RESEARCH

The GRACE Checklist for Rating the Quality of Observational Studies of Comparative Effectiveness: A Tale of Hope and Caution

Nancy A. Dreyer, PhD, MPH; Priscilla Velentgas, PhD; Kimberly Westrich, MA; and Robert Dubois, MD

ABSTRACT

BACKGROUND: While there is growing demand for information about comparative effectiveness (CE), there is substantial debate about whether and when observational studies have sufficient quality to support decision making.

OBJECTIVE: To develop and test an item checklist that can be used to qualify those observational CE studies sufficiently rigorous in design and execution to contribute meaningfully to the evidence base for decision support.

METHODS: An 11-item checklist about data and methods (the GRACE checklist) was developed through literature review and consultation with experts from professional societies, payer groups, the private sector, and academia. Since no single gold standard exists for validation, checklist item responses were compared with 3 different types of external quality ratings (N=88 articles). The articles compared treatment effectiveness and/or safety of drugs, medical devices, and medical procedures. We validated checklist item responses 3 ways against external quality ratings, using published articles of observational CE or safety studies: (a) Systematic Review–quality assessment from a published systematic review; (b) Single Expert Review–quality assessment made according to the solicited "expert opinion" of a senior researcher; and (c) Concordant Expert Review–quality assessments from 2 experts for which there was concordance. Volunteers (N=113) from 5 continents completed 280 article assessments using the checklist. Positive and negative predictive values (PPV, NPV, respectively) of individual items were estimated to compare testers' assessments with those of experts.

RESULTS: Taken as a whole, the scale had better NPV than PPV, for both data and methods. The most consistent predictor of quality relates to the validity of the primary outcomes measurement for the study purpose. Other consistent markers of quality relate to using concurrent comparators, minimizing the effects of bias by prudent choice of covariates, and using sensitivity analysis to test robustness of results. Concordance of expert opinion on the quality of the rated articles was 52%; most checklist items performed better.

CONCLUSIONS: The 11-item GRACE checklist provides guidance to help determine which observational studies of CE have used strong scientific methods and good data that are fit for purpose and merit consideration for decision making. The checklist contains a parsimonious set of elements that can be objectively assessed in published studies, and user testing shows that it can be successfully applied to studies of drugs, medical devices, and clinical and surgical interventions. Although no scoring is provided, study reports that rate relatively well across checklist items merit in-depth examination to understand applicability, effect size, and likelihood of residual bias.

The current testing and validation efforts did not achieve clear discrimination between studies fit for purpose and those not, but we have identified a critical, though remediable, limitation in our approach. Not specifying a specific granular decision for evaluation, or not identifying a single study objective in reports that included more than one, left reviewers with too broad an assessment challenge. We believe that future efforts will be more successful if reviewers are asked to focus on a specific objective or question.

Despite the challenges encountered in this testing, an agreed upon set of assessment elements, checklists, or score cards is critical for the maturation of this field. Substantial resources will be expended on studies of real-world effectiveness, and if the rigor of these observational assessments cannot be assessed, then the impact of the studies will be suboptimal. Similarly, agreement on key elements of quality will ensure that budgets are appropriately directed toward those elements. Given the importance of this task and the lessons learned from these extensive efforts at validation and user testing, we are optimistic about the potential for improved assessments that can be used for diverse situations by people with a wide range of experience and training. Future testing would benefit by directing reviewers to address a single, granular research question, which would avoid problems that arose by using the checklist to evaluate multiple objectives, by using other types of validation test sets, and by employing further multivariate analysis to see if any combination or sequence of item responses has particularly high predictive validity.

J Manag Care Pharm. 2014;20(3):301-08

What is already known about this subject

- While there is growing demand for information about comparative effectiveness (CE), there is little understanding about when noninterventional studies are good enough for decision support.
- Several expert reports have been issued listing criteria that are believed to be important in determining the quality of observational CE studies, yet there have been no systematic, published evaluations of whether or how such criteria actually perform.

What this study adds

- We developed the GRACE checklist, an objective 11-item checklist about the key attributes of high-quality noninterventional CE studies, a checklist that evaluates data and methods, but not motives, conflicts of interest, or interpretation. We then conducted several validation efforts using a large number of raters with diverse training and experience to determine how those individual elements performed when applied to expert opinions on quality.
- This testing revealed that the most consistent predictors of quality relate to the validity of the primary outcomes for the study purpose.
- Other relatively consistent predictors of quality were related to use of concurrent comparators, whether important covariates were recorded and accounted for, and whether sensitivity analyses were shown to support robustness of the conclusion.

What this study adds *(continued)*

- On the whole, GRACE checklist items performed better than opinions from individual experts and concurrent expert opinions. Nonetheless, the checklist would benefit from further validation efforts, including directing reviewers to address a specific objective for each evaluation, finding additional validation test sets to evaluate the robustness of the checklist, and conducting more multivariate analyses to determine whether any combinations or sequences of responses can improve the ability of the checklist to discriminate studies of reasonably strong quality for the purpose at hand.

Developing a sustainable health system requires health care that is guided by reliable information about which medical diagnostics and treatments work best, for whom, and in what situations.[1] To meet the diverse needs of clinicians, policy makers, and those who decide about formularies, the full range of comparative effectiveness (CE) studies—randomized controlled trials, observational research (also referred to as noninterventional research since treatments are not assigned by protocol), and meta-analyses—are needed. Observational studies are particularly useful because they often provide information about diverse populations, practitioners, and settings in a timely and cost-effective manner.

Recent calls for using the full range of high-quality CE research to inform decisions about medical diagnostics and interventions have brought forth a spate of consensus offerings about recognizing quality in observational CE studies and meta-analysis.[2-15] These papers have face validity and largely appear reasonable, but there is little, if any, evidence that any of these recommendations can actually distinguish studies of sufficient quality to merit serious evaluation for a particular clinical or payment decision. For example, some guidelines address potential conflicts of interest by calling for full disclosure, a standard journal practice that relies on individual assessment of potential conflicts. Some insist that, like clinical trials, only hypotheses that were specified in advance of collecting any data have validity. Others omit the criterion about prespecified hypotheses, instead giving more weight to the value of descriptive data for filling gaps and shaping subsequent research. One very practical, high-level description of good practice, published in this journal by Willke and Mullins in 2011,[9] focused on good research practices for the conduct and reporting of CE research using real-world data with nonrandom assignment of treatments. They offer "Ten Commandments for improving the systematic use of principles that are aimed at achieving the goals of developing credible and germane CE research studies using real world data."[9] We support that goal and have attempted to further it with the development of the Good ReseArch for Comparative Effectiveness (GRACE) checklist, which has been tested for its clarity and ability to distinguish sufficient quality work according to study purpose.

