

The Predictive Validity of Quality of Evidence Grades for the Stability of Effect Estimates was Low: A Meta-Epidemiological Study

Prepared for:

Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

This information is distributed solely for the purposes of predissemination peer review. It has not been formally disseminated by the Agency for Healthcare Research and Quality. The findings are subject to change based on the literature identified in the interim and peer-review/public comments and should not be referenced as definitive. It does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services (AHRQ) determination or policy.

Contract No.

Prepared by:

<Prepared by>
<City, State>

Investigators:

<Author, Degrees>

AHRQ Publication No. xx-EHCxxx

<Month Year>

This report is based on work conducted by an Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. <xxx-xxxx-xxxx>). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help clinicians, employers, policymakers, researchers, and others make informed decisions about the provision and research of health care services. This report is intended as a reference and not as a substitute for clinical judgment or good scientific practices.

This report may be, in whole or in part, as the basis for research design or funding opportunity announcements. AHRQ or the U.S. Department of Health and Human Services endorsement of such derivative products or actions may not be stated or implied.

This work was funded by the Agency for Healthcare Research and Quality under contract number <Contract_Number> to <EPC> and by internal funds provided by <NAME>. The authors of this report are responsible for its content. Statements in this manuscript should not be construed as endorsement by the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services.

Suggested citation:

<Authors>. The Predictive Validity of Quality of Evidence Grades for the Stability of Effect Estimates was Low: A Meta-Epidemiological Study. (Prepared by the <EPC> Evidence-based Practice Center under Contract No. <Contract_Number>.) AHRQ Publication No. <xx-EHCxxx>. Rockville, MD: Agency for Healthcare Research and Quality. <Month Year>. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-Based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC Program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.gov.

Richard G. Kronick, Ph.D.
Director
Agency for Healthcare Research and Quality

David Meyers, M.D.
Acting Director, Center for Evidence and
Practice Improvement
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director, EPC Program
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Christine Chang, M.D., M.P.H.
Task Order Officer
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Acknowledgments

To be added after peer review

Peer Reviewers

Prior to publication of the white paper, we sought input from independent Peer Reviewers without financial conflicts of interest. However, the conclusions and synthesis of the scientific literature presented in this report does not necessarily represent the views of individual reviewers.

Peer Reviewers must disclose any financial conflicts of interest greater than \$10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential non-financial conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any potential non-financial conflicts of interest identified.

The list of Peer Reviewers follows:

To be added after peer review

Abstract

Objective

We sought to determine the predictive validity of the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach by examining how reliably GRADE can predict the likelihood that treatment effects remain stable as new studies emerge.

Study Design and Setting

Based on 37 Cochrane reports with outcomes graded as high quality of evidence (QOE), we prepared 160 documents representing different levels of QOE. We randomly assigned these documents to professional systematic reviewers from seven academic centers in Austria, Canada, and the United States who dually graded the QOE using guidance for the U.S. Evidence-based Practice Centers. We determined the proportion of effect estimates that remained stable as new studies are added to the evidence base and linked the observed proportions from our sample with the expected proportions for each grade of QOE from an international survey. To determine the predictive validity we used the Hosmer-Lemeshow test to assess calibration and the C (concordance)- index to assess discrimination.

Results

Overall, the predictive validity of GRADE for the stability of effect estimates was limited. Except for moderate QOE, the expected and observed proportions of stable effect estimates differed considerably. Estimates graded as high QOE were less likely to remain stable than expected by producers and users of systematic reviews. By contrast, estimates graded as low or insufficient (very low) QOE were substantially more likely to remain stable than expected. In this sample, GRADE could not reliably predict the likelihood that individual bodies of evidence remain stable as new evidence becomes available. Depending on the definition used, C-indices ranged between 0.56 (95% CI 0.47-0.66) and 0.58 (95% CI 0.50-0.67) indicating a low discriminatory ability of GRADE.

Conclusion

The limited predictive validity of the GRADE approach seems to reflect a mismatch between expected and observed changes in treatment effects as bodies of evidence advance from insufficient (very low) to high QOE. In addition, many low or insufficient (very low) grades appear to be too strict.

Introduction

Despite the enormous amount of new information that medical research generates every year, uncertainty plays a major role in health care decisionmaking. The challenging task for clinical and health policy decisionmakers is to balance considerations about evidence, values, preferences, and resources, all of which are frequently fraught with uncertainty and conflicting perspectives.¹

GRADE (Grading of Recommendations Assessment, Development and Evaluation) has evolved as a widely used approach to communicate certainties and uncertainties in systematic reviews to readers and other stakeholders.^{2,3} GRADE uses information about risk of bias, imprecision, inconsistency, indirectness, and reporting bias to categorize the degree of uncertainty concerning the correctness of findings into four grades of quality of evidence (QOE).

Some organizations—such as the Evidence-based Practice Center (EPC) program of the U.S. Agency for Healthcare Research and Quality (AHRQ)—have made small adaptations to the GRADE system to meet their specific needs.^{4,5} Guidance for EPCs defines QOE (which they refer to as “strength of evidence”) as the degree of confidence that estimates are close to the true effect and the likelihood that findings will remain stable over time (i. e., the likelihood that future studies will not have an important impact on the estimate of an effect).⁴ In this manuscript, we refer to both approaches collectively as the GRADE approach. Table 1 summarizes the EPC definitions of the four levels of QOE.

