



Do We Need Good Practice
Principles for Observational
Comparative Effectiveness Research?

The GRACE Initiative
www.graceprinciples.org
coordinator@graceprinciples.org



Good ReseArch for CComparative EEffectiveness OBSERVED

Nancy A. Dreyer, MPH, PhD, FISPE
ON BEHALF OF THE GRACE INITIATIVEⁱ

Summary

The growth in therapeutic choices for many conditions has sparked interest in evaluating the comparative effectiveness of these alternatives. Randomized clinical trials, although revered as offering the highest quality of evidence, rarely provide direct comparisons of many treatments and have limitations for certain conditions and population subgroups. Observational studies can provide data to fill the evidence gaps left by clinical trials, but methodological challenges for analysis and interpretation and the lack of accepted principles to assess quality have limited the practical use of observational research.

The GRACE principles describe a hierarchy of evidence for observational research on comparative effectiveness that can be used by decision-makers, as well as key elements of good practice including defining research questions and methods a priori; collecting valid, clinically relevant data; analyzing, interpreting and reporting data, including sensitivity analyses and alternative explanations for findings; and conducting these studies in accordance with accepted good practices. These principles are drawn from consultation with expert researchers and policy makers, from review of existing guidance for good pharmacoepidemiologic practice and good registry practice, and from recent recommendations for reporting and reviewing observational studies.

Introduction

Comparative effectiveness (CE) has been defined as generating “clinical information on the relative merits or outcomes of one intervention in comparison to one or more others; benefits and harms.”¹ In this context, interventions can refer to the use of medical treatments, including drugs, devices, and procedures. CE research is needed to support treatment decisions by patients and physicians as to what therapy to use and to guide formulary decisions.

It is widely accepted that the strongest form of evidence for the assessment of CE comes from randomized clinical trials (RCT).² Observational studies have been relegated to lower tiers in commonly used hierarchies of evidence, largely because of their heterogeneity, the potential for bias in the results, and the challenges involved in their conduct and interpretation. However, these traditional evidence hierarchies were created primarily for the purpose of evaluating studies of the intended effect of a treatment. It has been argued that a different hierarchy of evidence is

ⁱ For a list of GRACE collaborators and supporters, see www.graceprinciples.org

needed for the evaluation of studies that seek to explain how therapies work in various subgroups where the likelihood of the intended effect is difficult to predict.³ RCTs are often limited in their ability to ascertain the effectiveness of healthcare products and services as they are actually used in the real world.⁴ For example, effectiveness depends in large part on the complex decision-making process of prescribers selecting therapies, decisions to stop or switch treatments based on tolerability, and so forth. Effectiveness also depends on the decisions and actions of patients, including whether they accept a course of therapy, how well a treatment is tolerated, the use of concomitant therapies, or treatment compliance once initiated. The impact of these issues on effectiveness may be difficult or impractical to assess in a traditional RCT, in which treatment assignments are protocol-driven and full compliance is assumed or actively monitored. RCTs may also be limited in situations where treatments or treatment practices are changing rapidly or where real-world use is broader than the types of patients that are typically accessible for and agreeable to a clinical trial. In addition, evaluations of long-term or sustained effectiveness of a specific therapy, which may change based on changes in tolerance and disease status, are generally not feasible under the typical time constraints of an RCT.

Awareness of these issues has sparked interest in more inclusive designs to evaluate the effectiveness of treatments for broader segments of the population. The value of observational research models in these settings is appreciated by many regulatory and advisory bodies involved in healthcare resource decision-making, but the methodological challenges must be addressed to facilitate meaningful analyses and reliable interpretations.

The methodological challenges in observational studies of CE primarily stem from the lack of randomization to treatment. This lack of randomization leads to concern about bias (systematic error) and confounding (a mixing of effects). A considerable amount of control over known and unknown confounders may be obtained in an observational CE study by design (exclusion, matching, or restricting the study groups to new or incident drug users) and/or by analysis (restriction, stratification or mathematical modeling). However, the sometimes lack of clarity about the exact methodologies used, and the on-going debates within the field lead to concern that the results are unreliable.

