Published in final edited form as:

Heart Fail Clin. 2013 January ; 9(1): 29–36. doi:10.1016/j.hfc.2012.09.007.

Epidemiologic and Statistical Methods for Comparative Effectiveness Research

Mark A. Hlatky, MD^{a,*}, Wolfgang C. Winkelmayr, MD, MPH, ScD^b, and Soko Setoguchi, MD, DrPH^c

^aDepartment of Health Research and Policy, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA

^bDepartment of Medicine, Stanford University School of Medicine, Stanford, CA, USA

^cDepartment of Medicine, Duke University Medical Center, Durham, NC, USA

Keywords

Comparative effectiveness research; Randomized trials; Observational data; Statistical methods; Epidemiologic methods

Comparative effectiveness research (CER) seeks to “compare the benefits and harms of alternative methods to prevent, diagnose, treat and monitor a clinical condition or to improve the delivery of care.”¹ This broad scope can only be addressed by a diversified research portfolio that encompasses many types of studies, including randomized trials, observational studies, simulation models, and meta-analyses.² Although each of these designs has methodologic issues, observational studies to compare treatments have methodologic challenges that threaten the internal validity of their results.^{3,4} This article reviews several approaches to the analysis of observational data that are in common use, or that may have promise even though they are not yet often applied. As a starting point, the analysis of randomized trials is reviewed, because many of the observational data analysis methods seek to emulate some key features of randomized trials.

RANDOMIZED TRIALS

The randomized clinical trial (RCT) is the benchmark for research quality against which all other study designs are compared. What aspects of the RCT make it the reference standard for quality? Many of the strengths of an RCT flow from the fact that it is an experiment, and the procedures to be followed in recruitment of patients, data collection, intervention, outcomes, and analyses are determined in advance by investigators. The types of patients to be enrolled are prespecified by inclusion and exclusion criteria, and suitable clinical centers are enlisted. The important pieces of information about study patients are identified at the outset, and data forms are designed to collect this essential information. Procedure manuals are written to standardize data collection, and the optimum treatment regimens are identified and applied uniformly and expertly. All patients are followed closely, and predetermined endpoint data are collected completely. Finally, the data are analyzed according to a prespecified plan to test study hypotheses. It is readily apparent from this litany that a major reason RCTs are so highly regarded is the care and planning that goes into every step of this

process. It is also apparent that most of these same strengths are not specific to RCTs, and could be applied to other research studies, such as high-quality clinical data registries.

The sine qua non of the RCT is that treatment is assigned based on the value of a random number (a coin flip) rather than by the choice of the patient or treating physician. This crucial step means that patients assigned to alternative treatments are equivalent at the start of the study in all respects, within the play of chance. If the number of patients is “large,” it is very unlikely that any baseline characteristic will be seriously imbalanced between the treatment groups. For instance, if there are 20 patients with a particular characteristic, a 50:50 randomization might lead to as many as 5 of these patients in one group and 15 in the other, but with 200 patients a split more extreme than 85/115 is unlikely, and for 1000 patients a split more extreme than 468/532 is unlikely. Certain procedures, such as block randomization or adaptive randomization, can further reduce the likelihood that certain key characteristics are unbalanced among treatment groups.

Although the randomization process ensures balance between the study patient groups at baseline, differences between groups may develop after randomization in background therapy, frequency of monitoring, and so forth that could affect subsequent outcomes. Randomized trials use protocols to minimize postrandomization differences in treatment or monitoring between randomly assigned groups. Blinding is a key principle that should ensure that neither the patient nor the treating physician becomes aware of the actual exposure and change their behavior in response. Clearly, it is impossible to blind for certain exposures. An important third blinding mechanism is the independent adjudication of key study outcomes of specialists that remain unaware of each patient’s exposure status.

The statistical analysis of RCTs relies on the intention-to-treat principle to maintain the benefits of random assignment and avoid bias introduced by informative censoring and crossover of treatments. Thus, even if some patients assigned to Treatment A do not take it, or cross over to alternative Treatment B, their outcomes are attributed to Treatment A in the data analysis.