Decisionmakers who rely on the GRADE approach assume that estimates of effect that are graded as high QOE are “close to the true effect” and, therefore, will remain stable as new evidence emerges. By contrast, decisionmakers can interpret effect estimates that are graded as low QOE as quite likely to change as new evidence accrues. In a recent international survey, we determined that producers and users of systematic reviews associated each grade of QOE with a distinct likelihood that estimates of effect will remain stable as new evidence emerges (see Table 1).⁶

Table 1. Definitions of grades of quality of evidence of the EPC guidance

Grade	Definition	Expected proportions of stable effect estimates ^a
High	We are very confident that the estimate of effect lies close to the true effect for this outcome. The body of evidence has few or no deficiencies. We believe that the findings are stable, i.e., another study would not change the conclusions.	86% to 100%
Moderate	We are moderately confident that the estimate of effect lies close to the true effect for this outcome. The body of evidence has some deficiencies. We believe that the findings are likely to be stable, but some doubt remains.	61% to 85%
Low	We have limited confidence that the estimate of effect lies close to the true effect for this outcome. The body of evidence has major or numerous deficiencies (or both). We believe that additional evidence is needed before concluding either that the findings are stable or that the estimate of effect is close to the true effect.	34% to 60%
Insufficient (very low) ^b	We have no evidence, we are unable to estimate an effect, or we have no confidence in the estimate of effect for this outcome. No evidence is available or the body of evidence has unacceptable deficiencies, precluding reaching a conclusion.	0% to 33%

^a Expected proportions are based on an international survey of producers and users of systematic reviews⁶

^b The AHRQ category of insufficient is similar to the GRADE category of very low; *insufficient*, however, also includes outcomes without evidence. For the purpose of this study we did not consider situations without any evidence

To date, the predictive validity of the GRADE approach concerning the stability of effect estimates has not been tested. Predictive validity refers, in general terms, to the degree to which a score (e.g., such as the grades cited in Table 1) predicts an outcome on a criterion measure.⁷ For GRADE, predictive validity refers to the degree to which this approach, and specifically different QOE grades, reliably predicts the stability of an estimate of effect because it is close to the true effect.

A true effect can be viewed as the effect size that we would observe if a study had an infinitely large sample size (and thus no sampling error).⁸ Realistically, however, a true treatment effect can rarely be determined and utilized as a reference standard. For that reason, here we equate true effect with stability of effect as new studies emerge, a concept that can be measured. Given accurate predictive validity, a rating of “high QOE” would reliably predict that future studies will have a minor impact on the estimate of effect of a given outcome. Likewise, a rating of “low QOE” would reliably predict a high likelihood that future studies will have a substantial impact on the direction or magnitude of the estimate of effect of a given outcome.

The objective of our study was to determine the predictive validity of the GRADE approach based on a diverse sample of interventions. That is, we examined how reliably GRADE can predict the likelihood that treatment effects remain stable.

Methods

We used a meta-epidemiological approach based on large, systematically appraised bodies of evidence that authors of Cochrane reports had graded as high QOE. We used effect estimates of such bodies of evidence as reference points because a grade of high QOE implies that investigators were very confident that the estimate of effect is close to the truth and that new studies are unlikely to change conclusions. The basic assumption for our study was that these bodies of evidence had been graded correctly and can serve as a “gold standard” to determine the stability of effect estimates.

Assembling Empirical Data

We searched the Cochrane Library from 2010 onward to find Cochrane reports that: (1) include a body of evidence of more than eight randomized controlled trials (RCTs) on therapeutic interventions that had been graded as high QOE; (2) present meta-analytic outcomes that were reported as relative risks or odds ratios for binary outcomes or as weighted mean differences or standardized mean differences (SMDs) for continuous outcomes; and (3) provide data to reproduce the meta-analyses. We chose a threshold of eight RCTs so that we had enough studies to meta-analyze subsections of these bodies of evidence.

Overall, we drew information from 37 Cochrane reports on 50 bodies of evidence (Table 1).

Table 1. Cochrane reports and characteristics of bodies of evidence used to prepare summary documents

Cochrane Report	Intervention and Outcome	Number of Participants	Effect Estimate (Confidence Interval)
Amato et al., 2010	Benzodiazepines and adverse events	431	RR: 1.37 (0.79-2.38)
Amato et al., 2010	Benzodiazepines and dropouts	839	RR: 1.1 (0.75-1.62)
Amato et al., 2011	Psychosocial maintenance intervention and retention in treatment	3,050	RR: 1.03 (0.99-1.07)
Amato et al., 2013	Tapered methadone and completion of treatment	1,309	RR: 1.06 (0.96-1.18)
Buchleitner et al., 2012	Perioperative glycaemic control and mortality	1,224	RR: 1.19 (0.89-1.58)
Chauhan et al., 2014	Long acting beta agonists and exacerbations	5,494	RR: 0.89 (0.78-1.02)
Chin et al., 2013	Infraclavicular block and adequate surgical anaesthesia	1,011	RR: 1.02 (0.95-1.1)
Chin et al., 2013	Infraclavicular block and tourniquet pain	556	RR: 0.75 (0.55-1.03)
Chin et al., 2013	Infraclavicular block and need for supplemental local anesthetic blocks or systemic analgesia	821	RR: 0.92 (0.61-1.4)
Chong et al., 2013	Phosphodiesterase-4- inhibitors and exacerbations	4,828	RR: 0.83 (0.78-0.88)
Chong et al., 2013	Phosphodiesterase-4- inhibitors and gastrointestinal side effects	5,842	RR: 2.7 (2.24-3.25)
Clifford et al., 2012	Autologous adult stem cells and left- ventricular ejection fraction	839	SMD: 0.27 (-0.01-0.54)
Feagan et al., 2012	Oral 5-aminosalicylic acid and failure to maintain remission	1,598	RR: 1.12 (0.98-1.28)
Fernandes et al., 2013	Systemic or inhaled glucocorticoids and rate of hospital admission	1,717	RR: 0.92 (0.79-1.08)
Fernandes et al., 2013	Systemic or inhaled glucocorticoids and length of hospital stay	614	SMD: -0.14 (-0.37-0.09)
Gafter et al., 2012	Antibiotic prophylaxis and mortality	219	RR: 0.69 (0.56-0.86)
Gowing et al., 2009	Buprenorphine and completion of withdrawal treatment	409	RR: 1.69 (1.35-2.1)
Griffiths et al., 2013	Inhaled anticholinergic drugs and hospital admission	1,967	RR: 0.74 (0.64-0.85)

