GRACE Principles: Recognizing High-Quality Observational Studies of Comparative Effectiveness

Nancy A. Dreyer, PhD; Sebastian Schneeweiss, MD; Barbara J. McNeil, MD; Marc L. Berger, MD; Alec M. Walker, MD; Daniel A. Ollendorf, MPH; and Richard E. Gliklich, MD; for the GRACE Initiative

Nonrandomized comparative effectiveness studies contribute to clinical and biologic understanding of treatments by themselves, via subsequent confirmation in a more targeted randomized clinical trial, or through advances in basic science. Although methodological challenges and a lack of accepted principles to assess the quality of nonrandomized studies of comparative effectiveness have limited the practical use of these investigations, even imperfect studies can contribute useful information if they are thoughtfully designed, well conducted, carefully analyzed, and reported in a manner that addresses concerns from skeptical readers and reviewers. The GRACE (Good Research for Comparative Effectiveness) principles have been developed to help healthcare providers, researchers, journal readers, and editors evaluate the quality inherent in observational research studies of comparative effectiveness. The GRACE principles were developed by experienced academic and private sector researchers and were vetted over several years through presentation, critique, and consensus building among outcomes researchers, pharmacoepidemiologists, and other medical scientists and via formal review by the International Society of Pharmacoepidemiology. In contrast to other documents that guide systematic review and reporting, the GRACE principles are high-level concepts about good practice for nonrandomized comparative effectiveness research. The GRACE principles comprise a series of questions to guide evaluation. No scoring system is provided or encouraged, as interpretation of these observational studies requires weighing of all available evidence, tempered by judgment regarding the applicability of the studies to routine care.

(Am J Manag Care. 2010;16(6):467-471)

Comparative effectiveness (CE) has been defined as "the conduct and synthesis of research comparing the benefits and harms of different interventions and strategies to prevent, diagnose, treat, and monitor health conditions in 'real world' settings."[1] As the demand for data to support decision making escalates, there is a growing recognition that randomized clinical trials alone will not fill the information gaps. Critics have characterized nonrandomized studies as having inferior quality of evidence because of limited internal validity, analytic challenges posed by a heterogeneous mix of patients with complex medical histories, and the lack of accepted guidance to distinguish more reliable studies.[2,3] Nevertheless, observational studies are often a rich resource for meaningful information about treatment adherence, tolerance, use of concomitant therapies, and the decision-making processes and consequences of selecting or switching treatments. Real-world studies sometimes provide the only information about sensitive populations,[4] sustained therapeutic effectiveness, and health services–related issues such as how the type of practitioner affects the choice of medical device.[5] Noninterventional studies also can provide important information about treatment effectiveness, sometimes with surprising results, such as the lack of benefit from some types of cardiac rehabilitation in nontrial settings.[6]

Although the International Society for Pharmacoeconomics and Outcomes Research,[7,8] the International Society of Pharmacoepidemiology,[9] and the Agency for Healthcare Research and Quality have recommended good practices for observational studies and registries, they have not promulgated simple high-level principles to guide users in design and evaluation. The STROBE (Strengthening the Reporting of Observational Studies) guidelines[10] and others[11] address reporting, not quality. Tools such as GRADE (Grading of Recommendations Assessment, Development and Evaluation) that address quality generally rank all nonrandomized studies as "low quality," regardless of the study quality.[12]

The GRACE (Good Research for Comparative Effectiveness) principles were created to guide practitioners, researchers, journal readers, and editors in evaluating the quality of observational CE studies. The active contributors are experienced academic and private sector researchers with different perspectives on the creation and use of observational CE data.[13] The GRACE principles were

In this article
Take-Away Points / p468
www.ajmc.com
Full text and PDF

For author information and disclosures, see end of text.

