doi:10.1111/j.1365-2753.2006.00615.x

Evaluating disease management programme effectiveness: an introduction to instrumental variables

Ariel Linden DrPH Ms[1,2] and John L. Adams PhD[3]
[1]President, Linden Consulting Group, Portland, OR, USA
[2]Clinical Assistant Professor, Department of Preventive Health/Preventive Medicine, School of Medicine, Oregon Health and Science University, Portland, OR, USA
[3]Senior Statistician, RAND Corporation, Santa Monica, CA, USA

Correspondence
Ariel Linden
Linden Consulting Group
6208 NE Chestnut Street
Hillsboro OR 97124
USA
E-mail: alinden@lindenconsulting.org

Keywords: disease management, instrumental variables, observational studies, unobserved covariates

Accepted for publication:
22 March 2005

Abstract

This paper introduces the concept of instrumental variables (IVs) as a means of providing an unbiased estimate of treatment effects in evaluating disease management (DM) programme effectiveness. Model development is described using zip codes as the IV. Three diabetes DM outcomes were evaluated: annual diabetes costs, emergency department (ED) visits and hospital days. Both ordinary least squares (OLS) and IV estimates showed a significant treatment effect for diabetes costs ($P = 0.011$) but neither model produced a significant treatment effect for ED visits. However, the IV estimate showed a significant treatment effect for hospital days ($P = 0.006$) whereas the OLS model did not. These results illustrate the utility of IV estimation when the OLS model is sensitive to the confounding effect of hidden bias.

Introduction

Disease management (DM), as defined by the Disease Management Association of America (2004) is a system of coordinated interventions and communications for populations with conditions in which patient self-care efforts are significant. DM programmes were initially developed under the assumption that by augmenting the traditional episodic medical care system with services and support between doctor visits, the overall cost of health care could be reduced. For many chronic diseases, such as diabetes, asthma and congestive heart failure, there is much opportunity to improve the quality and consistency of care. DM is meant to assist doctors and their patients in identifying and closing those gaps in care.

DM programmes are typically evaluated using observational study designs that are susceptible to various biases that threaten the validity of study findings (Linden *et al*. 2003). Most biases in DM are rooted in allowing individuals to self-select into the programme. As a result, the evaluator is unable to differentiate between a true programme effect and the impact of unobserved differences in characteristics between participants and non-participants that may have led to the differences noted between study groups in outcomes. Recently the propensity scoring technique (Linden *et al*. 2005a), coupled with a sensitivity analysis (Linden *et al*. 2005b) has been suggested as a suitable means of estimating the magnitude and mitigating the effects of these unobserved characteristics in DM effectiveness studies.

This paper introduces the concept of instrumental variables (IVs) as another approach to providing an unbiased estimate of a DM programme treatment effect. An IV model well suited to DM was developed by the authors and will be presented with discussion so that this technique can be easily replicated in DM programme evaluations. For those organizations that purchase DM services, this paper will provide a substantive background with which to discuss the inclusion of IVs as an alternative evaluation possibility with their contracted vendors.

Instrumental variables

The concept of IV was borne out of the *structural equation modelling* literature and has been an integral component in the field of econometrics since the 1920s (Wright 1928). In simple terms, an IV is a variable (*Z*) that is correlated with the DM programme intervention (*X*), but not associated with unobserved confounders of programme outcome (*Y*). Therefore, *Z* can only impact *Y* through *X*. A consequence of this requirement is that *Z* not be correlated with any unobserved covariates (*U*) that affect the relationship between *X* and *Y*. Figure 1 shows these relationships using a DM programme model. Note that all elements below the line are components of any standard observational study design that may be biased by the influence of *U*. The introduction of *Z* however, provides an unbiased estimate of the causal effect of *X* on *Y* (by remaining independent of *U* or *Y*, *Z* controls the effect of *U* on the relationship of *X* and *Y*). The logic for this is similar to that of a random controlled trial (Linden *et al*. 2005c). As indicated earlier, the application of the IV is by way of a structural equation model that uses a two-stage regression analysis as follows (Angrist *et al*. 1996):

