Chapter 10. Considerations for Statistical Analysis

Patrick G. Arbogast, Ph.D. (deceased)
Kaiser Permanente Northwest, Portland, OR

Tyler J. VanderWeele, Ph.D.
Harvard School of Public Health, Boston, MA

Abstract

This chapter provides a high-level overview of statistical analysis considerations for observational comparative effectiveness research (CER). Descriptive and univariate analyses can be used to assess imbalances between treatment groups and to identify covariates associated with exposure and/or the study outcome. Traditional strategies to adjust for confounding during the analysis include linear and logistic multivariable regression models. The appropriate analytic technique is dictated by the characteristics of the study outcome, exposure of interest, study covariates, and the underlying assumptions underlying the statistical model. Increasingly common in CER is the use of propensity scores, which assign a probability of receiving treatment, conditional on observed covariates. Propensity scores are appropriate when adjusting for large numbers of covariates and are particularly favorable in studies having a common exposure and rare outcome(s). Disease risk scores estimate the probability or rate of disease occurrence as a function of the covariates and are preferred in studies with a common outcome and rare exposure(s). Instrumental variables, which are measures that are causally related to exposure but only affect the outcome through the treatment, offer an alternative to analytic strategies that have incomplete information on potential unmeasured confounders. Missing data in CER studies is not uncommon, and it is important to characterize the patterns of missingness in order to account for the missing data in the analysis. In addition, time-varying exposures and covariates should be accounted for to avoid bias. The chapter concludes with a checklist including guidance and key considerations for developing a statistical analysis section of an observational CER protocol.

Introduction

Comparative effectiveness research utilizing observational data requires careful and often complex analytic strategies to adjust for confounding. These can include standard analytic strategies, such as traditional multivariable regression techniques, as well as newer, more sophisticated methodologies, such as propensity score matching and instrumental variable analysis. This chapter covers data analysis strategies from simple descriptive statistics to more complex methodologies. Also covered are important considerations such as handling missing data and analyzing time-varying exposures and covariates.

While this chapter provides a high-level summary of considerations and issues for statistical analysis in observational CER, it is not intended to be a comprehensive treatment of considerations and approaches. We encourage the reader to explore topics more fully by referring to the references provided.

Descriptive Statistics/ Unadjusted Analyses

Appropriate descriptive statistics and graphical displays for different types of data have been presented in numerous textbooks.[1] These include measures of range, dispersion, and central tendency for continuous variables, number and percent for categorical variables, and plots for evaluating data distributions. For comparative effectiveness research (CER), it is important to consider useful and informative applications of these descriptive statistics. For instance, for a cohort study, describing study covariates stratified by exposure levels provides a useful means to assess imbalances in these measures. For a propensity–matched-pairs dataset, summarizing study covariates by exposure group aids in detecting residual imbalances.

Univariate or unadjusted hypothesis testing, such as two-sample t-tests, can be conducted to identify covariates associated with the exposure and/or the study outcome. Since CER studies will need to consider potential confounding from a large number of study covariates, the descriptive statistics should provide a broad picture of the characteristics of the study subjects.

Adjusted Analyses

Traditional Multivariable Regression

Regression analysis is often used in the estimation of treatment effects to control for potential confounding variables.[2] In general, control is made for pretreatment variables that are related to both the treatment of interest and the outcome of interest. Variables that are potentially on the pathway from treatment to outcome are not controlled for, as control for such intermediate variables could block some of the effect of the treatment on the outcome. See chapter 7 (Covariate Selection) for further discussion. Traditional multiple regression, in which one uses regression models to directly adjust for potential confounders and effect modification, has long been used in observational studies and can be applied in CER. When applying regression modeling, careful attention must be paid to ensure that corresponding model assumptions are met.[3] For example, for linear regression, the assumption that the mean of the outcome is a linear function of the covariates should be assessed. Whether regression techniques or other approaches are preferred also depends in part on the characteristics of the data. For logistic regression, as long as the number of outcome events per covariate included in the regression model is sufficient (e.g., a rule of thumb is 10 or more) and the exposure of interest is not infrequent, traditional multiple regression is a reasonable strategy and could be considered for the primary analysis.[4-5] However, when this is not the situation, other options should be considered. Regression methods also have the disadvantage that they may extrapolate to regions where data are not available; other techniques such as propensity scores (discussed below) more easily diagnose this issue.

