Appropriate Econometric Methods for Pharmacoeconometric Studies of Retrospective Claims Data: An Introductory Guide

The Gianfrancesco et al.[1] study in the April 2005 issue of *JMCP* provides a good opportunity to examine some of the data measurement and statistical issues that should be considered in studies evaluating differences in drug costs or other treatment outcomes using retrospective insurance claims data. Now that large health care claims databases are readily available from government programs (e.g., Medicare, Medicaid, Veterans Affairs), commercial insurers (e.g., WellPoint, Pharmetrics, I3 Magnifi, Medstat, etc.), and other sources, it is easy to "crunch the numbers" with extremely large national patient data samples and derive conclusions about treatment cost differences, often with very high levels of parameter significance. Whether these conclusions are valid and reliable depends on the robustness of the econometric methods used in the analysis.

Jerry Avorn's recent book *Powerful Medicines* eloquently and engagingly describes the strengths and limitations of retrospective data analysis for pharmacoepidemiology and pharmacoeconomics.[2] There have been a number of situations where retrospective data analyses have provided invaluable insights into drug treatment outcomes, including studies of phentermine-fenfluramine,[3] troglitazone,[4] and rofecoxib.[5,6] There are also situations where retrospective data analyses have failed to detect important confounders and have generated biased assessments of drug effects, most notably with estrogen replacement therapy.[7]

Data Measurement Issues

Data measurement accuracy is the foundation of appropriate statistical inference. One basic data measurement issue raised by Gianfrancesco et al. concerns the use of billed charges rather than amounts allowed by the third-party payers. Provider-billed charges may exceed amounts allowed by insurance plans by 60% or more.[8] Allowed amounts are a much more accurate measure of actual medical costs since they reflect the portion of the bill for which the payer and patient are contractually obligated. For insured patients, billed charges are essentially fictional since the insurance plan decides, based on plan characteristics and provider contracts, what the actual transaction amount will be. While there is a problem with the reliability of allowed amounts when a family has more than one source of insurance coverage, this problem is not resolved by simply using provider charges. Concern about capitated provider benefits or about patients with multiple family insurance plan coverage benefits can be better addressed by isolating such cases from the main patient data and analyzing them separately.

Second, adjusting actual drug claims into dose equivalent drug costs is highly misleading. Dose adjustments for age, gender, or other characteristics that influence how much medication patients receive should be accomplished with propensity score or instrumental variables estimation methods (see below) not by artificially altering the observed drug costs particularly since this is the key outcome variable in the analysis.

Econometric Methods

Econometric analysis is a highly technical field that is highly relevant to inferring the impact of drugs on health care treatment costs. This presentation is made in a relatively nontechnical fashion, to give the reader a flavor of the motivation behind these technical topics. While most details are skipped over for brevity, references to several relevant technical articles are provided.

Retrospective data analyses using administrative databases such as health insurance claims histories have a number of advantages over prospective randomized controlled trials (RCTs), particularly their larger sample sizes and lower data collection costs. There are many clinical issues that can only be addressed through retrospective data analysis because of time or budget constraints. Retrospective data also have higher external validity since patients and providers do not need to consent to join the study, are not generally aware that their decisions are under study, and thus act in a "real-world" fashion, which is often very hard to replicate in prospective RCTs.

Some clinical trialists pejoratively categorize all retrospective statistical analyses as "exploratory" or "hypothesis-generating" as opposed to RCTs that "reveal causation" and are "hypothesis-testing." However, each study design has its strengths and weaknesses, and both are equally valuable. Many types of patients are excluded from, or will never sign up for, RCTs. For ethical reasons no one is going to conduct an RCT on whether smoking causes cancer, stress causes cardiovascular disease, or on other interventions with significant perceived a priori risks. For financial reasons, innumerable other crucial medical decisions will never be subjected to prospective RCT study. As Avorn aptly points out, retrospective data analysis plays an important counterbalancing "yin" to the RCT "yang."

If a prospective RCT is well-executed, statistical analysis of the results is easy to prespecify and often is no more complex than an independent sample *t* test. On the other hand, appropriate statistical analysis of retrospective data, particularly for health care claims histories, can be exceedingly complicated and can tax the most sophisticated econometric methodologies. The obvious concern with retrospective data is precisely what distinguishes it from RCTs—patients are not randomized to treatment in retrospective claims data analyses; rather, they are assigned to treatment because their physician and/or other provider selected the chosen therapy based on any number of observable and unobservable factors, including (but not limited to) the patients' preferences, medical history, and other characteristics.