For all these reasons, RCTs are considered to have the highest internal validity and trials alone are considered to be capable of producing the highest level of evidence. A frequent downside of RCTs, however, is the highly selected nature of the enrolled subjects, which may limit the generalizability of the findings to the more general population of interest, or to specific subgroups of patients. In addition, trials are usually powered to detect important differences in desired outcomes, but may not be sufficiently large to compare the safety of the comparators of interest.

DATA QUALITY

Unlike RCTs, treatment in observational studies is not determined by randomization but by clinical or nonclinical reasons (indications). Confounding by indication and selection bias are the biggest threats to the internal validity of observational CER, and epidemiologic and statistical methods are needed to minimize their effects on study results. The most basic and important way to strengthen the validity of observational CER is to use high-quality data about the patients, their treatment, and their subsequent outcomes.

The quality of data should be assessed by at least three aspects: (1) kinds of information (types of variables); (2) accuracy of that information (specificity and sensitivity); and (3) amount of information (completeness). A clinical registry of patients with heart failure could, for instance, collect data using standardized forms and data definitions, which would ensure that the key information about the patients (especially about their heart failure characteristics and other cardiovascular comorbidities) is available for analysis. However,

the completeness and accuracy of information can vary considerably depending on the method of collection and incentives of providers. Also, the follow-up information in many clinical registries is often short (mostly up to 6 months) and incomplete. Claims data, such as Medicare Part A, B, and D files, provide longitudinal data on a large real world representative population. For major clinical outcomes, such as hospitalizations and death, claims data provide accurate and complete information. However, billing diagnoses submitted to Medicare or private insurers are less accurate than diagnoses in registries or electronic health records. Furthermore, claims data do not contain important clinical information, such as New York Heart Association functional class; findings from the physical examination; or the results of laboratory or imaging examinations (eg, ejection fraction). Solutions to overcome limitations of various data sources include linking existing databases (eg, link a registry with claims data, link electronic health records or laboratory records to claims data)^{5,6} and collecting additional information to supplement existing databases by surveying patients or reviewing medical records (a more comprehensive review is provided elsewhere in this issue).

NEW-USER DESIGNS

Another advantage of randomized trials is that “time zero” is well defined by the date of randomization, and treatments are initiated at “time zero.” The paramount importance of mimicking this design in observational CER, where it is called “new-user design,” has long been recognized.⁷ Inclusion of patients who have been on a treatment for some time before “time zero” for a study may lead to biased associations. Prevalent users (patients who started a drug and remained on a drug) are a selected subset, because patients who cannot tolerate the drug, patients who develop complications, and patients who are not adherent have all dropped out, and the remaining users are more likely to be adherent, have favorable health behaviors, and be tolerant of the study drug. Also, including prevalent users in a treatment group can underestimate the event rate in the treatment group if events early in the course of treatment are not captured. Comparing prevalent users with other patients is likely to give a biased picture of the treatment’s safety and efficacy.

The new-user design attempts to minimize these problems by identifying a cohort of patients who are newly started on a treatment.⁷ The implementation of the design requires a look-back period of some duration (typically 6–12 months in claims data) to ensure that the patients were previously not on a treatment of interest. A related strategy is to identify patients with new onset of the disease who are then started on the study drug, because newly diagnosed individuals are unlikely to have been previously treated. For example, patients with new onset of heart failure may be less likely to have been treated with multiple prior drug regimens, and more likely to be new users of heart failure drugs. However, many drugs used to treat heart failure can also be used to treat hypertension (eg, angiotensin-converting enzyme inhibitors, β -blockers) or other cardiac conditions (eg, digoxin for atrial fibrillation), so merely identifying patients with new-onset heart failure does not necessarily identify new users of these drugs. By contrast, patients with newly diagnosed diabetes are most unlikely to have been treated previously with metformin, proglitazone, or insulin, because these drugs are used solely to treat diabetes.

The new-user design is strongly recommended to control biases introduced by including prevalent users (survivor bias and selection bias), but in and of itself does not fully control for confounding bias. After a suitable cohort of new users is identified, other epidemiologic and statistical techniques are still needed to control for confounding, especially for confounding by indication.

