ELSEVIER

Journal of Clinical Epidemiology 66 (2013) 736−742

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses

Reem A. Mustafa[a], Nancy Santesso[a], Jan Brozek[a,b], Elie A. Akl[a,c], Stephen D. Walter[a], Geoff Norman[a], Mahan Kulasegaram[a], Robin Christensen[d], Gordon H. Guyatt[a,b], Yngve Falck-Ytter[e], Stephanie Chang[f], Mohammad Hassan Murad[g], Gunn E. Vist[h], Toby Lasserson[i], Gerald Gartlehner[j], Vijay Shukla[k], Xin Sun[l], Craig Whittington[m], Piet N. Post[n], Eddy Lang[o], Kylie Thaler[p], Ilkka Kunnamo[q], Heidi Alenius[q], Joerg J. Meerpohl[r], Ana C. Alba[a,s], Immaculate F. Nevis[a], Stephen Gentles[a], Marie-Chantal Ethier[a,t], Alonso Carrasco-Labra[a,u], Rasha Khatib[a,v], Gihad Nesrallah[a,w], Jamie Kroft[x], Amanda Selk[y], Romina Brignardello-Petersen[a,u], Holger J. Schünemann[a,b,*]

[a]*Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton, Ontario L8S 4K1, Canada*
[b]*Department of Medicine, McMaster University, Hamilton, Ontario L8S 4K1, Canada*
[c]*Department of Medicine, State University of New York at Buffalo, Buffalo, NY, USA*
[d]*Musculoskeletal Statistics Unit, The Parker Institute, Copenhagen University Hospital, Frederiksberg, Denmark*
[e]*Louis Stokes Cleveland VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA*
[f]*Agency for Healthcare Research and Quality, Rockville, MD 20850, USA*
[g]*Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, MN, USA*
[h]*Norwegian Knowledge Centre for Health Services, Oslo 0130, Norway*
[i]*Cochrane Editorial Unit, London, UK*
[j]*RTI International, Durham, NC, USA*
[k]*Canadian Agency for Drugs and Technology in Health (CADTH), Ottawa, ON, K1S 5S8, Canada*
[l]*Center for Health Research, Kaiser Permanente Northwest, Portland, OR 97227, USA*
[m]*National Collaborating Centre for Mental Health, Centre for Outcomes Research and Effectiveness, Research Department of Clinical, Educational & Health Psychology, London, UK*
[n]*Post Voor Zorg, Delft, The Netherlands*
[o]*Department of Emergency Medicine, University of Calgary, Calgary, Alberta, 2Nt 2T9, Canada*
[p]*Austrian Cochrane Branch, Department for Evidence-Based Medicine and Clinical Epidemiology, Danube University Krems, Austria*
[q]*EBM Guidelines, Duodecim Medical Publications Ltd, Finland*
[r]*Institute of Medical Biometry and Medical Informatics, German Cochrane Centre, University Medical Center Freiburg, Germany*
[s]*Toronto General Hospital, University Health Network, Toronto, ON, M5G 2C4, Canada*
[t]*Hospital for Sick Children, Toronto, ON, M5G 1X8, Canada*
[u]*Evidence-Based Dentistry Unit, Faculty of Dentistry, Universidad de Chile, Chile*
[v]*Population Health Research Institute, Hamilton, ON, L8L 2X2, Canada*
[w]*Humber River Regional Hospital, Toronto, ON, M5B 1Z2, Canada*
[x]*Sunnybrook Health Sciences Centre, Toronto, ON, M4N 3M5, Canada*
[y]*Women's College Hospital, Toronto, ON, M5S 1B2, Canada*

Accepted 11 February 2013; Published online 23 April 2013

Abstract

Objective: We evaluated the inter-rater reliability (IRR) of assessing the quality of evidence (QoE) using the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) approach.

Study Design and Setting: On completing two training exercises, participants worked independently as individual raters to assess the QoE of 16 outcomes. After recording their initial impression using a global rating, raters graded the QoE following the GRADE approach. Subsequently, randomly paired raters submitted a consensus rating.

* Corresponding author. Department of Clinical Epidemiology & Biostatistics, McMaster University, HSC Room 2C15, 1280 Main Street, West Hamilton, Ontario L8S 4K1, Canada. Tel.: +1-905-525-9140x22296; fax: +1-905-522-9507.