Unfortunately, most retrospective claims data are limited in the extent of availability of patient or provider characteristics to use for statistical adjustment. In particular, medical history information is generally limited to relatively crude patient disease severity measures such as the Charlson Index, Chronic Disease Score, or comorbidity indicators.[9] Even items that should be straightforward to calculate are often unavailable, such as the provider's typical prescribing patterns with similar patients. This means that, in most retrospective data analyses, important characteristics that are correlates of treatment choice, patient outcomes, and treatment costs are unobservable confounders. Such confounders will create biased estimates if ignored.

Researchers have understood for decades that such treatment selection bias needs to be accounted for in econometric estimates.[10,11] James Heckman shared the 2000 Nobel Prize in economics for his pathbreaking 1974 work on econometric methods to detect and correct for treatment selection bias.[12] More recently, sample selection bias estimation methods based on the propensity score[13,14] and instrumental variables techniques[15,16] have been further refined and generalized to improve precision and robustness.

Most of these methods start with an equation that estimates treatment choice as a binary (or multinomial) index function of observable explanatory variables, usually using probit (probability unit) or logit regression.[17] This fitted equation generates the "propensity score" for a patient to be assigned to a given treatment. The propensity score is simply an estimated probability, based on observable characteristics in the data, that a specific patient would receive the treatment in question. Propensity score methods are often used to match patients receiving one treatment with others at the same level of propensity, controlling for all observable explanatory factors, and to determine whether treatment costs are similar or different in these propensity-matched cohorts. Alternatively, the estimated propensity score, p_i, itself can be inc luded as an explanatory variable in a regression, with treatment costs or other patient outcomes as the dependent variable, and treatment assignment, w_i, as another explanatory factor. Extensions of this approach include adding additional explanatory variables consisting of various low-order polynomial and power terms of p_i recentered at the sample mean of p_i and interacted with w_i. Wooldridge outlines conditions under which such propensity score estimators are consistent and asymptotically efficient.[18]

The main problem with propensity score methods is that they are based on the assumption of "strong ignorability," which requires that, given the observable explanatory variable x, the estimated treatment effect is unbiased. As an example, in looking at treatment costs for lower back pain in a Michigan managed care plan, I found the drug-specific cost estimates to be very different, depending on whether or not one adjusted for the impact of pain severity on medication choice. If back pain severity truly impacts drug choice and one ignores this by leaving that important variable out of the estimation equation, the key assumption of the propensity score method would fail, and one would get a biased estimate of the drug treatment costs. If, on the other hand, one could be confident that back pain and other observed variables (e.g., age, gender, ethnicity, medical history), adequately explained drug choice, then the p ropensity score method would eliminate drug treatment selection bias from the cost estimates.

Because of the ignorability assumption, propensity score methods preclude the possibility that there are important unobservable characteristics (error components) that affect both treatment choice and treatment costs or outcomes after adjusting for all available explanatory factors. Since such an ignorability assumption is often untenable with retrospective claims data, it is usually preferable to use instrumental variables methods to adjust for treatment selection bias in this context. Instrumental variables methods often also start with the estimated propensity score, p_i, but they explicitly allow for unobservables that are correlated with treatment choice as well as with treatment costs or other outcomes.

An instrumental variable is any exogenous variable that is correlated with the choice of treatment but not correlated with the unobservables that impact treatment outcomes and treatment costs. Certainly, the estimated propensity score, p_i, fits such criteria since it is a function (e.g., a logit or probit regression function) only of observable exogenous factors that predict treatment choice. Low-order polynomial and power transformations of p_i are also valid instrumental variables. It is perfectly acceptable to have a larger number of instruments than potentially biased treatment effects in the estimation equation. Such cases are referred to as "over-identified." However, problems can arise if the instruments are highly collinear with each other or with other explanatory variables in the treatment cost equation. Multicollinearity statistics and diagnostic tests should be examined to ensure that this is not a serious issue in each specific situation.[19]