Related to the new-user design is the concept of studying only “treatment eligible” patients. For example, if one wished to compare two drugs, one of which is contraindicated in patients with chronic kidney disease, it would bias the results of a comparison study to include any patients with chronic kidney disease. The patients with chronic kidney disease are likely to have worse outcomes than other patients, and if they were predominantly treated with one drug, their worse outcomes would bias the study results against that drug. So, when comparing two drugs, patients with contraindications to one or the other drug are best excluded, much as they would be in a randomized trial of the two drugs. Furthermore, there is no clinical question about which drug to choose when one of the drugs is contraindicated.

OVERVIEW OF STATISTICAL METHODS TO CONTROL CONFOUNDING

Confounding and selection bias are the biggest threats to the validity in observational CER. In observational CER studies, it is necessary to use epidemiologic and statistical methods that attempt to achieve the balance in prognostic characteristics of patients to avoid confounding without randomization.⁸ Various methods exist to control for potential confounding (Fig. 1). Although the importance of classic design and methods, such as restriction or stratification, should not be underestimated, this article focuses on the concept and use of methods that are more recent and increasing in their popularity. First discussed are methods to adjust for measured confounding: propensity score-based approaches and marginal structural models as representatives of so-called causal methods. Next reviewed are the concept and use of instrumental variable (IV) analysis as a method to control for measured and unmeasured confounding and the importance of sensitivity analyses in assessment of unmeasured confounding.

PROPENSITY SCORES

In the real world, treatments are determined by physicians and patients based on various factors including patient characteristics, clinical presentations, availability of treatment options, and patients’ and physicians’ preferences, but not by randomization. When assessing the effect of treatments, controlling for confounding is essential because these factors influencing the choice of treatments often affect the outcome of interest. The propensity score method attempts to control for confounding by balancing the factors influencing the treatment choice. The first step for propensity score analysis is to develop a model predicting the treatment based on measured factors (covariates), which yields a probability that any individual patient will receive Treatment A instead of Treatment B. This exposure propensity score model is typically developed using a multivariable logistic regression modeling in which the dependent variable is a dichotomous variable indicating the received treatment, and the independent (predictor) variables are a set of variables for measured patient characteristics and other factors potentially influencing the treatment choice:

$$P(T_{ij}) = f(X_i, Y_j)$$

where

T_{ij} = the treatment given to patient “i” by provider “j”

X_i = patient-level characteristics

Y_j = provider-level characteristics

For instance, the treatment might be prescribed to a patient with heart failure on the basis of the patient’s age, gender, cause of heart failure, functional class, left ventricular ejection fraction, and QRS duration on the electrocardiogram. An exposure propensity score model

estimates a set of regression coefficients for all included variables for the study population. Because each patient's characteristics are (ideally) known ("covariate vector"), this information can then be combined with the regression coefficients to calculate each patient's predicted probability of receiving Treatment A, which is the propensity score. Because it is a probability, the propensity score is bound by [0, 1].

The second step of propensity score analysis, after the propensity score is estimated, is to achieve a balance in the propensity scores between the two treatment groups (Treatment A and B in this example). Three different approaches based on propensity scores are currently used and capable of achieving balanced characteristics between treatment groups (Box 1) and controversy remains about the optimal approach. One key prerequisite in selecting and justifying an optimal approach is to compare the distribution of estimated propensity scores between patients who received Treatment A and Treatment B. In the unlikely event that these two distributions have no or very little overlap (indicated by almost perfect predictive capability of the model [eg, measured by a concordance (*c*) statistic of the exposure propensity score model very close to 1]), the conclusion would be that patients getting the alternative treatments are so completely different that they may not be validly compared. For example, if all the men got Treatment A and all the women got Treatment B, it would be impossible to disentangle the effects on outcome of treatment and of gender.

As long as the distributions of propensity scores overlap, Treatment A might be compared with Treatment B by propensity score matching, stratification, or direct adjustment in regression analysis, with or without trimming (see Box 1). Matching patients on propensity score is the method suggested by Rubin⁸ and ensures that nonoverlapping patients are excluded from the analyses. This method also has intuitive appeal and face validity. There are various algorithms for matching groups of patients on propensity score that range from selecting an exact match to closest match with various fineness of the caliper. The goal of propensity score analyses is to achieve the balance in measured covariates in the two treatment groups. By matching on propensity score, the study population has similar distribution of measured covariates, which is seen in two groups with random treatment assignment. However, it is crucial to recognize that propensity score analysis, in theory and in general, only achieves balance for measured factors that are included in the propensity score model, whereas randomization achieves balance for measured and unmeasured factors. Other propensity score-based approaches can be thought of as extensions of the matching process by relaxing the matching caliper (see Box 1). Stratification on propensity score involves dividing the entire population into quintiles (or deciles) of propensity score, and within each quintile comparing the outcomes of patients who received Treatment A with the outcomes of patients who received Treatment B. The patients in each quintile should all have sufficiently "similar" propensity scores, even though they were not matched.