Table 1. Cochrane reports and characteristics of bodies of evidence used to prepare summary documents (continued)

Cochrane Report	Intervention and Outcome	Number of Participants	Effect Estimate (Confidence Interval)
Gurion et al., 2012	Colony stimulating factors and mortality	3,017	RR: 1.01 (0.96-1.06)
Hauser et al., 2013	Serotonin and noradrenaline reuptake inhibitors and 50% pain reduction in fibromyalgia	1,677	RR: 1.48 (1.34-1.63)
Hauser et al., 2013	Serotonin and noradrenaline reuptake inhibitors and withdrawals due to adverse events	1,734	RR: 1.84 (1.52-2.22)
Hemmingsen et al., 2013	Intensive glycaemic control and hypoglycaemia	27,974	RR: 1.98 (1.36-2.86)
Hodson et al., 2013	Antiviral prophylaxis and cytomegalovirus infections	1,005	RR: 0.43 (0.35-0.52)
Hodson et al., 2013	Antiviral prophylaxis and cytomegalovirus disease	1,028	RR: 0.44 (0.33-0.59)
Howe et al., 2011	Exercise and change in bone mineral density	766	SMD: 0.22 (0.05-0.40)
Katalinic et al., 2010	Stretch interventions and joint mobility	109	SMD: 0.27 (-0.1-0.64)
Lai et al., 2013	Antimicrobial impregnation, coating or bonding and mortality	1,517	RR: 0.87 (0.73-1.03)
Lai et al., 2013	Antimicrobial impregnation, coating or bonding and adverse effects	2,954	RR: 1.08 (0.93-1.25)
Law et al., 2013	Sumatriptan plus naproxen and pain after 2 hours	2,819	RR: 2.62 (2.18-3.14)
Law et al., 2013	Sumatriptan plus naproxen and pain after 24 hours	2,820	RR: 2.85 (2.26-3.6)
Lemiengre et al., 2012	Antibiotics and cure from rhinosinusitis	1,552	RR: 1.07 (0.99-1.16)
Lemiengre et al., 2012	Antibiotics and treatment failure	1,983	RR: 0.55 (0.41-0.76)
Lewis et al., 2013	Nonsteroidal anti-inflammatory drugs and vomiting	883	RR: 0.73 (0.57-0.94)
Liakopoulos et al., 2012	Statins and atrial fibrillation	801	SMD: 0.55 (0.44-0.69)
Liakopoulos et al., 2012	Statins and length of stay in hospital	837	RR: -0.35 (-0.61—0.1)
Main et al., 2013	Hormone therapy and stroke	30,434	RR: 1.30 (1.12-1.50)
Moja et al., 2012	Trastuzumab and congestive heart failure	8,477	RR: 5.43 (2.42-12.17)
Musini Vijaya et al., 2009	Pharmacotherapy and Cardiovascular morbidity and mortality	23,013	RR: 0.75 (0.66-0.85)
Nannini et al., 2013	Long-acting beta2- agonist+inhaled corticosteroid and mortality	1,769	RR: 1.02 (0.58-1.79)
Nelson et al., 2012	Surgical therapy of anal fissure and- healing	955	RR: 0.24 (0.15-0.4)
Nüesch et al., 2010	Opioids and withdrawal because of adverse events	119	RR: 4.21 (3.03-5.84)
Pandian et al., 2013	Double embryo transfer and live birth rate	1,411	RR: 1.52 (1.32-1.76)
Pandian et al., 2013	Double embryo transfer and multiple pregnancy rate	1,411	RR: 6.94 (2.39-20.13)
Pani et al., 2011	Antidepressant medication and alcohol abstinence	922	RR: 1.25 (0.88-1.79)
Paul et al., 2013	Antibiotic therapy and death in cancer patients with neutropenia	1,614	RR: 0.89 (0.75-1.05)
Paul et al., 2013	Antibiotic therapy and nephrotoxicity in cancer patients with neutropenia	4,793	RR: 0.53 (0.41-0.68)
Perez et al., 2009	Angiotensin converting enzyme inhibitors and mortality	84,273	RR: 0.93 (0.88-0.98)
Perez et al., 2009	Beta-blockers and mortality	71,369	RR: 0.95 (0.89-1.02)
Rehman et al., 2011	Traditional suburethral sling procedures and incontinence	292	RR: 1 (0.81-1.24)
Wilhelmus et al., 2010	Antiviral therapies and healing of herpes simplex virus keratitis	331	RR: 2.1 (1.44-3.08)

Preparing “Gradeable” Documents

From each of the 50 included bodies of evidence, we used subsets of studies to prepare 160 documents (which we called “gradeable” documents) of different QOE categories. Sample size calculations indicated that 130 documents would provide 80 percent power for a 4 x 2 chi-square test of QOE (high, medium, low, insufficient [the AHRQ category *insufficient* is similar to the GRADE category of *very low; insufficient*, however, also includes outcomes without evidence. For the purpose of this study we did not consider situations without any evidence]) by stability of results (stable vs. not stable) for a medium-sized effect (Cohen’s *d* of 0.3).

We re-analyzed each body of evidence using cumulative meta-analyses. In general, a cumulative meta-analysis shows how the body of evidence evolves over time as new studies accrue. Likewise, the QOE changes (or can be expected to change) over time as new studies contribute to the body of evidence. Based on the cumulative meta-analyses, an independent investigator (who was not involved in the subsequent grading of the QOE) used subsets of bodies of evidence to create gradeable documents. The aim was to create approximately 40 documents for each category of QOE with sufficient information for the project’s investigators to grade the QOE. These documents included information on the objective of the Cochrane review, the PICO (population-intervention-control-outcome), study characteristics and risk of bias ratings of included trials as presented in the Cochrane report, a forest plot of a random effects meta-analysis, information about minimal important differences for continuous outcomes, and information about reporting bias (funnel plot, Kendall’s tau, Egger’s regression intercept, and Fail-Safe N). We relied on judgments of the Cochrane authors regarding risk of bias of individual trials. We pilot-tested the format and content of the gradeable documents and revised them based on feedback from investigators.