When there are many covariates, one approach has been to develop more parsimonious models using methods such as stepwise regression. However, this may involve subjective decisions such as the type of variable selection procedure to use, whether to base selection upon p-values or change in exposure parameter estimates, and where to set numeric cutoffs (e.g., p=0.05, 0.10, 0.20) for variable inclusion and retention in the model. For covariates that confer relatively modest increases in disease risk, some variable selection procedures, such as stepwise regression, may exclude important covariates from the final model.

Furthermore, stepwise regression has limitations that can lead to underestimation of standard errors for exposure estimates.[6] Other analytical strategies which have become more common in recent years include using summary variables, such as propensity scores and disease risk scores, which are described below. Propensity scores often perform better than logistic regression when the outcome is relatively rare (e.g., fewer than 10 events per covariate as noted above), whereas logistic regression tends to perform better than propensity score analysis when the outcome is common but the exposure is rare.[7]

Choice of Regression Modeling Approach

The forms of the study outcome, exposure of interest, and study covariates will determine the regression model to be used. For independent, non–time-varying exposures and study covariates, generalized linear models (GLMs) such as linear or logistic regression can be used. If the study outcome is binary with fixed followup and is rare, Poisson regression with robust standard errors can be used to estimate relative risks and get correct confidence intervals.[8-9] For count data, Poisson regression can also be used but is susceptible to problems of overdispersion, wherein the variance of the outcomes is larger than what is given by the Poisson model. Failure to account for this can lead to underestimation of standard errors. A negative binomial regression model can help address the issue of overdispersion.[10] If the value 0 occurs more frequently than is predicted by the Poisson or negative binomial model, the zero-inflated Poisson and zero-inflated negative binomial models can be used.[11]

In CER studies in which data are correlated, regression models should be specified that take this correlation into account. Examples of correlated data include repeated measures on study subjects over time, patients selected within hospitals across many hospitals, and matched study designs. There are a number of analysis options that can be considered, which depend on the study question and particulars of the study design. Repeated measures per study subject can be collapsed to a single summary measure per subject. Generalized estimating equations (GEE) are a frequently used approach to account for correlated data. Random effects models such as generalized linear mixed models (GLMM) are another suitable analytical approach to handle repeated measures data. Approaches for such longitudinal data are described in detail in a number of textbooks.[12-13] For matched study designs (e.g., case-controlled designs), models such as conditional logistic regression may be considered.

Time-to-event data with variable followup and censoring of study outcomes are commonly investigated in CER studies. Cox proportional hazards regression is a common methodology for such studies. In particular, this approach can easily handle exposures and study covariates whose values vary over time as described in detail below. When time-varying covariates are affected by time-varying treatment, marginal structural models (described below) may be required. A number of excellent textbooks describe the analysis of time-to-event data.[14-15]

A high-level overview of modeling approaches in relation to the nature of the outcome measure and followup assessments is shown in Table 10.1.

Table 10.1. Summary of modeling approaches as a function of structure of outcome measure and followup assessments

Number of Followup Measures and Time Intervals				
	Single Measure			
Outcome Measure	*No clustering*	*Clustering (e.g., multi-site study)*	**Repeated Measure, Fixed Intervals**	**Repeated Measure, Variable Intervals**
Dichotomous	Logistic regression	Multilevel (mixed) logistic regression, GLMM, GEE, conditional logistic regression	Repeated measures ANOVA (MANOVA), GLMM, GEE	GLMM, GEE
Continuous	Linear regression	Multilevel (mixed) linear regression, GLMM, GEE	Repeated measures ANOVA (MANOVA), GLMM, GEE	GLMM, GEE
Time to event	Cox proportional hazards regression	Variance-adjusted Cox model or shared frailty model		
Time to event (aggregate or count data)	Poisson regression	Multilevel (mixed) Poisson regression		