Finally, propensity scores can be used as independent predictor variables in outcome models (typically a Cox or logistic model) to adjust for the propensity to receive Treatment A versus Treatment B.

The stratification and regression approaches to using the propensity score have been criticized for including patients in the tails of the distribution of scores who are extremely unlikely to receive the alternate therapy and have no "match." This criticism can be overcome to some extent by restricting the analysis to patients who fall within the overlapping regions of the distribution of propensity scores (ie, by trimming the distribution, then comparing outcomes of the remaining patients) (see Box 1).⁹

Theoretically, the propensity score method has the advantage of reducing all the factors affecting treatment decisions into a single number. When comparing the results from

multivariable models with propensity score adjusted models in studies that reported both, no meaningful differences in effect estimates were observed in almost all studies,^{10,11} suggesting that propensity score adjustment adds little compared with using traditional multivariable adjustment for the same factors in the outcome model. Only in situations where outcomes are too scarce for adjustment of all relevant variables in a multivariable model because of the risk of model overfitting, propensity scores may provide a methodologic alternative. In addition to its use in situations of rare outcomes, another advantage of propensity score approach over traditional multivariate analyses is the possibility to remove extreme cases of treatment assignment from the dataset. Specifically, it is desirable to remove patients that have extremely low propensity for receiving Treatment A, but still received it, or patients with extremely high propensity for Treatment A, who still received Treatment B. These individuals usually reside in the area of nonoverlap of the two propensity score distributions. Frequently, these patients also often have either particularly bad or particularly good prognosis, which makes them influential observations. Propensity matching is an excellent method to eliminate such observations, as is “trimming,” removal of observations in the extremes of the propensity score distributions or the areas of nonoverlap. Using either of these methods increases internal validity. There may be some loss of external validity, however, but this is often a useful compromise.

Once again, the fundamental limitation to any propensity score approach is that the method does not adjust for unmeasured confounding. Some factors that were important in the treatment decision and were associated with the outcome of interest may not have been recorded in the data, and hence cannot be included in the propensity score model and cannot be adjusted. For instance, a clinician’s gestalt of patients who “look older than their stated age” or who seem to have “poor protoplasm” may be important in the choice of treatment and the patient’s subsequent outcome. Even well-designed heart failure registries do not capture this kind of clinician’s impression, so residual confounding caused by this unmeasured factor would remain even after exact matching on propensity score. Again, this problem of residual confounding caused by unmeasured factors does not occur in a randomized trial, because patients with “poor protoplasm” who are randomized will still be equally likely to be assigned to Treatment A or Treatment B.

MARGINAL STRUCTURAL MODELS

Heart failure can be treated with devices and procedures (eg, cardiac surgery) that are applied once and have long-lasting effects. These exposures are relatively straightforward to measure. In contrast, pharmacologic treatments for heart failure are supposed to be chronic treatment regimens, but they may be changed over time or discontinued because of side effects or poor adherence. It is much more complicated to assess the effectiveness of these time-varying treatments, because accurate identification of exposure and assessment and adjustment of time-varying confounding are challenging. To adjust for time-varying confounding, highly sophisticated approaches have been developed.

Marginal structural models, a member of the family of g-estimation methods, were initially developed to assess the efficacy of antiretroviral treatment for AIDS.¹² The HIV treatment regimens were often changed in response to results of laboratory tests, such as CD4 cell counts, that were also prognostic markers and therefore became confounders that were time-varying in studies assessing the effectiveness of HIV treatment regimen. In this situation, the probability of treatment at any time t_i depended on the patient’s condition at time t_i , and on prior treatment at time t_{i-1} . In a marginal structural model, the outcomes of patients who receive treatment A at time t_i are weighted by the inverse probability that they received Treatment A, and the outcomes of patients who received Treatment B are weighted by the inverse probability of receiving Treatment B (if there are only two alternatives $P_i[B] = 1 - P_i$

[A]). The outcomes of Treatment A and B are then compared in a weighted model. Note that the probability of receiving a given treatment is akin to a propensity score, although these probabilities are used in the analysis as weights.