Take-Away Points

The GRACE (Good Research for Comparative Effectiveness) principles are intended to help healthcare providers, researchers, and managed care decision makers evaluate the quality inherent in noninterventional (observational) studies of comparative effectiveness. The GRACE principles comprise 3 questions that can be used to characterize studies and provide guidance about what constitutes higher quality and about how to evaluate areas of uncertainty. The GRACE principles address the following:

- What belongs in a study plan.
- Key elements for good conduct and reporting.
- Ways to assess the accuracy of comparative effectiveness inferences for a population of interest.

tested and modified through presentations and critique,[14,15] including formal review by the International Society of Pharmacoepidemiology.

The GRACE principles can be used to guide the design and evaluation of studies that are based on new data collection, use existing data, and are consistent with good pharmacoepidemiologic practice[9] and the Agency for Healthcare Research and Quality's handbook on *Registries for Evaluating Patient Outcomes*.[16] The GRACE principles also may be useful for CE reviews following the Cochrane principles[17] or the Agency for Healthcare Research and Quality's *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*.[18]

The following 3 questions comprise the GRACE principles for evaluating nonrandomized studies of CE. Although many examples refer to drugs, the GRACE principles also apply, in large, to medical devices, procedures, complex clinical strategies, and other elements. No scoring system is proposed or encouraged, as evidence must be weighed and the interpretation tempered in light of all available evidence. Adaptations and augmentation are anticipated as science develops.

Were the study plans (including research questions, main comparisons, outcomes, etc) specified before conducting the study?

A good study plan describes the research questions and documents the study design, target population, and intended methods for conducting the primary analyses of effectiveness and safety. The study plan also defines the diseases and conditions, patient characteristics, comparators, treatment regimens, and outcomes of interest. Creating a study plan at the outset helps assure skeptics that comparisons were not conducted iteratively until support for a preconceived conclusion was found.

The study should include clinically meaningful outcomes that would assist health professionals and patients with treatment decisions or policymakers with decisions about allocations of resources. For example, decreases in a biomarker may not affect the risk of development of clinically apparent disease, but differences in survival after invasive diagnostic procedures for acute myocardial infarction could be used to justify increasing the availability of cardiac catheterization laboratories.[19] Intermediate end points can be useful when there are good data that link those end points to the long-term outcomes and when evaluation of the long-term outcomes is not feasible because of time or cost constraints. Quantitative evaluations of outcomes that are standardized, reproducible, and independently verifiable are preferable to clinical impressions or other measurements that have not been validated or have substantial interobserver variation.

Was the study conducted and analyzed in a manner consistent with good practice and reported insufficient detail for evaluation and replication?

Observational studies of CE should be conducted, reported, and evaluated in accord with generally accepted good practices for nonrandomized research.[16,20,21] Meaningful data can be collected or assembled from several sources. The challenge to their successful utilization is to understand what is recorded and why. To evaluate the validity of conclusions drawn from their analysis, data that are collected specifically for the purposes of the study (primary data) and data that were collected for other purposes (secondary data) require an understanding of the purpose and method by which they were assembled, enrollment and coverage factors, pathways to care, quality assurance, and other factors that may have affected the quality of the data. For example, insurance claims data may not be reflective of the actual clinical condition and may be coded inaccurately, imprecisely (eg, diagnosis-related groups), inconsistently, or under different constraints (eg, a treatment such as migraine medicines might be subject to pharmacy prescription limits). For prospective data collection, studies should not create an incentive for physicians to recommend specific treatments to fill recruitment quotas, and to promote retention study procedures should not be overly burdensome on patients or physicians. For primary and secondary data collection, it is important to assess and report which data are missing, whether their absence seems to be systematic or random, and what is their potential effect on the overall results.

For developing the outcomes under study, it is important to compare persons with similar risks and, for drugs, to consider focusing on new users (inception cohorts), as this design avoids studying only subjects who tolerate a given treatment well enough to continue treatment.[22] Groups of persons whose

risk for treatment or for the outcomes of interest differ may be examined using stratification and multivariable modeling techniques such as propensity scoring,[23] disease risk scores,[24] and instrumental variables.[25] Evaluations are enhanced when adherence and compliance are accounted for.

