assumptions and ignoring those assumptions can result in severe bias and inefficiency of the estimators.

A valid IV, which helps determine whether an individual is treated but does not determine other factors that affect outcome of interest, can overcome using the method of OLS. Current literature clarifies how to interpret estimated treatment effects using IVs. Because it is not possible to estimate the treatment effect for each individual, researchers rely on average treatment effect (ATE), which is the average of the individual treatment effects across the whole population of interest. When the treatment being evaluated has the same effect for everyone, any valid instrument will identify the ATE. However, when responses to treatment vary, different instruments measure different effects. Under this more realistic assumption, the only effect we can be

Table 5 Second stage estimation using generalized linear model with log link and gamma family

	No IV		With IV-1		With IV-2		With IV-1 and IV-2	
Variables	Coefficient	STD	Coefficient	STD	Coefficient	STD	Coefficient	STD
Explanatory Variables								
Age	0.02	0.00	0.02	0.00	0.01	0.00	0.01	0.00
(%) Female	−0.07	0.03	−0.06	0.03	−0.11	0.02	−0.11	0.02
Race								
White	−0.41	0.25	−0.40	0.25	−0.19	0.25	−0.18	0.25
Black	−0.45	0.28	−0.46	0.28	−0.72	0.27	−0.72	0.27
Other	−0.72	0.89	−0.75	0.89	−0.15	0.64	−0.16	0.64
Geographic Regions								
Northeast	0.10	0.11	0.10	0.11	0.19	0.08	0.19	0.08
North-Central	−0.12	0.05	−0.14	0.06	−0.25	0.05	−0.25	0.05
West	−0.19	0.09	−0.20	0.09	−0.01	0.08	−0.02	0.08
Year of Patient Identification								
1996	−0.05	0.03	−0.05	0.03	−0.14	0.03	−0.14	0.03
1998	0.02	0.03	0.02	0.03	−0.09	0.03	−0.09	0.03
1999	0.05	0.10	0.06	0.11	−0.23	0.13	−0.23	0.13
Member Type								
Employee	0.10	0.08	0.10	0.08	0.30	0.10	0.30	0.10
Spouse	0.14	0.08	0.14	0.08	0.43	0.11	0.43	0.11
Dependents								
12–18 years	0.21	0.05	0.21	0.05	0.22	0.10	0.22	0.10
Others	0.37	0.07	0.37	0.07	0.96	0.10	0.96	0.10
County Mean Household Income per $10,000	−0.01	0.02	−0.01	0.03	0.06	0.02	0.06	0.02
Number of Major Diagnosis Categories	0.22	0.01	0.22	0.01	0.18	0.01	0.18	0.01
Charlson Comorbidity Index	0.12	0.01	0.13	0.02	0.12	0.01	0.12	0.01
Asthma-Specific Comorbidities								
Allergic Rhinitis	0.04	0.03	0.07	0.04	−0.17	0.04	−0.17	0.04
Migraine	0.08	0.06	0.08	0.06	0.01	0.03	0.01	0.03
Depression	0.34	0.05	0.34	0.05	0.12	0.03	0.12	0.03
Gastrointestinal Disorder	0.17	0.03	0.17	0.03	0.19	0.03	0.19	0.03
Sinusitis	0.10	0.03	0.10	0.03	0.11	0.02	0.11	0.02
Anxiety	0.24	0.09	0.24	0.09	0.07	0.05	0.07	0.05
Treatment Indicator								
Controller	0.11	0.03	0.12	0.29	0.38	0.28	0.25	0.30
First Stage Residual								
Residual	N/A	N/A	−0.18	0.12	−0.02	0.12	−0.12	0.11

Standard errors are adjusted for first stage estimation.
IV, instrumental variable; STD, standard deviation.

sure the IV method estimates is the ATE among those who alter their treatment status because they react to the instrument. This is called local average treatment effect (LATE). When patients do not make decisions to reach the instrument based on factors that also determine treatment gains, the LATE equals the ATE among those exposed to the treatment.

In our application, the consequence of instruments with little explanatory power is increasing bias in the estimated IV coefficients and worsening the large sample approximations to the finite sample distributions. With weak instruments, the large sample bias of the IV estimator is the same as that of the OLS estimator, and IV becomes inconsistent and nothing is gained from instrumenting. One recommendation when faced with a weak instrument is to be parsimonious in the choice of instruments because if we use redundant instruments, even for the cases where the identification is not a problem, final estimates are inefficient.

