Obervational studies are used to inform health care policy and decision making when comparable data from randomized controlled trials (RCTs) are inadequate or unavailable due to ethical reasons, practical considerations, and other logistical issues. The need for evidence from observational studies is particularly relevant in comparative effectiveness research (CER) given the large evidence gaps that exist regarding the comparative effectiveness and value of a broad array of treatments. Furthermore, CER may require data from different sources, including RCTs, nonrandomized studies, and systematic reviews.1

It is generally accepted that RCTs are the gold standard for generating evidence pertaining to the benefits and risks of medical treatments. A major advantage of RCTs is that by design, the experimenter is able to control for selection bias. The assignment of study subjects through a random mechanism ensures comparability of the treatment groups with respect to both known and unknown confounding factors. This implies that any difference between groups before randomization is attributable to chance alone. The latter in turns permits the application of standard inferential procedures to draw conclusion about treatment efficacy in the trial population.2 However, results of RCTs often do not provide evidence of comparative effectiveness because clinically important active comparators were not selected by the study designers or because of other design limitations.

Even when there are comparative effectiveness data from RCTs, the data may be inadequate to address all relevant decisions. The conditions under which the trials are conducted may not reflect the real-world setting or important subpopulations. Under these circumstances, it may be necessary to rely on nonrandomized studies to inform medical decision making.

Use of observational studies, however, requires a careful consideration of important conceptual and practical issues. From a design perspective, the absence of random assignment of subjects to treatments almost always introduces selection bias that confounds the relationship between treatments and outcomes. More specifically, in the absence of randomization, study subjects use treatments dictated by factors, other than chance, that have the potential to confound outcomes. This problem results in imbalances with regard to known and unknown confounding factors that may influence the outcome of interest. For measured covariates, there are statistical approaches to mitigate the bias introduced by the imbalances. However, the problem is more challenging for important covariates that may not exist in the dataset. Thus, the standard inferential procedures are likely to lead to invalid conclusions, if applied uncritically to such data.3

With the growing awareness of the importance of data from nonrandomized studies in making critical health care decisions, considerable progress has been made in recent years in establishing guidance for best practices in the design, analysis, and reporting of observational studies.4-9 In this paper, we consider some of the major statistical issues that arise in the analysis of data from observational studies, with particular reference to the limitations of existing approaches, and recent methodological developments aimed at addressing bias introduced by unmeasured or latent confounders.

Bias in Nonrandomized Studies
There are several ways in which bias may arise in nonrandomized studies. Bias can arise as a consequence of systematic measurement error or misclassification of subjects on 1 or more of the explanatory or response variables. Another important type of bias is one that is intrinsic to observational studies, often referred to as selection or channeling bias. Since assignment to treatment is not random, the channeling of individuals into treatments results in imbalance with respect to relevant attributes. From a methodological perspective, the bias that results from imbalance of known and unknown risk factors is of particular interest, and will be the focus of the next 2 sections.

In the absence of randomization, differences in apparent treatment effects may be attributable to pretreatment differences in risk factors among subjects receiving the intervention groups being studied. For overt biases emanating from known covariates, there are established methodological approaches aimed at removing bias through appropriate matching and regression analysis. When the bias is hidden (i.e., caused by risk factors that have not been measured), the problem is generally complex, and the analytical procedures are not as well developed.

Although there has been considerable methodological progress in addressing both overt and hidden biases in observational studies, all the available techniques have certain limitations that require careful assessment to ensure the validity of the results for particular applications. In the next section, we review some of the commonly used approaches and highlight their limitations and other relevant features. It is essential for each investigator to carefully and thoroughly assess the potential biases in each proposed study and tailor the methods or combination of methods to best address these biases, while recognizing the general limitations of observational research relative to RCTs.

Traditional Analytical Approaches
In this section, we consider adjustment techniques, including
matching, stratification, and analysis of covariance, generally employed for overt biases, and instrumental variable procedures that are typically used for hidden biases. The emphasis will be on nontechnical aspects of the procedures, without delving into their mathematical formulations (see Johnson et al. for a review of such techniques9).