Enough information should be presented to allow others to replicate the analyses in another database or to test alternative methods of analysis in the same or a similar data set. Replication of CE in different populations and the use of alternative analytic methods can strengthen the conclusions that may be drawn from nonrandomized studies. It may also be useful to report the results of observational studies of CE in the context of how well they support existing clinical trials data.[26] When the results of observational CE studies are inconsistent with those of a randomized clinical trial for similar patient subgroups, plausible explanations must be sought to avoid uncertainty about how to interpret the results of either type of study.

How valid is the interpretation of CE for the population of interest, assuming sound methods and appropriate follow-up?

A key challenge to interpreting CE studies is understanding how determinants of treatment choice are related to the expected outcomes. The highest-quality evidence comes from nonrandomized studies with the least potential for bias, especially for treatment assignment. For example, a direct way to obtain unbiased evidence about drugs would be to compare groups with similar levels of insurance coverage in which treatment decisions are driven largely by differences in benefit design and less by patient characteristics or physician preferences, as the choice of insurance (or residence, for national insurance plans) is generally unrelated to formulary decisions and treatment outcomes. The challenge in using instrumental variables like these, which have the promise of approximating randomization, is the lack of complete assurance that the variable is unrelated to outcomes directly or through patient characteristics.

Offering almost as high quality is evidence derived from situations in which various treatments are commonly used and there is no good evidence favoring one treatment over another or in situations where a reliable understanding of the drivers for physician treatment preferences and treatment determinants is independent of patient characteristics. As an illustration, consider when differences in hospital formularies discourage physicians from using a product in one hospital but promote its use in another hospital. It is unlikely that patients would choose a hospital because of its formulary, so contrasting the outcomes of similar patients treated in hospitals with different coverage for the product of interest would be unbiased.[27]

The lowest-evidence quality comes from small studies and those that are less rigorous in the quality of data collected or that require assumptions about the causal inference chain that may be open to dispute. Nevertheless, such studies can identify important previously unrecognized benefits that bear further investigation such as the reduction in suicide from using clozapine,[28] a finding that was confirmed in a trial[29] and led to approval of a new indication. Studies that fall into this evidence tier may reduce some uncertainty about the magnitude of treatment effects, although it may be unclear to what extent unknown confounding factors could have artificially affected the apparent benefit of one treatment compared with another.

Generally, unless an effect is observed that is much larger than would be expected or larger than could reasonably be explained by bias or that provides new information where none was available, studies in this category of lowest evidence quality are less likely to contribute meaningfully to clinical decision making. Although there is no unanimity about how large a relative benefit is needed to be worthy of consideration as evidence for decision making, some investigators suggest that analyses showing a doubling (or more) of the relative benefit should be given serious consideration,[12] while others set the bar higher.[30]

Alternative explanations should be considered, as accurate interpretation depends on understanding the extent to which bias (systematic error stemming from factors that are related both to the decision to treat and to the outcomes of interest) has distorted the results. For example, selective prescribing (confounding by indication) results when persons with more severe disease or those who are resistant to other treatments are more likely to receive newer treatments. Misclassification can result from a patient's incorrect recall of a dose or drug (a particular issue for medications used as needed) or poor treatment adherence. Investigations that rely on pharmacy benefits data also present challenges for studying medications used on an as-needed basis (eg, migraine medications) and those dispensed in liquid, cream, or inhalable forms, as well as for over-the-counter medications, biologic agents, and medical devices that do not have unique product codes or that use other distribution systems. Other types of bias include detection bias[17] (when comparison groups are assessed at different points in time or by different methods), selective loss to follow-up by which patients of interest (eg, the sickest) are more likely to drop out of one treatment group than another, and performance bias in which systematic differences in care unrelated to the intervention under study affect outcomes (eg, a public health initiative promoting healthy lifestyles directed at patients receiving a particular class of treatments).

Because the potential for bias is present to some extent in all observational studies, the critical question is not "Is there bias?" but instead "What are the most likely sources of bias in this study and how much could they have distorted the results?" 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.[31] Based on structural assumptions and limited empirical data on systematic errors (including residual confounding, misclassification of exposure or outcomes, and misspecification of the exposure risk window), a range of plausible effect estimates should be computed, and such analyses should be routine.