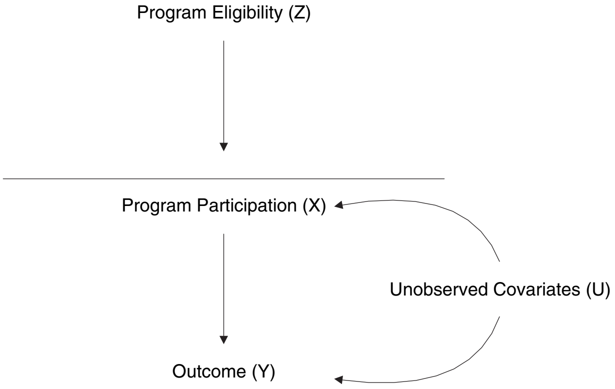

Figure 1 An illustration of the various elements of a study design that incorporates the use of an instrumental variable (IV). *Z* (the IV) can impact *Y* only through *X*. Hence, it is possible to estimate the causal relationship of *X* to *Y* in relation to *Z*.

$$\hat{X} = \alpha_0 + \alpha_1 \times Z_i + v_i \qquad (1)$$

$$Y = \beta_0 + \beta_1 \times \hat{X}_i + \varepsilon_i \qquad (2)$$

In equation (1), Z represents the IV that is used to estimate $\hat{X}$. $\hat{X}$ (the predicted value of X) is then 'plugged-in' to equation (2) instead of the actual X variable. To bring more meaning to this set of equations, we assume that Z is the IV found to be a significant predictor of who is likely to enrol in the DM programme $\hat{X}$. If the individual's actual X was used (indicating whether the individual actually enrolled or not) the result may be confounded by selection bias. However, using $\hat{X}$ given Z allows for an unbiased estimate of the impact of programme enrolment on an outcome (e.g. hospitalization, cost, etc.), because Z predicts X in equation (1) but remains independent of the X–Y relationship (here we assume that Z is uncorrelated with the unobserved covariates between X and Y). It is useful to think of equation (1) as 'purging' X of potentially confounding influences. Both of these equations may, and typically do, include other covariates.

Methods

The data used in this analysis represent a 1-year experience of a diabetes DM programme. The programme utilized an 'opt-out' enrolment process in which all eligible individuals were automatically included in the programme and those who chose not to participate had to specifically request to be excluded from the programme. In total, 1952 participants were continuously enrolled in the programme for the entire year. A control group was comprised of the 582 remaining diabetics in the population who chose not to participate in the programme but remained insured with the health plan throughout the entire programme period. The fact that the control group was only one-quarter the size of the programme group is another interesting and realistic phenomenon facing DM programme evaluators. As DM is a population-based programme, it is likely that fewer concurrent controls will be available for comparison or matching (as most eligible persons will have enrolled in the programme). This will require researchers to be thoughtful in their approach to evaluating these programmes' effectiveness.

Zip codes as instrumental variables

Finding a variable in DM that meets the specific criteria to be used as an IV is almost an impossible task. One must consider (i) what variable is predictive of an individual's enrolment in the DM programme, but (ii) is not associated with any of the potential unobserved covariates that influence that outcome. Moreover, given the limited data sets available (usually only claims and enrolment files are accessible) there are few variables to choose from.

We hypothesized that zip codes would make good instruments because: (i) an individual living in a DM covered service area would make them eligible for programme participation (assuming they met all the diagnosis and insurance benefit criteria), but not necessarily ensure that they would enrol, and (ii) living in a given zip code area may be independent of specific unobserved covariates (had we used a simple variable indicating whether a person lived in a rural or urban setting, we may have violated this assumption, as it has been demonstrated that people living in rural areas have less access to care and usually have poorer health outcomes than those living in urban centres) (Adams *et al*. 2001; Vargas *et al*. 2003; Glover *et al*. 2004; Zulkowski & Coon 2004). To further clarify the underlying assumptions of this model, we assume two zip codes: Zip Code-1 (high programme enrolment rate) and Zip Code-2 (low programme enrolment rate). The natural experiment that this instrument tries to model is how the difference in outcomes of these two zip codes is related to the difference in participation rates. Although the assumption is not testable, the hope is that within a given zip code the correlation between unmeasured confounders and programme participation is much smaller than that relationship between zip codes. Part of the motivation for using zip codes as an IV is the belief that geographic proximity will make participants and non-participants more similar on unmeasured confounders. This is certainly true for measured demographics. Therefore, the effect of zip code on a given outcome (e.g. hospitalization) is indirect, interceded by the probability of programme enrolment.