This article describes the development and approaches to validation of an item checklist that can be used to identify observational CE studies sufficiently rigorous in design and execution for decision support. We focused on relatively objective criteria that can be assessed through review of published study reports.

Methods

We drafted the initial checklist from the GRACE principles for observational CE studies, developed in collaboration with the International Society of Pharmacoepidemiolgy.[3] The checklist was fine-tuned for content validity by consultation with experts and extensive literature review, including reports from the Agency for Healthcare Research and Quality on rating the strength of scientific research findings,[16-18] the Grading of Recommendations Assessment, Development and Evaluation process,[19,20] reporting guidelines, and other tools for assessing clinical and observational study quality.[21-26] Senior scientists from academia, industry, and payers were also consulted about item selection and scoring, some of whom also served as expert raters. User instructions and response levels for the refined list of questions were developed by the authors.

Checklist testers were recruited via emails and personal requests and also through the website www.graceprinciples.org. Volunteers (N = 113) from North and South America, Europe, Asia, and Africa conducted a total of 280 assessments of 88 articles. Testers included clinicians, academics, and representatives from industry, health departments, and other nonprofit agencies. They reported a wide range of training and experience with epidemiologic and statistical methods. The construct validity of the checklist was assessed using a variation on the "Extreme Groups" approach[27] by applying the checklist to 3 "validation sets" of observational CE research studies. We compared checklist item responses 3 different ways with external quality ratings, using published articles of observational CE or safety studies: (a) Systematic Review–quality assessment from a published systematic review; (b) Single Expert Review–quality assessment made according to the solicited "expert opinion" of a senior researcher; and (c) Concordant Expert Review–quality assessments from 2 experts for which there was concordance. The first version of the checklist was used for the Systematic Review validation test. It was then fine tuned for subsequent testing in the Single Expert and Concordant Expert Reviews.

In the first test, a sample of articles was drawn from published systematic reviews that listed the articles considered for inclusion, along with their quality assessments (articles listed in Appendix, available in online article).[28-33] Articles were considered "good" if they met quality criteria required for inclusion

in the systematic review and were considered to be of insufficient quality if they were excluded from the review. For testing, authorship was blinded by redaction to avoid biasing quality determinations, and testers were asked not to try to identify the authors through other means. Of 48 articles, 21 were considered "good" and 27 "not good enough"; 172 completed assessments were received from 58 testers, with each article receiving an average of 4 reviews (range, 1-9 reviews).

In the second set of tests (Expert Reviews), the experts received directions explicit to the use of the articles for "decision support" and were asked to decide whether each observational CE study was of sufficient quality to support a formulary decision. Ten senior academic and industry experts were asked to rate 4 or more published observational CE articles as either "sufficient quality to be used to support a formulary decision" or "sufficiently flawed to make interpretation unreliable." The Single Expert Review consisted of 40 articles: 23 that experts rated as sufficient and 17 that were rated as too flawed to be useful for this purpose.

For the third set of tests (Concordant Expert Reviews), 5 experts reviewed 23 of the 40 articles to assess concordance. The articles used for testing are listed in Appendix B (available in online article), and the 14 experts are listed in the acknowledgements (10 participated in the Single Expert Review; 1 of those 10 plus 4 others reviewed articles in the Concordant Expert Review). Fifty-five additional volunteer testers applied the checklist to 2 articles each in this validation, completing a total of 108 assessments, with each article receiving an average of 2.7 reviews (range, 2-7 reviews). One item was dropped after the first round of testing when we learned that none of the articles reviewed stated whether the hypotheses had been specified before the study began. Checklist items were also revised before subsequent testing to improve clarity. In addition, user instructions were clarified after review by 2 authors (Dreyer and Velentgas) to accommodate better evaluation of studies of medical devices and procedures as well as drugs.

Question response levels in the checklist were mapped to dichotomized categories of "sufficient (good enough for decision support)" or "insufficient." Responses that indicated "not enough information in article" were treated as "insufficient," since this lack of information could be viewed as a negative aspect of study quality. Responses of "not applicable" were classified as "sufficient" so that an article would not be rated negatively if a specific question item was not relevant to its objective. Blank responses were treated as missing values. Positive and negative predictive values were estimated for each checklist item to describe how well a reviewer's assessments, using the checklist, compared with an expert's assessment of study quality (in this case, the best available "gold standard" for assessment of study quality). For each article, a single review from a tester, randomly selected from the multiple reviews per article, was compared with the "gold standard." This comparison was done twice to ensure that results were not highly dependent on the random subset selected. Results from both analysis subsets are presented in the Results section. All analyses presented were conducted using SAS 9.2 (Cary, NC).

Results

The GRACE checklist, as modified through this testing process, is shown in Table 1. Questions are grouped into those relating to data and methods, and the guide to scoring reflects clarifications and revisions based on feedback from raters and journal reviewers. Table 2 presents predictive values, comparing testers' assessments of checklist items to experts' overall quality assessments. This comparison was done for 2 sample reviews for each of the 3 validations (6 samples total), stratified by positive predictive value (PPV) and negative predictive value (NPV).

Taken as a whole, the checklist showed better NPV than PPV, with 31 individual items scoring at least 0.67 for NPV versus only 20 items for PPV. A similar trend was evident when looking at both data and methods questions; 20 versus 11 data items scored ≥0.67, and 11 versus 9 methods items NPV and PPV, respectively. Each of the 11 items showed some potential for NPV (using the ≥0.67 criterion), and 9 of the 11 questions also showed some potential for their PPV. The single question that most consistently showed strong NPV and PPV addressed the validity of the primary outcomes (D4, Table 1). For PPV, the other question that most consistently scored relatively high was whether a sensitivity analysis had been conducted (M5, Table 1). The 2 most frequently identifiable predictors of negative quality were the absence of a concurrent comparator group (M2, Table 1) and the lack of adequate details on outcomes (D2, Table 1), followed by not using appropriate clinical outcomes where applicable (D3 and D4, Table 1).

Discussion

The GRACE checklist was designed as an initial evaluation tool to broadly screen the quality of observational CE studies to select those worth in-depth consideration. We focused on 11 checklist elements, 6 relating to data and 5 relating to methods. Using an arbitrarily selected cut-point of 0.67 to indicate relatively strong predictive value, checklist questions about data generally showed better predictive value than questions about methods. Two of the most consistent predictors of quality appropriate for purpose related to (1) valid outcomes and (2) use of concurrent comparators, both factors with important design, analytic, and budgetary ramifications. Our small test of concordance among expert reviewers revealed an unsettling lack of agreement about what "good" looks like through consensus. There was agreement on quality only for 12 of 23 articles (52%) rated by 2 experts—hardly an endorsement for pure reliance on expert assessments.

