

Good Research Practices for Comparative Effectiveness Research: Analytic Methods to Improve Causal Inference from Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part III

Michael L. Johnson, PhD,¹ William Crown, PhD,² Bradley C. Martin, PharmD, PhD,³ Colin R. Dormuth, MA, ScD,⁴ Uwe Siebert, MD, MPH, MSc, ScD⁵

¹University of Houston, College of Pharmacy, Department of Clinical Sciences and Administration, Houston, TX, USA; Houston Center for Quality of Care and Utilization Studies, Department of Veteran Affairs, Michael E. DeBakey VA Medical Center, Houston, TX, USA; ²Innovus, Waltham, MA, USA; ³Division of Pharmaceutical Evaluation and Policy, College of Pharmacy, University of Arkansas for Medical Sciences, Little Rock, AR, USA; ⁴Department of Anesthesiology, Pharmacology & Therapeutics, University of British Columbia; Pharmacoepidemiology Group, Therapeutics Initiative, Vancouver, BC, Canada; ⁵Department of Public Health, Information Systems and Health Technology Assessment, UMIT, University of Health Sciences, Medical Informatics and Technology, Hall i.T., Austria; Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA, and Department of Health Policy and Management, Harvard School of Public Health, Boston, MA, USA

ABSTRACT

Objectives: Most contemporary epidemiologic studies require complex analytical methods to adjust for bias and confounding. New methods are constantly being developed, and older more established methods are yet appropriate. Careful application of statistical analysis techniques can improve causal inference of comparative treatment effects from nonrandomized studies using secondary databases. A Task Force was formed to offer a review of the more recent developments in statistical control of confounding.

Methods: The Task Force was commissioned and a chair was selected by the ISPOR Board of Directors in October 2007. This Report, the third in this issue of the journal, addressed methods to improve causal inference of treatment effects for nonrandomized studies.

Results: The Task Force Report recommends general analytic techniques and specific best practices where consensus is reached including: use of

stratification analysis before multivariable modeling, multivariable regression including model performance and diagnostic testing, propensity scoring, instrumental variable, and structural modeling techniques including marginal structural models, where appropriate for secondary data. Sensitivity analyses and discussion of extent of residual confounding are discussed.

Conclusions: Valid findings of causal therapeutic benefits can be produced from nonrandomized studies using an array of state-of-the-art analytic techniques. Improving the quality and uniformity of these studies will improve the value to patients, physicians, and policymakers worldwide.

Keywords: causal inference, comparative effectiveness, nonrandomized studies, research methods, secondary databases.

Background to the Task Force

In September 2007, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Health Science Policy Council recommended that the issue of establishing a Task Force to recommend Good Research Practices for Designing and Analyzing Retrospective Databases be considered by the ISPOR Board of Directors. The Council's recommendations concerning this new Task Force were to keep an overarching view toward the need to ensure internal validity and improve causal inference from observational studies, review prior work from past and ongoing ISPOR task forces and other initiatives to establish baseline standards from which to set an agenda for work. The ISPOR Board of Directors approved the creation of the Task Force in October 2007. Task Force leadership and reviewer groups were finalized by December 2007, and the first teleconference took place in January 2008.

Address correspondence to: Michael L. Johnson, University of Houston, College of Pharmacy, Department of Clinical Sciences and Administration, Houston, TX 77030, USA. E-mail: mikejohnson@uh.edu
10.1111/j.1524-4733.2009.00602.x

Task Force members were experienced in medicine, epidemiology, biostatistics, public health, health economics, and pharmacy sciences, and were drawn from industry, academia, and as advisors to governments. The members came from the UK, Germany, Austria, Canada, and the United States.

Beginning in January 2008, the Task Force conducted monthly teleconferences to develop core assumptions and an outline before preparing a draft report. A face-to-face meeting took place in October 2008, to develop the draft, and three forums took place at the ISPOR Meetings to develop consensus for the final draft reports. The draft reports were posted on the ISPOR website in May 2009 and the task forces' reviewer group and ISPOR general membership were invited to submit their comments for a 2 week reviewer period. In total, 38 responses were received. All comments received were posted to the ISPOR website and presented for discussion at the Task Force forum during the ISPOR 12th Annual International Meeting in May 2009. Comments and feedback from the forum and reviewer and membership responses were considered and acknowledged in the final reports. Once consensus was reached, the manuscript was submitted to *Value in Health*.

Introduction

We proceed from the assumption that proper statistical analysis of study data is dependent upon the research question to be answered, and the study design that led to the collection of data, in this case secondarily, to be analyzed [1,2]. We also assume that the data to be analyzed has been appropriately measured, validated, defined, and selected. Many authors have described good research practices in these fundamental areas, including other ISPOR task force groups, and we seek to build upon their work, not reproduce it [3,4]. Recognizing that new methods are constantly being developed, and older more established methods are yet appropriate, we intend to offer a review of the more recent developments in statistical control of confounding.

Stratification

Stratified analysis is a fundamental method in observational research. It involves placing data into subcategories, called strata, so that each subcategory can be observed separately. Its many uses in observational studies include standardization, control of confounding, subgroup analysis in the presence of effect-measure modification, and to address selection bias of the type that occurs in matched case control studies. When a cohort is categorized by follow-up time, stratification can also prevent bias from competing risks and losses to follow-up.

Like any analytical method, stratification has its strengths and limitations. Its strengths are that it is an intuitive and hands-on method of analysis, results are readily presented and explained, and it does not require restrictive assumptions. Its disadvantages include a potential for sparsely populated strata, which reduce precision, loss of information when continuous variables are split into arbitrarily chosen categories, and a tendency to become arduous when the number of strata is large.

Most contemporary observational studies use complex analytical methods such as multivariable regression analysis. Given the ubiquity of those methods, it is tempting to undervalue the role of stratified analysis, but to do so is a mistake because it has important uses that are not well served by other methods. Because stratified analysis can be applied “hands-on,” often in a simple spreadsheet program, it allows investigators to get closer to their data than they otherwise could by using more complex methods such as multivariable regression. Furthermore, stratified analysis in a cohort where observations are categorized by levels of the most influential covariates should provide results that are comparable to estimates from a rigorous multivariable model. If results from a stratified analysis are markedly different from estimates obtained from a regression model, then the discrepancy should serve as a warning that the investigator has possibly made a mistake. For this reason, it is important to do a stratified analysis in studies even when more complex analytical methods are finally used. It is advisable to conduct a stratified analysis prior to undertaking a more complex analysis because of its potential to provide important information on relevant covariates and how they could be optimally included in a model.

Stratified analysis can proceed by categorizing data into strata. In a spreadsheet, 10 to 20 strata should be manageable before the analysis becomes cumbersome. If the analysis requires more strata, then it may be helpful to perform the analysis in stages by first examining levels of a few variables, and then in subsets defined according to those first few variables, perform analyses on more variables. A staged approach can be time-consuming and does not lend itself easily to calculating summary or pooled estimates, but it can be useful for studying effect-measure modification.

Significant heterogeneity between strata suggests the presence of effect-measure modification. When this happens, stratum-specific estimates should be reported because effect-measure modification is a characteristic of the effect under study rather than a source of bias that needs to be eliminated. Pooled effect estimates can be calculated in the absence of effect-measure modification to obtain an overall estimate of an effect. A pooled estimate that is substantially different from stratum-specific estimates indicates the possible presence of confounding. Multiple methods are available for estimating pooled effects. The main difference between pooled estimators is how each assigns weights to strata. A simple but typically unsatisfactory method is to equally weight all strata regardless of the amount of information they contain. Person-time weights or inverse variance weights are better because they assign weights in proportion to the amount of information contained in each stratum. Another common method for estimating pooled effects is to use Mantel–Haenszel weights. Mantel–Haenszel estimators are accurate, widely recognized, and easy to calculate. For more information, a practical and comprehensive discussion of stratified analysis was written by Rothman and Greenland [5].