E-mail address: schuneh@mcmaster.ca (H.J. Schünemann).

http://dx.doi.org/10.1016/j.jclinepi.2013.02.004

Results: The IRR without using the GRADE approach for two individual raters was 0.31 (95% confidence interval [95% CI] = 0.21−0.42) among Health Research Methodology students (n = 10) and 0.27 (95% CI = 0.19−0.37) among the GRADE working group members (n = 15). The corresponding IRR of the GRADE approach in assessing the QoE was significantly higher, that is, 0.66 (95% CI = 0.56−0.75) and 0.72 (95% CI = 0.61−0.79), respectively. The IRR further increased for three (0.80 [95% CI = 0.73−0.86] and 0.74 [95% CI = 0.65−0.81]) or four raters (0.84 [95% CI = 0.78−0.89] and 0.79 [95% CI = 0.71−0.85]). The IRR did not improve when QoE was assessed through a consensus rating.

Conclusion: Our findings suggest that trained individuals using the GRADE approach improves reliability in comparison to intuitive judgments about the QoE and that two individual raters can reliably assess the QoE using the GRADE system.

Keywords: GRADE; Inter-rater reliability; Levels of evidence; Evidence-based medicine; Reproducibility; Validation studies

1. Background

The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) working group includes guideline developers, systematic reviewers, clinicians, public health officers, researchers, methodologists, and other health professionals from around the world [1]. The group has spent over a decade developing and refining a systematic, transparent, and explicit process for summarizing, grading, and presenting evidence, and for moving from evidence to health care recommendations. More than 65 international organizations (http://www.gradeworkinggroup.org/society/index.htm) have adopted the GRADE approach, which is becoming an international standard.

The GRADE approach assesses the quality of a body of evidence (QoE) defined as confidence in estimates of effects of alternative management strategies [2]. The process involves grading QoE for all individual outcomes deemed to be important or critical for decision making and subsequently, for use in health care recommendations, determining the overall QoE across critical outcomes [3].

In the GRADE system, evaluating the QoE for each outcome of interest begins with determining a consideration of the study design (randomized trials or observational studies) and then assessing eight additional domains: risk of bias [4], indirectness of evidence [5], inconsistency of evidence [6], imprecision of the estimated effect [7], likelihood of publication bias [8], the presence of a dose−response effect, magnitude of the estimated effect, and issues around residual confounding [9]. After assessing all the mentioned domains, QoE per outcome is categorized as: ⊕⊕⊕⊕ (high), ⊕⊕⊕◯ (moderate), ⊕⊕◯◯ (low), or ⊕◯◯◯ (very low) [10]. The overall QoE is determined by the QoE for each of the critical outcomes in that in most instances the overall QoE is based on the lowest QoE for any of the critical outcomes.

The final product of the GRADE process is frequently presented as a summary table [11]. The table shows the quality of the body of the available evidence for each outcome, judgments that bear on the quality rating, and effects of alternative management strategies. To enhance and ensure transparency, GRADE recommends the reporting and explicit justification of these judgments. This reporting is typically done in the form of mandatory footnotes for each judgment. The summary tables can be presented in different formats, which are referred to as GRADE evidence profiles or summary of findings tables (Appendix A and B; available on the journal's website at www.jclinepi.com).

The current practice of the users of the GRADE approach varies. Some raters assess the evidence without consulting with any other individuals. Some raters discuss it with others before reporting the QoE. There is no clear guidance about the effect of either of these two approaches on the reliability of assessing the QoE.

Much of the research regarding GRADE has thus far focused on transparency in reporting the judgments according to a systematic approach. In fact, there are limited published data on the reliability of the GRADE process when assessing the QoE. Based on the initial work of the GRADE working group, the only study that included an assessment of reliability focused on piloting the system [12] but that study was conducted when limited guidance was available for the judgments that must be made when evaluating evidence.

The GRADE approach is now supported by detailed and explicit guidance regarding these required judgments. An evaluation of the inter-rater reliability (IRR) is essential to inform the development of training materials and further refine the GRADE approach. Although more than 100 different systems to evaluate QoE exist in the literature [13], very few have published any formal testing of their IRRs. We aimed to answer the following questions: (1) What is the IRR of assessing QoE (high, moderate, low, and very low) using the GRADE approach by individuals with different levels of experience with this approach? (2) Does IRR improve when two raters report consensus QoE, compared with individual raters? (3) How do the four categories of QoE (high, moderate, low, and very low) in the GRADE approach compare to a global rating of the QoE on a visual analog scale (VAS)?