Grading Quality of Evidence

To grade the QOE, investigators could choose between GRADE or the EPC guidance for GRADE. All researchers involved in this study chose the EPC guidance for GRADE. Investigators took part in a calibration exercise and had access to a published guidance document.⁴

We randomly allocated 160 gradeable documents to 13 investigators from six U.S. and Canadian EPCs and Cochrane Austria. All are professional systematic reviewers, their experience with GRADE, however, varied. Three investigators (23 percent) stated that they had used the GRADE approach for more than 20 systematic reviews, three (23 percent) for 10 to 15 systematic reviews, one (8 percent) for 6 to 10 reviews, and 6 investigators (46 percent) declared that they had used GRADE for up to 5 systematic reviews.

A research associate at <TO BE PROVIDED AFTER PEER REVIEW> connected each participant with a unique identification number and emailed the gradeable documents. This research associate was not involved in the grading exercise or in the analysis of results. Two investigators, blinded to the results of the underlying Cochrane report (i.e., the reference standard), graded each body of evidence independently. Investigators were blinded to the second person grading the same body of evidence. In case grades differed, the research associate put investigators in contact. Investigators resolved conflicts by consensus or by involving a third, senior researcher.

Assessing the Stability of Effect Estimates

To determine the stability of effects, we compared effect estimates of the gradeable documents with the high QOE estimates from the Cochrane reports (the gold standard). To do so, we modified an approach developed to detect signals for updating systematic reviews.⁹

We used three definitions of stability (Table 2) which differed in the thresholds that determined whether the magnitude of treatment effects was similar. We deemed an estimate of effect as stable when (1) statistical significance did not change *and* (2) the magnitude of treatment effects remained similar compared to the high QOE estimate of the Cochrane report.

Table 2. Three definitions of stability of effect based on change in statistical significance and magnitude of effect

Stability of effect: definition 1 (strict definition)	
Change in statistical significance	Statistical significance does not change between graded effect and gold standard effect (changes within the range of p-values 0.04 to 0.06 are not counted as change).
Change in magnitude of effect	Difference in magnitude of effects is smaller than a relative risk change (increase or reduction) of 25 percentage points for dichotomous outcomes or 0.20 SMDs for continuous outcomes.
Stability of effect: definition 2 (lenient definition)	
Change in statistical significance	Same as definition 1
Change in magnitude of effect	Difference in magnitude of effects is smaller than a relative risk change of 50 percentage points for dichotomous outcomes or 0.50 SMDs for continuous outcomes.
Stability of effect: definition 3 (staggered definition)	
Change in statistical significance	Same as definition 1
Change in magnitude of effect	<ul style="list-style-type: none"> • For graded effects with small treatment effects (relative risk 0.5 to 2.00, or SMD <0.8): same as definition 1 • For graded estimates with large treatment effects (relative risk <0.5 and >2.00, or SMD >0.8): same as definition 2 • For outcomes that can be considered extremely patient-relevant (e.g., mortality, stroke, myocardial infarction): difference in magnitude of effects is smaller than relative risk change of less than 10 percentage points.

SMDs = standardized mean differences

To avoid counting trivial or ‘borderline’ changes in statistical significance, we required that at least one of the two results had had a p-value outside the range of 0.04 to 0.06. In other words, we did not consider cases in which a p-value changed statistical significance within this range. For example, neither a change from p = 0.041 to p = 0.059 nor a change from p = 0.059 to p = 0.041 counted as a change in statistical significance.

Conducting Statistical Analysis

To determine the predictive validity of the GRADE approach, we compared the expected proportion of stable effect estimates (presented in Table 1) with the observed proportion of stable effect estimates for different thresholds from our sample. We determined the calibration of the GRADE approach with the Hosmer-Lemeshow test¹⁰ and its discrimination with the concordance (C) index. Calibration is the ability to estimate correctly the likelihood of a future event (e.g., likelihood that estimates remain stable). Discrimination determines how well the grading system differentiates between bodies of evidence that will remain stable and those that will not remain stable.¹¹ Bodies of evidence that are stable should have higher expected likelihoods than those that are not stable. The C index compares the expected likelihoods from pairs of observations, in this case, stable vs. not stable bodies of evidence as shown below:¹²

$$C \text{ index} = \frac{\# \text{ of concordant pairs} + \frac{1}{2} (\# \text{ of tied pairs})}{\text{Total \# of pairs}}$$

Concordant pairs are pairs for which the expected likelihood for the stable body of evidence is higher than the expected likelihood for the nonstable body of evidence. *Tied pairs* are pairs where the stable and nonstable bodies of evidence have the same expected likelihood. Higher values for the C index indicate better discrimination. A C index of 0.50 would indicate no discrimination between stable and nonstable bodies of evidence. We conducted all statistical analyses with the `rcorr.cens` procedure in the `Hmisc` package in R¹³ or Microsoft Excel.

Results

Of 160 bodies of evidence, researchers dually graded 11 percent (n=17) as high, 42 percent (n=68) as moderate, 32 percent (n=51) as low, and 15 percent (n=24) as insufficient (very low) QOE.

Concordance Between Expected and Observed Proportions of Stable Effect Estimates

For each grade, we compared the expected proportions of stable effect estimates with the observed proportion from our sample using three different definitions of stability (see Methods and Table 2). Table 1 gave the proportions of estimates that producers and users of systematic reviews expect to remain stable for each QOE grade.