Regression

Numerous texts are available to teach regression methodology, and we do not attempt to summarize them here, nor cite a complete list [6–10]. Regression is a powerful analytical technique that can accomplish several goals at once. When more than a few strata are formed for stratified analysis, or when more than a few potential confounding factors need to be adjusted, multiple regressions can be used to determine the unique association between the treatment and the outcome, after simultaneously adjusting for the effects of all the other independent factors included in the regression equation. It is very common to see reports of the parameter estimates, rate ratio (RR), or odds ratio (OR), and their 95% confidence limits, for a given variable after adjustment for a long list of covariates. It is becoming less common to actually see the full regression model adjusted for all the covariates, but failure to present the full model may lead to concerns of a “black box” analysis that has other pitfalls [11,12]. Another very important use of regression is to use the regression equation to predict study outcomes in other patients. This is the primary use of multiple logistic regression when used for propensity scoring, which will be discussed later. Present the final regression model, not only the adjusted treatment effects. If journal or other publications limit the amount of information that can be presented, the complete regression should be made available to reviewers and readers in an appendix.

Variable Selection

One of the critical steps in estimating treatment effects in the observational framework is to adequately assess all the potential confounding variables that can influence treatment selection or the outcome. In order to capture all of the potentially confounding variables and any suspected effect modification or interactions, a thorough literature review should be conducted to identify measures that influence treatment selection and outcome measurement, and a table should be created detailing the expected associations. The analyst should identify those measures available in the data, or good proxies for them, and include them in the regression model irrespective of statistical significance at traditional significance levels. When using administrative data sources with limited clinical information, there are often instances when the analyst will not have access to meaningful

Table 1 Commonly used link functions and error distributions for outcomes with different types of data

Examples of outcomes	Data type	Link function	Error distribution
Trends in drug utilization or costs	Continuous	Identity	Gamma*
Predictors of treatment choice, death	Binary	Logistic	Binomial
Myocardial infarction, stroke, hospitalizations	Count	Log	Poisson [†]

*Assuming Gamma-distributed errors does not require log transformation of utilization and thus increases interpretability of results [13,14].

[†]Depending on the skewness of data, we may adjust for over dispersion using the scale parameter.

measures for some known or suspected confounders. In these instances when the model cannot include known or suspected confounders, the regression estimates of treatment effect will be biased leading to omitted variable or residual confounding bias. Because omitted variables can lead to biased estimates of treatment effect, it is important to identify all the known potential confounders as a starting point and to make every attempt to identify measures in the data and to incorporate them into the model. When known potential confounders cannot be included in the model, the analyst should acknowledge their missingness as a limitation and describe the anticipated directionality of the bias. For regressions where the magnitude of effect is weak or modest, omitted variable bias may lead to meaningful changes in conclusions drawn from the data. Even when all known confounders are included in the regression equation, unobserved or unobservable confounders may still exist resulting in omitted variable bias. To address the potential impact of omitted variable bias, there are techniques that can describe the characteristics of a confounder that would drive the results from significance to the null.

Model Selection

The form of the dependent variable generally drives the choice of a regression model. Continuously distributed variables are generally analyzed with ordinary least squares (OLS) regression, while dichotomous or binary outcomes (yes/no; either/or; dead/alive) can be modeled with logistic regression. In practice, common statistical software programs analyze both types of outcomes using maximum likelihood estimation and assuming the appropriate error distribution (normal, logistic, etc.). Logistic regression has become almost ubiquitous in the medical literature in the last 20 years, coinciding with advances in computational capacity and familiarity with the method. Linear and logistic regressions fall under a broader category of models known as generalized linear models (GLM). GLMs also include models with functional forms other than linear or log-linear to describe the relationship between the independent and dependent variables. Some commonly used link functions and error distributions are shown in Table 1.

Another valuable use of GLM models is their ability to incorporate different specifications of covariance structures when the assumption of the independence between observations is violated. Longitudinal analyses using data sets in which multiple measurements are taken on the same subject over time are common in comparative effectiveness studies. In such analyses, the observations are not independent, and any correlation must be accounted for to obtain valid and precise estimates of effects [15]. An analogous situation produces the same type of correlation when study subjects are sampled repeatedly from a single location, for example, within hospitals. Such a hospital has repeated measures taken from it, also called nested within or hierarchical, and these observations are likely correlated. Therefore, some studies that are conducted from a few or several sites may need to adjust for nested sampling. Situations in which

multiple observations are obtained from the same subjects can be estimated using generalized estimating equations (GEE) within the framework of GLM. In a GEE, the researcher chooses a functional form and link function as with any GLM, but then also chooses a covariance structure that adequately describes the known or suspected correlations between repeated observations. Commonly used types of covariance structures include exchangeable, autoregressive, and unstructured. When in doubt as to the true correlation structure, then an exchangeable matrix should be used.

Logistic regression analysis assumes that there is no loss to follow-up or competing risks in the study population. These are strong assumptions that are not necessary in another type of model known as Cox proportional hazards regression. In this model, the dependent variable becomes the time to the event rather than the probability of the event occurring over a specified period as in a logistic regression model. This difference is in terms of how long it takes until an outcome occurs, that is, modeling “when,” not just “if.” Cox regression models also allow for a very important advance, with the inclusion of time-varying or time-dependent covariates. These are independent variables, which are allowed to change their value over time (just as the dependent variable is changing its value over time). Because drug exposures can change over time as patients change to different drugs or doses, Cox regression is a powerful tool, which in some circumstances will more realistically model the exposure–outcome relationship in treatment effect studies.

Testing Model Assumptions

There are many statistical assumptions that underlie these regression techniques. For the linear and logistic models, the assumptions of normality and linear association are possibly the most important. Fortunately, these models are very robust to the assumption of normality, that is, the outcome variable has to be very non-normal (such as costs) to severely threaten parameter estimates. A more common problem is the use of continuous measures modeled as continuous variables without checking the assumption of linear association. For example, age is usually thought of as a continuous variable, ranging from maybe 18 to 90 or so in many studies of adults with medical conditions. Age in years is actually a categorical variable, with 72 different levels, in this case. It is essential to check the assumption of a linear relationship between continuous independent variables and study outcomes; if the independent variable does not have a linear relation, then nonlinear forms, such as categories should be used in modeling. Thus, it is essential to check and see if the association of age with the study outcome increases/decreases in a relatively constant amount as age increases/decreases, or the study results for age, and the control or adjustment for confounding by age, may not be valid. For example, if relatively few subjects are very old and also have a certain unfavorable response to a drug, and most of the other patients are younger and generally do not have unfavorable responses, a regression model may erroneously show that increased age is a risk factor

because it is assuming a linear relation with age that is not actually there. In this case, the variable should be modeled as a categorical variable. Another serious assumption that should be tested is the proportional hazards assumption for Cox regression. The Cox model is based on the assumption that the hazard rate for a given treatment group can change over time, but that the ratio of hazard rates for two groups is proportional. In other words, patients in two treatment groups may have different hazards over time, but their relative risks should differ by a more or less constant amount. It is not a difficult assumption to test, but if violated, study results using this technique become questionable. Thus, the proportional hazards assumption for treatment exposure should always be tested before conducting Cox proportional hazards regression, and if this assumption is violated, alternative techniques such as time-varying measures of exposure in extended Cox models should be implemented.

