

DIFFERENCE-IN-DIFFERENCE AND INSTRUMENTAL VARIABLES APPROACHES

AN ALTERNATIVE AND COMPLEMENT TO PROPENSITY SCORE MATCHING IN ESTIMATING TREATMENT EFFECTS

ZHUN CAO, PhD
XUE SONG, PhD
LAURIE HAMILTON, MA

An examination of results across different estimation methods is beneficial before a conclusion is drawn.

INTRODUCTION

As discussed in the Thomson Reuters CER Issue Brief in May 2011, observational data, such as data from retrospective healthcare claims, are a key resource for conducting comparative effectiveness research (CER). Many outcomes research studies compare outcomes (e.g., healthcare utilization and expenditures) between competing treatments using observational data. Endogeneity of treatment (also called confounding) is a common problem in retrospective claims data studies because patients in claims data were not assigned to treatment by randomization.

Propensity score matching (PSM) is commonly used in health economics and outcomes research literature to address the selection bias caused by endogeneity. PSM is designed to adjust for differences between two comparison cohorts and to generate matched cohorts that are similar to randomization of clinical trials. The other methods that address the endogeneity issue include difference-in-difference (DID) and instrumental variables (IV).

The study outlined in this Brief compared and contrasted the strengths, weaknesses, and the different types of information provided by DID, PSM, and IV methods empirically. DID, PSM, and IV were applied to estimate the impact of treatment with typical (first-generation antipsychotics such as chlorpromazine and haloperidol) versus atypical (second-generation antipsychotics such as risperidone and olanzapine) on healthcare expenditures in Medicaid patients with schizophrenia using the *Thomson Reuters MarketScan® Medicaid Database*.

ANALYTICAL APPROACHES

Propensity Score Matching: A logistic model was fitted to estimate the probability of a patient using atypical versus typical antipsychotic medication, controlling for demographic and clinical characteristics. Propensity scores were calculated as the predicted probability of using atypical antipsychotics. Nearest neighbor matching with 1:3 matching ratio was implemented. Generalized linear models (GLM) with log link and gamma variance function on costs (naive model) were estimated on the matched sample to control for any remaining imbalances.



Instrumental Variables Method: A two-stage residual inclusion model (2SRI) was employed to address the endogeneity of the treatment choice. Instrumental variables used in the first stage logistic regression were physician-level prescription pattern, including percentage of schizophrenia patients with the physician, percentage of schizophrenia patients using typical antipsychotics with the physician, and percentage of schizophrenia patients using atypical antipsychotics with the physician. The residual term from the first-stage regression was included in the second-stage GLM regression of all-cause and psychiatric-related costs to produce consistent estimates and to adjust for any unobservable confounding.

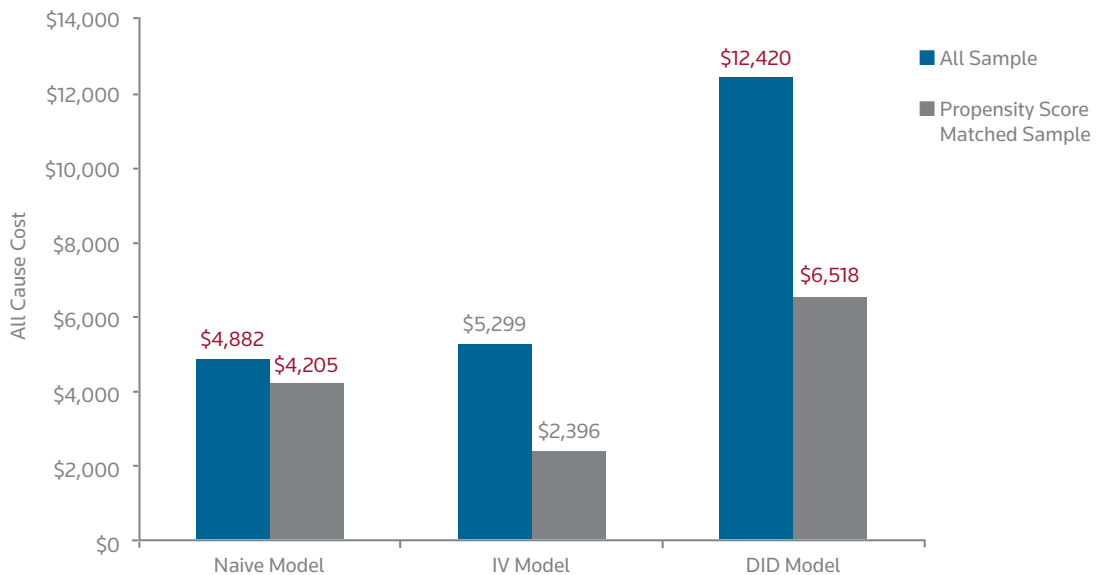
Difference-in-Difference: The pre-to-post period differences in all-cause costs and psychiatric-related costs were calculated using both prematching and matched samples. GLM regressions with log link and Gaussian variance functions were estimated on the pre-to-post difference in costs.