This set of principles is intended to guide good practice for observational studies of CE and is intended both for those who conduct these studies and for those who need to evaluate the rigor of such studies to inform decision-making regarding therapeutic alternatives. These principles are consistent with good pharmacoepidemiologic practice,⁵ the Agency for Healthcare Research and Quality (AHRQ) handbook on Registries for Evaluating Patient Outcomes,⁶ and the STROBE guidelines for reporting observational studies.⁷ These principles also should be useful for those performing CE reviews, following either the Cochrane principles⁸ or the AHRQ Guide for conducting comparative effectiveness reviews.⁹

Good Research Practice for Comparative Effectiveness

I. Identify evidence gaps and the potential value of an observational study according to the hierarchy of evidence for decision-makers.

Is there relevant information for the target population of interest with clinically relevant outcomes evaluated using sound methodology and with a reasonable length of follow-up? If little or no relevant trial data are available, how much value would an observational study of comparative effectiveness (CE) add, and at what cost?

The main challenge for observational studies of CE is to identify and disentangle systematic choices in prescribing that are related to the outcomes of interest. Various patient factors can lead to confounding (mixing of effects) between the treatment and outcome. A hierarchy of evidence can be applied to observational studies of CE to identify situations which can provide the strongest types of evidence, as well as other situations which may contribute useful information. The highest level of evidence from observational CE studies is that which is strong enough to support decision-making. Studies from other levels may assist in decision-making, depending on the depth of knowledge about prescribing practices, known risk factors for outcomes, and the strength of the associations that are detected.

The Evidence Hierarchy for Decision-Makers (ranked from strongest level to weaker levels)

1. Determinants of use are not related to determinants of outcomes

Treatment decisions are largely driven by reimbursements, such as different insurance plan formularies, rather than by patient characteristics or physician preferences. This situation enables an observational study to achieve an unbiased balance between comparisons groups because choice of insurance (or residence, for national insurance plans) is generally unrelated to formulary decisions and treatment outcomes.

2. No consistent determinants of treatment, or determinants of treatments are largely known

- a) Clinical equipoise: A variety of treatments are frequently used and there is no good evidence for one treatment over another. Evidence of clinical equipoise might come from journal debates about how a patient should be treated, as an example. Strongly held, widely differing recommendations would signal that there is little evidence to support one treatment over another.
- b) Determinants of treatment are largely known: A reliable understanding of the factors that drive physician treatment preferences exists and treatment determinants are independent of patient characteristics. In these situations, physicians always (or almost always) use the same product (e.g., anesthesia) and same course of treatment or same surgical approach. The experience of patients treated by physicians with differing strong, consistent practice patterns for these types of treatments could be compared.

3. Risk of toxicity from treatment is unlikely to be related to the outcome(s) of interest

For example, warfarin is prescribed as an anticoagulant based on strict, widely used treatment guidelines to limit the risk of bleeding (e.g., ulcers). A study of stroke in patients on and off warfarin would provide meaningful evidence of CE if the only reason patients are not prescribed warfarin is because of a contraindication (ulcers) unrelated to the outcome of interest (stroke). The strength of this level of evidence depends on the likelihood that the determinants of treatment are truly unrelated to the outcomes of interest.

4. Little relevant evidence available

The lowest level is reserved for those situations where an observational CE study may reduce some uncertainty and provide some useful information, at an affordable cost, but it is difficult to understand to what extent unknown confounding factors could have artificially inflated the apparent relative benefit of one treatment compared to another. There are still some substantial advantages for observational studies, such as comparisons of surgical versus medical treatments, where RCTs tend to overestimate the real benefits achievable in routine clinical practice because of the technical expertise of surgeons in trials, and the rapid onset of treatment.¹⁰ There are other areas in which there are no comparative effectiveness trials – pragmatic or otherwise – and any observational data derived from relatively rigorous studies with reasonable end-points and follow-up time should be considered. Consider autism for example. Like attention deficit disorder and other conditions in children, there is little research available about treatment effectiveness and great need for information.