However they may be judged, nonrandomized CE studies contribute to clinical and biologic understanding of treatments by themselves or through subsequent confirmation in a more targeted randomized clinical trial or via advances in basic science. Useful information can be obtained even from imperfect studies if they are thoughtfully designed, well conducted, carefully analyzed, and reported in a manner that addresses concerns from skeptical readers and reviewers.

Acknowledgments

We gratefully acknowledge thoughtful reviews and contributions by Jesse Berlin, ScD, Michael Cook, PhD, Ken Hornbuckle, PhD, Stephen Motsko, PhD, Daniel Singer, MD, Til Stürmer, MD, and the International Society of Pharmacoepidemiology.

Author Affiliations: From Outcome Sciences, Inc (NAD, REG), Cambridge, MA; Harvard Medical School (SS, BJM), Boston, MA; Eli Lilly and Company (MLB), Indianapolis, IN; World Health Information Science Consultants (AMW), Wellesley, MA; and the Institute for Clinical and Economic Review (DAO), Boston, MA.

Funding Source: Partial funding was provided by the National Pharmaceutical Council; the funder had no role in determining authorship or control of the editorial content.

Author Disclosures: Drs Dreyer and Gliklich are employees of Outcome Sciences, Inc, which received partial funding from the National Pharmaceutical Council for this research. Dr Schneeweiss reports serving as a paid advisor to HealthCore, Inc, RTI, and the World Health Information Science Consultants and has received research grants from the National Institutes of Health. Dr McNeil reports serving as a paid advisor and speaker for Genentech, Pfizer, and Wyeth. She also reports serving on the board of Edwards Lifesciences, a cardiovascular device company. Dr Berger is an employee of Eli Lilly and Company, and reports owning stock in the company. Dr Walker and Mr Ollendorf report no relationships or financial interests with any entity that would pose a conflict of interest with the subject matter of this article.

Authorship Information: Concept and design (NAD, SS, BJM, MLB, AMW, DAO, REG); acquisition of data (NAD); analysis and interpretation of data (NAD, AMW); drafting of the manuscript (NAD, SS, BJM, MLB, DAO, REG); critical revision of the manuscript for important intellectual content (NAD, SS, BJM, MLB, AMW, DAO, REG); obtaining funding (NAD); administrative, technical, or logistic support (NAD); and supervision (REG).

Address correspondence to: Nancy A. Dreyer, PhD, Outcome Sciences, Inc, 201 Broadway, Cambridge, MA 02139. E-mail: ndreyer@outcome.com.

REFERENCES

1. Federal Coordinating Council for Comparative Effectiveness Research. Report to the president and the Congress. June 30, 2009. http://www.hhs.gov/recovery/programs/cer/cerannualrpt.pdf. Accessed October 11, 2009.

2. Concato J, Shah N, Horowitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med.* 2000;342(25):1887-1892.

3. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med.* 2000;342(25):1878-1886.

4. Tilson H, Doi PA, Covington DL, Parker A, Schields K, White A. The Antiretrovirals in Pregnancy Registry: a fifteenth anniversary celebration. *Obstet Gynecol Survey.* 2007;62(2):137-148.

5. Curtis JP, Luebbert JJ, Wang Y, et al. Association of physician certification and outcomes among patients receiving an implantable cardioverter-defibrillator. *JAMA.* 2009;301(16):1661-1670.

6. Taylor RS, Bethell HJN, Brodie DA. Clinical trials versus the real world: the example of cardiac rehabilitation. *Br J Cardiol.* 2007;14(3): 175-178.

7. Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report: part I [published online ahead of print September 29, 2009]. *Value Health.* 2009;12(8):1044-1061.

8. Johnson ML, Crown W, Martin BC, Dormuch CR, Siebert U. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report: part III [published online ahead of print September 29, 2009]. *Value Health.* 2009;12(8):1062-1073.