Ideally, one would like to find an instrumental variable that only explains treatment choice and has no impact on treatment outcomes or treatment costs. A classic health care services research example of this was demonstrated in the McClellan et al. evaluation of cardiovascular surgical outcomes in which they used patients' residential proximity to certain types of hospital as an instrument.[20] Certain hospitals are more likely to utilize specific cardiovascular surgical techniques, so patients living closer to those hospitals are more likely to receive those treatments. But people don't choose their residences based on local hospital surg e rypre fe rences. Thus, while patient residential location is correlated with surgical treatment, it is clearly not correlated with the outcomes of surgery.[21]

Heckman's original (1974) selection bias regression correction method, which uses the inverse Mills ratio transformation of the propensity score as an additional regressor in the treatment cost regression, can be thought of as a propensity score method if one assumes that the inverse Mills ratio is the precise additional explanatory variable that preserves the propensity score framework. This is precisely the case when all the unobservables in the propensity equation and unobservables in treatment costs are joint-normally distributed. The Heckman method can be refashioned as an instrumental variables method if the inverse Mills ratio is used as an instrument for the selection-biased treatment effect rather than directly including it as a regressor in the treatment cost equation. The question of whether propensity score methods are better than instrumental variables methods or vice versa is not settled and probably varies from case to case. Any valid instrumental variable should be included in the propensity score estimation equation and also could be included as an additional regressor in the treatment cost equation. But adding such an instrumental variable to the treatment cost equation will only eliminate selection bias if the strong ignorability assumption holds.

Instrumental variables methods may be preferable in the context of retrospective claims data analyses since they explicitly allow for unobservable correlates of treatment choice and treatment costs. Also, any time one can estimate a propensity score, one also can use it (and/or its transformations) as a valid instrument. Moreover, because of their properties in estimating simultaneous equations as 2-stage least squares estimators, instrumental variables estimators create valid estimates of the structural (policy-relevant) parameters, unlike reduced-form parameters that conflate the higher-order impacts of exogenous variables amplified through all relevant equations. Propensity score estimators are better thought of as reduced-form parameter estimators. There are empirical examples where propensity score methods do better than instrumental variables methods and vice versa.[22,23] This debate directly parallels an earlier debate as to whether sample selection models are better than 2-part models in estimating health care cost functions since the econometric issues are very similar.[24,25]

Furthermore, Gianfrancesco et al. ignore issues that lead to bias in their reported regression *t*-statistics. If drug treatment selection bias exists, then correcting for it by using an estimated propensity score (or propensity score transformation) rather than the "true" propensity score will create biased *t*-statistics, just as using the estimated mean rather than the true mean biases the estimates of the standard deviation in the simple univariate case. Regardless of which estimation method is used, one should adjust the estimated model coefficient standard errors for this "heteroscedasticity" induced by including a fitted propensity score (and/or its transformations) in the treatment cost regression. The simplest and most accurate way to accomplish this is to bootstrap the entire estimation step sequence, using resampling with replacement to obtain a sufficient number of sample replicates (using the original sample size) to generate robust parameter confidence intervals.[26] Split-sample model validation (estimating the model on a random half of the data and measuring prediction error on the other half) is a very useful way to compare alternative estimation models, particularly when sample sizes are large, as is usually the case with retrospective claims data.

Another major item of concern in any estimation of treatment effects on health care costs relates to the fact that nearly all health care cost samples are highly skewed, with a small percentage of patients accounting for a large proportion of total costs. Logarithmic transformation, or some other power transformation of the cost variable(s) is often an appropriate correction, but this adjustment creates a number of additional complexities that are often inappropriately ignored. First, when one transforms costs to the log-scale, one cannot simply take the exponent of the estimated treatment coefficient to predict the treatment effect on the raw cost scale. As Duan and others have pointed out, log-transformations (or other power transformations) create a "retransformation bias" that must be accounted for in calculating treatment effects in the cost scale.[27,28] Gianfrancesco et al. do not take this retransformation bias into account in generating their estimates of the cost differences between drug treatment groups.