Generally, unless an effect is observed that is much larger than would be expected or larger than could be explained by bias, it is unlikely that the study will contribute meaningfully to clinical decision-making. Although there is no unanimity about how large a relative benefit needs to be in order to be worthy of serious consideration as evidence for decision-making, some suggest that “as-treated” analyses showing a relative benefit of two or more or more should be given serious consideration.¹¹ However, stronger effects, like five or more, are even more reliable.¹²

II. Prepare a study plan in advance of conducting a CE study

Should an observational CE study be warranted and feasible to conduct, a research plan should be developed before starting the study. This plan should describe a clinically relevant research question(s), and document the study design, target population, intended methods for conducting the primary analyses of effectiveness and safety, and the sensitivity analyses. This process describes key constructs that, in an ideal world, would be specified and evaluated. The study plan should be sufficiently detailed to allow replication of methodology. The following elements should be addressed:

1. What is the main purpose of the study?

Defining the research question/problem, including goals and objectives, at the outset of the study provides the context for structured data collection and a focused analysis plan. Objectives should describe the main outcomes of interest as well as the intended comparisons.

2. What conditions and treatments are of interest, what comparisons will be made, and in what populations?

Clearly define the diseases/conditions, the comparators, the treatment regimens, and the patient population of interest, with consideration of:

- a) Diseases/conditions – diagnostic certainty, severity, time since diagnosis, significant co-morbidities, treatment history, etc.
 - b) Comparators – comparison with one or many? Good CE research generally reflects the complexities of interventions as used in real practice settings. Comparisons to a number of real-world alternatives are preferable to a single comparator. Also consider to what extent the treatment and its therapeutic alternatives are already in use in the target population in sufficient numbers for meaningful analysis and interpretation.
 - c) Treatment regimens – for each comparator, consider brand, dosage, method of delivery, duration of use, whether for a labeled indication, therapeutic alternatives currently in use, the likelihood that the necessary information will be accurately recorded and accessible, etc.
 - d) Patient characteristics – age, sex, ethnicity, socioeconomic status, geography, opportunities to present for medical attention (e.g., health insurance coverage), opportunity to be prescribed all of the medications that are being compared, if their healthcare provider wanted to prescribe them, etc.
3. How will effectiveness be measured?
 - a. Is the outcome clinically relevant for all comparators?

Are the endpoints appropriate for evaluating effectiveness? Straight-forward clinical outcomes are preferable to intermediate and composite endpoints. Intermediate endpoints, however, can be useful when there are good data that link the intermediate outcome and the long-term outcome and studying the long-term outcome is not feasible due to time or cost constraints.¹³

Are the outcomes meaningful and valid for all comparators, regardless of the intervention's mechanism of action? This is especially important to consider when using surrogate markers of effectiveness (e.g., bone density as a surrogate for fracture).

- b. Is the expected difference in effectiveness clinically meaningful in terms of information that will assist health professionals with treatment decisions or policy-makers with decisions about allocations of resources? E.g., Differences in survival after invasive diagnostic procedures for acute myocardial infarction could be used to justify increasing the availability of cardiac catheterization labs,¹⁰ whereas decreases in a biomarker may not affect the risk of development of clinically apparent disease.
 - c. Are the treatments safe?
 - d. Are the treatments tolerable?
4. How will the sample size be determined and statistical power to detect a given difference in CE be estimated? What specific assumptions are being made and how are they supported?

