The GRACE Checklist: A Validated Assessment Tool for High Quality Observational Studies of Comparative Effectiveness

Nancy A. Dreyer, PhD, MPH; Allison Bryant, MPH; and Priscilla Velentgas, PhD

ABSTRACT

BACKGROUND: Recognizing the growing need for robust evidence about treatment effectiveness in real-world populations, the Good Research for Comparative Effectiveness (GRACE) guidelines have been developed for noninterventional studies of comparative effectiveness to determine which studies are sufficiently rigorous to be reliable enough for use in health technology assessments.

OBJECTIVE: To evaluate which aspects of the GRACE Checklist contribute most strongly to recognition of quality.

METHODS: We assembled 28 observational comparative effectiveness articles published from 2001 to 2010 that compared treatment effectiveness and/or safety of drugs, medical devices, and medical procedures. Twenty-two volunteers from academia, pharmaceutical companies, and government agencies applied the GRACE Checklist to those articles, providing 56 assessments. Ten senior academic and industry experts provided assessments of overall article quality for the purpose of decision support. We also rated each article based on the number of annual citations and impact factor of the journal in which the article was published. To identify checklist items that were most predictive of quality, classification and regression tree (CART) analysis, a binary, recursive, partitioning methodology, was used to create 3 decision trees, which compared the 56 article assessments with 3 external quality outcomes: (1) expert assessment of overall quality, (2) citation frequency, and (3) impact factor. A fourth tree looked at the composite outcome of all 3 quality indicators.

RESULTS: The best predictors of quality included the following: use of concurrent comparators, limiting the study to new initiators of the study drug, equivalent measurement of outcomes in study groups, collecting data on most if not all known confounders or effect modifiers, accounting for immortal time bias in the analysis, and use of sensitivity analyses to test how much effect estimates depended on various assumptions. Only sensitivity analyses appeared consistently as a predictor of quality in all 4 trees. When a composite outcome of the 3 quality measures was used, the GRACE Checklist showed high sensitivity and specificity (71.43% and 80.95%, respectively).

CONCLUSIONS: The GRACE Checklist stands out from other consensus-driven and expert guidance documents because of its extensive validation efforts. This most recent work shows that the checklist has strong sensitivity and specificity, increasing its utility as a screening tool to identify high-quality observational comparative effectiveness research worthy of in-depth review and applicability for decision support.

J Manag Care Spec Pharm. 2016;22(10):1107-13

Copyright© 2016, Academy of Managed Care Pharmacy. All rights reserved.

What is already known about this subject

- There are tremendous unmet needs for information on comparative treatment effectiveness, including treatment effect heterogeneity.
- The Good Research for Comparative Effectiveness (GRACE) Checklist is an 11-item tool designed to evaluate the quality of data and methods used in study design and analysis.
- Although many guidance documents have been developed for the evaluation of observational study quality, only the GRACE Checklist has been validated against external measures of quality.

What this study adds

- This study extends validation by using classification and regression tree analysis to examine predictive values of the GRACE Checklist for several external measures of study quality.
- Use of sensitivity analyses appeared as the single most consistent individual predictor of quality across all 4 quality measures.
- Using a composite outcome of 3 measures of quality, the GRACE Checklist showed high sensitivity (71.43%) and even higher specificity (80.95%) to identify high-quality studies.

As the need for information about the safety, effectiveness, and value of medical treatments escalates along with their cost, the demand for reliable evidence intensifies. In the quest for personalized medicine, patients and health care providers seek information about treatment effectiveness among similar patient populations. Health systems need effective and efficient approaches to health technology assessments. Countries, such as Japan, that have provided drug coverage through national health insurance are facing the pressures of increasing drug costs.¹ We see similar challenges in countries such as the United States, where cost is not explicitly part of drug coverage determination. For example, the Centers for Medicare & Medicaid Services (CMS) in the United States assigns 20% of medication costs under Part B to the patient through copayments without any cap on the amount of coinsurance, an increasingly insurmountable problem for new high-cost treatments for cancer and other conditions.² Recognizing these economic pressures, we see medical professions, such as cardiology and oncology, developing structured approaches to evaluating evidence.^{3,4}