TABLE 1 GRACE Checklist: Components and Response Guide

Components	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 sources? 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, batch, dose, route, and site of administration, etc.; for devices, 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 was adequately recorded in the data sources (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 was 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 measured objectively rather than subject to clinical judgment (e.g., opinion about whether the patient's condition has improved)?	(+) Yes, clinical outcome was measured objectively (e.g., hospitalization, mortality). (+) Not applicable (primary outcome not clinical, such as 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 PROs [such as 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; billing data were used to assess health resource utilization, etc). (-) No, *or* not enough information in article.
D5. Was the primary outcome measured or identified in an equivalent manner between the treatment/intervention group and the comparison groups?	(+) 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) prior to 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 was impossible for researchers to identify current users of older treatments or when a concurrent comparison group was 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"? (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.

ICD-9-CM = International Classification of Diseases, Ninth Revision, Clinical Modifications; PRO = patient-reported outcome.

TABLE 2 Predictive Values by Item for All Validation Test Sets

	Adequate Treatment	Adequate Outcomes	Objective Outcomes	Valid Outcomes	Similar Outcomes	Covariates Recorded	New Initiators	Concurrent Comparators	Covariates Accounted For	Immortal Time Bias	Sensitivity Analysis
	D1	D2	D3	D4	D5	D6	M1	M2	M3	M4	M5
Positive Predictive Values											
Systematic Review 1											
PPV	0.40	0.49	0.49	0.59	0.47	0.60	0.38	0.46	0.69	—[a]	0.75
N/D	10/25	20/41	21/43	16/27	20/43	12/20	6/16	18/39	11/16	—[a]	9/12
Systematic Review 2											
PPV	0.48	0.48	0.51	0.52	0.48	0.75	0.33	0.48	0.91	—[a]	0.70
N/D	15/31	19/40	20/39	17/33	20/42	9/12	7/21	20/42	10/11	—[a]	7/10
Single Review 1											
PPV	0.57	0.63	0.59	0.70	0.58	0.73	0.56	0.59	0.65	0.59	0.56
N/D	16/28	19/40	20/34	16/23	22/38	11/15	14/25	23/39	13/20	19/32	9/16
Single Review 2											
PPV	0.68	0.62	0.61	0.74	0.58	0.64	0.65	0.58	0.68	0.61	0.69
N/D	19/28	16/26	20/33	14/19	19/33	9/14	13/20	22/38	13/19	19/31	11/16
Concordant Review 1											
PPV	0.67	0.71	0.67	0.83	0.44	1.00	0.44	0.67	0.50	0.63	1.00
N/D	4/6	5/7	6/9	5/6	4/9	1/1	4/9	6/9	2/4	5/8	2/2
Concordant Review 2											
PPV	0.63	0.60	0.55	0.67	0.55	0.50	0.67	0.55	0.50	0.56	0.50
N/D	5/8	6/10	6/11	4/6	6/11	2/4	6/9	6/11	2/4	5/9	1/2
Number rated ≥0.67	2	1	1	4	0	3	1	1	3	0	4
Negative Predictive Values											
Systematic Review 1											
NPV	0.55	0.86	1.00	0.76	0.80	0.68	0.53	0.71	0.71	—[a]	0.67
N/D	12/22	6/7	5/5	16/21	4/5	19/28	17/32	5/7	22/31	—[a]	24/36
Systematic Review 2											
NPV	0.65	0.75	0.88	0.71	0.83	0.66	0.50	1.00	0.69	—[a]	0.62
N/D	11/17	6/8	7/8	10/14	5/6	23/35	13/26	4/4	25/36	—[a]	23/37
Single Review 1											
NPV	0.42	0.67	0.50	0.59	0.50	0.52	0.40	1.00	0.50	0.50	0.39
N/D	5/12	4/6	3/6	10/17	1/2	13/25	6/15	1/1	10/20	4/8	9/23
Single Review 2											
NPV	0.67	0.50	0.57	0.57	0.43	0.44	0.50	0.50	0.55	0.56	0.48
N/D	8/12	6/12	4/7	12/21	3/7	11/25	10/20	1/2	11/20	5/9	11/23
Concordant Review 1											
NPV	0.67	0.67	1.00	1.00	0.33	0.55	0.33	1.00	0.50	0.75	0.56
N/D	2/4	2/3	2/2	2/2	1/3	6/11	1/3	3/3	4/8	3/4	5/9
Concordant Review 2											
NPV	0.75	1.00	1.00	0.67	1.00	0.50	1.00	1.00	0.50	0.67	0.50
N/D	3/4	2/2	1/1	4/6	1/1	4/8	3/3	1/1	4/8	2/3	5/10
Number rated ≥0.67	3	5	4	4	3	1	1	5	2	2	1

[a]*Question not included in Systematic Review.*

D = denominator (total number of articles rated on quality by raters); N = numerator (number of articles in which raters and experts agreed on quality); NPV = negative predictive value; PPV = positive predictive value.

Limitations

Although the current testing and validation efforts did not achieve clear discrimination between studies fit for purpose and those not, we identified a critical but remediable limitation in our approach. By not specifying a specific granular decision for evaluation (e.g., "Is this study of sufficient quality to compare the relative safety of two drugs?") or identifying a single study objective in situations where reports included more than

1 objective (e.g., "Does this study demonstrate greater compliance with once per week vs. daily therapy?"), reviewers were left with too broad an assessment. We believe that future efforts will be more successful if reviewers are asked to focus on a specific objective or question.

The GRACE checklist also does not provide a single quantitative summary score or "pass/fail" result. Our experts counseled that a summary result from the checklist would not be broadly reflective of the numerous considerations that go into assessing the quality of a given study and whether it is sufficient for a specific purpose. Related efforts have concluded that a pass/fail score would require much more tailoring of a checklist to address specific issues and contexts, such as the types of decisions faced by pharmacy, payer, and other health care constituencies and specific therapeutic areas. Nonetheless, we conducted some preliminary analyses using CART software (Salford Systems, San Diego, CA) to create regression trees. Unfortunately, no consistently high-performing combination of checklist items was identified that would correctly classify studies as good or of insufficient quality. Since then, the checklist instructions and scoring have been improved through testing, and additional analyses may be more fruitful. In addition, by addressing the limitation discussed above and specifying the purpose of the review, an overall quantitative assessment may be feasible.