III. Collect the most valid, clinically relevant data needed to answer the study question as efficiently as possible.

1. Who will qualify for inclusion in your study?

To the extent feasible, it is desirable to focus observational CE studies on new users of the treatments of interest (also known as inception cohorts). These cohorts are more likely to provide a complete picture of the full benefits and risks of using new treatments since this design avoids the “healthy user” effect that can result from studying people who are not treatment naive.¹⁴ Beyond studying new users, it is quite acceptable to include a broad range of patients to the extent affordable in order to enhance the information yield and generalizability of studies.

2. What data will be collected or assembled, and what checks will be used to assure their validity?

Primary data collection and secondary data collection each have strengths and limitations that should be considered in the context of the study objectives. Both primary collection of data, i.e., data that are collected specifically for the purposes of the study, and secondary use of data, i.e., data that were collected for other purposes (such as administrative claims data and medical (health) records) require an understanding of how the data were collected, enrollment and coverage factors, pathways to care, quality assurance, and what other factors may have affected the quality of the data and the validity of conclusions that may be drawn from their analysis.

Many of the methods that assure good clinical practice for RCTs¹⁵ are appropriate for observational studies of CE. A main difference, however, is the trade-off between resources that are appropriate to devote to internal validity as compared with external validity. Whereas on-site monitoring to assure data quality is common for most, if not all, sites and patients in an RCT, observational CE studies generally have more sites and patients, and longer follow-up. Consequently a larger proportion of budget is devoted to

scope and duration to ensure broad generalizability, which enhances external validity, and less of the budget is devoted to individual data quality, which may diminish internal validity. Were budgetary constraints not a practical reality, this trade-off would not be required.

For prospective, primary data collection, studies should be designed to avoid affecting treatment patterns by creating an incentive for physicians to prescribe certain treatments to fill study recruitment quotas, and study procedures should not be overly burdensome on patients or physicians. To the extent feasible, scales and measures for patient-reported outcomes should have been validated in populations similar to those under study and for similar methods of administration whenever possible. It is important to identify off-label use and treatments that may be used outside of their indication and to consider the feasibility of collecting data on these uses.

When using secondary data, a strong understanding of the data source and how the data were collected is essential and will help minimize errors in interpretation. For example, billing codes may not be reflective of the actual clinical condition, may be recorded inaccurately (e.g., coding errors), imprecisely (e.g., DRGs),^{16,17} inconsistently, and/or under different constraints (e.g., one intervention might have been subject to pharmacy prescription limits, such as for migraine medicines).

3. What relevant and important data might be missing systematically?

Although all studies have missing data, some data related to exposures or outcomes of interest may be systematically missing (*not* at random). Some information may not be available in retrospectively collected data because the treatment or outcome of interest is not reimbursed by health insurance, is not accurately or completely recorded in sufficient detail for meaningful interpretation, or may be so sensitive that treatments are sought outside of the health system (e.g., treatment for drug addiction). Similarly, some drugs like IV antibiotics may not be recorded in prescription billing systems because they are dispensed in the doctor's office and are not distributed to patients through the pharmacy. Also, retrospective data rarely include information on over-the-counter products which may be important. For both retrospective and prospectively collected data, the accuracy and validity of the information from the patient's perspective also must be considered (e.g., self reported substance abuse or medication sharing).

IV. Analyze the data by comparing people who are similar in the characteristics that would cause them to receive the treatment and in their likelihood of benefiting from the treatment, and consider alternative explanations for the findings.

1. Is the actual use of medications studied or proxies like refills?

Intent-to-treat analyses can be used to assess prescribing practices, but true CE should examine actual use to the extent possible, including any means that can be used to quantify adherence and compliance.