ANOVA = analysis of variance; GEE = generalized estimating equation; GLMM = generalized linear mixed models; MANOVA = multivariate analysis of variance

Note: This high-level summary provides suggestions for selection of a regression modeling approach based on consideration of the outcome measure and nature of the followup measures or assessments. Many of these methods allow time-varying exposures and covariates to be incorporated into the model. Time-varying **confounding** may require use of inverse-probability-of-treatment-weighted (IPTW)/marginal structural model techniques.

Model Assumptions

All analytic techniques, including regression, have underlying assumptions. It is important to be aware of those assumptions and to assess them. Otherwise, there are risks with regards to interpretation of study findings. These assumptions and diagnostics are specific to the regression technique being used and will not be listed here. They are covered in numerous textbooks, depending on the methods being used. For example, if Cox proportional hazards regression is used, then the proportional hazards assumption should be assessed. If the validity of this assumption is questionable, then alternatives such as time-dependent covariates may need to be considered.

Time-Varying Exposures/Covariates

In most CER studies, it is unrealistic to assume that exposures and covariates remain fixed throughout followup. Consider, for example, HIV patients who may be treated with antiretroviral therapy. The use of antiretroviral therapy may change over time and decisions about therapy may in part be based on CD4 count levels, which also vary over time. As another illustration, consider a study of whether proton pump inhibitors (PPIs) prevent clopidogrel-related gastroduodenal bleeding. In this situation,warfarin may be started during followup. Should one adjust for this important potential confounder? Failure to account for the time-varying status of such exposures and confounders (i.e., by fixing everyone's exposure status at baseline) may severely bias study findings.

As noted above, for time-to-event study outcomes, time-dependent Cox regression models can be used to account for time-varying exposures and covariates. However, difficult issues arise when both treatment and confounding variables vary over time. In the HIV example, CD4 count may be affected by prior therapy decisions, but CD4 count levels may themselves go on to alter subsequent therapy decisions and the final survival outcome. In examining the effects of time-varying treatment, a decision must be made as to whether to control for CD4 count. A difficulty arises in that CD4 count is both a confounding variable (for subsequent therapy and final survival) and also an intermediate variable (for the effect of prior treatment). Thus, control for CD4 count in a time-varying Cox model could potentially lead to bias because it is an intermediate variable and could thus block some of the effect of treatment; but failure to control for CD4 count in the model will result in confounding and thus bias for the effect of subsequent treatment. Both analyses are biased. Such problems arise whenever a variable is simultaneously on the pathway from prior treatment and also affects both subsequent treatment and the final outcome.

These difficulties can be addressed by using inverse-probability-of-treatment weighting (IPTW),[16] rather than regression adjustment, for confounding control. These IPTW techniques are used to estimate the parameters of what is often called a marginal structural model, which is a model for expected counterfactual outcomes. The marginal-structural-model/IPTW approach is essentially a generalization of propensity-score weighting to the time-varying treatment context. The IPTW technique assumes that at each treatment decision, the effect of treatment on the outcome is unconfounded given the past covariate and treatment history. A similar weighting approach can also be used to account for censoring as well.[16] This marginal-structural-model/IPTW approach has been developed for binary and continuous outcomes,[16] time-to-event outcomes,[17] and repeated measures data.[18]

Another consideration for time-varying exposures is accounting for exposure effect (e.g., the effect of medication use) after the subject stopped receiving that exposure. One approach is to create another exposure level that is a carryover of a biologically plausible number of days after exposure use has ended and incorporate it as a time-varying exposure level in the analysis. Another approach is an intent-to-treat analysis in which exposure status (e.g., treatment initiation) is assumed throughout followup. Cadarette and colleagues (2008) used this approach in a study of fracture risk.[19] The motivation was that treatment adherence may be low and accounting for on-treatment status may result in information bias.