Six hundred and sixty-seven unique five-digit zip codes were identified for the given population. By collapsing these based on the first three digits of the zip code (which provided a contiguous expanded geographical area), 71 new categories of zip-code groups were established. Upon further scrutiny, it was determined that a category called 'other' should be created to include all zip codes with frequency counts less than nine. Figure 2 shows the frequency of the 25 three-digit zip-code category used as IVs.

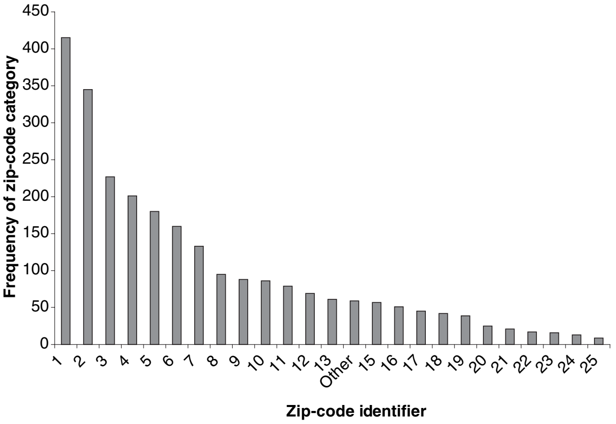

Figure 2 Frequency of three-digit zip-code categories used as instrumental variables.

First stage regression

A logistic regression equation (Cox 1970, 1972) was constructed using 24 dummy variables to represent the 25 three-digit zip-code categories. Other exogenous variables included: age (years), gender (1 = female, 0 = male) and a risk score (0.0–1.0). The risk score is a predicted value derived from a logistic regression model that includes covariates divided into four areas including: demographic, utilization, clinical and financial. The outcome variable was the probability of a given individual enrolling in the programme. This model was estimated based on the actual participation status, denoted as a dichotomous variable with 1 indicating programme participation and 0 designated as control.

The overall model was significant (log ratio $\chi^2 = 676.92$, DF = 24, $P < 0.0001$). Additionally a modified R^2 developed for logistic regression called the McFadden $R^2 = 0.25$, indicating that the model fit the data adequately. Somers' D and Goodman–Kruskal gamma measures were 0.61 and 0.65, respectively, indicating that the model has very good predictive ability.

Table 1 A comparison of group characteristics based on actual and predicted programme enrolment status

	Group	n	*Age (SE)*	*% Female (SE)*
Actual				
	Control	582	51.4 (0.22)	44 (1.0)
	Programme	1952	51.8 (0.24)	45 (1.0)
Predicted				
	Control	505	52.0 (0.24)	44 (1.0)
	Programme	2029	51.6 (0.23)	45 (1.0)

Table 1 compares characteristics of the programme participant and control groups, using the actual enrolment status and the predicted enrolment status (based on the first stage regression). As shown, the predicted enrolment model compared favourably to the actual enrolment data. While no significant differences were noted in the age or gender distributions, the predicted model tended to slightly over-predict programme participation. These results satisfy the assumption that Z must be associated with X.

Second stage regression

Ordinary least squares (OLS) regression models were constructed to estimate the impact of programme participation on three different outcome variables: (i) programme year hospital days, (ii) programme year emergency department (ED) visits and (iii) programme year diabetes-related medical costs (excluding pharmacy). Independent variables included those exogenous variables from the first stage regression: age, gender, risk score, as well as the 'plug-in' or predicted value of programme participation.