In addition, the GRACE checklist would benefit from further development using different validation sets, improving instructions to raters, and further analysis of results to see if any combination or sequence of item responses has particularly high predictive validity. The articles we selected from systematic reviews, for example, reflected publications that had been examined thoroughly and vetted by a group of experts. However, not all of these articles reflected use of modern methods, particularly as they relate to design and analysis of noninterventional CE studies, because by the time a systematic review had been conducted and published, the articles used were dated. Finding well-accepted standards against which to test checklist items to further refine the distinguishing aspects of quality remains an open question.[34-36]

When considering the GRACE checklist's limitations, it is important to keep in mind what alternative tools exist and their utility for this purpose. The well-recognized STROBE and CONSORT guidelines address how to report study results and were not designed to assess study quality; therefore, they would not be sufficient substitutes.[37] Tools not developed specifically for pharmacoepidemiology are unlikely to include the relevant elements critical for description, assessment of CE, and likelihood of bias.[38] Perhaps most importantly, to our knowledge, none of the other assessment guidelines or standards have been subjected to much, if any, testing. The developers of those tools employed consensus methods, which have face validity, but without evaluation of reliability and discriminant validity, it is uncertain how they would perform in a similar exercise.

Conclusions

Taken as a whole, the GRACE checklist can help as a screening tool to eliminate studies that do not meet the baseline quality requirements for observational studies of comparative effectiveness. We recommend that the GRACE checklist be used as a "first pass" to evaluate how a given study measures against each of the checklist items when applied to a specific study question. Those studies that appear to be fairly sound in design and methods in the context of the study purpose should be examined more closely to evaluate the comparability of the study population to the target population of interest, the appropriateness of the specific medical interventions and comparators for use in the target population, and the likelihood of intractable bias and relevance of the outcomes to patients and health care providers. Studies should also receive further review in the context of available evidence regarding relative risks and benefits and the required threshold for decision support, ideally by those with methodological and content area knowledge.

Despite the drawbacks in the GRACE checklist and other tools, having an agreed upon set of assessment elements, checklists, or score cards is critical for the maturation of the field. Substantial resources will be expended on studies of real-world effectiveness, and if the rigor of these observational assessments cannot be ascertained, then the impact of those studies will be suboptimal. Similarly, agreement on key elements of quality will ensure that budgets are appropriately directed toward those key elements of quality. Given the centrality of this task and the lessons learned from these extensive efforts at validation and user testing, we are optimistic about the potential for improved assessments. We believe that the necessary tools can be produced, enabling diverse types of assessments by people with a wide range of experience and training.

Authors

NANCY A. DREYER, PhD, MPH, is Global Chief of Scientific Affairs and Senior Vice President, and PRISCILLA VELENTGAS, PhD, is Senior Director, Epidemiology, Quintiles Real-World & Late Phase Research, Cambridge, Massachusetts. KIMBERLY WESTRICH, MA, is Director, Health Services Research, and ROBERT DUBOIS, MD, is Chief Science Officer, National Pharmaceutical Council, Washington, DC.

AUTHOR CORRESPONDENCE: Nancy A. Dreyer, PhD, MPH, Senior Vice President, Quintiles Real-World & Late Phase Research, 201 Broadway, Cambridge, MA 02139. Tel.: 617.715.6810; Fax: 617.621.1620; E-mail: nancy.dreyer@quintiles.com.

DISCLOSURES

This research was supported in part by a contract from the National Pharmaceutical Council. The sponsor has provided scientific collaboration and has rights to nonbinding review of manuscripts but does not have the right to choose authors or manuscript topics, nor does it have the right to final approval of the wording of any manuscripts. Dreyer had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Westrich and Dubois are employees of the National Pharmaceutical Council, which is funded by the pharmaceutical industry. This work was supported in part by a contract from the National Pharmaceutical Council.

Dreyer and Velentgas were jointly responsible for the design and conduct of the study, data collection, management, analysis, and interpretation of data, as well as drafting and reviewing the manuscript. Westrich and Dubois provided consultation at every stage of the project and contributed to the design of the study, as well as drafting and reviewing of the manuscript.

ACKNOWLEDGMENTS

Expert raters included Mark Berger (Pfizer), Nick Black (London School of Hygiene and Tropical Medicines), Alan Brookhart (University of North Carolina), Sarah Garner and Tarang Sharma (NICE, UK), Tobias Gerhard (Rutgers), Kathy Lohr and Nancy Berkman (RTI), Newell McElwee (Merck), Peter Neumann (Tufts University), Steve Pearson and Dan Ollendorf (ICER), Brian Sweet (AstraZeneca), and Noel Weiss (University of Washington). A full list of participating expert consultants and raters can be found at www.graceprinciples.org.

The authors also wish to acknowledge the support of April Duddy, MSc, who contributed to data collection and analysis, and Jaclyn Bosco, PhD, and Allison Bryant, MPH, who contributed to analysis.

REFERENCES

1. Greene JA, Podolsky SH. Reform, regulation, and pharmaceuticals—the Kefauver-Harris Amendments at 50. *N Engl J Med*. 2012;367(16):1481-83. Available at: http://www.nejm.org/doi/full/10.1056/NEJMp1210007. Accessed November 14, 2013.

2. Dreyer NA. Tunis SR, Berger M, Ollendorf D, Mattox P, Gliklich R. Why observational studies should be among the tools used in comparative effectiveness research. *Health Aff (Millwood)*. 2010;29(10):1818-25.

3. Dreyer NA, Schneeweiss S, McNeil B, et al. GRACE principles: recognizing high-quality observational studies of comparative effectiveness. *Am J Manag Care*. 2010;16(6):467-71.

4. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed November 14, 2013.

5. Agency for Healthcare Research and Quality. Assessing confounding, the risk of bias and precision of observational studies of interventions or exposures: further development of the RTI Item Bank. Evidence-Based Practice Center. Draft Methods Research Report. May 2012. Available at: http://www.effectivehealthcare.ahrq.gov/ehc/products/414/1272/Observational-studies-refinement_DraftReport_20120927.pdf. Accessed January 8, 2014.

6. Thomas L, Peterson ED. The value of statistical analysis plans in observational research: defining high-quality research from the start. *JAMA*. 2012;308(8):773-74.

7. Luce BR, Drummond MF, Dubois RW, et al. Principles for planning and conducting comparative effectiveness research. *J Comp Eff Res*. 2012;1(5):431-40.

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. Rockville, MD: Agency for Healthcare Research and Quality; January 2013.

9. Willke RJ, Mullins CD. "Ten Commandments" for conducting comparative effectiveness research using "real-world data." *J Manag Care Pharm*. 2011;17 (9 Suppl A):S10-S15. Available at: http://amcp.org/WorkArea/DownloadAsset.aspx?id=13714.

10. 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.

11. 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.

12. 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.

13. Berger ML, Dreyer N, Anderson F, Towse A, Sedrakyan A, Normand SL. Prospective observational studies to assess comparative effectiveness: the ISPOR Good Research Practices Task Force Report. *Value Health*. 2012;15(2):217-30.

14. Appendix D: research questions and PICO(TS): the effective health care program stakeholder guide. July 2011. Agency for Healthcare Research and Quality. Rockville, MD. Available at: http://www.ahrq.gov/research/findings/evidence-based-reports/stakeholderguide/appendixd.html. Accessed November 14, 2013.