Performance Measurement

Analyses using regression models should report model performance. For example, OLS regression models should report the coefficient of determination (R^2). This is so that the reader can determine if this regression equation has made any realistic explanation of the total variance in the data. In a large database study, the R^2 could be very small, but parameter estimates may be unbiased. Although valid, it may be questionable as to the value of intervening on such variables. In logistic regression, the c-statistic or area under the receiver operating characteristic curve (ROC) is a standard output in many statistical software packages to assess the ability of model to distinguish subjects who have the event from those who do not. Qualitative assessments of the area under the ROC curve have been given by Hosmer and Lemeshow 1999 [9]. Performance measures (R^2 , area under ROC curve) should be reported and a qualitative assessment of these measures should be discussed regarding the explanation of variance or discrimination of the model in predicting outcome. If the emphasis is on prediction, one might care more about model performance than if the emphasis is on multivariable control of several confounding factors.

Diagnostics

All statistical software packages also provide an array of measures to determine if the regression model has been properly produced. Plots of the observed values to predicted values, or the difference between observed and predicted (residual values) can be easily produced to examine for outlier observations that may be exerting strong influence on parameter estimates. With powerful computers in today's world, on the desktop, or in one's lap at a sidewalk cafe, it is very easy to rely on the computer to crank out all kinds of fascinating parameter estimates, RRs, ORs, confidence intervals and *P*-values that may be questionable if not actual garbage. It is very important for the analyst to check the results of regression modeling by at least checking plots of residuals for unexpected patterns. Regression diagnostics including goodness of fit should be conducted and reported.

Missing Data

One of the many challenges that face researchers in the analysis of observational data is that of missing data. In its most extreme form, observations may be completely missing for an important analytic variable (e.g., Hamilton depression scores in a medical claims analysis). However, the issue of missing data is more commonly that of missing a value for one or more variables across different observations. In multivariate analyses such as

regression models, most software packages simply drop observations if they are missing any values for a variable included in the model. As a result, highly scattered missing observations across a number of variables can lead to a substantial loss in sample size even though the degree of "missingness" might be small for any particular variable.

The appropriate approach for addressing missing data depends upon its form. The simplest approach is to substitute the mean value for each missing observation using the observed values for the variable. A slightly more sophisticated version of this approach is to substitute the predicted value from a regression model. However, in both instances, these approaches substitute the same value for all patients (or all patients with similar characteristics). As a result, these methods reduce the variability in the data. (This may not be a particularly serious problem if the pattern of missingness seems to be random and does not impact large numbers of observations for a given variable.) More sophisticated methods are available that preserve variation in the data. These range from hot deck imputation methods to multiple imputation [16,17]. The extent of missing data and the approach to handle it should always be reported.

Recommendations

- Conduct a stratified analysis prior to undertaking a more complex analysis because of its potential to provide important information on relevant covariates and how they could be optimally included in a model.
- Present the final regression model, not only the adjusted treatment effects. If journal or other publications limit the amount of information that can be presented, the complete regression should be made available to reviewers and readers in an appendix.
- Conduct a thorough literature review to identify all potential confounding factors that influence treatment selection and outcome. Create a table detailing the expected associations.
- When known potential confounders cannot be included in the model, the analyst should acknowledge their missingness as a limitation and describe the anticipated directionality of the bias.
- When in doubt as to the true correlation structure, then an exchangeable matrix should be used.
- Check the assumption of a linear relationship between continuous independent variables and study outcomes; if the independent variable does not have a linear relation, then nonlinear forms, such as categories should be used in modeling.
- The proportional hazards assumption for treatment exposure should always be tested before conducting Cox proportional hazards regression; if this assumption is violated, alternative techniques such as time-varying measures of exposure in extended Cox models should be implemented.
- Performance measures (R^2 , area under ROC curve) should be reported and a qualitative assessment of these measures should be discussed regarding the explanation of variance or discrimination of the model in predicting outcome.
- Regression diagnostics including goodness of fit should be conducted and reported.
- The extent of missing data and the approach to handle it should always be reported.

Propensity Score Analysis

The propensity score is an increasingly popular technique to address issues of selection bias, confounding by indication or

endogeneity commonly encountered in observational studies estimating treatment effects. In 2007, a Pubmed search identified 189 human subjects “propensity score” articles compared to just 2 retrieved in 1997. Propensity scoring techniques are now being used in observational studies to address a wide range of economic, clinical, epidemiologic, and health services research topics. The appeal of the propensity scoring techniques lies in an intuitive tractable approach to balance potential confounding variables across treatment and comparison groups. When propensity scores are utilized with a matching technique, the standard “Table 1” that compares baseline characteristics of treated and untreated subjects often resembles those obtained from randomized clinical trials where measured covariates are nearly equally balanced across comparison groups [18,19]. This transparent balancing of confounders facilitates confidence in interpreting the results compared to other statistical modeling approaches; however, unlike randomization, balance between unmeasured or unmeasurable factors cannot be assumed.

The propensity score is defined as the conditional probability of being treated given an individual’s covariates [20,21]. The more formal definition offered by Rosenbaum and Rubin for the propensity score for subject i ($i = 1, \dots, N$) is the conditional probability of assignment to a treatment ($Z_i = 1$) versus comparison ($Z_i = 0$) given observed covariates, x_i :

$$E(x_i) = \text{pr}(Z_i = 1 | X_i = x_i)$$

The underlying approach to propensity scoring uses observed covariates X to derive a “balancing score” $b(X)$ such that the conditional distribution of X given $b(X)$ is the same for treated ($Z = 1$) and control ($Z = 0$) [20,21]. When propensity scores are used in matching, stratification, or regression, treatment effects are unbiased when treatment assignment is strongly ignorable [21]. Treatment assignment is strongly ignorable if treatment groups, Z , and the outcome (dependent) variable are conditionally independent given the covariates, X . This independence assumption will not hold in situations where there are variables or at least good proxy measures not included as propensity score covariates that are correlated with outcome events and treatment selection. These situations are fundamentally the same issue associated with omitted variable bias encountered in more classical regression-based methods. The most common approach to estimate propensity scores are logistic regression models; however, other approaches such as probit models, discriminant analysis, classification and regression trees, or neural networks are possible [22,23]. A tutorial by D’Agostino provides a good description of how to calculate propensity scores including sample SAS code [21].

Once a propensity score has been developed, there are three main applications of using the propensity score: matching, stratification, and regression. Matching on the propensity score takes several approaches, but all are centered on finding the nearest match of a treated (exposed) individual to a comparison subject(s) based on the scalar propensity score [24]. Onur Baser described and empirically compared seven matching techniques (stratified matching, nearest neighbor, 2 to 1 nearest neighbor, radius matching, kernel matching, Mahalanobis metric matching with and without calipers) and found Mahalanobis metric matching with calipers to produce more balanced groups across covariates. This was the only method to have insignificant differences in the propensity score density estimates, supporting previous work demonstrating the better balance obtained with this matching technique [23,25]. By using calipers in the matching process, only treated control pairs that are comparable are retained. Persons in which treatment is contraindicated or rarely indicated from the control sample (low propensity) or in which

treatment is always indicated in the treatment sample (high propensity) are excluded, thus ensuring the desired feature of greater overlap of covariates. This restriction of ensuring overlap on important covariates is a relative strength of propensity score matched analysis; however, if large numbers of unmatched subjects are excluded, one should note the impact on generalizability and in the extreme case if nearly all subjects go unmatched, the comparison should probably not be made in the first place. A lack of overlap may go undetected using traditional regression approaches where the results may be overly influenced by these outliers. Ironically, one of the criticisms sometimes leveled at propensity score analysis is that it is not always possible to find matches for individuals in the respective treatment groups—this suggests that these individuals should not be compared in the first place!