Each model was estimated twice, once using the actual programme participation status, and a second time using the predicted participation status from the first stage regression. Running OLS twice, using the predicted value from the first stage as a plug-in variable in the second stage results in incorrect estimations of the residual sum of squares and their standard errors. Therefore, it is important that these analyses be performed using a two-stage least squares (2SLS) model that is available in most statistical and econometric software packages.

Results

Table 2 provides the OLS and IV estimates for the programme effect on the three outcome variables (cost, ED visits and hospital days).

Programme participation (denoted as 'Group') appeared to have a significant effect on costs ($P = 0.011$ and $P = 0.012$, in the OLS and IV models respectively). While the level of significance was similar in both models the magnitude of the cost reduction in the IV was nearly double that in the OLS ($2328 vs. $1288, for the IV and OLS respectively). As expected, the risk score also appeared to significantly impact costs ($P < 0.0001$ for both OLS and IV models). Somewhat surprising however, was the magnitude of the effect. A one-unit increase in the risk score was associated with an increase of nearly $36 000 per diabetic member during the programme year.

There appeared to be no significant programme effect on ED visits, under either model estimation ($P = 0.664$ and $P = 0.594$, for the OLS and IV models respectively). However, the coefficients changed directions from indicating a 0.003 increase in ED visits under the OLS to a 0.007 decrease in ED visits under the IV. Also of interest, age was the only covariate having a significant effect on ED visits ($P < 0.0001$).

Programme participation did not appear to impact hospital stays under the OLS model ($P = 0.081$); however, participation was shown to affect hospital stays significantly in the IV estimate ($P = 0.006$). The magnitude in the reduction of hospital stays also greatly differed between the models as illustrated by the coefficients (–0.231 vs. 0.678, for the OLS and IV respectively. Risk scores were significant predictors of hospital days ($P = 0.0001$), with a one-unit increase in score associated with roughly 5.5 extra hospital days per year.

The adjusted R^2 for the total cost model was 9.4% and for the hospital stays model was 3.4–3.0% (OLS and IV respectively). These values are similar to what is generally seen in the health services research literature. All model parameters were tested for co-linearity but none was detected.

To test the assumption that the IV estimates were not directly associated with any of the three outcome measures, residuals from each equation were regressed on the zip code instruments. No direct relationships were evident for any model ($P = 0.127$,

Table 2 Ordinary least squares (OLS) and instrumental variable (IV) estimates of the disease management programme effect on three outcomes: programme year total diabetes-related costs, programme year emergency department (ED) visits and programme year hospital days

		OLS			*IV*		
	Variable	*Coefficient*	*SE*	*Significance (P)*	*Coefficient*	*SE*	*Significance (P)*
Total diabetes costs							
	Intercept	2 329.966	1054.468	0.027	3101.28	1204.09	0.010
	Group	–1288.412	503.856	0.011	–2328.93	930.52	0.012
	Female	–559.134	428.535	0.192	–543.99	429.05	0.205
	Risk score	35 886.787	2284.228	<0.0001	35 652.06	2292.95	<0.0001
	Age	25.878	18.235	0.156	26.71	18.26	0.144
	Adjusted R^2	0.095			0.094		
	Hausman *F*-test						0.183
ED visits							
	Intercept	0.070	0.015	<0.0001	0.077	0.016	<0.0001
	Group	0.003	0.007	0.664	–0.007	0.013	0.594
	Female	0.001	0.006	0.915	0.0008	0.006	0.896
	Risk score	0.029	0.032	0.358	0.027	0.032	0.353
	Age	–0.001	0.000	<0.0001	–0.001	0.000	<0.0001
	Adjusted R^2	0.006			0.006		
	Hausman *F*-test						0.360
Hospital days							
	Intercept	0.325	0.277	0.241	0.656	0.317	0.038
	Group	–0.231	0.132	0.081	–0.678	0.245	0.006
	Female	–0.110	0.113	0.329	–0.104	0.113	0.359
	Risk score	5.581	0.600	<0.0001	5.480	0.604	<0.0001
	Age	0.002	0.005	0.670	0.002	0.005	0.618
	Adjusted R^2	0.034			0.030		
	Hausman *F*-test						0.029

$P = 0.812$ and $P = 0.209$, for the cost, ED and hospital day models respectively).

