



A Primer on Comparative Effectiveness Research Methods

Daniel C. Malone, PhD, RPh
Professor, University of Arizona



Comparative effectiveness research (CER) has been defined by the Federal Coordinating Council as studies that involve pragmatic comparative trials and/or synthesis of existing research that is applicable to real-world settings.¹ The council further states, "The purpose of this research is to improve health outcomes by developing and disseminating evidence-based information to patients, clinicians, and other decision-makers, responding to their expressed needs, about which interventions are most effective for which patients under specific circumstances."

The American Recovery and Reinvestment Act of 2009 directed the Institute of Medicine to develop a broad-based list of priority areas for CER.² The Institute of Medicine called for building a "broad and supportive infrastructure to carry out a sustainable national CER strategy."² The institute rated "health care delivery system" as the most important topic area, as it was included in 50 of 100 priority research areas as ranked by primary and secondary researchers. Clearly, one of the most critical steps for improving the country's health is to assist healthcare systems in implementing CER evidence.²

There has been much discussion in the literature over the past 24 months concerning CER and about as much speculation concerning the impact of CER. Comparative effectiveness research has 6 extant themes, including the "generation and synthesis of evidence" that is comparative.³ Ultimately, CER should help to inform decisions that improve health. Another defining component of CER is that it distinguishes between efficacy and effectiveness.⁴

In the pharmacy benefits world, it has long been recognized that randomized clinical trials involving placebo comparators, surrogate measures of disease improvement, strict inclusion/exclusion criteria, high medication adherence, and close monitoring of patients provide

estimates of the *efficacy* of medications. The *effectiveness* of medications is measured under real-world conditions, where medications are prescribed for a variety of patients rather than a specific studied population, often are used after a series of treatment failures or in combination with other agents that modify the disease process, and are taken less often than prescribed; also, patients are seen rather infrequently. Comparative effectiveness research is intended to reflect real-world use, and the corresponding benefits and risks.

It is important to note that CER involves multiple types of research activities. One could categorize CER into 3 different study types: (1) large simple randomized controlled trials; (2) analysis of administrative and other existing data; and (3) synthesis of multiple trials using meta-analysis and/or Bayesian techniques. In this commentary, I provide a brief description of each of these study designs, as well as a discussion of their advantages and limitations.

Large simple randomized trials such as ALLHAT (Anti-hypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial), ACCORD (Action to Control Cardiovascular Risk in Diabetes), and STAR (Study of Tamoxifen and Raloxifene) are extremely costly. The large number of patients (ALLHAT N = 42,418; ACCORD N = 10,251; STAR N = 19,747) required to make statements about treatment effectiveness drives the cost of these studies well into the millions of dollars.⁵ The advantage of these studies is that they evaluate final health outcomes such as heart attacks and mortality. In addition, they often involve multiple treatment regimens and subpopulations of interest. However, due to the time required to see final health outcomes and the large numbers of patients, these studies are very expensive to conduct. In addition, practice patterns often change over the course of the study,

and the treatments may not be relevant once the trial is completed.⁶

Another emerging type of CER study is evaluations of existing data such as healthcare claims and electronic medical records.⁷ These studies have become increasingly popular over the past 20 years because they are relatively inexpensive and can be conducted in a relatively short period of time. With millions of Americans enrolled in commercial and government programs that rely on the electronic exchange of healthcare data, finding sufficient numbers of patients who have been exposed to the treatment of interest is possible. Furthermore, retrospective studies have the advantage of examining the treatments in real-world settings, not in the artificial environment often created with phase III clinical trials. The major criticism of retrospective cohort studies is the inability to control for disease severity and confounding by indication. Administrative data often contain only diagnosis and procedural codes to measure disease severity. Diagnosis codes are well known for a multitude of issues with respect to disease severity. *International Classification of Diseases (ICD)* codes can be up to 5 digits, with the fourth and fifth digits occurring after a decimal point. For some diagnoses the absence of a fifth or fourth digit is not relevant; other times, it is. Few diagnosis codes provide any insight into the severity of the condition of interest. In addition, the lack of specificity may result in “rule-out” admissions being misclassified. Coding of conditions also is prone to misspecification, miscoding, incorrect sequencing, and clerical mistakes. *ICD* codes often fail to reflect whether the condition or etiology had a rapid onset. Procedural *ICD* codes may permit tracking of sequential procedures.