2. Are people with similar disease severity and similar opportunities for treatment compared?

Patients should be analyzed in meaningful subgroups, using techniques such as stratification of groups with common characteristics or by multivariate modeling techniques like propensity scoring to account for various relevant risk factors,¹⁸ prior events rate ratios,¹⁹ and instrumental variables.²⁰

3. How well have alternative explanations for the findings been considered and evaluated?

To what extent could bias (stemming from factors that are related both to the decision to treat and to the outcome(s) of interest) have distorted the results? For example,

- **Selection bias** refers to systematic differences among the groups being compared that arise from self-selection or physician-directed selection of treatments, or association of treatment assignments with other characteristics such as education, ethnicity, age, access to healthcare, etc.²¹ Selective prescribing, or confounding by indication, describes the situation in which people with more severe forms of the disease/condition, or those who are resistant to other treatments, are more likely receive newer treatments.
- **Misclassification** occurs when an exposure or outcome is incorrect or missing. Misclassification of drug exposure can result from the patient's incorrect recall of dose or poor adherence or treatment compliance. Studies using data sources that track prescription fills and refills are particularly vulnerable for treatments used on an as-needed basis (e.g., migraine medications) and for treatments dispensed in liquid or inhalable forms.²²
- **Detection bias**²³ applies to situations in which comparison groups are assessed at different points in time or using different methods or by assessors who may have knowledge of which treatment was used. Quantitative evaluations of outcomes that are standardized, reproducible, and independently verifiable are preferable to clinical impressions and/or other measurements that have not been validated or were not validated in the target study population.
- **Performance bias** refers to systematic differences in care other than the intervention under study. This bias gets at differences like situational factors that might affect adherence or persistence, or health practices such as diet, exercise, and smoking cessation. For example, a public health initiative promoting healthy lifestyles might be directed only at patients who have received one class of medical treatments, and the initiative, not the treatment, could be responsible for an observed benefit.
- **Attrition** refers to selective loss to follow-up. For example, if patients generally stop using a treatment with poor effectiveness but a small group of responders continue treatment, then the treatment could appear to have been more effective

than appropriate. The effect of attrition can be addressed by characterizing those who drop out of studies, at what point, and why.²⁴

4. Have sensitivity analyses been conducted and reported?

Sensitivity analyses can provide a framework for evaluating the extent to which assumptions and common sources of bias may have explained any apparent differential effectiveness.

VI. Conduct and report the study in a manner that adheres to accepted good practices of evidence quality for observational research and puts the findings in context with other good evidence.

It is important to place information about comparative effectiveness into the public domain whenever possible. All such study should be conducted,^{25,15} reported,^{7,24} and evaluated^{6,26} in accordance with generally accepted good practices for observational research, as described elsewhere.

The study report should have enough information to allow replication of analyses in another database or testing of alternative methods of analysis in the same or a similar data set. Replication of comparative effectiveness from different populations and through alternative analytic methods can strengthen the conclusions that may be drawn from observational studies.

It may be useful to report the results of observational studies of CE in the context of how well they support existing clinical trials data.^{27,11} This approach has been used to demonstrate the effectiveness of various analytic methods. However, when the results of observational CE studies are not consistent with those of RCT for subgroups not studied in RCT, it is not clear which interpretations are correct and which are not. Nonetheless, the reporting of observational CE studies, however they may be judged, may contribute to a better clinical and biological understanding of the disease, either by confirmation in a more targeted RCT or through advances in basic science.

¹ Institute of Medicine. 2007. Learning What Works Best: The Nations Need for Evidence on Comparative Effectiveness in Health Care. Retrieved November 28, 2007 from <http://www.iom.edu/ebm-effectiveness>.

² Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies and the hierarchy of research designs. *N Engl J Med.* 2000;342:1887–1892.

³ Vandembroucke J: Observational Research, Randomised Trials, and Two Views of Medical Science. *PLOS Medicine* 2008;5(3)e67:339-343.

⁴ Avorn J. In defense of pharmacoepidemiology – embracing the yin and yang of drug research. *N Engl J Med* 2007;357:2219-2221.

⁵ Guidelines for good pharmacoepidemiologic practice. *Pharmacoepidemiology and drug safety* 2005; 14: 589–595.