Propensity Scores

Propensity scores are an increasingly common analytic strategy for adjusting for large numbers of covariates in CER. The use of the propensity score for confounding control was proposed by Rosenbaum and Rubin.[20] The propensity score is defined as the probability of receiving treatment (or exposure) conditional on observed covariates, and it is typically estimated from regression models, such as a logistic regression of the treatment conditional on the covariates. Rosenbaum and Rubin showed that if adjustment for the original set of covariates suffices to control for confounding, then adjustment for just the propensity score also would suffice as well. This strategy is particularly favorable in studies having a common exposure and rare outcome or possibly multiple outcomes.[7] Propensity scores can be used in subclassification or stratification,[21] matching,[22] and weighting,[23] and further adjustment can be done using regression adjustment.[24] Stürmer and colleagues provide a review of the application of propensity scores.[25]

If adjustment using the propensity score is used, balance in study covariates between exposure groups should be carefully assessed. This can include, but is not limited to, testing for differences in study covariates by exposure group after adjusting for propensity score. Another common assessment of the propensity score is to visually examine the propensity score distributions across exposure groups. It has been demonstrated that if there is poor overlap in these distributions, there is a risk of biased exposure estimates when adjusting for the propensity score in a regression model.[26] One remedy for this is to restrict the cohort to subjects whose propensity score overlaps across all exposure groups.[27-28]

When feasible, matching on the propensity score offers several advantages. Matching subjects across exposure groups on propensity score ensures, through restriction, that there will be good overlap in the propensity score distributions. In addition, the presentation of a summary of subject characteristics by exposure groups in a propensity-matched design allows a reader to assess the balance in study covariates achieved by matchingin a similar manner to the comparison of randomized treatment groups from a randomzed clinical trial. This can be done graphically or by comparing standardized differences across groups. However, in a propensity-matched design, one can only ensure that measured covariates are being balanced. The consequences of unmeasured confounding will need to be assessed using sensitivity analysis. See chapter 11 for further details. Matching techniques for causal effects are described in detail in Rubin[29] and best practices for constructing a matched control group are provided by Stuart and Rubin.[30] Care must be taken when estimating standard errors for causal effects when using matching,[31-32] though software is now available that makes this task easier.[33]

A tradeoff between using regression adjustment on the full cohort and a propensity-matched design is that in the former there may still be imbalances in study covariates, and in the latter sample size may be reduced to the extent that some of the subjects cannot be matched. Connors and colleagues[34] used both analytic strategies in a cohort study of the effectiveness of right heart catheterization and reported similar findings from both analyses. Use of multiple analytic strategies as a form of sensitivity analysis may serve as a useful approach, drawing from the strengths of both strategies.

Brookhart and colleagues[35] investigated variable selection approaches and recommend that the covariates to be included in the propensity score model either be true confounders or at least related to the outcome; including covariates related only to the exposure has been shown to increase the variance of the exposure estimate.

Disease Risk Scores

The disease risk score (DRS) is an alternative to the propensity score.[36-37] Like the propensity score, it is a summary measure derived from the observed values of the covariates. However, the DRS estimates the probability or rate of disease occurrence as a function of the covariates. The DRS may be estimated in two ways. First, it can be calculated as a "full-cohort" DRS, which is the multivariate confounder score originally proposed by Miettinen in 1976.[38] This score was constructed from a regression model relating the study outcome to the exposure of interest and the covariates for the entire study population. The score was then computed as the fitted value from that regression model for each study subject, setting the exposure status to nonexposure. The

subjects were then grouped into strata according to the score and a stratified estimate of the exposure effect was calculated. The DRS may also be estimated as an "unexposed-only" DRS, from a regression model fit only for the unexposed population, with the fitted values then computed for the entire cohort.