Second, as described in Diehr et al., in situations where a subset of patients exhibit zero-levels of cost or spending and are therefore not measurable after a log-transformation, a common "trick" is to treat their spending as if it were a very small positive number (e.g., $1), and then include their observation with a dependent value of log ($1) = 0.[29] This trick actually creates a potentially arbitrary and serious bias in estimating all model coefficients. For example, if adding $1 to a zero-cost individual's spending is inconsequential, why not add $.01 instead or, even better, $0.000001. It is easy to demonstrate that all model coefficient estimates are highly sensitive to the choice of these arbitrary small spending amounts. Since none of the amounts are truly justified, the analyst will be inducing an arbitrary artificial bias into the estimation process. A more robust solution is to estimate a 2-part model or a sample selection model, splitting the patient sample into the subgroup with zero expenditures and the subgroup with positive expenditures and estimating a person's predicted costs in 2 parts—the probability that they have positive expenditures times the expected value of expenditures, given that they're positive.[25]

There are a number of additional important issues in estimating drug treatment effects on patient health care costs in retrospective claims database analyses. These include the fact that disease dates of diagnosis are often unknown, and diagnostic episodes are often either left-censored, right-censored, or both in retrospective claims data. For example, in looking at mental health patients, one seldom is interested in only the subset of cases that are initially diagnosed during the retrospective data observation period. Even if one were, the lack of health care utilization prior to the first observed diagnosis-related claim or service could reflect either a new diagnosis, the fact that the condition was in remission, or that the patient recently transferred to the observed insurance plan. All nonincident cases in the sample are termed "left-censored." Similarly all sample patients who are not "cured" or dead by the last date in the observation period are considered "right-censored." Statistical methods for dealing with these episode-censoring concerns include hazard functions, survival models, Cox proportional hazards models, and other multivariate extensions of the Kaplan-Meier survival curve.[18,30]

Also, the fact that one often has repeated observations on (a subset of) the same patients over time in retrospective claims data, allows the use of fixed-effect or random effects variance components models to adjust for patient-specific unobservables.[18,19] Gianfrancesco et al. did not take advantage of the additional precision that these repeated episodes per patient adds to the estimation process. Moreover, they treat each episode as an independent event, ignoring issues of treatment switching and censoring of episode events.

Finally, it is often suggested that statistical test adjustments such as the Bonferroni correction for multiple comparisons should be undertaken when evaluating the coefficients in multivariate regression analyses and with multiple patient subgroups.[31] Such statistical corrections are often much too drastic and substantially increase the risk of type II errors (failing to detect an effect that is truly there). As Ken Rothman, the editor of *Epidemiology*, stated:

> The theoretical basis for advocating a routine adjustment for multiple comparisons is the "universal null hypothesis" that "chance" serves as the first-order explanation for observed phenomena. This hypothesis undermines the basic premises of empirical research, which holds that nature follows regular laws that may be studied through observations. A policy of not making adjustments for multiple comparisons is preferable because it will lead to fewer errors of interpretation when the data under evaluation are not random numbers but actual observations on nature. Furthermore, scientists should not be so reluctant to explore leads that may turn out to be wrong that they penalize themselves by missing possibly important findings.[32]

All of these econometric issues increase the complexity of statistical inference dramatically when using retrospective claims data, particularly in comparison to statistical analysis of RCTs. The variety and complexity of statistical tools and alternatives that can plausibly be brought to bear on any given estimation problem create the additional concern that if one has many different estimation procedures at hand, it is easy and tempting to find some sequence of estimation methods that achieves the "predetermined" answer, while ignoring other estimation methods. Any retrospective database provides the overly enthusiastic analyst with ample opportunity to "torture the data until it reveals the truth."

However, bootstrapping and split-sample validation of model results provide a more balanced and robust assessment of model parameter precision and safeguard the estimates against such inappropriate "overfitting." In any case, with all of the recent econometric advances in health care cost analysis, it is unacceptable to simply ignore retrospective data analysis issues such as (1) treatment selection bias, (2) log-cost or other power transformation bias, (3) variance components models with repeated observations, or (4) data censoring issues. When application of alternative reasonable estimation approaches give similar answers, one can conclude that the results are robust. When different plausible estimators give very different answers, it is difficult to draw any conclusions from the estimates. Given all of these issues, it is never acceptable to present a single econometric model estimate, just as in cost-effectiveness analysis it would not be acceptable to present model point estimates without running appropriate model sensitivity analysis. The value of any statistical analysis is not to generate the "correct" answer but to show what range of parameters are plausible, given the available information.

Joel W. Hay, PhD
Associate Professor
Department of Pharmaceutical Economics and Policy
University of Southern California School of Pharmacy
1540 East Alcazar St., CHP 140
Los Angeles, CA 90033
jhay@usc.edu

DISCLOSURES

The author discloses no potential bias or conflict of interest relating to this editorial.