15. Patient-Centered Outcomes Research Institute. PCORI methodology standards. December 14, 2012. Available at: http://www.pcori.org/assets/PCORI-Methodology-Standards.pdf. Accessed November 14, 2013.

16. Agency for Healthcare Research and Quality. Rating the strength of scientific research findings. AHRQ Publication No. 02-P022. April 2002. Available at: http://archive.ahrq.gov/clinic/epcsums/strenfact.htm. Accessed November 14, 2013.

17. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions—Agency for Healthcare Research and Quality and the Effective Health-Care Program. *J Clin Epidemiol*. 2010;63(5):513-23.

18. West S, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center under Contract No. 290-97-0011). AHRQ Publication No. 02-E016. April 2002. Rockville, MD: Agency for Healthcare Research and Quality. April 2002. Available at: http://www.thecre.com/pdf/ahrq-system-strength.pdf. Accessed January 8, 2014.

19. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-26. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2335261/. Accessed November 14, 2013.

20. Sharma T, Nyong J, Shaw E. S22—Using and adapting GRADE methodology in an area of low-quality evidence: an example from a national guideline on ablative therapies for the treatment of Barrett's esophagus. *Otolaryngol Head Neck Surg*. 2010;143(1 Suppl 1):S21.

21. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA*. 2000;283(15):2008-12.

22. Wong WC, Cheung CS, Hart GJ. Development of a quality assessment tool for systematic reviews of observational studies (QATSO) of HIV prevalence in men having sex with men and associated risk behaviours. *Emerg Themes Epidemiol*. 2008;17(5):23. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2603000/. Accessed November 14, 2013.

23. Tseng TY, Breau RH, Fesperman SF, Vieweg J, Dahm P. Evaluating the evidence: the methodological and reporting quality of comparative observational studies of surgical interventions in urological publications. *BJU Int*. 2009;103(8):1026-31.

24. Von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol*. 2008;61(4):344-49.

25. Dubois RW, Kindermann SL. Demystifying comparative effectiveness research: a case study learning guide. National Pharmaceutical Council. November 2009. Available at: http://www.npcnow.org/publication/demystifying-comparative-effectiveness-research-case-study-learning-guide. Accessed November 14, 2013.

26. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials*. 1996;17(1):1-12.

27. Streiner DL, Norman GR. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4th ed. New York: Oxford University Press; 2008.

28. McDonagh MS, Peterson K, Carson S, Fu R, Thakurta S. Drug class review: atypical antipsychotic drugs. Final Update 3 Report [Internet]. Portland, OR: Oregon Health and Science University; July 2010.

29. Humphrey LL, Chan BK, Sox HC. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. *Ann Intern Med*. 2002;137(4):273-84.

30. Nelson HD, Humphrey LL, Nygren P, Teutsch SM, Allan JD. Postmenopausal hormone replacement therapy: scientific review. *JAMA*. 2002;288(7):872-81.

31. Ip S, Bonis P, Tatsioni A, et al. Comparative effectiveness of management strategies for gastroesophageal reflux disease. Comparative Effectiveness Review No. 1. (Prepared by Tufts-New England Medical Center Evidence-based Practice Center under Contract No. 290-02-0022). Rockville, MD: Agency for Healthcare Research and Quality. AHRQ Publication No. 06-EHC003-1 December 2005. Available at: http://effectivehealthcare.ahrq.gov/ehc/products/1/43/GERD%20Final%20Report.pdf. Accessed November 14, 2013.

32. Yank V, Tuohy CV, Logan AC, et al. Comparative effectiveness of in-hospital use of recombinant factor VIIa for off-label indications vs. usual care. Comparative Effectiveness Review No. 21. (Prepared by Stanford-UCSF Evidence-based Practice Center under Contract No. #290-02-0017). Rockville, MD: Agency for Healthcare Research and Quality. AHRQ Publication No. 10-EHC030-EF May 2010. Available at: http://www.effectivehealthcare.ahrq.gov/ehc/products/20/450/Final%20Report_CER21_Factor7.pdf. Accessed November 14, 2013.

33. Donahue KE, Gartlehner G, Jonas DE, et al. Comparative effectiveness of drug therapy for rheumatoid arthritis and psoriatic arthritis in adults. Comparative Effectiveness Review No. 11. (Prepared by RTI-University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016). Rockville, MD: Agency for Healthcare Research and Quality. November 2007. Available at: http://www.effectivehealthcare.ahrq.gov/ehc/products/14/68/RheumArthritisFinal.pdf. Accessed November 14, 2013.

34. Lang S, Kleijnen J. Quality assessment tools for observational studies: lack of consensus. *Int J Evid Based Healthc*. 2010;8(4):247.

35. Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int J Epidemiol*. 2007;36(3):666-76.

36. Dreyer NA. Using observational studies for comparative effectiveness: finding quality with GRACE. *J Comp Eff Res*. 2013;2(5):413-18.

37. Da Costa BR, Cevallos M, Altman DG, Rutjes AW, Egger M. Uses and misuses of the STROBE statement: bibliographic study. *BMJ Open*. 2011;1(1):e000048. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3191404/. Accessed November 14, 2013.

38. Neyarapally GA, Hammad TA, Pinheiro SP, Iyasu S. Review of quality assessment tools for the evaluation of pharmacopidemiological safety studies. *BMJ Open*. 2012;2(5):e001362. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3467649/. Accessed November 14, 2013.