The DRS is particularly favorable in studies having a common outcome and rare exposure or possibly multiple exposures. It is useful for summarizing disease risk and assessing effect modification by disease risk. Ray and colleagues[39] reported effect modification by cardiovascular disease risk, derived and summarized using DRS, in a study of antipsychotics and sudden cardiac death. Also, in the presence of a multilevel exposure in which some of the levels are infrequent, the DRS may be a good alternative to propensity scores.

Instrumental Variables

A limitation of study designs and analytic strategies in CER studies, including the use of traditional multiple regression, propensity scores, and disease risk scores, is incomplete information on potential unmeasured confounders. An alternative approach to estimate causal effects, other than confounding/covariate control, is the use of instrumental variables.[40] An "instrument" is a measure that is causally related to exposure but only affects the outcome through the treatment and is also unrelated to the confounders of the treatment-outcome relationship. With an instrument, even if there is unmeasured confounding of the treatment-outcome relationship, the effect of the instrument on the treatment, and the effect of the instrument on the outcome can together be used to essentially back out the effect of the treatment on the outcome. A difficulty of this approach is identifying a high-quality instrument.

An instrument must be unrelated to the confounders of the treatment and the outcome; otherwise, instrumental variable analyses can result in biases. An instrument also must not affect the outcome except through the treatment. This assumption is generally referred to as the "exclusion restriction." Violations of this exclusion restriction can likewise result in biases. Finally, the instrument must be related to the treatment of interest. If the association between the instrument and the treatment is weak, the instrument is referred to as a "weak instrument." Finite-sample properties of estimators using weak instruments are often poor, and weak instruments moreover tend to amplify any other biases that may be present.[41-44] If a variable is found that satisfies these properties, then it may be used to estimate the causal effect of treatment on the outcome. However, such a variable may be difficult or impossible to identify in some settings. Moreover, the assumptions required for a variable to be an instrument cannot be fully verified empirically.

Two-stage least squares techniques are often employed when using instrumental variables, though with a binary treatment, ratio estimators are also common.[40] For estimates to be causally interpretable, often a monotonicity assumption must also be imposed; that is, that the effect of instrument on the treatment only operates in one direction (e.g., that it is causative or neutral for all individuals). Assumptions of homogeneous treatment effects across individuals also are commonly employed to obtain causally interpretable estimates. When homogeneity assumptions are not employed, the resulting causal effect estimate is generally only applicable for certain subpopulations consisting of those individuals for whom the instrument is able to change the treatment status.[40] Such effects are sometimes referred to as "local average treatment effects." When the treatment is not binary, interpretation of the relevant subpopulation becomes more complex.[45] Moreover, when two-stage least squares procedures are applied to binary rather than continuous outcomes, other statistical biases can arise.[46]

Brookhart and colleagues[47] applied this approach in a study of COX-2 inhibitors with nonselective, nonsteroidal anti-inflammatory drugs (NSAIDs) on gastrointestinal complications. Their instrument was the prescribing physician's preference for a COX-2 inhibitor relative to an NSAID. The results of the instrumental variable analysis were statistically similar to results from two clinical trials, and contrary to the traditional multiple regression analysis that was also conducted.

Schneeweiss and colleagues[48] examined the use of aprotinin during coronary-artery bypass grafting and risk of death. Their primary analysis was a traditional multiple regression. In addition to the primary analysis, they also conducted a propensity score matched-pairs analysis as well as an instrumental variable analysis. All three analyses had similar findings. This methodology of employing more than one analytical approach may be worth consideration, since the propensity score matching does not rely on the exclusion restriction and other instrumental variable assumptions, whereas instrumental variable analysis circumvents the biases introduced by unmeasured confounders, provided a good instrument is identified. When results differ, careful attention needs to be given to what set of assumptions is more plausible.