In addition to matching techniques, propensity scores can be used in stratified analyses and regression techniques [21]. Propensity scores can be used to group treated and untreated subjects into quintiles, deciles, or some other stratification level based on the propensity score, and the effects of treatment can be directly compared within each stratum. Regression approaches commonly include the propensity score as a covariate along with a reduced set of more critical variables in a regression model with an indicator variable to ascertain the impact of treatment. It should be noted that, because the propensity score is a predicted variable, a correction should be made to the standard error of any propensity score variable included in a regression. This is not standard practice and, as a result, the statistical tests of significance for such variables are generally incorrect.

One of the main potential issues of propensity scoring techniques lies in the appropriate specification or selection of covariates that influence the outcome measure or selection of the treatment. The basis of selecting variables should be based on a careful consideration of all factors that are related to treatment selection and or outcome measures [26]. There is some empirical work to help guide the analyst in specifying the propensity models, but additional research in this area is warranted before variable specification recommendations can be made conclusively. One temptation may be to exclude variables that are only related to treatment assignment but have no clear prognostic value for outcome measures. Including variables that are only weakly related to treatment selection should be considered because they may potentially reduce bias more than they increase variance [23,27]. Variables related to outcome should be included in the propensity score despite their strength of association on treatment (exposure) selection. Because the coefficients of the covariates of the propensity score equation are not of direct importance to estimating treatment effects per se, parsimony is less important and all factors that are theoretically related to outcome or treatment selection should be included despite statistical significance at traditional levels of significance. This is why many propensity score algorithms do not use variable reduction techniques, such as stepwise regression, or use very liberal variable inclusion criteria such as $P < 0.50$.

One of the clear distinctions between observational data analyses using propensity scoring and large randomized experiments is the inability to balance unmeasured or unmeasurable factors that may violate the treatment independence assumption critical to obtain unbiased treatment estimates. To gain practical insights into the impact of omitting important variables, an empirical exercise compared two propensity models of lipid-lowering treatment and acute myocardial infarction (AMI). One model included 38 variables and 4 quadratic terms, including laboratory results (low density lipoprotein (LDL), high density

lipoprotein (HDL), triglycerides) commonly not available in claims data, and another full model which included 14 additional variables that are not routine measures incorporated in many analyses such as the number of laboratory tests performed [28]. The reduced propensity model had a very high level of discrimination (c-statistic = 0.86) and could have been assumed to be a complete list of factors; however, it failed to show any benefit of statin initiation while the full model showed a lower risk of AMI with statin therapy, a finding comparable to clinical trials. This case study demonstrates the importance of carefully selecting all possible variables that may confound the relationship and highlights the caution one should undertake when using data sets that have limited or missing information on many potentially influential factors, such as commonly encountered with administrative data. Stürmer et al. have proposed a method of using validation data that contains richer data on factors unmeasured in larger data sets to calibrate propensity score estimates [29]. This technique offers promise to address the omitted variable issue with administrative data but would be difficult to implement on a wide scale as more rich validation samples are not routinely available.

Because propensity scoring largely utilizes the same underlying covariate control as standard regression techniques, the benefits of building a propensity scoring scalar instead of directly using the same covariates in a standard regression technique may not be obvious. Empirical comparisons between regression and propensity score estimates have been reviewed and have shown that estimates of treatment effect do not differ greatly between propensity score methods and regression techniques. Propensity scoring tended to yield slightly more conservative estimates of exposure effect than regression [30]. Despite the lack of clear empirical differences between these approaches, there are several theoretical and practical advantages of propensity scoring [22]. Matching and stratification with propensity scores identifies situations in which there exist little overlap on covariates, and these situations are elucidated clearly with propensity scoring; in matched analyses these subjects are excluded from analysis whereas these differences in exposure would be obscured in regression analyses. Stratified analyses can also elucidate propensity score treatment interactions. Because parsimony is not a

consideration in the propensity scoring equation, many more covariates, more functional forms of the covariates, and interactions can be included than would be routinely considered in regression analyses. This issue is emphasized when there are relatively few outcome events where there are greater restrictions imposed on the number of covariates in regression techniques when using rules of thumb such as 8–10 events per covariate. One of the drawbacks of stratified or matched propensity scoring approaches relative to regression approaches is that the influence of other covariates (demographics, comorbidities) on the outcome measure is obscured unless additional analyses are undertaken. Overall, there is no clear superiority of regression or propensity score approaches, and ideally, both approaches could be undertaken.

When operating in the observational framework, omitting important variables because they are unavailable or are unmeasurable is often the primary threat to obtaining unbiased estimates of treatment effect. Propensity scoring techniques offer the analyst an alternative to more traditional regression adjustment and when propensity-based matching or stratification techniques are used, the analyst can better assess the “overlap” or comparability of the treated and untreated. However, the propensity score analyses in of themselves cannot address the issues of bias when there are important variables not included in the propensity score estimation. Instrumental variable (IV) techniques have the potential to estimate unbiased estimates, at least local area treatment effects in the presence of omitted variables if one or more instruments can be identified and measured. An empirical comparison between traditional regression adjustment, propensity scoring, and IV analysis in the observational setting was conducted by Stukel et al. that estimated the effects of invasive cardiac management on AMI survival [31]. The study found very minor differences between several propensity score techniques and regression adjustment with rich clinical and administrative covariates. However, there were notable differences in the estimates of treatment effect obtained with IVs (Fig. 1). The IV estimates agreed much more closely with estimates obtained from randomized controlled trials. This empirical example highlights one of the key issues with propensity scoring when there are strong influences directing treatment that are not observed in the data.

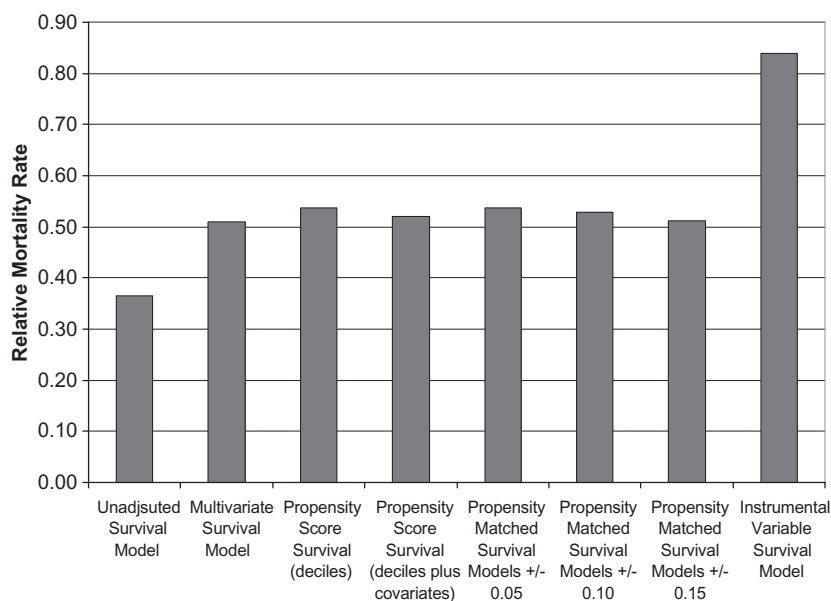


Figure 1 Effects of invasive cardiac management on AMI survival [31].