All evidence frameworks rely on high-quality studies to provide evidence of reasonable quality for the patients of interest. Traditionally, we have relied on evidence from randomized controlled clinical trials and systematic reviews of these trials, mainly because it is easier to draw causal inferences regarding the treatments being compared when the only difference between patients was random assignment of treatment.⁵ While such trials are tremendously valuable, they may not reflect the population for whom reimbursement is sought; often use surrogate clinical markers rather than true clinical endpoints; and may have insufficient follow-up for endpoints of real interest to patients, providers, payers, and health agencies. Increasingly, payers and those who conduct health technology assessments are turning to noninterventional studies for information about populations, subgroups, and comparators of interest and for granular information about treatment pathways, combinations, and sequencing to support decisions about step therapy.⁶

Recognizing this need, guidelines have been developed to promote the design and conduct of high-quality noninterventional studies.⁷⁻⁹ At the same time, consensus-based and evidence-based assessment tools have been developed to assist reviewers in determining which nonrandomized studies are sufficiently rigorous to contribute to a health technology assessment.¹⁰⁻¹³ Nonetheless, there is substantial variability in the guidelines and assessment tools with regard to which methodologic elements are included and the recommended approaches.¹⁴ When faced with a series of publications, all addressing treatments of interest, researchers may wonder which aspects of these guidance documents actually have practical importance.

The Good Research for Comparative Effectiveness (GRACE) Checklist is composed of 11 items that can be used to evaluate the quality of an observational study of comparative effectiveness in the context of any practical treatment or formulary decision. Six items evaluate the quality of the data, and 5 items address the methods used in study design and analysis. The checklist item descriptions and scoring guide are presented in Table 1. Description of the checklist development and initial validation testing have been published previously.¹¹⁻¹³

Methods

We tested each of the GRACE Checklist items against 3 different measures of quality. Twenty-two volunteers from 4 continents, consisting of professors, senior scientists, and researchers with training primarily in epidemiology and statistics, were recruited from academia, industry, and government agencies through emails and personal requests. All volunteers possessed a masters or doctorate degree and had on average 12 years of postdegree experience. The volunteers completed checklist assessments for 28 observational comparative effectiveness articles published from 2001 to 2010 that compared

treatment effectiveness and/or safety of drugs, medical devices, and medical procedures. These articles were selected from the expert-rated articles used in the previous GRACE Checklist validation testing.¹² Volunteers rated from 1 to 4 articles each, with the majority rating 3 articles each, for a total of 56 assessments. The external benchmarks, or “gold standards,” of quality to which the checklist assessments were compared were the following: (1) expert assessments of overall article quality, (2) number of citations per year, and (3) the impact factor of the journal in which the article was published. The expert assessments were conducted by 10 experts in observational research conduct recruited from academia, pharmaceutical companies, and payers based on lengthy record of publications in the area of observational research methods and/or previous participation in the GRACE Initiative and willingness to participate. Each expert was asked to review and rate the published articles in a pass/fail manner in terms of whether the study was of sufficient quality to be used to support a formulary decision. We examined the number of times these articles were cited per year and the impact factor of the journal in which each article was published, excluding self-citations, using the Web of Science in January 2015.¹⁵

We used classification and regression tree (CART) analyses (Salford Systems, San Diego, CA), a recursive partitioning methodology that creates decision trees based on binary splits of the outcome variable according to cut-offs of the predictor variables. CART analysis is a nonparametric method that can automatically detect complex interactions between predictor variables and uses sophisticated methods for handling missing variables. We created 3 decision trees based on the 56 article assessments to identify checklist items that were most predictive of the following outcomes: (a) expert assessment of overall quality (sufficient or insufficient); (b) frequency of article citations (≤ 2 vs. > 2 citations per year); and (c) higher impact factors (≤ 2.5 vs. > 2.5 journal impact factor). Cut-points for article citations and journal impact factor were determined based on the median of the distribution of values. A fourth tree was created based on the composite outcome of all 3 outcomes, with quality defined as an article with a journal impact factor > 2.5 , > 2 article citations per year, and an expert assessment of overall quality classified as “sufficient.”