Overall, except for moderate QOE, the stability differed considerably between expected and observed proportions regardless of the definition used. *Fewer* estimates graded as high QOE in our sample remained stable relative to the expectations of producers and users of systematic reviews, i.e., in our survey 208 experts expected high QOE outcomes to remain stable in at least 86 percent of the cases. In our sample the observed proportions of stable estimates for definitions 1, 2, and 3 were, respectively, 71 percent, 76 percent, and 76 percent. Conversely, substantially *more* low or insufficient (very low) QOE estimates than expected remained stable. Table 3 presents expected and observed proportions of stable effect estimates by grade of QOE for each of the three definitions of stability.

Table 3. Comparison of expected with observed proportions of stable effect estimates for different definitions of stability

Grade	Number of effect estimates	Expected proportions (%) ^a	Observed proportions (%) definition 1 (95% CI)	Observed proportions definition 2 (95% CI)	Observed proportions definition 3 (95% CI)
High	17	86-100	71 (43-88)	76 (48-92)	76 (48-92)
Moderate	68	61-85	71 (58-80)	75 (63-84)	72 (59-91)
Low	51	34-60	55 (41-68)	73 ^b (58-83)	59 (44-72)
Insufficient (very low)	24	0-33	54 ^b (33-74)	58 ^b (37-77)	58 ^b (37-77)

CI = confidence interval

^a Expected proportions are based on an international survey of producers and users of systematic reviews⁶

^b Statistically significantly different from the upper bound of expected stability

Figures 1, 2, and 3 illustrate the overlap of expected proportions of stable effects (black large boxes) and confidence intervals (CI) of observed proportions (yellow columns) for different grades of QOE and different definitions of stability. The dots in the columns reflect the point estimates. The y-axis delineates the proportion of estimates that remained stable; the x-axis presents the four grades of QOE. For insufficient (very low) QOE, for example, producers and users of systematic reviews expected 0 percent to 33 percent of estimates to remain stable as new studies are added to the evidence base. For definition 1, which was the most rigorous of the three definitions of stability, more than half (54 percent) of effect estimates graded as insufficient (very low) remained stable. The CIs ranged from 33 percent to 74 percent, which barely overlaps the expected range for insufficient (very low) QOE. For the less rigorous definitions 2 and 3, CIs did not overlap at all with the range that producers and users of systematic reviews expected from insufficient (very low) QOE grades. By

contrast, observed proportions of stable results for moderate QOE grades were concordant for all three definitions. Confidence intervals overlap widely with the range of expected proportions. Estimates graded as low QOE show some concordance for definitions 1 and 3 but little for definition 2. Estimates graded as low QOE show some concordance for definitions 1 and 3 but little for definition 2.

Figure 1. Comparison of expected proportions of stable effect estimates with confidence intervals of observed proportions for different definitions of stability—Definition 1

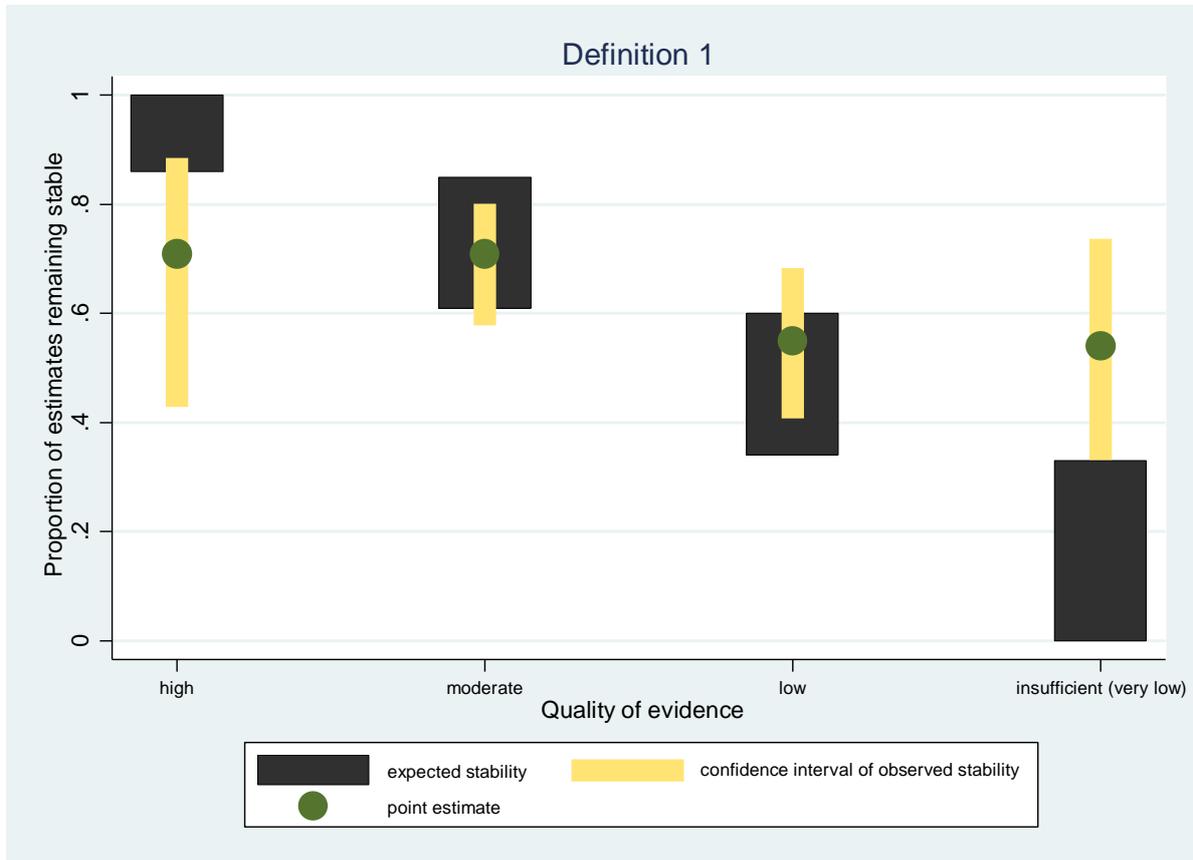


Figure 2. Comparison of expected proportions of stable effect estimates with confidence intervals of observed proportions for different definitions of stability—Definition 2