The Hausman *F*-test (Hausman 1978) was performed on all three outcome models to assess whether the results obtained between the OLS and IV estimates were significantly different. If not, then typically the OLS model becomes the default as the estimates are more efficient in terms of coefficients and SE. In both the cost and ED visit models, the Hausman test failed to find differences between the OLS and IV models. However, the hospital days model showed a significant difference ($P = 0.029$) between the OLS and IV estimates.

Discussion

The results of these analyses provide excellent examples of the various possible outcomes of using the IV method in comparison to the standard OLS model to assess treatment effect. In relation to annual costs, both the OLS and IV models showed a significant treatment effect of programme participation. In this case, one might conclude that hidden bias was not a factor that influenced costs, as the IV method did not add much to results. The magnitude of the cost savings estimate in the IV method was nearly double that in the OLS. However, the Hausman test indicated that the estimate between the two models was not significantly different, and thus we default to the results produced by the OLS model.

In the case of ED visits, the IV estimate did not improve the results of the OLS. In both cases programme participation did not impact ED visit rates significantly. However, an interesting finding was that in the OLS model, being a programme participant increased ED visits by 0.003 while the IV model showed a more intuitive result of a decrease in ED visits by 0.007 associated with programme participation. While these results were not statistically significant, the direction of change noted in the IV estimate would lead us to believe that this model was more appropriate than the OLS estimation. However, since the Hausman test showed that there was no statistically significant difference between the model estimates, the determination of which model should be used is left to the judgment of the analyst. In this scenario, the decrease in ED visits estimated by the IV model makes intuitive sense.

In the case of yearly hospital days, the IV model showed a significant treatment effect while the OLS model did not. It appears that there may have been unobserved biases impacting outcomes in the OLS model that are controlled using the unbiased estimate of the IV model.

As expected, the risk score was a significant contributor to both the cost and hospital stay models. The predictive model incorporates up to 150 independent variables as a means of explaining the following year's cost. As hospital stays are the primary drivers of cost, we would expect to see this variable reach significance level in these two models. However, it was not a significant predictor of ED visits. This finding has been noted in other studies as well (Linden *et al*. 2005a, 2005b). The Hausman test was significant ($P = 0.029$) and thereby supported the other differences noted between the IV and OLS estimates for this model.

While not performed in the current study, sensitivity analyses can be provided as a means of estimating bounds around X. For a comprehensive discussion on estimating bounds, the reader is referred to studies by Manski (1990), by Heckman & Vytlacil (1999) and by Linden *et al*. (2005c).

Limitations in the use of the IV method

While the IV method presents an excellent alternative, providing an unbiased estimate of DM programme effect, there are several limitations that may impede its widespread adoption in DM programme evaluation. First, as suggested earlier, finding suitable IVs is problematic in DM because of the limited data available for analysis and the general structure of DM programmes. For example, McClellan *et al*. (1994) in one of the few health care-related studies using IVs, used patients' differential distance to the hospital as an IV to assess whether intensive treatment after acute myocardial infarction reduces mortality. A measure of distance would not work as an IV in DM because most programme interventions rely on telephonic, Internet or postal-based communications with participants. However, in this study zip codes were used successfully as an IV. That said, zip codes may not be generalizable to other programmes.

Second, IV estimates will only work if certain assumptions are met. However, some of these assumptions do not lend themselves to testing and thus it may never be known if those assumptions were violated. For example, the assumption that Z is not correlated with unobserved covariates U cannot be definitively proven. Another assumption is that one person's outcome is not related to another person's treatment assignment (Rubin 1990). Theoretically this assumption could be violated if two members of a given household had the same diagnoses (e.g. diabetes) but one was participating in the programme while the second was not. It is possible that the non-participant would change behaviour that would impact outcomes based on their observation of the intervention that the other house member was receiving.