For analysis, the volunteers’ checklist question responses that were not already mapped to the categories of sufficient (good enough for decision support) or insufficient were coded as follows: (a) responses that indicated “not enough information in article” were treated as insufficient, since this lack of information was viewed as a negative aspect of study quality; (b) responses of “not applicable” were classified as sufficient so that an article would not be rated negatively if a specific question was not relevant to its objective; and (c) blank responses were treated as missing values.

TABLE 1 GRACE Checklist: Components and Response Guide

Component Item	Scoring as Fit for Purpose: Sufficient (+), Insufficient (-)
Data	
D1. Were treatment and/or important details of treatment exposure adequately recorded for the study purpose in the data source(s)? Note: not all details of treatment are required for all research questions.	(+) Yes, reasonably necessary information to determine treatment or intervention was adequately recorded for study purposes (e.g., for drugs, sufficient detail on dose, days supplied, route, or other important data. For vaccines, consider the importance of batch, dose, route, and site of administration, etc. For devices, consider type of device, placement, surgical procedure used, serial number, etc.). (-) No, data source clearly deficient, or not enough information in article.
D2. Were the primary outcomes adequately recorded for the study purpose (e.g., available in sufficient detail through data sources)?	(+) Yes, information to ascertain outcomes were adequately recorded in the data source (e.g., if clinical outcomes were ascertained using ICD-9-CM diagnosis codes in an administrative database, the level of sensitivity and specificity captured by the codes were sufficient for assessing the outcome of interest). (-) No, data source clearly deficient (e.g., the codes captured a range of conditions that was too broad or narrow, and supplementary information such as that from medical charts was not available), or not enough information in article.
D3. Was the primary clinical outcome(s) measured objectively rather than subject to clinical judgment (e.g., opinion about whether the patient's condition has improved)?	(+) Yes, clinical outcomes were measured objectively (e.g., hospitalization, mortality). (+) Not applicable; primary outcome not clinical (e.g., PROs). (-) No (e.g., clinical opinion about whether patient's condition improved) or not enough information in article.
D4. Were primary outcomes validated, adjudicated, or otherwise known to be valid in a similar population?	(+) Yes, outcomes were validated, adjudicated, or based on medical chart abstractions with clear definitions (e.g., a validated instrument was used to assess patient-reported outcomes [e.g., SF-12 Health Survey]; a clinical diagnosis via ICD-9-CM code was used, with formal medical record adjudication by committee to confirm diagnosis or other procedures to achieve reasonable sensitivity and specificity; and billing data were used to assess health resource utilization). (-) No, or not enough information in article.
D5. Was the primary outcome(s) measured or identified in an equivalent manner between the treatment/intervention group and the comparison group?	(+) Yes. (-) No, or not enough information in article.
D6. Were important covariates that may be known confounders or effect modifiers available and recorded? Important covariates depend on the treatment and/or outcome of interest (e.g., body mass index should be available and recorded for studies of diabetes; race should be available and recorded for studies of hypertension and glaucoma).	(+) Yes, most if not all important known confounders and effect modifiers available and recorded (e.g., measures of medication dose and duration). (-) No, at least 1 probable known confounder or effect modifier not available and recorded (as noted by authors or as determined by user's clinical knowledge), or not enough information in article.
Methods	
M1. Was the study (or analysis) population restricted to new initiators of treatment or those starting a new course of treatment? Efforts to include only new initiators may include restricting the cohort to those who had a washout period (specified period of medication nonuse) before the beginning of study follow-up.	(+) Yes, only new initiators of the treatment of interest were included in the cohort, or for surgical procedures and devices, including only patients who never had the treatment before the start of study follow-up. (-) No, or not enough information in article.
M2. If 1 or more comparison groups were used, were they concurrent comparators? If not, did the authors justify the use of historical comparison groups?	(+) Yes, data were collected during the same time period as the treatment group (concurrent), or historical comparators were used with reasonable justification (e.g., when it is impossible for researchers to identify current users of older treatments or when a concurrent comparison group is not valid, as when uptake of new product is so rapid that concurrent comparators differ greatly on factors related to the outcome). (-) No, historical comparators used without being scientifically justifiable, or not enough information in article.
M3. Were important confounding and effect-modifying variables taken into account in the design and/or analysis? Appropriate methods to take these variables into account may include restriction, stratification, interaction terms, multivariate analysis, propensity score matching, instrumental variables, or other approaches.	(+) Yes, most if not all important covariates that would be likely to change the effect estimate substantially were accounted for (e.g., measures of medication dose and duration). (-) No, some important covariates were available for analysis but not analyzed appropriately, or at least 1 important covariate was not measured, or not enough information in article.
M4. Is the classification of exposed and unexposed person-time free of "immortal time bias," i.e., "immortal time" in epidemiology refers to a period of cohort follow-up time during which death (or an outcome that determines end of follow-up) cannot occur.	(+) Yes. (-) No, or not enough information in the article.
M5. Were any meaningful analyses conducted to test key assumptions on which primary results are based (e.g., were some analyses reported to evaluate the potential for a biased assessment of exposure or outcome, such as analyses where the impact of varying exposure and/or outcome definitions was tested to examine the impact on results)?	(+) Yes, and primary results did not substantially change. (-) Yes, and primary results changed substantially. (-) None reported, or not enough information in article.