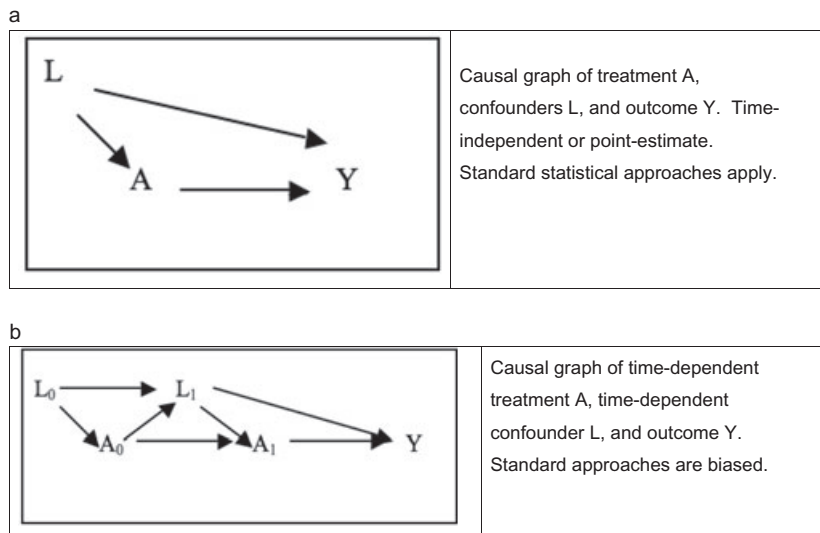


Figure 2 Simplified causal diagram for time-independent and time-dependent confounding.

Marginal Structural Models

Standard textbook definitions of confounding and methods to control for confounding refer to independent risk factors for the outcome, that are associated with the risk factor of interest, but that are not an intermediate step in the pathway from the risk factor to disease. The more complicated (but probably not less common) case of time-varying confounding refers to variables that simultaneously act as confounders and intermediate steps, that is, confounders and risk factors of interest mutually affect each other.

Standard methods (stratification, regression modeling) are often adequate to adjust for confounding except for the important situation of time-varying confounding. In particular, confounding by indication is often time varying, and therefore, an additional concern common to pharmacoepidemiologic studies. In the presence of time-varying confounding, standard statistical methods may be biased [32,33], and alternative methods such as marginal structural models or G-estimation should be examined.

Marginal structural models using inverse probability of treatment weighting (IPTW) have been recently developed and shown to consistently estimate causal effects of a time-dependent exposure in the presence of time-dependent confounders that are themselves affected by previous treatment [34,35]. The causal relationship of treatment, outcome, and confounder can be represented by directed acyclic graphs (DAGs) [36–38].

In the Figure 2a above, A represents treatment (or exposure), Y is the outcome, and L is a (vector of) confounding factor(s).

In the case of pharmacoepidemiologic studies, drug treatment effects are often time dependent, and affected by time-dependent confounders that are themselves affected by the treatment. An example is the effect of aspirin use on the risk of myocardial infarction (MI) and cardiac death [39]. Prior MI is a confounder of the effect of aspirin use on risk of cardiac death because prior MI is associated with (subsequent) aspirin use, and is associated with (subsequent) cardiac death. However, (prior) aspirin use is also associated with (protective against) the prior MI. Therefore, prior MI is both a predictor of subsequent aspirin use, and predicted by past aspirin use, and hence is a time-dependent

confounder affected by previous treatment. This is depicted in the DAG graph in Figure 2b above. Aspirin use is treatment A, and prior MI is confounder L.

In the presence of time-dependent covariables that are themselves affected by previous treatment, $L(t)$, the estimates of the association of treatment with outcome is unbiased, but it is a biased estimate of the causal effect of a drug of interest on outcome. This bias can be reduced or eliminated by weighting the contribution of each patient i to the risk set at time t by the use of stabilized weights (Hernan et al. 2000 [35]). The stabilized weights, $sw_i(t) =$

$$\prod_{k=0}^{int(t)} \frac{pr(A(k) = a_i(k) | A^*(k-1) = a_i^*(k-1), V = v_i)}{pr(A(k) = a_i(k) | A(k-1) = a_i^*(k-1), L^*(k) = l_i^*(k))}$$

These stabilized weights are used to obtain an IPTW partial likelihood estimate. Here, $A^*(k-1)$ is defined to be 0. The $int(t)$ is the largest integer less than or equal to t , and k is an integer-valued variable denoting days since start of follow-up. Because by definition each patient’s treatment changes at most once from month to month, each factor in the denominator of $sw_i(t)$ is the probability that the patient received his own observed treatment at time $t = k$, given past treatment and risk-factor history L^* , where the baseline covariates V are now included in L^* . The factors in the numerator are interpreted the same, but without adjusting for any past time-dependent risk factors (L^*).

Under the assumption that all relevant time-dependent confounders are measured and included in $L^*(t)$, then weighting by $sw_i(t)$ creates a risk set at time t , where 1) $L^*(t)$ no longer predicts initiation of the drug treatment at time t , that is, $L^*(t)$ is not a confounder), and 2) the association between the drug and the event can be appropriately interpreted as a causal effect (association equals causation).

Standard Cox proportional hazards software does not allow subject-specific weights if they are time-dependent weights. The approach to work around this software limitation is to fit a weighted pooled logistic regression, treating each person-month as an observation [35,40]. Using the weights, $sw_i(t)$, the model is: $logit pr[D(t) = 1 | D(t-1) = 0, A^*(t-1), V] = \beta_0(t) + \beta_1 A(t-1) + \beta_2 V$. Here, $D(t) = 0$ if a patient was alive in month t and 1 if the patient died in month t . In an unweighted case, this model is

equivalent to fitting an unweighted time-dependent Cox model because the hazard in a given single month is small [40]. The use of weights induces a correlation between subjects, which requires the use of generalized estimating equations [15]. These can be estimated using standard software in SAS by the use of Proc GENMOD, with a “repeated” option to model the correlation between observations. Results are obtained in terms of the usual log-odds of the event. The final practical problem to solve is actual estimation of the weights. This is accomplished by essentially estimating the probability of treatment at time t from the past covariable history, using logistic regression to estimate the treatment probabilities in the numerator (without time-dependent confounders) and in the denominator (with time-dependent confounders) [41]. The method is related to propensity scoring, where the probability of treatment is p_i , given covariables [20,42,43]. The IPTW-stabilized weight, $sw_i(t)$, is the inverse of the propensity score for treated subjects, and the inverse of $1 - p_i$, for untreated subjects [39].

IVs Analysis

Sources of Bias in Treatment Effects Estimates

There are a variety of sources of bias that can arise in any observational study. For example, bias can be generated by omitted variables, measurement error, incorrect functional form, joint causation (e.g., drug use patterns lead to hospitalization risk and vice versa), sample selection bias, and various combinations of these problems. One or more of these problems nearly always exist in any study involving observational data. It is useful to understand that, regardless of the source, bias is always the result of a correlation between a particular variable and the disturbance or error term of the equation. Economists refer to this problem as endogeneity, and it is closely related to the concept of residual confounding.

Unfortunately, the researcher never knows how big the endogeneity problem is in any particular study because the disturbance term is unobserved and, as a consequence, so is the extent of the correlation between the disturbance term and the explanatory variable. Given its importance, it is not surprising that the topic of endogeneity has long been an important topic in the econometrics literature. The method of IV is the primary econometric approach for addressing the problem of endogeneity. The IV approach relies on finding at least one variable that is correlated with the endogenous variable but uncorrelated with the outcome. IV approaches for addressing the problem of endogeneity date to the 1920s—although the identity of the inventor remains in doubt and will probably never be established for certain [44]. With more than nine decades to accumulate, the theoretical and applied literature on IVs estimation is vast. IVs and endogeneity are described in all of the major econometrics texts [45,46].

The IVs Approach

In outcomes research applications, endogeneity often raises its head in the form of sample selection bias. This is the case of nonrandom selection into treatment being due to unmeasured variables that are also correlated with the error term of the outcome equation. Sample selection bias methods developed to address this problem [47] are closely related to IVs. For the purposes of simplifying the discussion, we will consider them to be synonymous.