APPENDIX A Articles Tested from Systematic Reviews

Article Reference	Quality for Purpose
Advokat C, Dixon D, Schneider J, Comaty JE. Comparison of risperidone and olanzapine as used under "real-world" conditions in a state psychiatric hospital. *Prog Neuropsychopharmacol Biol Psychiatry.* 2004;28(3):487-95.	Poor
Bond GR, Kim HW, Meyer PS, et al. Response to vocational rehabilitation during treatment with first- or second-generation antipsychotics. *Psychiatr Serv.* 2004;55(1):59-66.	Poor
de Haan L, Beuk N, Hoogenboom B, Dingemans P, Linszen D. Obsessive-compulsive symptoms during treatment with olanzapine and risperidone: a prospective study of 113 patients with recent-onset schizophrenia or related disorders. *J Clin Psychiatry.* 2002;63(2):104-107.	Poor
Dolder CR, Lacro JP, Dunn LB, Jeste DV. Antipsychotic medication adherence: is there a difference between typical and atypical agents? *Am J Psychiatry.* 2002;159(1):103-108.	Poor
Garcia-Cabeza I, Gomez JC, Sacristan JA, Edgell E, Gonzalez de Chavez M. Subjective response to antipsychotic treatment and compliance in schizophrenia. A naturalistic study comparing olanzapine, risperidone and haloperidol (EFESO Study). *BMC Psychiatry.* 2001;1:7.	Poor
Joyce AT, Harrison DJ, Loebel AD, Ollendorf DA. Impact of atypical antipsychotics on outcomes of care in schizophrenia. *Am J Manag Care.* 2005;11(8 Suppl):S254-61.	Poor
Kasper S, Jones M, Duchesne I. Risperidone olanzapine drug outcomes studies in schizophrenia (RODOS): health economic results of an international naturalistic study. *Int Clin Psychopharmacol.* 2001;16(4):189-96.	Poor
Koro CE, Fedder DO, L'Italien GJ, et al. An assessment of the independent effects of olanzapine and risperidone exposure on the risk of hyperlipidemia in schizophrenic patients. *Arch Gen Psychiatry.* 2002;59(11):1021-26.	Poor
Meyer JM. A retrospective comparison of weight, lipid, and glucose changes between risperidone- and olanzapine-treated inpatients: metabolic outcomes after 1 year. *J Clin Psychiatry.* 2002;63(5):425-33.	Poor
Pelagotti F, Santarlasci B, Vacca F, Trippoli S, Messori A. Dropout rates with olanzapine or risperidone: a multi-centre observational study. *Eur J Clin Pharmacol.* 2004;59(12):905-09.	Poor
Soholm B, Lublin H. Long-term effectiveness of risperidone and olanzapine in resistant or intolerant schizophrenic patients. A mirror study. *Acta Psychiatr Scand.* 2003;107(5):344-50.	Poor
Voruganti L, Cortese L, Owyeumi L, et al. Switching from conventional to novel antipsychotic drugs: results of a prospective naturalistic study. *Schizophr Res.* 2002;57(2-3):201-08.	Poor
Askling J, Fored CM, Brandt L, et al. Risk and case characteristics of tuberculosis in rheumatoid arthritis associated with tumor necrosis factor antagonists in Sweden. *Arthritis Rheum.* 2005;52(7):1986-92.	Good
Brody DL, Aiyagari V, Shackleford AM, Diringer MN. Use of recombinant factor VIIa in patients with warfarin-associated intracranial hemorrhage. *Neurocrit Care.* 2005;2(3):263-67.	Poor
Fernando HC, Schauer PR, Rosenblatt M, et al. Quality of life after antireflux surgery compared with nonoperative management for severe gastroesophageal reflux disease. *J Am Coll Surg.* 2002;194(1):23-27.	Poor
Flendrie M, Creemers MC, Welsing PM, den Broeder AA, van Riel PL. Survival during treatment with tumour necrosis factor blocking agents in rheumatoid arthritis. *Ann Rheum Dis.* 2003;62(Suppl 2):ii30-33.	Poor
Gelsomino S, Lorusso R, Romagnoli S, et al. Treatment of refractory bleeding after cardiac operations with low-dose recombinant activated factor VII (NovoSeven): a propensity score analysis. *Eur J Cardiothorac Surg.* 2008;33(1):64-71.	Good
Hallevi H, Gonzales NR, Barreto AD, et al. The effect of activated factor VII for intracerebral hemorrhage beyond 3 hours versus within 3 hours. *Stroke.* 2008;39(2):473-75.	Poor
Harrison TD, Laskosky J, Jazaeri O, et al. "Low-dose" recombinant activated factor VII results in less blood and blood product use in traumatic hemorrhage. *J Trauma.* 2005;59(1):150-54.	Poor
Holzman MD, Mitchel EF, Ray WA, Smalley WE. Use of health care resources among medically and surgically treated patients with gastroesophageal reflux disease: a population-based study. *J Am Coll Surg.* 2001;192(1):17-24.	Poor
Hyrich KL, Symmons DP, Watson KD, et al. Comparison of the response to infliximab or etanercept monotherapy with the response to cotherapy with methotrexate or another disease-modifying antirheumatic drug in patients with rheumatoid arthritis: results from the British Society for Rheumatology Biologics Register. *Arthritis Rheum.* 2006;54(6):1786-94.	Good
Isolauri J, Luostarinen M, Viljakka M, et al. Long-term comparison of antireflux surgery versus conservative therapy for reflux esophagitis. *Ann Surg.* 1997;225(3):295-99.	Poor
Kalicinski P, Markiewicz M, Kaminski A, et al. Single pretransplant bolus of recombinant activated factor VII ameliorates influence of risk factors for blood loss during orthotopic liver transplantation. *Pediatr Transplant.* 2005;9(3):299-304.	Poor
Khaitan L, Ray WA, Holzman MD, et al. Health care utilization after medical and surgical therapy for gastroesophageal reflux disease: a population-based study, 1996 to 2000. *Arch Surg.* 2003;138(12):1356-61.	Poor
Niemann CU, Behrends M, Quan D, et al. Recombinant factor VIIa reduces transfusion requirements in liver transplant patients with high MELD scores. *Transfus Med.* 2006;16(2):93-100.	Poor
Setoguchi S, Solomon DH, Weinblatt ME, et al. Tumor necrosis factor alpha antagonist use and cancer in patients with rheumatoid arthritis. *Arthritis Rheum.* 2006;54(9):2757-64.	Good
Tran T, Spechler SJ, Richardson P, et al. Fundoplication and the risk of esophageal cancer in gastroesophageal reflux disease: a Veterans Affairs cohort study. *Am J Gastroenterol.* 2005;100(5):1002-08.	Poor
Wetscher GJ, Gadenstaetter M, Klingler PJ, et al. Efficacy of medical therapy and antireflux surgery to prevent Barrett's metaplasia in patients with gastroesophageal reflux disease. *Ann Surg.* 2001;234(5):627-32.	Poor