**Methods for Overt Bias**

**Matching.** A common approach to adjust for overt biases is matching, which involves comparing each individual in the treated group with 1 or more subjects in a comparison cohort with respect to observed covariates that are known to confound the relationship between treatment and outcomes. When performed properly (e.g., with appropriate and adequate matching criteria), the procedure has the dual advantage of improving the precision of estimators as well as reducing the overt bias.10

**Propensity Score.** One way of achieving balance among the treated and comparison groups with regard to the distributions of observed covariates is through propensity score analysis, which involves quantifying the conditional probability, given the covariates, that a subject receives the treatment rather than the control.11,12 It has been long established that when interest is in balancing treatment groups on all observed covariates, it is sufficient to balance on the propensity scores.13 The propensity score is particularly useful when the number of covariates is large and matching is not practical. However, matching or adjusting for propensity scores does not solve the problem of hidden biases. Further, the validity of the propensity score matching is heavily dependent on the adequacy of the model used to estimate the scores. It is, therefore, necessary to check whether balance has been achieved in the distributions of observed covariates, and to update the model, as appropriate, through inclusion of interaction or other higher order terms in the logit model.14,15 In a recent study, Basu et al. showed that in moderate sample sizes, balancing on estimated propensity scores may fail to balance higher-order moments and covariances among covariates and that the usual inverse-probability weighting in propensity scores may be sensitive to misspecification of the model for estimating propensity scores.16

Implementation of propensity score methods in the medical literature has been a subject of some scrutiny that can be illuminating—in a “what not to do” sense—for those planning to use such methods; see Weitzen et al.17 and Austin18 for critical reviews. D’Agostino provides a useful tutorial and some basic SAS (SAS Institute Inc., Cary, NC) code for creating propensity scores.19 Baser provides an interesting overview and empirical comparison of 7 different methods of creating propensity scores.20

**Stratification.** Stratification attempts to create balance between control and study drug subjects by matching subjects as groups rather than pairs. Stratification may be achieved based on 1 or more known covariates. When there are several covariates, suitable cut-off points (e.g., quintiles) of a propensity score may be employed to define strata. Optimal stratification strategies are available to ensure that subjects in a given stratum are as similar as possible.21 In general, stratification is known to reduce bias and enhance precision of estimates considerably.22 However, the value of stratification is reduced by the often arbitrary way strata are defined. The available approaches to determine an optimal stratification are not commonly used in routine applications.

**Model-Based Approaches.** An alternative to matched sampling and stratification is use of suitable models, such as analysis of covariance, to estimate treatment effects adjusting for observed covariates and/or propensity scores. The performance of model-based adjustments is, of course, dependent on the accuracy of the model and validity of model assumptions. In fact, when there is significant departure from model assumptions, the procedure may increase bias rather than reducing it.23,24 Accordingly, a combination of matching and model-based adjustments may be preferred.

**Methods for Hidden Biases**

Without randomization, hidden biases might result from imbalances between treatment groups with respect to important covariates that were not observed by the investigator. Such hidden biases are likely to distort the conclusions of observational studies. While traditional propensity scoring can only condition on observed confounders and cannot deal with unobserved confounding, traditional instrumental variable methods can not only condition on observed confounders but also average over unobserved confounders, thereby addressing hidden selection biases in observational data. Below, we discuss some of the measures that may be taken to mitigate consequences of hidden biases.

**Instrumental Variables.** A method that is borrowed from econometrics is instrumental variable analysis, which involves identifying 1 or more variables (instruments) that are highly correlated with treatment but are unassociated with other confounders and have no direct effect on the response variable.25 In RCTs, an obvious instrument is the randomization mechanism. In observational studies, common instruments include prescriber preference and the distance a patient has to travel to a hospital or site of care.26

Suppose E[Y|Z = z] is the average value of the response Y for all subjects with values for an instrument Z = z. A measure of the effect of treatment X on Y may be given by:

$$\beta = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[X|Z = 1] - E[X|Z = 0]}$$