The first step in the estimation of a sample selection model mirrors that of the propensity score approach [48,49]. A model of treatment selection is estimated (generally using a probit

model, rather than logit). Once estimated, this model can be used to predict the probability of selecting treatment A as a function of observable variables, and these predicted probabilities can be compared to the patient’s actual status to calculate a set of empirical residuals. In the second step, the empirical residuals (or, more specifically, a function of these residuals known as the inverse mills ratio) are included as an additional variable. If no endogeneity bias is present, the parameter estimate on the inverse mills ratio will be statistically insignificant. However, if, for example, there are important unmeasured variables that are correlated with both treatment selection and outcomes, the included residuals will not be randomly distributed, and the variable will be either positively or negatively correlated with the outcome variable. Thus, sample selection bias models provide a test of the presence of endogeneity due to nonrandom selection into treatment due to unobserved variables that are correlated with the error term of the outcome equation. Even better, if such endogeneity is present, it is now confined to the IV—like magic the problem is solved!

Sounds Good but Is IV Really the Holy Grail?

Despite the appeal of sample selection or IV methods for addressing the many variants of endogeneity that commonly arise in the analysis of observational data, researchers have raised concerns over the performance of IV and parametric sample selection bias models—noting, in particular, the practical problems often encountered in identifying good instruments. It is remarkably difficult to come up with strong instruments (i.e., variables that are highly correlated with the endogenous variable) that are uncorrelated with the disturbance term. As a result, instruments tend to be either weakly correlated with the variable for which they are intended to serve as an instrument, correlated with the disturbance term, or both. As a consequence, researchers tend to gravitate toward the use of weak instruments to reduce the chance of using an instrument that is itself endogenous. Unfortunately, several studies have shown that weak instruments may lead not only to larger standard errors in treatment estimates but may, in fact, lead to estimates that have larger bias than OLS [50–53].

Staiger and Stock [51] note that empirical evidence on the strength of instruments is sparse. In their review of 18 articles published in the *American Economic Review* between 1988 and 1992 using two-stage least squares, none reported first stage F -statistics or partial R^2 s measuring the strength of identification of the instruments. In several applications of IV to outcomes research problems, however, researchers have reported on the strength of their instruments [48,49,54–56]. This is good practice and should always be done to allow the reader to assess the potential strengths and weaknesses of the evidence presented.

Most recently, Crown et al. [Crown W, Henk H, Van Ness D, unpubl. ms.] have conducted simulation studies that show that even in the presence of significant endogeneity problems and when the researcher has a strong instrument, OLS analysis often leads to less estimation error than IVs. This is because even low correlations between the instrument and the error term introduce more bias than it takes away. Given the tendency to identify weak instruments in the first place, it seems unlikely that IV will actually outperform OLS in most applied situations.

This suggests that, despite the appeal of IV methods, researchers would be well advised to focus their efforts on reducing the sources of bias (omitted variables, measurement error, etc.), rather than wishing for a “magic bullet” from an IV. Among others, these methods include propensity score matching

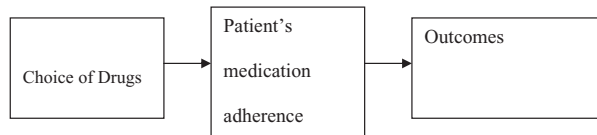


Figure 3 Simplified conceptual framework for path diagram of drug choice to patient outcome.

methods, structural equation approaches, nonlinear modeling, and many of the other methods described elsewhere in this document. That said, researchers should always test for endogeneity using standard specification tests such as one of the many variants of the Hausman test [45,46]. In instances where it is possible to identify strong, uncontaminated instruments, IV methods will yield treatment estimates that are unbiased even when endogeneity is present. For excellent introductions and summaries of the IV literature, the reader may wish to consult Murray [56], Brookhart et al. [Brookhart MA, Rassen JA, Schneeweiss S, unpubl. ms.], and Basu et al. [57].

Structural Equation Modeling

In all the statistical methods discussed thus far, dummy variables are generally used to evaluate treatment effects. Although multivariate models attempt to control for other observable (and in the case of IVs, unobservable) variables, they ultimately measure an expected mean difference in the dependent variable between treatment groups. Structural models enable much more detail about the treatment effects to be elicited.

To illustrate this, consider that pharmaceutical treatment for an illness may generally be characterized by three behavioral processes and associated outcomes: 1) the choice of the drug; 2) the subsequent realization of the patient's medication adherence behavior or drug use patterns; and 3) outcomes (e.g., mortality, survival time, relapse, tumor progression). The conceptual framework that links medical outcomes to drug choice can be represented in the form of a path analysis diagram as follows (Fig. 3).

As seen in the path diagram, we envisage choice of pharmaceutical treatment as having an effect or impact on the patient's compliance behavior or drug use patterns. The arrow that goes from the first to the second box in the path diagram captures this effect. In turn, we expect that patient's medication adherence will impact outcomes. The arrow between the second and third boxes in the path diagram captures this relationship.

The relationships sketched in Figure 3 may be summarized in a general way as follows:

Drug choice (D)	$D = f_0(X, T, Z, H_0)$
Medication Adherence (A)	$A = f_1(D, X, T, Z, H_0)$
Outcomes (O)	$O = f_2(D, A, X, T, H_0)$

Note that the major concepts of interest to us (drug choice, medication adherence, and outcomes) appear on the left, and the relationships among these concepts and their predictors are summarized on the right.

With this notation, f_0 - f_2 refers to the relationships among drug choice, medication adherence, outcomes, and their predictors. Some of these relationships may be linear and others may be nonlinear, as described below. X refers to a vector of explanatory variables that include patient characteristics such as demographic variables (e.g., gender, age, region dummies, diagnosis dummies). T refers to the vector of treatment patients

received in the prior period (e.g., number of psychotherapy visits in the prior period) and baseline health conditions. Z refers to a vector of variables measuring provider characteristics. H_0 refers to baseline health characteristics of the person. In this example the structure is assumed to be recursive in nature (i.e., it is sequential). Furthermore, while a recursive relationship is plausible among the major concept areas described in Figure 3, it is also likely that some of these are determined jointly rather than sequentially. This, along with the potential that some of the equations may be nonlinear, presents a variety of interesting estimation challenges in the statistical modeling of these relationships.

In particular, the drug selection or choice of pharmacotherapy occurs first in the sequence of events. After the drug selection decision, patients generate medication adherence patterns that in turn influence the observed outcomes (i.e., probability of relapse). It is also possible, however, that outcomes can feed back on drug use patterns. For example, a patient hospitalized for mental illness is likely to experience a medication change as a result. Moreover, both drug use patterns and outcomes may be influenced by unobserved factors. As discussed earlier, such patterns of time-varying and omitted variables can lead to biased parameter estimates. Finally, drug use patterns and observed outcomes may be correlated with unobserved variables associated with drug choice.

To illustrate the issues involved in modeling outcomes associated with alternative pharmaceutical treatments, consider the outcomes associated with a decision to treat depression-related illness with an selective serotonin reuptake inhibitor (SSRI) antidepressant versus an serotonin/norepinephrine (SNRI). Drug use patterns are considered as an intermediate outcome that may have a significant effect on costs of treatment. In this analysis, antidepressant use patterns will be defined using a dichotomous variable that identifies antidepressant use as stable (4 or more 30-day prescriptions for the initial antidepressant within the first 6 months) or some other pattern of use.

These two relationships may be expressed in the following equations, which are analogous to f_0 and f_1 above:

$$D = X_1 B_1 + \epsilon_1 \tag{1}$$

$$U = X_2 B_2 + D\pi + \epsilon_2 \tag{2}$$

where D is an indicator of initial SSRI versus SNRI antidepressant selection; U is an indicator of the subsequent antidepressant use pattern that is realized; X_1 and X_2 , are sets of explanatory variables (not mutually exclusive); B_1 , B_2 , and π are parameters to be estimated.

Equation (1) models the selection of the initial antidepressant as a function of explanatory variables that include patient demographics, baseline health conditions, and provider characteristics. Similar explanatory variables appear in the use patterns equation, Equation (2), which also includes the indicator for the class of drug initially prescribed for the patient.