2. Method

2.1. Design

Participants initially worked independently as individual raters assessing the QoE. Once individual raters submitted their judgments about QoE, we randomly paired them with another rater. We asked each pair to discuss their independent ratings and resolve discrepancies by discussion before

What is new?

Key points

- Using the GRADE approach improves reliability in comparison to intuitive judgments about the quality of a body of evidence.
- Two individual raters can reliably assess the QoE using the GRADE system.
- The inter-rater reliability did not improve when QoE was assessed through a consensus rating by pairs of raters.

submitting their final consensus judgment. Both individual raters and pairs of raters worked independently from other raters who evaluated the same evidence. This design allowed us to test the effect of requiring independent assessment of QoE in duplicate before resolving disagreement on IRR.

Each rater completed a questionnaire (Appendix C; available on the journal's website at www.jclinepi.com) to assess his/her baseline characteristics, familiarity, and expertise with evaluating QoE and producing GRADE summary tables. Each participant had access to the help file of GRADEpro (version 3.6; McMaster University, Hamilton, ON, Canada) as discussed later and was encouraged to read it before starting the study. All participants were required to complete two calibration exercises designed to review the GRADE process in assessing QoE and identify any technical difficulties with the study. In this exercise, raters evaluated the QoE for 10 outcomes from two systematic reviews [14,15]. After all raters submitted their assessment, we provided general feedback to the whole group.

Subsequently, independent raters and pairs evaluated the body of evidence published in four Cochrane systematic reviews [16−20]. Initially, raters provided an assessment of the QoE based on a global rating in the form of a VAS. Raters received clear instructions that they should only base this first assessment of the QoE on overall intuition and not GRADE criteria. The raters were instructed that they should assign a score of 100 when they were 100% confident that the pooled effect estimates are the actual true effect estimates. The raters were also instructed to assign a score of zero when the rater had no confidence in the pooled effect estimates at all. After documenting and submitting their rating using the VAS, they rated the QoE per outcome based on their assessment of the eight domains determining QoE in the GRADE system. Then they determined the overall QoE across outcomes based on their assessment of the QoE per outcome. We provided all raters with standardized judgments about the importance of the outcomes. All included outcomes in this study were critical or important for decision making.

All responses were anonymous. One of the research associates in the Department of Clinical Epidemiology and Biostatistics at McMaster University generated and saved a list connecting each participant to a random unique identification number. This research associate was not involved with the design, execution, or analysis of the study. She e-mailed each of the participants their unique identification number. The investigators, data analysts, and other participants in the study had no access to this list. We used this list to link responses from the survey to the corresponding ratings of QoE.

2.2. Participants

Volunteers from the GRADE working group [1] and from the Health Research Methodology (HRM) graduate program in McMaster University [21] participated in this study. We chose these two groups to ensure that participants had at least minimal exposure to the GRADE approach. The second group, however, allowed us to explore effects of increasing familiarity on reliability. The GRADE working group is an international group of more than 300 members. It includes methodologists, clinicians, public health officers, other health care providers, and researchers involved in conducting systematic reviews and providing support to health care practice guideline panels. There are no inclusion criteria or prerequisites to becoming a member of the GRADE working group. Members usually sign up because they are interested in contributing to, discussing, or expanding their knowledge in the area of critical appraisal of the body of evidence and health care practice guideline development. The expertise of those involved in the group varies from methodological experts who lead the field to those with less experience in preparing GRADE evidence summaries. We recruited participants from the group using two methods. First, we announced the idea of the study during GRADE working group meetings in January 2010, October 2010, and January 2011, respectively. Interested members signed up voluntarily to participate as raters for this study. Second, we sent an electronic invitation to the entire working group to participate in this study. In this invitation, we explained the objectives of the study with a clear description of expected workload and anticipated time line.

The HRM program in the Faculty of Health Sciences at McMaster University provides training at the MSc and PhD levels. The students in this program are exposed to the GRADE methods in assessing QoE during one of the required courses for students in the clinical epidemiology stream. We recruited graduate students by sending an electronic invitation to the students' list serve. We required students to have completed a systematic review course during which they are required to use the GRADE approach to assess QoE.