⁶ Glicklich RE, Dreyer NA (eds): Registries for Evaluating Patient Outcomes: A User's Guide. (Prepared by Outcome DEcIDE Center [Outcome Sciences, Inc. dba Outcome] under Contract No. HHS29020050035I TO1.) AHRQ Publication No. 07-EHC001-1. Rockville, MD: Agency for Healthcare Research and Quality. April 2007.

⁷ Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandembroucke JP: The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for reporting observational studies. *Ann Intern Med* 2007;147:573-577.

⁸ The Cochrane Collaboration. Retrieved November 27, 2007 from <http://www.cochrane.org/index.htm>

-
- ⁹ AHRQ Guide for conducting comparative effectiveness reviews. October 10, 2007. Draft for public comment
- ¹⁰ Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ: Analysis of Observational studies in the presence of treatment selection bias. *JAMA* 2007;297:278-285.
- ¹¹ GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ* 2004; 328:1490-1498.
- ¹² Tannen RL, MG Weiner, D Xie. Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication. *Pharmacoepidemiology and Drug Safety* 2008;17:671-685.
- ¹³ Ramsey S, Wilcke R, Briggs A, Brown R, Buxton M, Chawla A, Cook J, Glick H, Liljas B, Pettitte D, Reed S: Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA task force report. March 1, 2005. Unpublished.
- ¹⁴ Ray W: Evaluating Medication Effects Outside of Clinical Trials: New-User Designs. *Am J Epidemiol* 2006; 158:915-920.
- ¹⁵ Guidance for Industry: E6 Good Clinical Practice: Consolidated Guidance. April 1996. Available online at: <http://www.fda.gov/cder/guidance/959fnl.pdf>.
- ¹⁶ Jollis JG, Ancukiewicz, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. *Ann Intern Med* 1993;119:844-850.
- ¹⁷ Lanza L, Walker AM, Bortnichak E, Dreyer NA: Peptic Ulcer and Gastrointestinal Hemorrhage Associated With Nonsteroidal Anti-inflammatory Drug Use in Patients Younger Than 65 Years A Large Health Maintenance Organization Cohort Study. *Ann Intern Med* 1995;125:1371-13.
- ¹⁸ Glynn RJ, Schneeweiss S, Stürmer T. Indications for Propensity Scores and Review of Their Use in Pharmacoepidemiology. *Basic Clin Pharmacol Toxicol* 2006; 98(3): 253-9.
- ¹⁹ Tannen RL, Weiner MG, Xie D: Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication. *Pharmacoepidemiology and Drug Safety* 2008; 17: 671-685
- ²⁰ Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T : Variable Selection for Propensity Score Models *Am J Epidemiol* 2006;163:1149-1156.
- ²¹ AHRQ Guide for conducting comparative effectiveness reviews. October 10, 2007. Draft for public comment, page 32
- ²² Strom, BL. Methodologic Challenges to Studying Patient Safety and Comparative Effectiveness. *Medical Care* 2007;45:S13-S15.
- ²³ Higgins J, S Green (2005). The Cochrane Collaboration. The Cochrane handbook for systematic reviews of interventions, 2006. Retrieved November 27, 2007 from <http://www.cochrane.org/resources/handbook/handbook.pdf>
- ²⁴ Tooth L, Ware R, Bain C, et al. Quality of reporting of observational longitudinal research. *Am J Epidemiol*. 2005;161:280-288.
- ²⁵ Guidance for Industry: Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment. U.S. Food and Drug Administration. March 2005. Available online at: <http://www.fda.gov/cder/guidance/6359OCC.htm>.
- ²⁶ Deeks JJ, Dines J, D'Amico R et al. Evaluating non-randomised intervention studies. *Health Technol Assess*. 2003; 7(27).
- ²⁷ Schneeweiss S, AR Patrick, T Stürmer, A Brookhart, J Avorn, M Maclure, KJ Rothman, and RJ Glynn: Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Medical Care* 2007;45 (10):S131-S142.