In RCTs where Z is an indicator of random assignments,
a Wald estimator of $\beta$ corresponds to an intention-to-treat (ITT) estimator, while when $Z$ is an instrumental variable in observational studies, it corresponds to the instrumental variable estimator. A common approach to instrumental variable estimation involves 2-stage least squares, in which 1 model (generally probit or ordinary least squares [OLS] regression) is specified for the treatment assignment process that depends on the instrument and potential confounding variables, and a second for the outcome that includes the predicted probability of treatment from the first stage and the additional covariates that are included in $Y$.27

A major drawback of instrumental variable techniques is that suitable variables frequently are not available. Even when such variables are available, it is often difficult to assess the validity of the underlying assumptions. For example, if the instrument is weakly correlated with treatment, the resulting treatment effect estimate may be biased.28,29 In addition, the estimators may be inefficient relative to OLS when the instrument is redundant.30 For further discussion about instrumental variable techniques, see references 25 and 31-33.

One should note that even with the successful implementation of instrumental variable methodology, the interpretation of the results is limited to what is called the local average treatment effect.31 This local average could apply to a small portion of the study population, so-called marginal patients who are defined as the subset of patients whose treatment choices vary with the instrument. In the case where the instrumental variable is a binary indicator of distance from a hospital offering a particular treatment or procedure, this local average treatment effect pertains only to the comparison between patients who received the treatment because they lived relatively close to a hospital offering the treatment and those who lived further away but would have received the treatment had they lived close by. If one were to use a different instrumental variable, the resulting treatment effect would be different because it would apply to a different group of marginal patients.

**Sensitivity Analysis.** A general approach to assessing the impact of unobserved confounders involves sensitivity analyses that attempt to quantify the degree to which hidden bias would explain any observed association. For example, if the instrument is weakly correlated with treatment, the resulting treatment effect estimate may be biased.31 In addition, the estimators may be inefficient relative to OLS when the instrument is redundant.30 For further discussion about instrumental variable techniques, see references 25 and 31-33.

One should note that even with the successful implementation of instrumental variable methodology, the interpretation of the results is limited to what is called the local average treatment effect.31 This local average could apply to a small proportion of the study population, so-called marginal patients who are defined as the subset of patients whose treatment choices vary with the instrument. In the case where the instrumental variable is a binary indicator of distance from a hospital offering a particular treatment or procedure, this local average treatment effect pertains only to the comparison between patients who received the treatment because they lived relatively close to a hospital offering the treatment and those who lived further away but would have received the treatment had they lived close by. If one were to use a different instrumental variable, the resulting treatment effect would be different because it would apply to a different group of marginal patients.

**Pattern Specificity.** Pattern specificity is a technique employed to detect hidden biases or to reduce sensitivity to hidden biases, and is based on the fact that observational studies are variable in terms of their sensitivity to hidden bias. Typically, latent biases tend to leave “visible traces” in observed data37 and the approach involves distinguishing real treatment effects from hidden biases.37-40

### Recent Developments and Future Directions

**Individualization in CER.** Basu discusses the need to individualize comparative effectiveness research.43 Although a rich array of biomarkers is usually required to generate individual-level treatment effects, Basu proposes 2 methods that can be used to learn about treatment effect heterogeneity even in the absence of such biomarkers.43,45 Both methods estimate treatment effect heterogeneity conditional on individual level confounders, some of which are observed in the data and the remaining are unobserved. The first is a method of local instrumental variable (LIV) that addresses limitations of traditional instrumental variable approaches.43-45 LIV methods attempt to leverage this selection and allow for unobserved confounders to be moderate treatment effects. Therefore, they can be used to estimate marginal treatment effects that are conditional on both observed and unobserved confounders. Such marginal treatment effects can also be estimated using a second method that uses latent factors to proxy for the unobserved confounding.41 The data requirements are different for the alternative methods, and usually careful nonparametric identification is required to make sure that the methods are estimating the relevant parameters.