Combined Methods: The IV and DID approaches were also applied to the propensity score matched sample to adjust for bias caused by both observable and unobservable characteristics.

RESULTS

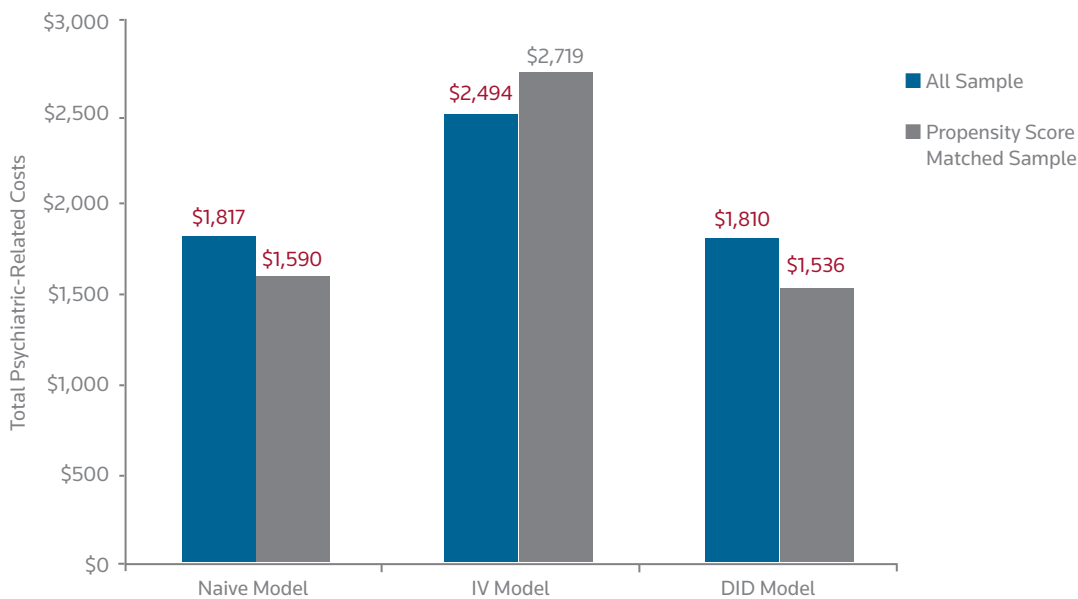
Six regression models were estimated – three different estimate methods (GLM, IV, and DID) and two samples (matched sample and unmatched sample). The effect size varied across methods since different model specifications were used. In the naive model, only the demographic and clinical factors were included as covariates in the regression. However, in the IV model, a residual term was included in addition to the covariates. In the DID model, the dependent variables were the pre-post differences in costs. Although different effect sizes were found using different methods, the conclusion was consistent in general. Four out of the six models (with the exception of the IV model on the matched and unmatched samples) found that total costs of atypical medication users were significantly higher than those of typical medication users. Five out of the six models found atypical medication users had significantly higher psychiatric-related costs than typical medication users. The exception was the IV model on the matched sample. The multivariate regression adjusted marginal effects are reported in Figures 1 and 2.

Figure 1: Adjusted Total All-Cause Costs



NOTE: Numbers in red indicate p values less than 0.05.

Figure 2: Adjusted Total Psychiatric-Related Costs



NOTE: Numbers in red indicate p values less than 0.05.

DISCUSSION

Compared with the PSM method where control patients are selected for case patients, DID estimator uses the same individual as their own control, and thus may be less biased when the list of confounding factors is incomplete in the claims data. While PSM may lose patients for whom an appropriate match cannot be selected, DID keeps all patients in the sample. In addition, the DID method may also address bias due to unobservable characteristics, assuming the unobservable characteristics are constant or at least similar over time. However, DID may lead to misleading estimates if utilization rates from the groups being compared are very different in the baseline period; in which case, the DID estimator may be just measuring regression to the mean. The DID method may also result in bias when factors contributing to utilization outcomes change over time but are not controlled for.

The IV method addresses the endogeneity of the treatment choice due to unobservable characteristics and keeps all patients in the sample. However, the consistency of the estimator is dependent on the strength and validity of the instruments. In this study, the coefficient for the residual term was found to be insignificant in the second-stage regression, which may be considered as evidence that the unobservable may not have a significant impact on the outcomes.