Marginal structural models are in principle a powerful tool for dissecting the effect of treatments that vary over time, where time-varying confounding threatens the internal validity, such as drugs or monitoring strategies. These models depart from the “intention-to-treat” paradigm for analyzing the effects of drug treatment in an RCT, however, because periods of time when a patient is off Treatment A are not counted against Treatment A in a marginal structural model, but they are in an RCT analyzed under the intention-to-treat principle. The extension of the method can be used to model and adjust for informative censoring or discontinuation. Despite the attractiveness of the idea and extensive research on the method, it has not been used extensively.¹³ One of the reasons and also the biggest challenge in operationalizing and effectively making use of the method is lack of detailed data on the time-varying exposure (treatment) and time-varying confounders in most data sources. Without precise data on the exposure and confounders that are measured in a time-varying manner, the method does add beyond other traditional methods. In addition, the method can still not account for any confounders, time-dependent or not, which are unrecorded.

INSTRUMENTAL VARIABLES

Economists rarely do experimental studies and consequently have developed many powerful tools to analyze observational data and draw causal inferences. Health economists have applied some of these econometric techniques to compare the effect of treatments on clinical outcomes. IV methods are particularly attractive for this purpose. The distinct characteristics of the IV method, as opposed to propensity scoring and marginal structural models, is its theoretical ability to adjust for unmeasured confounding.^{14,15} The method has been applied in several prominent studies in cardiovascular medicine.^{16,17}

Conceptually, an “instrument” is a variable that predicts the treatment used, and does not directly or indirectly affect subsequent outcome (ie, is only correlated with outcome to the extent that it is correlated with choice of treatment). Randomization is actually an example of a perfect “instrument”: it is highly correlated with choice of treatment, and only affects outcome through the choice of treatment, not by itself. Identification of a good instrument is not straightforward, however, and usually requires a clever insight into the specific treatment at hand. For instance, formulary committees at different hospitals may select different drugs for use in patients with heart failure, facilitating observational comparisons among similar patient populations.¹⁸ Some tests and procedures may be available on weekdays but not on weekends, leading to different management of otherwise similar patients admitted on different days of the week. A new therapy may be introduced on a certain date, and patients treated before that date could not receive it, whereas patients treated later could receive the new therapy. Identification of such natural experiments can be exploited to compare treatment outcomes even though patients were not randomized.

If a suitable instrument can be found, an IV analysis basically compares patients who have different exposures to the instrument, but are otherwise similar:

$$O(X_i) = f(X_i, IV_i)$$

where

X_i = patients characteristics known to affect outcome, and

IV_i = value of the instrumental variable for patient “i”

If the IV shifts 20% of patients from Treatment A to Treatment B and outcomes change by 2%, the IV approach would estimate that a 100% shift from Treatment A to Treatment B would lead to a 10% change in outcome. This estimate of the treatment effect can be sharpened by additional statistical controls for differences in baseline clinical characteristics.

The major limitation of the IV approach is the difficulty in identifying a suitable instrument, because good “natural experiments” are haphazard. If the instrument is weak and shifts in treatment are very small (eg, from 40% use to 42% use of a drug or device), it is less convincing that any outcome differences are truly because of differences in treatment, and not due to differences in other clinical factors. Also, even if a suitable or reasonable instrument can be identified, the assumptions that are necessary for the IV to be valid cannot be tested or validated fully. It is usually possible to judge whether observed baseline characteristics are balanced across levels of the instrument. If that is the case, one may be assured that the instrument may also balance unobserved characteristics. If the instrument does not balance observed characteristics, one can certainly include such observed information in formal IV analysis; however, it is less plausible that unobserved characteristics are balanced across levels of the instrument. The judgment on the validity of the IV analysis will be left to theoretical and substantive discussion.