Table 6 Comparison of differences in outcome measures between the controller-only users and reliever-only users (standard errors are in parentheses)

Cost	Difference
Unadjusted	$1,471 ($114)
OLS Estimate	$260 ($75)
IV Estimate with Weak IV	$270 ($613)
IV Estimate with Strong IV	$894 ($611)
IV Estimate with Redundant IV	$601 (618)

IV, instrumental variable; OLS, ordinary least squares.

In the past, ingenious instruments have been proposed and methods produced closer to "true" estimates than standard risk adjustment models. In a recent article, Stukel compared four analytic methods to remove the effects of selection bias in observational studies: multivariable model risk adjustment, propensity score risk adjustment, propensity-based matching, and IV analysis [56]. She concluded that the IV method produced closer results to the results from RCTs, which balances both measurable and immeasurable factors.

If we go back to our original question and answer: Do we have a method to control for both observed and unobserved bias? The answer is "theoretically YES" but practical application is very limited because of the difficulty in finding the right instrument. Researchers should test whether their instruments satisfy the two key assumptions. Application of the instrument without prior tests may produce inconsistent and inefficient results, which is worse than applying simple OLS or propensity score matching. Therefore, the challenge for outcomes researchers remains to find and adopt the right instruments for outcomes research studies. Otherwise, the cure can be worse than the disease.

The empirical work for this article was completed when the author was an employee of Thomson-Reuters and presented at the ISPOR conference.

Source of financial support: None.