Suppose the research objective was to estimate rates of relapse for patients using SSRIs versus patients using SNRIs. The outcome models would have the general form:

$$Y^t = X_3 B^t + U\theta^t + \lambda^t \Gamma^t + \epsilon_3^t \tag{3a}$$

$$Y^s = X_3 B^s + U\theta^s + \lambda^s \Gamma^s + \epsilon_3^s \tag{3b}$$

where, Y^t and Y^s are outcome variables (i.e., probability of relapse) for SNRI and SSRI patients respectively; X_3 are sets of explanatory variables; B_3 and θ are parameters to be estimated; Γ

and Γ^* are inverse mills ratios, λ^i and λ^s are the associated parameter estimates; and ϵ_1^i and ϵ_1^s are residuals.

Equations (3a) and (3b) specify observed outcomes as a function of patient and provider characteristics and according to the use pattern achieved by the patient with the study antidepressants. Also included in this specification of Equations (3a) and (3b) is an inverse mills ratio, λ , to test for the possibility of unobserved variables that may be correlated with both initial drug selection and time to relapse [47–49,58]. This is the IVs (sample selection bias) approach described in the previous section.

The outcome models are estimated using a variety of statistical techniques, depending upon the nature of the dependent variable. For example, time to relapse could be estimated using Cox Proportional Hazard models. Separate outcome equations for patients treated with SNRIs and SSRIs allow for structural differences in the relationships between observed outcomes and the observable characteristics of patients receiving each type of drug (i.e., different coefficient signs and/or significance of variables in the outcome models for each drug). Including the sample selection terms accounts for the potential influence of the covariance between the residuals of the antidepressant choice (1) and the outcome Equations (3a) and (3b). However, correlation between the residuals of the drug choice Equation (1) and use pattern Equation (2), as well as the correlation of the residuals of the use pattern and outcome Equations (3a) and (3b) may also affect the standard errors and bias of parameter estimates in the various equations. Moreover, the use of IV methods with non-linear outcomes equations is straightforward only for a small number of specific functional forms. For example, although our conceptual structural equation model calls for the use of IV and Cox Proportional Hazard models in combination, the reader will not find this estimator to be available in any statistical software packages.

The appropriate estimation method for the above system of equations critically depends on the structure of the covariance in the error terms across Equations (1)–(3). If the error terms are uncorrelated across equations, each equation can be estimated independently of the others. Often, however, there is reason to believe that the error covariances across each of the three equations may be nonzero. If so, parameter estimates and standard errors may be biased if the equations are estimated independently. If the equations are interrelated, any bias such as that resulting from unobserved variables may be transferred to the other equations as well.

Deriving Treatment Effects from Structural Equation Models

The major challenge with the use of structural equation models is that they do not contain a simple dummy variable providing the magnitude, sign, and statistical significance of the estimated treatment effect. In particular, when separate outcome models are estimated for each treatment cohort, decomposition methods are required to construct the treatment effect estimate. This is done by estimating separate outcome equations for each treatment cohort as in the above example. The coefficients in each equation show the structural relationship between the explanatory variables and the outcome variable within each cohort. In addition to these structural effects, the variables within each treatment cohort may well have different distributions (e.g., different distributions on age, gender, race, medical comorbidities). By substituting the distributions of one treatment cohort through the estimated equation of another cohort, it is possible to estimate

the expected value of the outcome holding both the structural and distributional effects constant. Standard errors for the differences in expected values across treatment groups can then be generated using bootstrapping methods.

Regression-based decomposition methods have not been widely used in outcomes research but have seen considerable use in labor economics to investigate wage disparities by gender and race [59]. Recently, this approach has been used to examine racial disparities in access to health care.

Recommendations

- Include variables that are only weakly related to treatment selection because they may potentially reduce bias more than they increase variance.
- Variables related to outcome should be included in the propensity score despite their strength of association on treatment (exposure) selection.
- All factors that are theoretically related to outcome or treatment selection should be included despite statistical significance at traditional levels of significance.
- In the presence of time-varying confounding, standard statistical methods may be biased, and alternative methods such as marginal structural models or G-estimation should be examined.
- Researchers should always report on the strengths of their instruments to allow the reader to assess the potential strengths and weaknesses of the evidence presented.
- Researchers would be well advised to focus their efforts on reducing the sources of bias (omitted variables, measurement error, etc.), rather than wishing for a “magic bullet” from an IV.
- Residual confounding should be assessed, and approaches to estimating its effect, including sensitivity analyses, should be included.

Residual Confounding

Residual confounding refers to confounding that has been incompletely controlled, so that confounding effects of some factors may remain in the observed treatment-outcome effect. Residual confounding is often only discussed qualitatively without trying to quantify its effect. Yet, methods are available to attempt to assess the magnitude of residual confounding after adjusted effects have been obtained [60,61]. Residual confounding should be assessed and approaches to estimating its effect, including sensitivity analyses, should be included.

Sensitivity Analyses Related to Residual Confounding

The basic concept of these sensitivity analyses is to make informed assumptions about potential residual confounding and quantify its effect on the relative risk estimate of the drug-outcome association [62]. Several approaches are available to obtain a quantitative estimate in the presence of assumed imbalance of the confounder prevalence in the exposure or outcome groups. The array approach varies the confounder prevalence in the exposed versus the unexposed and the magnitude of the confounder–disease association and obtains different risk estimates over a wide range of parameter constellations [63].

Another approach is directed to the question on how strong a single confounder would have to be to move the observed study findings to the null (rule-out approach). This method allows us to rule out confounders that would not be strong enough to bias our results. A limitation of this method is that it is constrained to one

binary confounder and that it does not address the problem of the effect of several unmeasured confounders.

Approaches to reduce residual confounding from unmeasured factors include:

- case-crossover study designs;
 - different time periods with same patient serving as case and control.
- clinical details in a subsample;
 - additional clinical information obtained on a subset of patients to adjust main results.
- proxy measures;
 - measured confounders may be correlated with unmeasured confounders. High dimension propensity scoring may represent unmeasured covariate matrix.
- other methods.
 - IVs.

Conclusions

The analysis of well-designed studies of comparative effectiveness is complex. However, the methods reviewed briefly in this section are relatively well established, in the case of stratification and regression, and/or rapidly on their way to becoming so, in the case of propensity scoring and IV analysis. Other methods, such as marginal structural models and structural equation modeling may not be as common yet in pharmaceutical outcomes research, but we expect these to become more so in the near future. Indeed, one may predict that longitudinal data analysis with time-varying measures of exposure will be almost a requirement of good observational research of treatment effects in the near future. Many other techniques such as multinomial or ordered logit or probit modeling, parametric survival analysis, transition modeling, nested models, G-estimation, and many others could not be treated at all in our report. The use of all of these methods requires extensive training, careful implementation, and appropriate balanced interpretation of findings.

Careful framing of the research question with appropriate study design and application of statistical analysis techniques can yield findings with validity, and improve causal inference of comparative treatment effects from nonrandomized studies using secondary databases.

Source of financial support: This work was supported in part by the Department of Veteran Affairs Health Services Research and Development grant HFP020-90. The views expressed are those of the authors and do not necessarily reflect the views of the Department of Veteran Affairs.