The DID and IV methods may be combined with propensity score adjustment. The combined methods have the advantages in addressing bias caused by unobservable characteristics in a more balanced sample. On the other hand, the combination approaches may be subject to the loss of generalizability compared with DID or IV method on the entire sample, and may be less efficient compared with the PSM method alone. In this case study, GLM on the matched sample and DID on the matched sample both resulted in significantly higher psychiatric-related and total costs for patients with atypical antipsychotic medication treatment. The IV methods based on the matched sample suggested no significant association between atypical antipsychotic use and psychiatric-related cost, however, and this may be caused by the fact that the IV estimator is usually less efficient.

As pointed out in the May CER Issue Brief, a good study design is very important in conducting CER using observational data. By defining the appropriate comparison groups, we can reduce the potential loss of generalizability in PSM, minimize the risk of measuring regression to mean in DID method, and offset the effect of the unobservable characteristics.

CONCLUSION

While each method has its advantages and disadvantages, there is no simple conclusion that one method is better than the other. Researchers need to assess the potential causes of endogeneity and evaluate the tradeoff between the methods before deciding which method to use. Sensitivity analysis with alternative models is recommended when selection bias is considered a problem in a study. In this specific study, the DID model combined with the propensity score matching seems to be most reliable, given that the finding is consistent with most of other models. The use of DID offsets the potential bias caused by patient-level unobservable characteristics, while the use of a propensity score matched sample provides more comparable groups and minimizes the risk of measuring regression to mean.

ABOUT THE AUTHORS

Dr. Cao is a senior economist in the Outcomes Research Group at Thomson Reuters. She has extensive experience conducting health economic and outcomes research for pharmaceutical, biotechnology, and medical device companies. Her research interests include statistical methodology in outcomes research and econometric modeling using retrospective data. She has published more than 20 peer-reviewed journal articles. Dr. Cao received her PhD and MA degrees in economics from Boston University.

Dr. Song is a senior research leader in the Outcomes Research Group at Thomson Reuters. She has 10 years of experience conducting health economics and outcomes research work for pharmaceutical, biotechnology, and medical device companies and for governmental agencies. She has extensive experience conducting research in various therapeutic areas including oncology, diabetes, cardiovascular disease, infectious diseases, respiratory diseases, and osteoarthritis, and has published in *Diabetes Care*, *Value in Health*, *Annals of Oncology*, *Annals of Pharmacotherapy*, *Journal of Occupational and Environmental Medicine*, *PharmacoEconomics*, and other journals. She holds a PhD and MA from Johns Hopkins University, both in economics.

Ms. Hamilton is a senior programmer/analyst in the Outcomes Research Group at Thomson Reuters. She has more than 23 years of experience in advanced analytics and statistical programming, design of automated statistical analysis systems, and management of SAS programming teams. She specializes in methodologies for managing and analyzing very large databases, with a current focus on pattern recognition in medical administrative claims data. She holds a Master's degree in zoology from the University of Florida.

CONTACT INFORMATION

To contact the authors, please email us at cer@thomsonreuters.com.

REFERENCES

Abadie A. Semiparametric difference-in-difference estimators. *The Review of Econometric Studies* 2005; 72: 1-19.

Athey S, Imbens GW. Identification and inference in nonlinear difference-in-differences models. *Econometrica* 2006; 74: 431-97.

Fu AZ, Dow WH, Liu GG. Propensity score and difference-in-difference methods: A study of second-generation antidepressant use in patients with bipolar disorder. *Health Services and Outcomes Research Methodology* 2006; 7: 23-38.

Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health* 2006; 9(6): 377-85.

Mandrekar JN, Mandrekar SJ. An introduction to matching and its application using SAS®. Accessible at <http://www2.sas.com/proceedings/sugi29/208-29.pdf>.

Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 2007; 15:199-236.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70: 41-55.

Terza JV, Basu A, Rathouz PJ. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 2008; 3:531-43.

ABOUT THOMSON REUTERS

Thomson Reuters is the world's leading source of intelligent information for businesses and professionals. We combine industry expertise with innovative technology to deliver critical information to leading decision makers in the financial, legal, tax and accounting, healthcare and science and media markets, powered by the world's most trusted news organization. With headquarters in New York and major operations in London and Eagan, Minnesota, Thomson Reuters employs 55,000 people and operates in over 100 countries.

healthcare.thomsonreuters.com/pharma

Thomson Reuters
777 E. Eisenhower Parkway
Ann Arbor, MI 48108 USA

©2011 Thomson Reuters.
All rights reserved.
H PAY PHARM 1108 10189 MC



THOMSON REUTERS™