References

1 Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. Value Health 2006;9:377–85.
2 Baser O. Choosing propensity score matching over regression adjustment for causal inference: when, why and how it makes sense. J Med Econ 2007;10:379–91.
3 McClellan M. Uncertainty, health-care technologies, and healthcare choices. Am Econ Rev 1995;38–44.
4 Stürmer T, Joshi M, Glynn RJ, et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol 2006;59:437.e1–e24.
5 Heckman JJ. Econometric causality. Int Stat Rev 2008;76:1–27.
6 Ashenfelter O. Estimating the effect of training programs on earnings. Rev Econ Stat 1978;60:47–57.
7 Ashenfelter O, Card D. Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. Working Papers 554, Princeton University, Department of Economics, Industrial Relations Section, 1984.
8 Heckman J, Robb R. Alternative methods for evaluating the impact of interventions: an overview. J Econom 1985;30:239–67.
9 LaLonde R. Evaluating the econometric evaluations of training programs with experimental data. Am Econ Rev 1986;76:604–20.
10 Fraker T, Maynard R. The adequacy of comparison group designs for evaluations of employment-related programs. J Hum Resour 1987;22:194–227.
11 Card D, Sullivan DG. Measuring the effect of subsidized training programs on movements in and out of employment. Econometrica 1988;56:497–530.
12 Manski CF, Institute for Research on Poverty, University of Wisconsin–Madison. Identification of Endogenous Social Effects: The Reflection Problem. Madison, WI: University of Wisconsin–Madison, Institute for Research on Poverty, 1993.
13 Imbens G, Wooldridge J. Recent Developments in the Econometrics of Program Evaluation. J Econ Lit 2009;47:5–86.
14 Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. Annu Rev Public Health 1998;19:17–34.
15 Clever SL, Jin L, Levinson W, Meltzer DO. Does doctor-patient communication affect patient satisfaction with hospital care? Results of an analysis with a novel instrumental variable. Health Serv Res 2008;43:1505–19.
16 Johnston MK, Gustafson P, Levy AR, Grootendorst P. Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to empidemiological research. Stat Med 2008;27:1539–56.
17 Grootendorst P. A review of instrumental variables estimation of treatment effects in the applied health sciences. Health Serv Outcomes Res Methodol 2007;7:159–79.
18 Basu A, Heckman JJ, Navarro-Lozano S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. J Health Econ (Chichester) 2007;16:1133.
19 Eisenberg D, Quinn BC. Estimating the effect of smoking cessation on weight gain: an instrumental variable approach. Health Serv Res 2006;41:2255–66.
20 Wooldridge JM. Introductory Econometrics: A Modern Approach (3rd ed.). Mason, OH: Thomson/South-Western, 2006.
21 Wooldridge JM. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press, 2002.
22 Bowden RJ, Turkington DA. Instrumental Variables. Cambridge: Cambridge University Press, 1984.
23 Kelejian HH, Prucha IR. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. J Real Estate Finance Econ 1998;17:99–121.
24 Amemiya T. The nonlinear two-stage least-squares estimator. J Econom 1974;2:105–10.
25 Rivers D, Vuong QH. Limited information estimators and exogeneity tests for simultaneous probit models. J Econom 1988;39:347–66.
26 Mátyás L. Generalized Method of Moments Estimation. Cambridge: Cambridge University Press, 1999.
27 Dow WH, Norton EC. Choosing between and interpreting the heckit and two-part models for corner solutions. Health Serv Outcomes Res Methodol 2003;4:5–18.
28 Fisher ES, Wennberg DE, Stukel TA, et al. The implications of regional variations in medicare spending. Part 1: the content, quality, and accessibility of care. Ann Intern Med 2003;138:273–87.
29 Altonji JG, Elder TE, Taber C. An evaluation of instrumental variable strategies for estimating the effects of catholic schools. J Human Res 2005;4:791–821.
30 Rodrik D. Getting institutions right. CESifo DICE Rep 2004;2:10–15.
31 Miguel E, Satyanath S, Sergenti E. Economic shocks and civil conflict: an instrumental variables approach. J Polit Econ 2004;112:725–53.
32 Fisher ES, Wennberg DE, Stukel TA, et al. The implications of regional variations in medicare spending. Part 2: health outcomes and satisfaction with care. Ann Intern Med 2003;138:288–98.
33 Witzke H. Determinants of the US wheat producer support price: do presidential elections matter? Public Choice 1990;64:155–65.
34 Revelli F. Local taxes, national politics and spatial interactions in English district election results. Eur J Polit Econ 2002;18:281–99.
35 Breitung J, Meyer W. Testing for unit roots in panel data: are wages on different bargaining levels cointegrated? Appl Econ 1994;26:353–61.
36 Angrist JD, Krueger AB. Instrumental variables and the search for identification: from supply and demand to natural experiments. J Econ Perspect 15:69–85.
37 Hansen LP, Singleton KJ. Generalized instrumental variables estimation of nonlinear rational expectations models. Econometrica 1982;50:1269–86.
38 Howard D. The impact of waiting time on liver transplant outcomes. Health Serv Res. 2000;35(5 Pt 2):1117.
39 Sacerdote B. Peer effects with random assignment: results for Dartmouth roommates. Q J Econ 2001;116:681–704.
40 Miller WB, Pasta DJ. The psychology of child timing: a measurement instrument and a Model 1. J Appl Soc Psychol 1994;24:218–50.
41 Hedges LV. Modeling publication selection effects in meta-analysis. Stat Sci 1992;7:246–55.
42 Kinal TW. The existence of moments of k-class estimators. Econometrica 1980;48:241–50.
43 Staiger D, Stock JH. Instrumental variables regression with weak instruments. Econometrica (Evanston, Ill) 1997;65:557–86.
44 Chao JC, Swanson NR. Consistent estimation with a large number of weak instruments. Econometrica 2005;73:1673–92.
45 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.
46 Shea J. Instrument relevance in multivariate linear models: a simple measure. Rev Econ Stat 1997;79:348–52.
47 Anderson TW. An Introduction to Multivariate Analysis. New York: Wiley, 1958.
48 Hall AR, Peixe FPM. A consistent method for the selection of relevant instruments. Econom Rev 2003;22:269–87.
49 Cragg JG, Donald SG. Inferring the rank of a matrix. J Econom 1997;76:223–50.
50 Stock JH, Yogo M. Testing for Weak Instruments in Linear IV Regression. NBER Working Paper T0284, 2002.
51 Crown WH, Berndt ER, Baser O, et al. Benefit plan design and prescription drug utilization among asthmatics: do patient copayments matter? Front Health Policy Res 2004;7:95–127.
52 Hausman J. Specification tests in econometrics. Econometrica 1978;46:1251–71.
53 Hausman J, Taylor W. Panel data and unobservable individual effects. Econometrica 1981;49:1377–98.

54 Manning WG, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ 2001;20:461–94.
55 Shea D, Terza J, Stuart B, Briesacher B. Estimating the effects of prescription drug coverage for medicare beneficiaries. Health Serv Res 2007;42(3 Pt 1):933.
56 Stukel TA, Fisher ES, Wennberg DE, et al. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007;297: 278–85.