REFERENCES

1. Gianfrancesco F, Pesa J, Wang R. Comparison of mental health resources used by patients with bipolar disorder treated with risperidone, olanzapine, or quetiapine. *J Manag Care Pharm.* 2005;11(3):220-30.

2. Avorn J. *Powerful Medicines: The Benefits, Risks, and Costs of Prescription Drugs.* New York: Knopf; 2004.

3. Abenheim L, Moride Y, Brenot F, et al. Appetite-suppressant drugs and the risk of primary pulmonary hypertension. *N Engl J Med.* 1996;335:609-16.

4. Gale EAM. Lessons from the glitazones: a story of drug development. *Lancet.* 2001;357:1870-75.

5. Solomon DH, Schneeweiss S, Glynn RJ, et al. Relationship between selective cyclooxygenase-2 inhibitors and acute myocardial infarction in older adults. *Circulation.* 2004;109:2068-73.

6. Graham DJ, Campen D, Hui R, et al. Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study. *Lancet.* 2005;365(9458):475-81.

7. Roussouw J, Anderson G, Prentice R, et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative. *JAMA.* 2002;288(9):321-33.

8. Hay J. Hospital cost drivers: an evaluation of 1998-2001 state-level data. *Am J Manag Care.* 2003;9(S1):S14-S24.

9. Schneeweiss S, Wang PS, Avorn J, Maclure M, Levin R, Glynn R. Consistency of performance ranking of comorbidity adjustment scores in Canadian and U.S. utilization data. *J Gen Int Med.* 2004;19(5, pt 1):444-50.

10. Roy AD. Some thoughts on the distribution of earnings. *Oxford Econ Pap.* 1951;3:135-46.

11. Gronau R. Wage comparisons—a selectivity bias. *J Polit Economy*. 1974;82:1119-43.

12. Heckman J. Shadow prices, market wages, and labor supply. *Econometrica.* 1974;42:679-94.

13. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc.* 1984;79:516-24.

14. Hirano K, Imbens G, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003;71(4):1161-89.

15. Heckman J. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. *J Human Res.* 1997;32:441-62.

16. Heckman J, Vytlacil E. Instrumental variables methods for the correlate random coefficient model. *J Human Res.* 1998;33:974-87.

17. UCLA Department of Education. Education2-31C. Applied categorical and nonnormal data analysis: probit regression models. Available at: http://www.gseis.ucla.edu/courses/ed231c/notes3/probit1.html. Accessed April 13, 2005.

18. Wooldridge J. *Econometric Analysis of Cross Section and Panel Data.* New York: MIT Press; 2002.

19. Greene W. *Econometric Analysis.* 5th ed. New York: FT Prentice Hall; 2002.

20. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction reduce mortality? *JAMA.* 1994;272:859-66.

21. Newhouse JP, McClellan M. Econometrics in outcomes research: the use of instrumental variables. *Annu Rev Public Health.* 1998;19:17-34.

22. Obenchain B, Melfi C. Propensity score and Heckman adjustments for treatment selection bias in database studies. Proceedings of the Biopharmaceutical Section, American Statistical Association. 1997:297-306.

23. Hay J, Leahy M. Cost effectiveness of oral antihistamines in the California Medi-Cal program. *Value Health.* In press.

24. Hay J, Olsen R. Let them eat cake: a note on comparing alternative models of the demand for medical care. *J Bus Econ Stat.* 1984;2(3):279-82.

25. Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *J Health Econ.* 1998;17:247-81.

26. Chernick MR. *Bootstrap Methods: A Practitioner's Guide.* New York: Wiley Interscience; 1999.

27. Duan, N. Smearing estimate: a nonparametric retransformation method. *J Am Stat Assoc.* 1983;78:605-10.

28. Manning W, Mullahy J. Estimating log models: to transform or not to transform? *J Health Econ.* 2001;20(4):461-94.

29. Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY. Methods for analyzing health care utilization and costs. *Annu Rev Public Health.* 1999;20:125-44.

30. Keifer N. Economic duration data and hazard functions. *J Econ Lit.* 1988;28(2):646-79.

31. Motheral M, Brooks J, Clark MA, et al. Checklist for retrospective database studies—report of the ISPOR task force on retrospective databases. *Value Health.* 2003;6(2):90-97.

32. Rothman K. No adjustments are needed for multiple comparisons. *Epidemiology.* 1990;1(1):43-46.