Source: Dreyer NA, Valentgas P, Westrich K, Dubois R. The GRACE checklist for rating the quality of observational studies of comparative effectiveness: a tale of hope and caution.¹² GRACE = Good Research for Comparative Effectiveness; ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modification; PRO = patient-reported outcomes.

FIGURE 1 CART Trees for GRACE Checklist Questions that Predict Article Quality, Article Citations, and Journal Impact Factor

Note: N is the total cases in each node; % is the proportion of cases classified as having sufficient quality, >2 article citations per year, or >2.5 journal impact factor. An article was considered to have sufficient quality for the composite outcome if it had >2 article citations per year, >2.5 journal impact factor, and sufficient expert-rated quality. CART=classification and regression tree; GRACE=Good Research for Comparative Effectiveness.

Results

Four approaches to predicting an assessment of sufficient quality are shown in Figure 1. When using expert assessment as the measure of quality, the strongest predictors of quality, listed in rank order, were as follows: (1) use of concurrent comparators, (2) equivalent measurement of outcomes in study groups, (3) collection of data on most if not all known confounders and effect modifiers, and (4) reporting sensitivity analyses. When the number of article citations per year was used as the marker of quality, 3 items emerged as important predictors of quality: (1) limiting

the study to new initiators of the study drug, (2) reporting sensitivity analyses, and (3) avoiding immortal time bias in the analysis. When the journal impact factor was the marker of quality, reporting sensitivity analyses emerged as the only predictor that differentiated impact factors over 2.5 from lower impact factors. Also, reporting of sensitivity analyses was the only predictor of article quality when the composite outcome of higher journal impact factor, frequent article citations, and sufficient expert assessment was used as the marker of quality. The decision tree based on the composite outcome as shown in Figure 1 had higher overall predictive ability than for each outcome examined

TABLE 2 Tree Performance According to Quality Outcome Measure

"Gold Standard" Quality Outcome Measure	Articles Meeting Quality Standard %	Sensitivity ^a %	Specificity ^b %
Expert assessment	57.14 (n=16 of 28)	78.13	58.33
Article citations per year	46.43 (n=13 of 28)	50.00	63.33
Journal impact factor	50.00 (n=13 of 26) ^c	66.67	85.29
Composite outcome	26.92 (n=7 of 26) ^c	71.43	80.95

^aSensitivity indicates the proportion of articles with sufficient gold standard quality outcomes that were deemed sufficient by the volunteer raters.

^bSpecificity indicates the proportion of articles with insufficient gold standard quality outcomes that were deemed insufficient by the volunteer raters.