9. International Society of Pharmacoepidemiology (ISPE). Guidelines for good pharmacoepidemiology practices (GPP). *Pharmacoepidemiol Drug Saf.* 2008;17(2):200-208.

10. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med.* 2007;147(8):573-577.

11. Vandenbroucke Jan P. STREGA, STROBE, STARD, SQUIRE, MOOSE, PRISMA, GNOSIS, TREND, ORION, COREQ, QUORUM, REMARK… and CONSORT: for whom does the guideline toll? *J Clin Epidemiol.* 2009;62(6):594-596.

12. Atkins D, Best D, Briss PA, et al; GRADE Working Group. Grading quality of evidence and strength of recommendations. *BMJ.* 2004;328(7454):e1490.

13. GRACE Web site. GRACE principles: Good Research for Comparative Effectiveness. http://www.graceprinciples.org. Accessed May 15, 2010.

14. Dreyer NA, Rubino A, L'Italien GJ, Schneeweiss S. Developing good practice guidance for non-randomized studies of comparative effectiveness: a workshop on quality and transparency. *Pharmacoepidemiol Drug Saf.* 2009;18:S123.

15. Dreyer NA, Berger M, Sullivan S, LeLorier J. Are good practices principles for observational comparative effectiveness studies needed? Paper presented at: International Society for Pharmacoeconomics and Outcomes Research 13th Annual International Meeting; May 6, 2008; Toronto, Ontario, Canada.

16. Gliklich 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 HHSA29020050035I TO1. Rockville, MD: Agency for Healthcare Research and Quality; April 2007. AHRQ publication 07-EHC001-1.

17. Cochrane Collaboration Web site. Higgins JPT, Green S, eds. *Cochrane Handbook for Systematic Reviews of Interventions 5.0.2.* Updated September 2009. www.cochrane-handbook.org. Accessed May 27, 2010.

18. Agency for Healthcare Research and Quality. AHRQ Methods Guide for Effectiveness and Comparative Effectiveness Reviews [draft for public comment]. Rockville, MD: Agency for Healthcare Research and Quality; October 10, 2007.

19. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA.* 2007;297(3):278-285.

20. Deeks JJ, Dines J, D'Amico R, et al; International Stroke Trial Collaborative Group, European Carotid Surgery Trial Collaborative Group. Evaluating non-randomised intervention studies. *Health Technol Assess.* 2003;7(27):iii-x, 1-173.

21. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). *Checklist of Methodological Research Standards for ENCePP Studies* [draft for public consultation]. London, England: European Network of Centres for Pharmacoepidemiology and Pharmacovigilance; November 16, 2009. Document reference EMEA/540136/2009.

22. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. *Am J Epidemiol.* 2006;158(9):915-920.

23. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol.* 2006;98(3):253-259.

24. Stürmer T, Schneeweiss S, Brookhart MA, Rothman KJ, Avorn J, Glynn RJ. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. *Am J Epidemiol.* 2005;161(9):891-898.

25. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Instrumental variable analysis of secondary pharmacoepidemiologic data. *Epidemiology.* 2006;17(4):373-374.

26. Schneeweiss S, Patrick AR, Stürmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Med Care.* 2007;45(10) (suppl 2):S131-S142.

27. Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinen during coronary-artery bypass grafting and risk of death. *N Engl J Med.* 2008;358(8):771-783.

28. Walker AM, Lanza LL, Arellano FA, Rothman KJ. Mortality in current and former users of clozapine. *Epidemiology.* 1997;8(6):671-677.

29. Alphs L, Anand R, Islam MZ, et al. The International Suicide Prevention Trial (InterSePT): rationale and design of a trial comparing the relative ability of clozapine and olanzapine to reduce suicidal behavior in schizophrenia and schizoaffective patients. *Schizophr Bull.* 2004;30(3):577-586.

30. 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. *Pharmacoepidemiol Drug Saf.* 2008;17(7):671-685.

31. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data.* New York, NY: Springer New York; 2009. ■