APPENDIX A Articles Tested from Systematic Reviews *(continued)*

Article Reference	Quality for Purpose
Zink A, Listing J, Kary S, et al. Treatment continuation in patients receiving biological agents or conventional DMARD therapy. *Ann Rheum Dis*. 2005;64(9):1274-79.	Good
Bush TL, Barrett-Connor E, Cowan LD, et al. Cardiovascular mortality and noncontraceptive use of estrogen in women: results from the Lipid Research Clinics Program Follow-up Study. *Circulation*. 1987;75(6):1102-09.	Good
Cauley JA, Seeley DG, Browner WS, et al. Estrogen replacement therapy and mortality among older women. The study of osteoporotic fractures. *Arch Intern Med*. 1997;157(19):2181-87.	Good
Ettinger B, Friedman GD, Bush T, et al. Reduced mortality associated with long-term postmenopausal estrogen therapy. *Obstet Gynecol*. 1996;87(1):6-12.	Good
Grodstein F, Manson JE, Colditz GA, et al. A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. *Ann Intern Med*. 2000;133(12):933-41.	Good
Grodstein F, Stampfer MJ, Colditz GA, et al. Postmenopausal hormone therapy and mortality. *N Engl J Med*. 1997;336(25):1769-75.	Good
Henderson BE, Paganini-Hill A, Ross RK. Decreased mortality in users of estrogen replacement therapy. *Arch Intern Med*. 1991;151(1):75-78.	Poor
Lafferty FW, Fiske ME. Postmenopausal estrogen replacement: a long-term cohort study. *Am J Med*. 1994;97(1):66-77.	Poor
Persson I, Yuen J, Bergkvist L, et al. Cancer incidence and mortality in women receiving estrogen and estrogen-progestin replacement therapy—long-term follow-up of a Swedish cohort. *Int J Cancer*. 1996;67(3):327-32.	Good
Sidney S, Petitti DB, Quesenberry CP, Jr. Myocardial infarction and the use of estrogen and estrogen-progestogen in postmenopausal women. *Ann Intern Med*. 1997;127(7):501-08.	Good
Varas-Lorenzo C, Garcia-Rodriguez LA, Perez-Gutthann S, et al. Hormone replacement therapy and incidence of acute myocardial infarction. A population-based nested case-control study. *Circulation*. 2000;101(22):2572-78.	Good
Cauley JA, Seeley DG, Ensrud K, et al. Estrogen replacement therapy and fractures in older women. Study of Osteoporotic Fractures Research Group. *Ann Intern Med*. 1995;122(1):9-16.	Good
Colditz GA, Hankinson SE, Hunter DJ, et al. The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. *N Engl J Med*. 1995;332(24):1589-93.	Good
Folsom AR, Mink PJ, Sellers TA, Hong CP, Zheng W, Potter JD. Hormonal replacement therapy and morbidity and mortality in a prospective study of postmenopausal women. *Am J Public Health*. 1995;85(8 Pt 1):1128-32.	Good
Grodstein F, Stampfer MJ, Falkeborn M, et al. Postmenopausal hormone therapy and risk of cardiovascular disease and hip fracture in a cohort of Swedish women. *Epidemiology*. 1999;10(5):476-80.	Good
Paganini-Hill A, Henderson VW. Estrogen replacement therapy and risk of Alzheimer disease. *Arch Intern Med*. 1996;156(19):2213-17.	Poor
Schairer C, Gail M, Byrne C, et al. Estrogen replacement therapy and breast cancer survival in a large screening study. *J Natl Cancer Inst*. 1999;91(3):264-70.	Good
Sellers TA, Mink PJ, Cerhan JR, et al. The role of hormone replacement therapy in the risk for breast cancer and total mortality in women with a family history of breast cancer. *Ann Intern Med*. 1997;127(11):973-80.	Good
Acute myocardial infarction and combined oral contraceptives: results of an international multicentre case-control study. WHO Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception. *Lancet*. 1997;349(9060):1202-09.	Good
Willis DB, Calle EE, Miracle-McMahill HL, et al. Estrogen replacement therapy and risk of fatal breast cancer in a prospective cohort of postmenopausal women in the United States. *Cancer Causes Control*. 1996;7(4):449-57.	Good

APPENDIX B Articles Tested from Published Studies with Quality Assessment by Experts

Article Reference	Quality	
	Single Expert Review	Concordant Review
Solomon DH, Massarotti E, Garg R, et al. Association between disease-modifying antirheumatic drugs and diabetes risk in patients with rheumatoid arthritis and psoriasis. *JAMA*. 2011;305(24):2525-31.	Poor	Good
Ray WA, Chung CP, Stein CM, et al. Risk of peptic ulcer hospitalizations in users of NSAIDs with gastroprotective cotherapy versus coxibs. *Gastroenterology*. 2007;133(3):790-98.	Good	Good
Kim SY, Schneeweiss S, Katz JN, et al. Oral bisphosphonates and risk of subtrochanteric or diaphyseal femur fractures in a population-based cohort. *J Bone Miner Res*. 2011;26(5):993-1001.	Good	Not reviewed
Srivastava R, Downey EC, O'Gorman M, et al. Impact of fundoplication versus gastrojejunal feeding tubes on mortality and in preventing aspiration pneumonia in young children with neurologic impairment who have gastroesophageal reflux disease. *Pediatrics*. 2009;123(1):338-45.	Good	Not reviewed
Shishehbor MH, Hawi R, Singh IM, et al. Drug-eluting versus bare-metal stents for treating saphenous vein grafts. *Am Heart J*. 2009;158(4):637-43.	Good	Not reviewed
Silverman SL, Watts NB, Delmas PD, et al. Effectiveness of bisphosphonates on nonvertebral and hip fractures in the first year of therapy: the risedronate and alendronate (REAL) cohort study. *Osteoporos Int*. 2007;18(1):25-34.	Good	Not reviewed
Applegate RJ, Sacrinty MT, Kutcher MA, et al. Comparison of drug-eluting versus bare metal stents on later frequency of acute myocardial infarction and death. *Am J Cardiol*. 2007;99(3):333-38.	Poor	Poor
Mayor M, Malik AZ, Minor RJ Jr, et al. One-year outcomes from the TAXUS express stent versus cypher stent. *Am J Cardiol*. 2009;103(7):930-36.	Good	Good
Patel MR, Pfisterer ME, Betriu A, et al. Comparison of six-month outcomes for primary percutaneous revascularization for acute myocardial infarction with drug-eluting versus bare metal stents (from the APEX-AMI study). *Am J Cardiol*. 2009;103(2):181-86.	Good	Not reviewed
Perdue DG, Freeman ML, DiSario JA, et al; ERCP Outcome Study ERCOST Group. Plastic versus self-expanding metallic stents for malignant hilar biliary obstruction: a prospective multicenter observational cohort study. *J Clin Gastroenterol*. 2008;42(9):1040-46.	Good	Poor
Kamat SA, Gandhi SK, Davidson M. Comparative effectiveness of rosuvastatin versus other statin therapies in patients at increased risk of failure to achieve low-density lipoprotein goals. *Curr Med Res Opin*. 2007;23(5):1121-30.	Poor	Not reviewed
Wu E, Greenberg PE, Yang E, et al. Comparison of escitalopram versus citalopram for the treatment of major depressive disorder in a geriatric population. *Curr Med Res Opin*. 2008;24(9):2587-95.	Poor	Good
Friedman HS, Rajagopalan S, Barnes JP, et al. Combination therapy with ezetimibe/simvastatin versus statin monotherapy for low-density lipoprotein cholesterol reduction and goal attainment in a real-world clinical setting. *Clin Ther*. 2011;33(2):212-24.	Poor	Good
Kubiak DW, Bryar JM, McDonnell AM, et al. Evaluation of caspofungin or micafungin as empiric antifungal therapy in adult patients with persistent febrile neutropenia: a retrospective, observational, sequential cohort analysis. *Clin Ther*. 2010;32(4):637-48.	Poor	Not reviewed
Ramanath VS, Brown JR, Malenka DJ, et al; Dartmouth Dynamic Registry Investigators. Outcomes of diabetics receiving bare-metal stents versus drug-eluting stents. *Catheter Cardiovasc Interv*. 2010;76(4):473-81.	Good	Good
Delea TE, Taneja C, Moynahan A, et al. Valsartan versus lisinopril or extended-release metoprolol in preventing cardiovascular and renal events in patients with hypertension. *Am J Health Syst Pharm*. 2007;64(11):1187-96.	Poor	Poor
Platts-Mills TF, Campagne D, Chinnock B, et al. A comparison of GlideScope video laryngoscopy versus direct laryngoscopy intubation in the emergency department. *Acad Emerg Med*. 2009;16(9):866-71.	Poor	Not reviewed
Diaz-Guzman E, Mireles-Cabodevila E, Heresi GA, et al. A comparison of methohexital versus etomidate for endotracheal intubation of critically ill patients. *Am J Crit Care*. 2010;19(1):48-54.	Good	Poor
Konstance RP, Eisenstein EL, Anstrom KJ, et al. Outcomes of second revascularization procedures after stent implantation. *J Med Syst*. 2008;32(2):177-86.	Good	Poor
Hannan EL, Racz M, Walford G, Holmes DR, et al. Drug-eluting versus bare-metal stents in the treatment of patients with ST-segment elevation myocardial infarction. *JACC Cardiovasc Interv*. 2008;1(2):129-35.	Good	Poor
Lee MS, Tarantini G, Xhaxho J, et al. Sirolimus- versus paclitaxel-eluting stents for the treatment of cardiac allograft vasculopathy. *JACC Cardiovasc Interv*. 2010;3(4):378-82.	Poor	Not reviewed
Shorr AF, Sarnes MW, Peeples PJ, et al. Comparison of cost, effectiveness, and safety of injectable anticoagulants used for thromboprophylaxis after orthopedic surgery. *Am J Health Syst Pharm*. 2007;64(22):2349-55.	Good	Good
Allen RH, Kumar D, Fitzmaurice G, et al. Pain management of first-trimester surgical abortion: effects of selection of local anesthesia with and without lorazepam or intravenous sedation. *Contraception*. 2006;74(5):407-13.	Good	Poor
Carragee EJ, Spinnickie AO, Alamin TF, et al. A prospective controlled study of limited versus subtotal posterior discectomy: short-term outcomes in patients with herniated lumbar intervertebral discs and large posterior anular defect. *Spine (Phila Pa 1976)*. 2006;31(6):653-57.	Poor	Poor
Chu WW, Kuchulakanti PK, Torguson R, et al. Comparison of clinical outcomes of overlapping sirolimus- versus paclitaxel-eluting stents in patients undergoing percutaneous coronary intervention. *Am J Cardiol*. 2006;98(12):1563-66.	Poor	Not reviewed
Cordera F, Long KH, Nagorney DM, et al. Open versus laparoscopic splenectomy for idiopathic thrombocytopenic purpura: clinical and economic analysis. *Surgery*. 200;134(1):45-52.	Poor	Poor