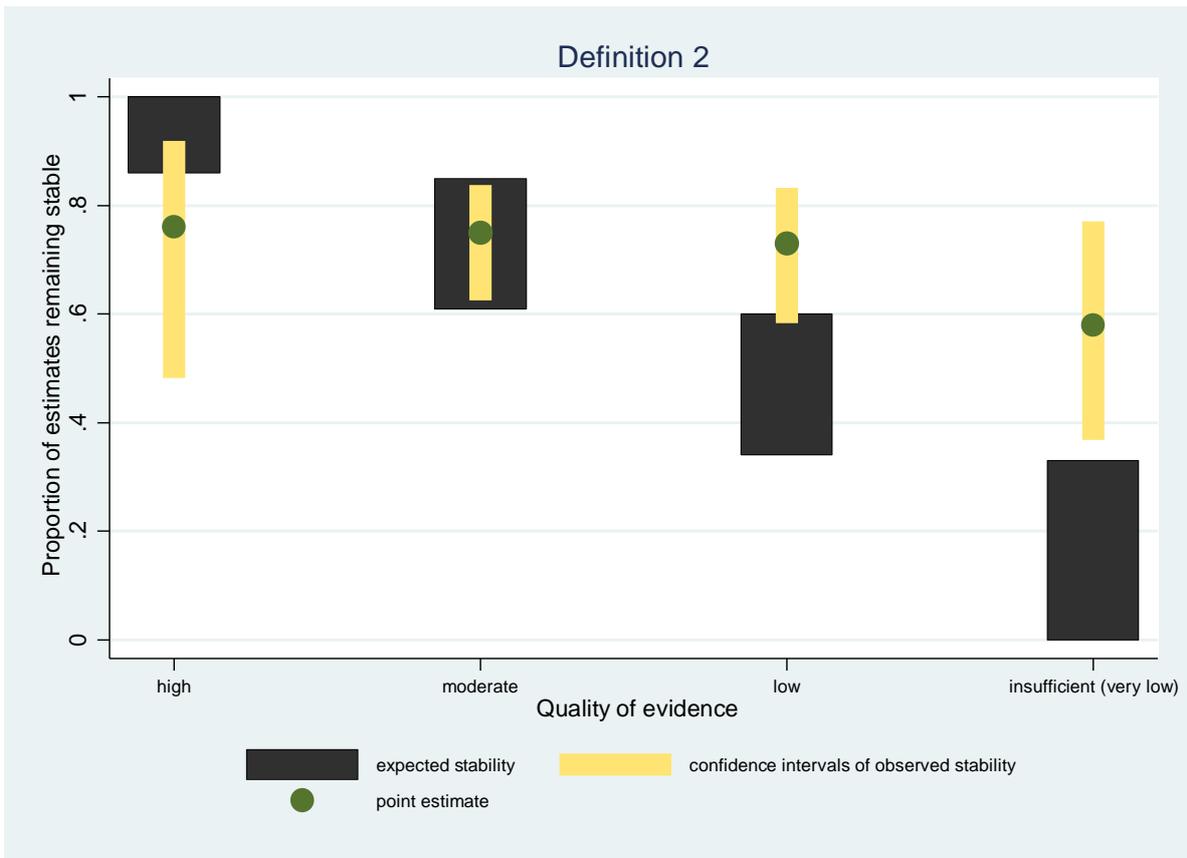
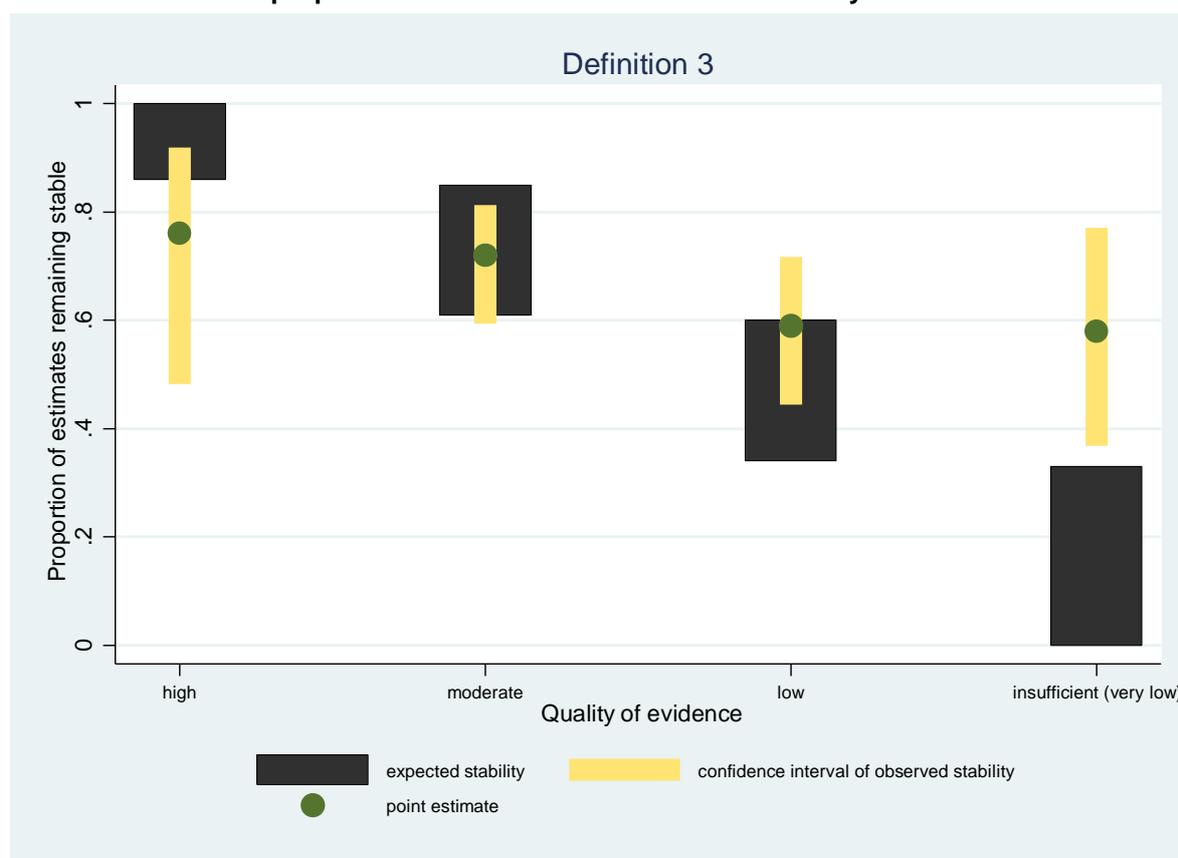