Missing Data Considerations

It is not uncommon in CER to have missing data. The extent of missingness and its potential impact on the analysis needs to be considered. Before proceeding with the primary analyses, it is important to characterize the patterns of missingness using exploratory data analyses. This step can provide insights into how to handle the missing data in the primary analysis.

For the primary analysis, a common analytical approach is to analyze just those subjects who have no missing data—called a complete-case analysis. However, an initial limitation of this approach is that sample size is reduced, which affects efficiency even if data are missing completely at random. If subjects with missing data differ from subjects with complete data, then exposure estimates may be biased. For example, suppose blood pressure is a potential confounder, and it is missing in very ill subjects. Then, excluding these subjects can bias the exposure estimate.

Little and Rubin's textbook describes several analytic approaches for handling missing data.[49] One common approach to filling in missing data when they are "missing completely at random" or "missing at random" is imputation, which the book describes in detail. In chapter 3 of Harrell's textbook, he describes missing data and imputation and also provides some guidelines for handling such data.[50] Inverse-probability-weighting techniques, described below, can also be employed to address issues of missing data.

Conclusion

This chapter has provided a brief overview of statistical methods, as well as suggestions and recommendations to address the complex challenges of analyzing data from observational CER studies. Both traditional approaches such as multivariable regression and novel but established methods such as propensity scores and instrumental variable approaches may be suitable to address specific data structures, under certain assumptions. Thoughtful application of these approaches can help the investigator improve causal inference.

Checklist: Guidance and key considerations for developing a statistical analysis section of an observational CER protocol

Guidance	Key Considerations	Check
Describe the key variables of interest with regard to factors that determine appropriate statistical analysis.	– Should discuss independent variables (when they are measured, whether they are fixed or time-varying; e.g., exposures, confounders, effect modifiers). – Should discuss dependent variables or outcomes (continuous or categorical, single or repeated measure, time to event). – Should state if there will be a "multilevel" analysis (e.g., an analysis of effects of both practice-level and patient-level characteristics on outcome).	☐
Propose descriptive analysis or graph according to treatment group.	– Should include the available numbers per group, number missing for all key covariates, distributions or graphs that are needed to decide if transformation of data is needed or to determine an accurate functional form of the final model. – Should include all potential confounders and effect modifiers to assess initial covariate balance by study group.	☐
Propose the model that will be used for primary and secondary analysis objectives.	– Should take into account the design (independent vs. dependent observations, matched, repeated measurement, clustered), objectives, functional form of model, fixed/time-varying followup period, fixed and time-varying exposure and other covariates, assessment of effect modification/heterogeneity, type of outcome variables (categorical, ordinal, or continuous), censored data, and the degree of rarity of outcome and exposure. – Should propose a suitable approach for adjusting for confounding (e.g., multiple regression model, propensity scores, instrumental variable [as secondary or main analysis]).	☐

References

1. Pagano M, Gauvreau K. Principles of Biostatistics. 2nd edition. Pacific Grove, CA: Duxbury; 2000.
2. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd edition. Philadelphia: Lippincott, Williams & Wilkins; 2008.
3. McCullagh P, Nelder JA. Generalized Linear Models. 2nd edition. London: Chapman & Hall; 1989.
4. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–9.
5. Harrell FE, Lee KL, Matchar DB, et al. Regression models for prognostic prediction: advantages, problems and suggested solutions. Cancer Treatment Reports. 1985;69(10):1071–7.
6. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. Stat Med. 1989;8(7):771-83.
7. Cepeda MS, Boston R, Farrar JT, et al. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol. 2003;158:280-7.
8. Zou G. A modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol. 2004;159:702-6.