SUMMARY

Observational methods are evolving in response to the widespread availability of data from clinical registries, electronic health records, and administrative databases. These approaches will never eliminate the need for randomized trials, but clearly have a role in evaluating the effect of therapies in unselected populations treated in routine practice.

REFERENCES

1. Institute of Medicine. Initial national priorities for comparative effectiveness research. Washington: The National Academies Press; 2009.
2. Hlatky MA, Douglas PS, Cook NL, et al. Future directions for cardiovascular disease comparative effectiveness research. Report of a workshop sponsored by the National Heart, Lung, and Blood Institute. *J Am Coll Cardiol*. 2012; 60:569–580. [PubMed: 22796257]
3. Normand SL. Some old and some new statistical tools for outcomes research. *Circulation*. 2008; 118:872–884. [PubMed: 18711024]
4. Johnson ML, Crown W, Martin BC, et al. 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. *Value Health*. 2009; 12:1062–1073. [PubMed: 19793071]
5. Setoguchi S, Glynn RJ, Avorn J, et al. Statins and the risk of lung, breast, and colorectal cancer in the elderly. *Circulation*. 2007; 115:27–33. [PubMed: 17179016]
6. Hammill BG, Hernandez AF, Peterson ED, et al. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. *Am Heart J*. 2009; 157:995–1000. [PubMed: 19464409]
7. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol*. 2003; 158:915–920. [PubMed: 14585769]
8. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007; 26:20–36. [PubMed: 17072897]
9. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol*. 2006; 163:262–270. [PubMed: 16371515]
10. Shah BR, Laupacis A, Hux JE, et al. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol*. 2005; 58:550–559. [PubMed: 15878468]

11. Stürmer T, Joshi M, Glynn RJ, et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol.* 2006; 59:437–447. [PubMed: 16632131]
12. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000; 11:550–560. [PubMed: 10955408]
13. Mehrotra R, Chiu YW, Kalantar-Zadeh K, et al. Similar outcomes with hemodialysis and peritoneal dialysis in patients with end-stage renal disease. *Arch Intern Med.* 2011; 171:110–118. [PubMed: 20876398]
14. Rassen JA, Brookhart MA, Glynn RJ, et al. Instrumental variables I: instrumental variables exploit natural variation in nonexperimental data to estimate causal relationships. *J Clin Epidemiol.* 2009; 62:1226–1232. [PubMed: 19356901]
15. Brookhart, MA.; Rassen, JA.; Schneeweiss, S. Effective health care research report No. 22 (Prepared by Brigham and Women’s Hospital DEcIDE Center Under Contract No. 290-2005-00161 T03). Rockville (MD): Agency for Healthcare Research and Quality; 2010. Instrumental variable methods in comparative safety and effectiveness research. Available at: <http://effectivehealthcare.ahrq.gov/reports/final.cfm>.
16. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *JAMA.* 1994; 272:859–866. [PubMed: 8078163]
17. Stukel TA, Fisher ES, Wennberg DE, et al. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA.* 2007; 297:278–285. [PubMed: 17227979]
18. Johnston SC, Henneman T, McCulloch CE, et al. Modeling treatment effects on binary outcomes with grouped-treatment variables and individual covariates. *Am J Epidemiol.* 2002; 156:753–760. [PubMed: 12370164]

KEY POINTS

- Methods for observational comparative effectiveness research are evolving in response to the widespread availability of data from clinical registries, electronic health records, and administrative databases.
- To conduct valid observational studies evaluating the effect of therapies in unselected populations treated in routine practice, researchers need to understand the strengths and limitations of rigorous design options and statistical methods and use them appropriately.

Box 1**Methods for propensity score adjustment**

Matching	Match on the value of propensity score. The most commonly used matching method is greedy matching.
Strafitification	Stratify by quintiles ~ deciles of propensity score. Can be done with or without trimming.
Propensity score adjustment in outcome regression model	Include propensity score as independent variable in a regression outcome model. Propensity score can be included as a continuous variables or indicator variables for quintiles or deciles. Can be done with or without trimming.
Trimming	Exclude patients with extreme values for propensity scores to avoid inclusion of nonoverlapping or less overlapping patients.

Fig. 1.

Methods for controlling confounding. (*Adapted from Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiol Drug Saf 2006;15(5):291–303; with permission.*)