Confounding by indication is another critical issue that affects retrospective studies treatments. Varying levels of disease severity can result in the use of different treatments or combinations of treatments. Recently, a number of statistical methods have been developed to overcome this limitation. These methods include propensity scores and instrumental variables.⁸⁻¹⁰

Propensity scores are used to help select control subjects who are similar to the treated subjects, thereby allowing differences in outcome to be attributed to the treatment or intervention. Propensity scoring, essentially a more complex approach to matching, is becoming widely accepted and is considered to be a valid approach in evaluating treatments, especially as the treatment relates to harm. See, for example, Johannes et al and Seeger et al.^{11,12}

Because unmeasured factors such as extent of disease could affect both the outcome and use of interacting

therapies, instrumental-variables methods can be used. Numerous instrumental-variables studies have assessed the effects of different treatment rates across patients grouped by instrumental variables or “instruments.” The use of instrumental variables is appropriate in situations where it is not possible to assess causal relationships through other types of experiments. Instrumental variables are strongly related to the exposure (ie, treatment), but not related directly to the outcome.¹⁰ The selection of instrumental variables requires solid statistical and clinical knowledge. Therefore, this technique currently is less common than propensity score methods.

The third study design relevant to CER is evidence synthesis. The vast majority of these analyses are meta-analyses, which combine multiple studies that use similar constructs or outcomes to evaluate a particular therapy. Critics of meta-analyses often complain of combining apples and oranges.¹³ These differences can be due to differences in severity of disease across trials, differences in inclusion/exclusion criteria, or approaches to measuring the outcomes of interest, to name a few. To determine whether this problem exists, it is recommended that the analyst assess and report the degree of homogeneity across the studies. Meta-analyses are limited by the trials that have been conducted, as well as the manner in which they are reported. A common issue with reported studies is the failure of the authors to include measures of variability, such as standard deviation or standard error, for the outcome measures. Another limitation of meta-analyses is that because they are based on existing studies with placebo arms, there often is little information about comparative effectiveness. Because of this issue, traditional meta-analyses may fall short of providing useful comparative effectiveness data.

An alternative to the traditional meta-analysis is the Bayesian indirect treatment comparison approach.¹⁴ Although many clinicians and decision makers may not fully understand the details of this approach, the results from such an analysis are informative. The concept of indirect comparisons has been in the medical literature since 1994, when O’Brien et al evaluated enoxaparin versus warfarin prophylaxis for deep vein thrombosis after hip replacement.¹⁵ Bayesian indirect analysis is similar to other statistical methods in which differences in effect between treatments are evaluated using multiple independent-variable regression models. Bucher et al provide a simple illustration of this approach.¹⁶ The validity of this type of analysis appears to depend on the comparability of the original clinical trials.^{17,18} Indirect comparisons and meta-analyses share this limitation. The key advantage

of the Bayesian approach is that the results can be rank-ordered.¹⁴ Those treatments ranked first would have the highest likelihood of achieving treatment success. This approach is appealing because it provides decision makers with knowledge that can be acted on directly. It is important to keep in mind that Bayesian analyses are not static, meaning that as new information becomes available (ie, new studies are published), the analysis can be rerun and the results evaluated to determine whether a different decision should be made.

Certainty in healthcare is a rare commodity. Therefore, health professionals are forced to make decisions about treatment selection that may or may not benefit the patient. There also is no assurance that patients will experience positive outcomes. Not every treatment works in every patient, but some treatments may have a higher probability of success. At the population level, CER should help inform policies and incentives that result in improvements in health and minimize risk. However, CER is not a panacea for all healthcare decision making, because results may differ depending on the methodology used to answer the question.¹⁹ Rather, CER should be thought of as a tool to assist and inform decision making.

Author Affiliation: From University of Arizona College of Pharmacy, Tucson, AZ.