References

- 1 Berger M, Mamdani M, Atkins D, et al. Good research practices for comparative effectiveness research: defining, reporting and interpreting non-randomized studies of treatment effects using secondary data sources. ISPOR TF Report 2009—Part I. *Value Health* 2009; doi: 10.1111/j.1524-4733.2009.00600.x.
- 2 Cox E, Martin B, van Staa T, et al. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of non-randomized studies of treatment effects using secondary data sources. ISPOR TF Report 2009—Part II.
- 3 Mothel B, Brooks J, Clark MA, et al. A checklist for retrospective database studies—report of the ISPOR task force on retrospective database studies. *Value Health* 2003;6:90–7.
- 4 Iezzoni LI. The risks of risk adjustment. *JAMA* 1997;278:520.
- 5 Rothman KJ, Greenland S, Lash TJ. *Modern Epidemiology* (3rd ed.). Philadelphia, PA: Lippincott Williams and Wilkins, 2008.
- 6 Kleinbaum D, Kupper L, Morgenstern H. *Epidemiologic Research: Principles and Quantitative Methods*. New York: John Wiley and Sons, 1982.
- 7 Kleinbaum D, Kupper L, Muller K, Nizam A. *Applied Regression Analysis and Other Multivariable Methods* (3rd ed.). Pacific Grove, CA: Duxbury Press, 1998.
- 8 Harrell FE. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer, 2001.
- 9 Hosmer DW, Lemeshow S. *Applied Logistic Regression* (2nd ed.). New York: John Wiley and Sons, 2000.
- 10 Hosmer DW, Lemeshow S, May S. *Applied Survival Analysis. Regression Modeling of Time to Event Data* (2nd ed.). New York: John Wiley and Sons, 2008.
- 11 Kuykendall DH, Johnson ML. Administrative databases, case-mix adjustments and hospital resource use: the appropriateness of controlling patient characteristics. *J Clin Epidemiol* 1995;48:423–30.
- 12 Iezzoni LI. *Risk Adjustment for Measuring Health Care Outcomes* (3rd ed.). Chicago: Health Administration Press, 2003.
- 13 Diehr P, Yanez D, Ash A, et al. Methods for analyzing health care utilization and costs. *Annu Rev Public Health* 1999;20:125–44.
- 14 Thompson SG, Barber JA. How should cost data in pragmatic randomised trials be analysed? *BMJ* 2000;320:1197–200.
- 15 Diggle PJ, Liang K-Y, Zeger SL. *Analysis of Longitudinal Data*. New York: Oxford University Press, 1996.
- 16 Montaquila JM., Ponikowski CH. An evaluation of alternative imputation methods. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1995.
- 17 Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, 1987.
- 18 McWilliams JM, Meara E, Zaslavsky AM, Ayanian JZ. Use of health services by previously uninsured Medicare beneficiaries. *N Engl J Med* 2007;357:143–53.
- 19 Fu AZ, Liu GG, Christensen DB, Hansen RA. Effect of second-generation antidepressants on mania- and depression-related visits in adults with bipolar disorder: a retrospective study. *Value Health* 2007;10:128–36.
- 20 Rosenbaum PR, Rubin DB. The Central role of propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
- 21 D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998;17:2265–81. PMID: 9802183.
- 22 Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol* 2006;98:253–9.
- 23 Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. *Value Health* 2006; 9:377–85.
- 24 Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med* 2007;27:2037–49.
- 25 Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat* 1985;39:33–8.
- 26 Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol* 2006;163:1149–56.
- 27 Rubin DB, Thomas N. Combining propensity score matching with additional adjustment for prognostic covariates. *J Am Stat Assoc* 2000;95:573–85.
- 28 Seeger JD, Kurth T, Walker AM. Use of propensity score technique to account for exposure-related covariates: an example and lesson. *Med Care* 2007;45(10 Suppl. 2):S143–8.
- 29 Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am J Epidemiol* 2005;162:279–89.
- 30 Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005;58:550–9.

- 31 Stukel TA, Fisher ES, Wennberg DE, et al. Analysis of observational studies in the presence of treatment selection bias. *JAMA* 2007;297:278–85.
- 32 Robins JM. Marginal Structural Models. 1997 Proceedings of the Section on Bayesian Statistical Science. Alexandria, VA: American Statistical Association, 1998.
- 33 Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: Halloran E, Berry D, eds. *Statistical Models in Epidemiology: The Environment and Clinical Trials*. New York: Springer-Verlag, 1999.
- 34 Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11:550–60.
- 35 Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000;11:561–70.
- 36 Pearl J. Causal diagrams for empirical research. *Biometrika* 1995;82:669–88.
- 37 Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiological research. *Epidemiology* 1999;10:37–48.
- 38 Robins JM. Data, design, and background knowledge in etiologic inference. *Epidemiology* 2001;11:313–20.
- 39 Cook NR, Cole SR, Hennekens CH. Use of a marginal structural model to determine the effect of aspirin on cardiovascular mortality in the Physicians' Health Study. *Am J Epidemiol* 2002;155:1045–53.
- 40 D'Agostino RB, Lee M-L, Belanger AJ. Relation of pooled logistic regression to time-dependent Cox regression analysis: the Framingham Heart Study. *Stat Med* 1990;9:1501–15.
- 41 Mortimer KM, Neugebauer R, van der Laan M, Tager IB. An application of model-fitting procedures for marginal structural models. *Am J Epidemiol* 2005;162:382–8.
- 42 Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997;127:757–63.
- 43 Johnson ML, Bush RL, Collins TC, et al. Propensity score analysis in observational studies: outcomes following abdominal aortic aneurysm repair. *Am J Surg* 2006;192:336–43.
- 44 Stock JH, Trebbi F. Who invented instrumental variables regression? *J Econ Perspect* 2003;17:177–94.
- 45 Greene WH. *Econometric Analysis* (3rd ed.). New York: Maxwell-MacMillan, 1997.
- 46 Wooldridge J. *Econometric Analysis of Cross-Section and Panel Data*. Cambridge: MIT Press, 2002.
- 47 Heckman J. The common structure of statistical models of truncation, sample selection, and limited dependent variables and an estimator for such models. *Ann Econ Soc Meas* 1976;5:475–92.
- 48 Crown W, Hylan T, Meneades L. Antidepressant selection and use and healthcare expenditures. *Pharmacoeconomics* 1998a;13:435–48.
- 49 Crown WR, Obenchain L, Englehart T, et al. Application of sample selection models to outcomes research: the case of evaluating effects of antidepressant therapy on resource utilization. *Stat Med* 1998b;17:1943–58.
- 50 Bound JD, Jaeger A, et al. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *JASA* 1995;90:443–50.
- 51 Staiger D, Stock JH. Instrumental variables regression with weak instruments. *Econometrica* 1997;65:557–86.
- 52 Hahn J, Hausman J. A new specification test for the validity of instrumental variables. *Econometrica* 2002;70:163–89.
- 53 Kleibergen F, Zivot E. Bayesian and classical approaches to instrumental variables regression. *J Econom* 2003;114:29–72.
- 54 Brooks J, Chrischilles E. Heterogeneity and the interpretation of treatment effect estimates from risk adjustment and instrumental variable methods. *Med Care* 2007;45(Suppl. 2):S123–30.
- 55 Hadley P, Weeks MM. An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a medicare population. *Health Econ* 2003;12:171–86.
- 56 Murray M. Avoiding invalid instruments and coping with weak instruments. *J Econ Perspect* 2007;20:111–32.
- 57 Basu A, Heckman J, 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. *Health Econ* 2007;16:1133–57.
- 58 Heckman J. Sample selection as a specification error. *Econometrica* 1979;47:153–61.
- 59 Oaxaca R., Ransom M. On discrimination and the decomposition of wage differentials. *J Econometrics* 1994;61:5–21.
- 60 Psaty BM, Koepsell TD, Lin D, et al. Assessment and control for confounding by indication in observational studies. *J Am Geriatr Soc* 1999;47:749–54.
- 61 Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology* 2003;14:451–8.
- 62 Schneeweiss S. Developments in post-marketing comparative effectiveness research. *Clin Pharmacol Ther* 2007;82:143–56.
- 63 Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf* 2006;15:291–303.