9. Lumley T, Kronmal R, Ma S. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. UW Biostatistics Working Paper Series. University of Washington. Paper 293;2006.
10. Lawless, Jerald F. Negative binomial and mixed Poisson regression. Can J Statistics. 1987;15: 209-25.
11. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics. 2000: 56S:1030-9.
12. Diggle PJ, Heagerty P, Liang K-Y, et al. Analysis of Longitudinal Data. 2nd edition. New York: Oxford University Press; 2002.
13. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis. New Jersey: Wiley; 2004.
14. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer-Verlag; 1997.
15. Hosmer DW, Lemeshow S, May S. Applied Survival Analysis. 2nd edition. New Jersey: Wiley; 2008.
16. Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiol. 2000;11:550-60.
17. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiol. 2000;11:561-70.
18. Hernán MA, Brumback B, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. Stat Med. 2002;21:1689-709.
19. Cadarette SM, Katz JN, Brookhart MA, et al. Relative effectiveness of osteoporosis drugs for preventing nonvertebral fracture. Ann Intern Med. 2008;148:637-46.
20. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41-55.
21. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc. 1984;79: 516-524.
22. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat. 1985;39:33-8.
23. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health. 2006;60:578-86.
24. Reinisch J, Sanders S, Mortensen E, et al. In-utero exposure to phenobarbital and intelligence deficits in adult men. JAMA. 1995;274:1518-25.
25. 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;59:437-47.
26. 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;163:262-70.
27. Joffe MM, Rosenbaum PR. Invited commentary: Propensity scores. Am J Epidemiol. 1999;15: 327-33.
28. Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. Pharmacoepidemiol Drug Saf. 2010;19:858-68.
29. Rubin DB. Matched Sampling for Causal Effects. Cambridge: Cambridge University Press; 2006.
30. Stuart EA, Rubin DB. Best practices in quasi-experimental designs: matching methods for causal inference. In: Best Practices in Quantitative Methods. Ed. J. Osborne. Thousand Oaks, CA: Sage Publications; 2008;155-76.
31. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006;74:235-67.
32. Abadie A, Imbens GW. On the failure of the bootstrap for matching estimators. Econometrica. 2008;76:1537-57.
33. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J Stat Softw. 2011;42(7):1-52.
34. Connors AF, Speroff T, Dawson NV, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. JAMA. 1996;276:889-97.
35. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149-56.

36. Arbogast PG, Kaltenbach L, Ding H, et al. Adjustment for multiple cardiovascular risk factors using a summary risk score. Epidemiol. 2008;19(1):30-7.

37. Arbogast PG, Ray WA. Use of disease risk scores in pharmacoepidemiologic studies. Stat Methods Med Res. 2009;18(1):67-80.

38. Miettinen OS. Stratification by a multivariate confounder score. Am J Epidemiol. 1976;104(6):609–20.

39. Ray WA, Meredith S, Thapa PB, et al. Antipsychotics and the risk of sudden cardiac death. Arch Gen Psychiatry. 2001;58:1161-7.

40. Angrist JD, Imbens, GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). J Am Stat Assoc. 1996;91: 444-72.

41. Nelson CR, Startz R. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. J Bus. 1990;63(1):S125-40.

42. Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. J Am Stat Assoc. 1995;90(430): 443-50.

43. Stock JH, Yogo M. Testing for Weak Instruments in Linear IV Regression. Cambridge, MA: National Bureau of Economic Research; November 2002.

44. Stock JH, Wright JH, Yogo M. A survey of weak instruments and weak identification in generalized method of moments. J Bus Econ Stat. 2002;20(4):518-29.

45. Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. J Am Stat Assoc. 1995;90:431-42.

46. Rassen JA, Schneeweiss S, Glynn RJ, et al. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. Am J Epidemiol. 2009;169(3): 273-84.

47. Brookhart MA, Wang PS, Solomon DH, et al. Evaluating short-term drug effects using a physician-specific prescribing preferences as an instrumental variable. Epidemiol. 2006;17: 268-75.

48. Schneeweiss S, Seeger JD, Landon J, et al. Aprotinin during coronary-artery bypass grafting and risk of death. New Eng J Med. 2008;358: 771-83.

49. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd edition. Hoboken, NJ: John Wiley & Sons; 2002.

50. Harrell FE. Regression Modeling Strategies. New York: Springer-Verlag; 2001.