APPENDIX B Articles Tested from Published Studies with Quality Assessment by Experts *(continued)*

Article Reference	Quality	
	Single Expert Review	Concordant Review
Hecht HS, Harman SM. Comparison of the effects of atorvastatin versus simvastatin on subclinical atherosclerosis in primary prevention as determined by electronbeam tomography. *Am J Cardiol.* 2003;91(1):42-45.	Poor	Poor
Klemme WR, Owens BD, Dhawan A, et al. Lumbar sagittal contour after posterior interbody fusion: threaded devices alone versus vertical cages plus posterior instrumentation. *Spine (Phila Pa 1976).* 2001;26(5):534-37.	Poor	Not reviewed
Konstantinou K, Baddam K, Lanka A, et al. Cefepime versus ceftazidime for treatment of pneumonia. *J Int Med Res.* 2004;32(1):84-93.	Poor	Poor
Lazarus HM, Pérez WS, Klein JP, et al. Autotransplantation versus HLA-matched unrelated donor transplantation for acute myeloid leukaemia: a retrospective analysis from the Center for International Blood and Marrow Transplant Research. *Br J Haematol.* 2006;132(6):755-69.	Good	Not reviewed
London MJ, Moritz TE, Henderson WG, et al; Participants of the Veterans Affairs Cooperative Study Group on Processes, Structures, and Outcomes of Care in Cardiac Surgery. Standard versus fiberoptic pulmonary artery catheterization for cardiac surgery in the Department of Veterans Affairs: a prospective, observational, multicenter analysis. *Anesthesiology.* 2002;96(4):860-70.	Good	Not reviewed
Rha SW, Kuchulakanti PK, Pakala R, et al. Bivalirudin versus heparin as an antithrombotic agent in patients who undergo percutaneous saphenous vein graft intervention with a distal protection device. *Am J Cardiol.* 2005;96(1):67-70.	Good	Poor
Schneeweiss S, Solomon DH, Wang PS, et al. Simultaneous assessment of short-term gastrointestinal benefits and cardiovascular risks of selective cyclooxygenase 2 inhibitors and nonselective nonsteroidal antiinflammatory drugs: an instrumental variable analysis. *Arthritis Rheum.* 2006;54(11):3390-98.	Good	Not reviewed
Solomon DH, Avorn J, Katz JN, et al. Immunosuppressive medications and hospitalization for cardiovascular events in patients with rheumatoid arthritis. *Arthritis Rheum.* 2006;54(12):3790-98.	Poor	Not reviewed
Villareal RP, Lee VV, Elayda MA, et al. Coronary artery bypass surgery versus coronary stenting: risk-adjusted survival rates in 5,619 patients. *Tex Heart Inst J.* 2002;29(1):3-9.	Good	Good
Volicer L, Lane P, Panke J, et al. Management of constipation in residents with dementia: sorbitol effectiveness and cost. *J Am Med Dir Assoc.* 2005;6(3 Suppl):S32-34.	Good	Not reviewed
Weinstein JN, Lurie JD, Tosteson TD, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT) observational cohort. *JAMA.* 2006;296(20):2451-59.	Good	Poor
Wolfe F, Michaud K, Stephenson B, et al. Toward a definition and method of assessment of treatment failure and treatment effectiveness: the case of leflunomide versus methotrexate. *J Rheumatol.* 2003;30(8):1725-32.	Good	Poor
Zimmerman S, Hawkes WG, Hudson JI, et al. Outcomes of surgical management of total HIP replacement in patients aged 65 years and older: cemented versus cementless femoral components and lateral or anterolateral versus posterior anatomical approach. *J Orthop Res.* 2002;20(2):182-91.	Good	Good