Author correspondence to: Daniel C. Malone, RPh, PhD, Professor, University of Arizona College of Pharmacy, 1295 N Martin, Drachman Hall B307F, Tucson, AZ 85721-0202. E-mail: malone@pharmacy.arizona.edu.

REFERENCES

1. Federal Coordinating Council for Comparative Effectiveness Research. Draft definition, prioritization criteria, and strategic framework for public comment. Updated 2009. <http://www.hhs.gov/recovery/programs/cer/draftdefinition.html>. Accessed December 12, 2009.
2. Iglehart JK. Prioritizing comparative-effectiveness research—IOM recommendations. *N Engl J Med*. 2009;361(4):325-328.
3. Sox HC. Defining comparative effectiveness research: the importance of getting it right. *Med Care*. 2010;48(6 suppl):S7-S8.
4. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in clinical trials. *J Chronic Dis*. 1967;20(8):637-648.
5. AcademyHealth. A First Look at the Volume and Cost of Comparative Effectiveness Research in the United States. June 2009. http://www.academyhealth.org/files/FileDownloads/AH_Monograph_09FINAL7.pdf. Accessed May 21, 2010.
6. Luce BR, Kramer JM, Goodman SN, et al. Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change. *Ann Intern Med*. 2009;151(3):206-209.
7. Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med*. 2009;151(3):203-205.
8. Seeger JD, Kurth T, Walker AM. Use of propensity score technique to account for exposure-related covariates: an example and lesson. *Med Care*. 2007;45(10 suppl 2):S143-S148.
9. Seeger JD, Williams PL, Walker AM. An application of propensity score matching using claims data. *Pharmacoepidemiol Drug Saf*. 2005;14(7):465-476.
10. Brooks JM, Chrischilles EA. Heterogeneity and the interpretation of treatment effect estimates from risk-adjustment and instrumental variable methods. *Med Care*. 2007;45(10 suppl 2):S123-S130.
11. Johannes CB, Koro CE, Quinn SG, Cutone JA, Seeger JD. The risk of coronary heart disease in type 2 diabetic patients exposed to thiazolidinediones compared to metformin and sulfonylurea therapy. *Pharmacoepidemiol Drug Saf*. 2007;16(5):504-512.
12. Seeger JD, Walker AM, Williams PL, Saperia GM, Sacks FM. A propensity score-matched cohort study of the effect of statins, mainly fluvastatin, on the occurrence of acute myocardial infarction. *Am J Cardiol*. 2003;92(12):1447-1451.
13. Moayyedi P. Meta-analysis: can we mix apples and oranges? *Am J Gastroenterol*. 2004;99(12):2297-2301.
14. Berry SM, Ishak KJ, Luce BR, Berry DA. Bayesian meta-analysis for comparative effectiveness and informing coverage decisions. *Med Care*. 2010;48(6 suppl):S137-S144.
15. O'Brien BJ, Anderson DR, Goeree R. Cost-effectiveness of enoxaparin versus warfarin prophylaxis against deep-vein thrombosis after total hip replacement. *CMAJ*. 1994;150(7):1083-1090.
16. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*. 1997;50(6):683-691.
17. Collins R, Fenwick E, Trowman R, et al. A systematic review and economic model of the clinical effectiveness and cost-effectiveness of docetaxel in combination with prednisone or prednisolone for the treatment of hormone-refractory metastatic prostate cancer. *Health Technol Assess*. 2007;11(2):iii-iv, xv-xviii, 1-179.
18. Bucher HC, Griffith L, Guyatt GH, Opravil M. Meta-analysis of prophylactic treatments against *Pneumocystis carinii* pneumonia and toxoplasma encephalitis in HIV-infected patients. *J Acquir Immune Defic Syndr Hum Retrovirol*. 1997;15(2):104-114.
19. Teutsch SM, Berger ML, Weinstein MC. Comparative effectiveness research: asking the right questions, choosing the right method. *Health Aff (Millwood)*. 2005;24(1):128-132. [ajpb](https://doi.org/10.1186/1475-2875-24-132)