Most candidate IVs are subject to some criticism. The criticism here would be that somehow patients with higher or lower levels of unmeasured severity choose to live in certain zip codes and that this pattern would not be explained by their risk scores. But the possibility that unmeasured severity is related to programme participation in a way that is not explained by risk scores is a much more immediate threat to the validity of other methods. We suggest that IV method with whatever instruments available can provide some insight into these potential sources of bias even if IVs are less than perfect. There has been an evolution in the interpretation of IV model in the econometric literature. One emerging way to think about an IV like ours is that it is the average of a large collection of small studies comparing the results from one zip code to another.

In the absence of randomization there are always potential unmeasured confounders that could invalidate an instrument. However, as assumptions made in IV cannot be tested, there is typically no empirical information to refute them. Criticism of the IV method is usually a criticism aimed at the use of OLS regression in general. However, even a potentially flawed instrument is worth considering if it is thought to be better than naïve analyses with known biases.

Conclusions

This paper introduced the concept of IVs as another approach to providing an unbiased estimate of a DM programme treatment effect. This method uses an IV model well-suited to DM using zip codes. This method is particularly suitable to DM programme analysis (where the use of randomized control groups is generally not practical) because IVs may reduce many of the biases typically inherent to observational studies, most notably selection bias. An important fact to keep in mind when using the IV method is that there are several assumptions that must be met in order for this method to produce valid estimates of a treatment effect. Therefore, any discussion on the results achieved through the analysis must describe how the assumptions were tested and if any violations were noted. Nonetheless, because of its simplicity and utility, IV estimates should be considered as an alternative procedure for use with current non-experimental designs in evaluating DM programme effectiveness.

References

Adams C.E., Michel Y., DeFrates D. & Corbett C.F. (2001) Effect of locale on health status and direct care time of rural versus urban home health patients. *Journal of Nursing Administration* **31**, 244–251.

Angrist J.D., Imbens G.W. & Rubin D.B. (1996) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**, 444–455.

Cox D.R. (1970) *The Analysis of Binary Data*. Methuen, London.

Cox D.R. (1972) The analysis of multivariate binary data. *Applied Statistics* **21**, 113–120.

Disease Management Association of America (DMAA) (2004) *Definition of Disease Management.* Available at: http://www.dmaa.org/definition.html (last accessed 23 June 2004).

Glover S., Moore C.G., Samuels M.E. & Probst J.C. (2004) Disparities in access to care among rural working-age adults. *Journal of Rural Health* **20**, 193–205.

Hausman J.A. (1978) Specification tests in econometrics. *Econometrica* **46**, 1251–1271.

Heckman J.J. & Vytlacil E.J. (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Science* **96**, 4730–4734.

Linden A., Adams J. & Roberts N. (2003) An assessment of the total population approach for evaluating disease management program effectiveness. *Disease Management* **6**, 93–102.

Linden A., Adams J. & Roberts N. (2005a) Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management and Health Outcomes* **13**, 107–127.

Linden A., Adams J. & Roberts N. (2005b) Strengthening the case for disease management effectiveness: unhiding the hidden bias. *Journal of Evaluation in Clinical Practice* doi:10.1111/j.1365-2753.2005.00612.x

Linden A., Adams J. & Roberts N. (2005c) Evaluating disease management program effectiveness: an introduction to the regression-discontinuity design. *Journal of Evaluation in Clinical Practice* doi:10.1111/j.1365-2753.2005.00573.x

McClellan M., McNeil B.J. & Newhouse J.P. (1994) Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *Journal of the American Medical Association* **272**, 859–866.

Manski C.F. (1990) Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings* **80**, 319–323.

Rubin D.B. (1990) Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* **5**, 472–480.

Vargas C.M., Yellowitz J.A. & Hayes K.L. (2003) Oral health status of older rural adults in the United States. *Journal of the American Dental Association* **134**, 479–486.

Wright S. (1928) Appendix. In *The Tariff on Animal and Vegetable Oils* (ed. P.G. Wright). Macmillan, New York.

Zulkowski. K. & Coon P.J. (2004) Comparison of nutritional risk between urban and rural elderly. *Ostomy Wound Management* **50**, 46–48, 50, 52 passim.