^cJournal impact factor was not available for 2 articles.

separately, with a sensitivity of 71.43% and specificity of 80.95% (Table 2). Looking across the 4 types of quality assessment, only sensitivity analyses appeared in all 4 trees.

Discussion

The results of this analysis showed that sensitivity analyses, a basic epidemiologic tool, stands out among other factors in its consistent prediction of quality. What is it about the use and reporting of sensitivity analyses in a journal article that makes it such a powerful differentiator for noninterventional studies?

Recognizing that epidemiologic research is "an exercise in measurement," sensitivity analyses allow quantitative or semiquantitative estimates of how much a study's results are dependent on any key assumptions.¹⁶ For example, the Registry in Glaucoma Outcomes Research (RiGOR) study, a prospective observational study of the comparative effectiveness of treatment strategies for open-angle glaucoma in the United States, based its primary analyses of treatment effectiveness as measured by a 15% decrease in intraocular pressure at 12 months. Sensitivity analyses were conducted where patients who discontinued before 12 months were considered treatment failures. Results from the sensitivity analyses assuming all patients without a 12-month visit were treatment failures were reassuringly similar for both the comparison of laser surgery with additional medication (primary analysis odds ratio [OR]=1.11, 95% confidence interval [CI]=0.88-1.39; sensitivity analysis OR=1.07, 95% CI=0.87-1.32) and that of incisional surgery versus additional medication (primary analysis OR=2.26, 95% CI=1.68-3.05; sensitivity analysis OR=2.23, 95% CI=1.72-2.89).¹⁷ A straightforward approach such as this can be useful in helping to exclude an alternative explanation for the observed study results.

It is also possible that the identification of sensitivity analyses as a predictor of article quality is "confounded" to some degree by the tendency of higher impact journals to favor or promote the inclusion of sensitivity analyses results in published manuscripts. This cannot be directly assessed by the

available data on published manuscripts, since researchers may have conducted sensitivity analyses that were not mentioned in the published article because of length restrictions and other considerations made by the authors and editors.

Other strong indicators of quality that emerged include use of concurrent comparators (the strongest predictor of expert assessment of quality) and restriction to new initiators of treatment (strongest predictor of number of article citations per year). Both of these are well established in the pharmacoepidemiology methods literature as important considerations in the design of observational studies to support causal inference and have received renewed attention in the context of comparative effectiveness research.^{8,18}

The GRACE Checklist, unlike solely consensus-driven guidelines, shown in Table 3, uses a set of items based on consensus of experts that were then tested by users from a wide variety of backgrounds and training, representing perspectives from 4 continents.¹² Validation of checklist items against multiple external measures of article quality with high sensitivity and specificity provides additional support for its use as an assessment tool in evaluating the quality of observational comparative effectiveness research studies.

Limitations

There is no single accepted "gold standard" for article quality against which assessments such as the GRACE Checklist or other instruments may be validated. One of the most intuitively appealing measures, an overall assessment of quality by recognized experts in the design and conduct of observational studies, showed relatively low agreement (52%) between the reviewers for the articles that received multiple expert assessments.¹² This variability may also reflect the difficulty of making an assessment of article quality without the context of a specific real-world decision and its consequences, although the experts received instructions as to the hypothetical context of the decision they were asked to make ("a formulary decision"). Sample size, in this case the number of published comparative effectiveness research studies for which both checklist assessments and external measures of "quality" were available, also constrained the number of analyses that could be conducted and the identification of a larger number of possibly significant predictors through CART analysis. The limitations of publication bias, with regard to whether the results of completed observational studies ever are published and whether specific components of study quality assessed by the GRACE Checklist are reported in a given article, would be limitations to the assessment of the body of evidence for a specific comparative effectiveness question by any method and are not specific to the GRACE Checklist's attributes. Finally, the GRACE Checklist can only be used to qualify methodologic rigor, not applicability to a given decision. For example, comparators may not include all those of interest in a given country or reimbursement program.