**Bias Adjustment through Prior Event Ratio.** Tannen et al. introduced a technique they dubbed prior event rate ratio (PERR) to adjust for hidden confounders in the analysis of data from electronic medical record databases.45 The adjustment involves knowledge of event rates in the 2 groups prior to initiation of the interventions. While the technique worked reasonably well to identify and reduce the effects of unmeasured confounding when applied to cardiovascular outcomes considered in the study, the procedure requires strong assumptions about constant temporal effects, absence of confounder-by-treatment interaction, and nonterminal events as outcome. However, these issues are present to some degree in other estimators, and the PERR technique can provide a useful alternative approach in CER estimation.

**Bayesian Inference for Observational Data.** Despite the growing body of literature on the role of Bayesian statistics in the analysis of observational studies, the potential is not fully realized among practitioners. The application may range from sensitivity analysis for unmeasured confounding in observational studies46 to covariate adjustment based on a Bayesian propensity score.47 Additional information may be found in references 48-50.

**Meta-Analysis of Observational Studies.** In addition to the known issues with the synthesis of data from RCTs, meta-analysis of observational studies requires a careful assessment of problems peculiar to such studies.51,52 Accordingly there have been efforts to establish good practices for the
data is the preservation of privacy. Accordingly, there are laws and regulations such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA, Title II)\(^5\) that govern the transmission and use of such data. For pooling de-identified data on the patients from alternative sources, probabilistic record linkage\(^5\) or similar machine learning techniques\(^5\) may be used. The techniques typically are computationally intensive and involve identification of similar groups of records, relative to predefined criteria, and then evaluating the likelihood that the records belong to the same patient. As an integral part of the methodological consideration, parallel efforts must also be exerted to enhance other aspects of the studies, including sound design, pre-specification of analytical strategy, high-quality data, and appropriate reporting of results.

Discussion

Well-conducted observational studies are useful for CER. When RCTs are inadequate for decision making, observational databases can provide relevant information from the real-world setting in a timely manner. However, effective use of data from nonrandomized studies requires overcoming significant conceptual and technical issues. In this paper, we highlighted some of the available statistical methods that can be used to mitigate the effects of overt and hidden biases, with emphasis on limitations of the approaches and opportunities for further research. A major issue with the analysis of observational data is the preservation of privacy. Accordingly, there are laws and regulations such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA, Title II)\(^5\) that govern the transmission and use of such data. For pooling de-identified data on the patients from alternative sources, probabilistic record linkage\(^5\) or similar machine learning techniques\(^5\) may be used. The techniques typically are computationally intensive and involve identification of similar groups of records, relative to predefined criteria, and then evaluating the likelihood that the records belong to the same patient. As an integral part of the methodological consideration, parallel efforts must also be exerted to enhance other aspects of the studies, including sound design, pre-specification of analytical strategy, high-quality data, and appropriate reporting of results.

Conclusions

When RCTs are inadequate or unavailable, observational studies may play useful roles in addressing major health care questions. However, the validity of analytic results from observational studies is adversely impacted by biases that may be introduced due to lack of randomization. In this paper, we reviewed some of the methodological challenges that arise in the analysis of data from nonrandomized studies, with particular emphasis on the limitations of traditional approaches and potential solutions from recent methodological developments.

DISCLOSURES

This supplement was funded by Pfizer. Alemayehu, Alvir, and Willke are Pfizer employees. Jones was a Pfizer employee during the production of the manuscript.

The 4 authors contributed equally to writing and revision of the manuscript.

ACKNOWLEDGEMENT

The authors are grateful to C. Daniel Mullins, PhD, for his constructive comments on a draft version of the manuscript.

REFERENCES


Authors

DEMISSE ALEMAYEHU, PhD, is Executive Director, OR and Disease Area Statistics Head; JOSE MA. J. ALVIR, DrPH, is Senior Director, OR Statistics; BYRON JONES, PhD, is currently Biometrical Fellow at Novartis, formerly Senior Director, Statistical Research Consulting Center; and RICHARD J. WILLKE, PhD, is Head, Global Health Economics & Outcomes Research, Global Market Access, Primary Care, Pfizer Inc., New York, New York.