Figure 3. Comparison of expected proportions of stable effect estimates with confidence intervals of observed proportions for different definitions of stability—Definition 3



Predictive Validity of the GRADE Approach

To determine the predictive validity of the GRADE approach, we assessed the calibration (i.e., how accurately GRADE can predict the likelihood that effect estimates will remain stable as new evidence evolves) and the discrimination (i.e., how accurately GRADE can differentiate between effect estimates that will remain stable and those that will substantially change). In theory, an ideal predictive tool would reliably identify estimates with a high likelihood of remaining stable and always grade them as high QOE. Conversely, effect estimates with a very low likelihood of remaining stable would always be graded as very low. Such an ideal tool would have high calibration and a C index of 1.

Overall, regardless of the definition used, the calibration of GRADE was suboptimal. When we compared observed proportions of stable effect estimates with lower, middle, and upper values of the ranges of expected proportions, eight of nine comparisons were statistically significantly different based on Hosmer-Lemeshow test ($p < 0.05$), indicating a lack of calibration. Likewise, the C indices for the GRADE approach were low with values close to that expected by chance (i.e., C index = 0.50). For definitions 1, 2, and 3, the C indices were 0.57 (95% CI, 0.50-0.67), 0.56 (95% CI, 0.47-0.66), and 0.58 (95% CI, 0.50-0.67), respectively. C indices for definitions 1 and 3 reached statistical significance (CIs did not cross 0.5). Taking the uncertainty of the confidence intervals into consideration, results mean that in the worst case (lower limit of CIs) the GRADE approach has no discriminatory ability when it comes to distinguish between effect estimates with a low or high likelihood of remaining stable. In the best case (upper confidence limits), the GRADE approach can accurately distinguish between effect estimates with a low or high likelihood of remaining stable in 67 percent of cases.

The low overall predictive validity, however, is primarily caused by the discordance of expected and observed proportions of stable effect estimates for high and insufficient (very low) QOE. In sensitivity analyses we chose proportions within the expected ranges (see Table 1) that were closest to the observed proportions of stable effect estimates. Using expected proportions of 86 percent for high (lower end of expected range), 71 percent for moderate, 60 percent for low, and 33 percent for insufficient QOE (both upper end of expected range), we found that the GRADE approach achieved satisfactory calibration for definitions 1 and 3.

Discussion

To our knowledge, our study was the first attempt to determine the predictive validity of the GRADE approach. To be considered useful in practice, any tool that conveys certainties and uncertainties of estimates of effect should have a high ability to discriminate between estimates that will remain stable in the future and those that will substantially change; it should also be able to associate respective likelihoods with an expected outcome. Our research indicates that the GRADE approach only partly fulfilled these qualities of predictive validity: only moderate QOE had satisfactory predictive validity. In the following sections we discuss possible reasons for these findings and potential starting points for improving the predictive validity.

A predictive model, in general, is a mathematical equation describing the relationship between a prognostic marker (here, a grade of QOE) and a given outcome (stability of effect estimates).¹² In our study, three main factors determined the predictive validity of the GRADE approach:

1. The definition of stability,
2. The likelihood of expected stability associated with each grade of QOE (the prognostic marker), and
3. The operationalization of the prognostic tool (the GRADE approach) to achieve the most appropriate prognostic marker (i.e., the grade of QOE).

With respect to the first factor, the definition of stability, our study showed that strict or lenient definitions of stability had minimal impact on the predictive validity of GRADE. Therefore, the other two factors appear to be the reasons for the low predictive validity and could serve as starting points for future improvements.

To determine the proportion of stable estimates that users and producers of systematic reviews associate with each grades of QOE, we recently conducted an international survey which we used as the basis of the comparison between expected and observed proportions of stable results.⁶ Except for moderate QOE, the expectations of survey participants did not match results from our sample. Expectations were too optimistic for high QOE and too pessimistic for low and insufficient (very low) QOE. Current definitions of different grades of QOE, however, employ vague terminology to forecast certainty—such as “likely,” “very likely,” or “may be substantially different” which might contribute to the low predictive validity. Psychological research has demonstrated that perceptions of certainty can vary substantially among individuals and that interpretation of qualitative certainty expressions also differ depending on the context in which they are used and on baseline event rates. Adding numerical predictions such as likelihoods to the definitions of the individual grades of QOE seems to be one solution that could reduce unwarranted variation in interpretations.