TABLE 3 Guidance Documents and Assessment Tools for Designing and Assessing the Quality of Observational Studies

Documents	Category	Method of Development
AHRQ Registries for Evaluating Patient Outcomes, 2014 ⁷	Guidance	Consensus based
AHRQ Developing a Protocol for Observational Comparative Effectiveness Research, 2013 ⁸	Guidance	Consensus based
ISPOR Prospective Observational Studies to Assess Comparative Effectiveness, 2012 ⁹	Guidance	Consensus based
The PCORI Methodology Report, 2013 ¹⁹	Guidance	Consensus based
ENCePP Guide on Methodologic Standards in Pharmacoepidemiology, rev. 5, 2016 ²⁰	Guidance	Consensus based
Good Research Practices for Comparative Effectiveness Research (ISPOR), 2009 ²¹⁻²³	Guidance	Consensus based
GRACE Principles, 2010 ¹¹	Guidance	Consensus based
GRACE Checklist, 2014 ¹²	Assessment tool	Consensus based, with validation against external measures of quality
CER Collaborative Questionnaire, 2014 ¹⁰	Assessment tool	Consensus based, with validation against quality assessments by task force members

AHRQ=Agency for Healthcare Research and Quality; CER=comparative effectiveness research; ENCePP=European Network of Centres for Pharmacoepidemiology and Pharmacovigilance; GRACE=Good Research for Comparative Effectiveness; ISPOR=International Society for Pharmacoeconomics and Outcomes Research; PICORI=Patient-Centered Outcomes Research Institute.

Conclusions

When the GRACE Checklist for assessment of observational study quality was applied to 28 articles and compared with 4 external measures of article quality (expert assessment, number of citations/year, journal impact factor, and a composite of all 3 measures) using CART analysis, reporting sensitivity analyses showing how much effect estimates depended on various assumptions was shown to be a strong predictor of each of the 4 measures of quality. Other strong predictors of quality that appeared in individual trees were use of concurrent comparators, limiting the study to new initiators of the study drug, equivalent measurement of outcomes in study groups, collection of data on most if not all confounders and effect modifiers, and accounting for immortal time bias in the analysis. These validation efforts distinguish the GRACE Checklist from other guidance documents that were either solely consensus driven or have not been compared with external assessments of article quality.¹⁰ This most recent work shows that the GRACE Checklist has strong sensitivity and specificity, increasing its utility as a screening tool to identify high-quality observational comparative effectiveness research worthy of consideration in decision making.

Authors

NANCY A. DREYER, PhD, MPH; ALLISON BRYANT, MPH; and PRISCILLA VELENTGAS, PhD, Quintiles Real-World & Late-Phase Research, Cambridge, Massachusetts.

AUTHOR CORRESPONDENCE: Nancy A. Dreyer, PhD, MPH, Quintiles Real-World & Late-Phase Research, 201 Broadway, Cambridge, MA 02139. Tel.: 617.715.6810; E-mail: nancy.dreyer@quintiles.com.

DISCLOSURES

No outside funding supported this research. All authors are full-time employees of Quintiles, which provides research and consulting services to the biopharmaceutical industry. The authors have no other disclosures to report.

Two of the 3 CART trees were presented at the International Society of Pharmacoepidemiology in 2015 ("Article Citations per Year" and "Journal Impact Factor"). The original validation study was published in the March 2014 issue of the *Journal of Managed Care & Specialty Pharmacy*. The checklist questions and scoring were included using a table that was originally published by this journal in 2014.

Study concept and design were primarily contributed by Dreyer and Velentgas, along with Bryant. Bryant took the lead in data collection and analysis, along with Dreyer and Velentgas, and data interpretation was performed by Dreyer, Velentgas, and Bryant. The manuscript was written and revised primarily by Dreyer, along with Bryant and Velentgas.

ACKNOWLEDGMENTS

Statistical consultation was provided by Zhaohui Su, PhD.