Finally, the GRADE approach, or the way systematic reviewers operationalize it, appears to be too strict. More than half of estimates graded as insufficient (very low) (defined as “we have no confidence in the estimate of effect for this outcome”) remained stable; this indicates that GRADE too often leads to low or insufficient (very low) grades of QOE. Possible reasons could be: a) systematic reviewers use GRADE too mechanistically, b) recommended thresholds for downgrading in guidance documents are too strict, or c) a tool with four levels of QOE is not granular enough to categorize uncertainty. Adding a fifth category, e.g. by using GRADE *very low* for bodies of evidence in which systematic reviewers still have some confidence (albeit little confidence) and AHRQ *insufficient* for bodies of evidence that have truly unacceptable deficiencies that preclude reaching a conclusion, would allow for more granularity.

Our study has several limitations. First, we relied on risk of bias assessments and QOE grades of Cochrane authors. Because author groups differed across these systematic reviews, some heterogeneity in approaches regarding QOE grades is likely. Nevertheless, such heterogeneity reflects a real-world situation because most guideline developers or other decisionmakers who use Cochrane reports to support decisions would not reassess QOE. In addition, Cochrane reports go through rigorous international peer review, and the methodological quality usually is high.

Second, how representative our sample is remains unclear. Because we wanted to use a reference standard for which researchers had high confidence that effect estimates are correct (close to the true effect), we focused on high QOE evidence. A remaining question is whether our findings are generalizable to bodies of evidence that will never progress to high QOE. In addition, our sample was limited to RCTs and findings are likely not generalizable to research based on non-randomized studies.

Third, systematic reviewers grading the QOE had access to guidance documents but they did not use a tool such as the GRADEpro Guideline Development Tool (www.guidelinedevelopment.org) to guide them through the grading exercises in a standardized manner. Using such a tool could increase inter-rater reliability and might reduce the number of grades of QOE that are too strict. In situations with conflicting grades, strong personalities (maybe with a tendency to strict grades) often dominate the consensus process. Increasing inter-rater reliability¹⁴ would reduce the number of situations that require systematic reviewers to reach a consensus.

Finally, elements of the GRADE approach itself can be criticized. GRADE links QOE grades to the degree of confidence that estimates are close to the true effect. This concept can be criticized from an epistemological perspective because quantifiable entities (grades of QOE) are linked to an abstract concept (the truth) that can never be verified. Nevertheless, we purposely took GRADE definitions at face value. GRADE is used by more than 70 international organizations; most decisionmakers conceivably accept and rely on GRADE assessments and their current definitions.

Over the past decade GRADE has evolved as a widely used approach to convey the certainties and uncertainties inherent in research. Its conceptual framework uses information about factors that most researchers would intuitively consider when assessing the confidence in findings based on a body of evidence. Compared with other approaches, GRADE has clear advantages because it makes decisions about the QOE transparent and explicit.¹⁵

The lack of predictive validity, therefore, is probably not grounded in the concept of GRADE but rather in the way the instrument is operationalized which, overall, appears too strict. The GRADE Working Group, as well as organizations such as the Evidence-based Practice Centers need to reflect on how to reduce unwarranted variation in the interpretation of the definitions of individual grades of QOE and how to avoid overly strict grades. Future research needs to confirm or refute our findings and explore which domains may lead to a too strict operationalization and influence the predictive validity of the GRADE approach.

References

1. Atkins D, Siegel J, Slutsky J. Making policy when the evidence is in dispute. *Health Affairs (Millwood)*. 2005 Jan-Feb;24(1):102-13. Epub: 2005/01/14. PMID: 15647220.
2. Schunemann HJ, Hill SR, Kakad M, et al. Transparent development of the WHO rapid advice guidelines. *PLoS Med*. 2007;4(5):e119.
3. Qaseem A, Forland F, Macbeth F, et al. Guidelines International Network: Toward International Standards For Clinical Practice Guidelines. *Ann Intern Med*. 2012 Apr 3;156(7):525-31. Epub: 2012/04/05. PMID: 22473437.
4. Berkman ND, Lohr KN, Ansari M, et al. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update. RTI-UNC Evidence-based Practice Center; 2013 <http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=1752>. Accessed Contract No. HHS-290-2007-10056-I-EPC3, Task Order #5.
5. 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 May;63(5):513-23. PMID: 19595577.
6. Gartlehner G, Sommer I, Swinson Evans T, et al. Grades for quality of evidence are associated with distinct likelihoods that treatments effects will remain stable. *J Clin Epidemiol*. 2014: doi: 10.1016/j.jclinepi.2014.09.018. [Epub ahead of print].
7. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychological Bulletin*. 1955 Jul;52(4):281-302. PMID: 13245896.
8. Borenstein M, Hedges LV, Higgins JPT, et al. *Introduction to Meta-Analysis*. John Wiley & Sons Ltd; 2009.
9. Shojania KG, Sampson M, Ansari MT, et al. *Updating Systematic Reviews*. Technical Review No. 16. Rockville (MD): Agency for Healthcare Research and Quality. AHRQ Publication No. 07-0087.; 2007.
10. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. in *Communications in Statistics - Theory and Methods*; 1980.
11. McGeechan K, Macaskill P, Irwig L, et al. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. *Arch Intern Med*. 2008 Nov 24;168(21):2304-10. Epub: 2008/11/26. PMID: 19029492.
12. Tripepi G, Jager KJ, Dekker FW, et al. Statistical methods for the assessment of prognostic biomarkers (Part I): discrimination. *Nephrology, Dialysis, Transplantation* : official publication of the European Dialysis and Transplant Association - European Renal Association. 2010 May;25(5):1399-401. Epub: 2010/02/09. PMID: 20139066.
13. Harrell J, F.E. Package 'Hmisc'. 2014 <http://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf>. Accessed December 19, 2014.
14. Mustafa RA, Santesso N, Brozek J, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. *J Clin Epidemiol*. 2013 Jul;66(7):736-42 e5. PMID: 23623694.
15. Schunemann HJ, Best D, Vist G, et al. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. *CMAJ*. 2003;169(7):677-80.