REFERENCES

- Reich MR, Shibuya K. The future of Japan's health system—sustaining good health with equity at low cost. *N Engl J Med*. 2015;373(19):1793-97.
- Bach PB. New math on drug cost-effectiveness. *N Engl J Med*. 2015; 373(19):1797-99.
- Young RC. Value-based cancer care. *N Engl J Med*. 2015;373(27):2593-95.
- Neuman PJ, Cohen JR. Measuring the value of prescription drugs. *N Engl J Med*. 2015;373(2):2595-97.
- Barton S. Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. *BMJ*. 2000;321(7256):255-56.
- Faria R, Alava MH, Manca A, Wailoo AJ. NICE DSU technical support document 17: the use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data. May 2015. Available at: <http://www.nicedsu.org.uk/TSD17%20-%20DSU%20Observational%20data%20FINAL.pdf>. Accessed August 15, 2016.

7. Gliklich R, Dreyer N, Leavy M, eds. Registries for evaluating patient outcomes: a user's guide. Third edition. Two volumes. (Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc., a Quintiles company] under Contract No. 290 2005 00351 TO7.) AHRQ Publication No. 13(14)-EHC111. April 2014. Agency for Healthcare Research and Quality. Rockville, MD. Available at: <http://www.ncbi.nlm.nih.gov/books/NBK208616/>. Accessed August 15, 2016.
8. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, eds. Developing a protocol for observational comparative effectiveness research: a user's guide. AHRQ Publication No. 12(13)-EHC099. January 2013. Agency for Healthcare Research and Quality. Rockville, MD. Available at: www.effectivehealthcare.ahrq.gov/ehc/products/440/1166/User-Guide-to-Observational-CER-1-10-13.pdf. Accessed August 15, 2016.
9. Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand, SL. Prospective observational studies to assess comparative effectiveness: ISPOR good research practices task force report. *Value Health*. 2012;15(2):217-30.
10. Berger ML, Martin BC, Husereau D, et al. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC good practice task force report. *Value Health*. 2014;17(2):143-56.
11. Dreyer NA, Schneeweiss S, McNeil BJ, et al. GRACE principles: recognizing high-quality observational studies of comparative effectiveness. *Am J Manag Care*. 2010;16(6):467-71.
12. Dreyer NA, Velentgas P, Westrich K, Dubois R. The GRACE checklist for rating the quality of observational studies of comparative effectiveness: a tale of hope and caution. *J Manag Care Spec Pharm*. 2014;20(3):301-08. Available at: <http://www.jmcp.org/doi/10.18553/jmcp.2014.20.3.301>.
13. Dreyer NA. Using observational studies for comparative effectiveness: finding quality with GRACE. *J Comp Eff Res*. 2013;2(5):413-18.
14. Morton SC, Costlow MR, Graff JS, Dubois RW. Standards and guidelines for observational studies: quality is in the eye of the beholder. *J Clin Epidemiol*. 2016;71:3-10.
15. Garfield E. The Thomson Reuters impact factor. Web of Science. Available at: <http://wokinfo.com/essays/impact-factor/>. Accessed August 15, 2016.
16. Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health*. 2005;95(Suppl 1):S144-50.
17. Coleman A, Lum F, Velentgas P, Su Z, Gliklich R. Impact of treatment strategies for open angle glaucoma on intraocular pressure: the RiGOR study. *J Comp Eff Res*. 2016;5(1):87-98.
18. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nature Rev Rheumatol*. 2015;11(7):437-41.
19. Patient-Centered Outcomes Research Institute Methodology Committee. The PCORI methodology report. November 2013. Available at: <http://www.pcori.org/assets/2013/11/PCORI-Methodology-Report.pdf>. Accessed August 15, 2016.
20. European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. ENCePP guide on methodological standards in pharmacoepidemiology. Revision 5. EMA/95098/2010 Rev. 5. July 2016. Available at: http://www.encepp.eu/standards_and_guidances/methodologicalGuide.shtml. Accessed August 15, 2016.
21. 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. *Value Health*. 2009;12(8):1044-52.
22. Cox E, Martin BC, Van Staa T, Garbe E, Siebert U, Johnson ML. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II. *Value Health*. 2009;12(8):1053-61.
23. Johnson ML, Crown W, Martin BC, Dormuth 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. *Value Health*. 2009;12(8):1062-73.