2.3. Masking

All raters in addition to the investigators had a chance to review an early draft of the protocol and give feedback. They were blinded to the final analysis plan that included the need to stratify outcomes based on QoE. We did that to reduce bias

in making balanced judgments about the QoE. The final protocol was reviewed and approved by eight investigators who remained unblinded to the analysis plan (R.A.M., J.B., N.S., E.A.A., S.D.W., R.C., G.H.G., and H.J.S.).

2.4. Data sources: systematic reviews and outcomes

The investigators selected four well-conducted and well-reported Cochrane systematic reviews, based on assessment using the AMSTAR tool [22]. We chose our sample of systematic reviews to ensure variability in clinical and public health areas and variability in the QoE across the eight domains to enhance external validity of our findings. We selected four critical or important outcomes from each review. We decided to use more than one outcome from each systematic review to simulate real-life situations. In addition, we originally thought that this would allow us to assess the overall QoE that is important for decision making. We were unable to calculate IRR of overall QoE as our final sample did not have enough variability in the overall QoE, that is, the judgments across outcomes.

On the other hand, we stratified our sample of outcomes from the systematic reviews by QoE (high, moderate, low, and very low) to guarantee that we have a representative sample that includes all four categories of QoE.

2.5. GRADE profiler (GRADEpro) software and GRADE summary tables

The raters used computer software (GRADEpro, version 3.6) specifically designed to assess the QoE and create GRADE evidence profiles and summary of findings tables and record the required judgments. The investigators provided each of the raters with partially completed GRADEpro files. These files included the outcomes of interest and numerical information about the effects estimated in the systematic reviews. We asked the raters to assess each of the eight domains determining QoE and provide justifications for each of the judgments in footnotes.

2.6. Statistical analysis

We considered four systematic reviews with four different outcomes in each review, feasible for each of the reviewers to evaluate. The raters judged the QoE for each of these 16 (4×4) outcomes to be (1) very low QoE, (2) low QoE, (3) moderate QoE, or (4) high QoE. The primary statistical analysis for IRR was based on intraclass correlation coefficient (ICC) statistics. We decided a priori to interpret the results using Landis and Koch [23] guidelines (values of 0−0.20 represent slight agreement, 0.21−0.40 fair agreement, 0.41−0.60 moderate agreement, 0.61−0.80 substantial agreement, and greater than 0.80 almost perfect agreement). The secondary analysis for IRR was based on assessing ICC statistics for the VAS ratings. We compared IRR of individual raters vs. consensus rating for the GRADE approach. We performed the analysis separately for the HRM students and members of the GRADE working group.

We used generalizability theory and G-String-IV software (http://www.fhs.mcmaster.ca/perd/download; Bloch & Norman, Hamilton, ON, Canada) to estimate the different variance components for our study. We used these variance components to calculate ICC. We used a crossed design where all raters evaluated all outcomes in all systematic reviews. The outcomes are the facet of differentiation, and they are nested in systematic reviews. We calculated descriptive statistics of baseline characteristics of the raters using SPSS version 17 software (SPSS Inc, Chicago, IL).

2.7. Sample size estimation

Anticipating that the GRADE raters will have a substantial agreement, our sample size was based on an ICC of 0.7 with a standard error of 0.05, corresponding to an approximate 95% confidence interval (95% CI) ranging from 0.6 to 0.8. We estimated that we would need to include 96 replicates of observation (ie, 96 mutually independent pairs). Thus, based on the experimental design and requirements for power in our per outcome analysis, we needed each outcome to be assessed 12 times independently—in order for us to have six consensus GRADE ratings for each of the 16 outcomes.

2.8. Ethical consideration

This study satisfied the ethical and scientific requirements of the Hamilton Health Sciences/Faculty of Health Sciences Research Ethics Board at McMaster University, and the requirement for individual consent was waived. We did not anticipate any direct or indirect harm to the participants in the study. We kept all ratings and presentations of the results anonymous. The goal of this work is to advance the knowledge about the reliability of the GRADE approach in assessing QoE and is not intended to evaluate individual abilities and judgments.

3. Results

Twenty-seven members of the GRADE working group and 10 students from the McMaster University HRM graduate program agreed to participate in the study. Fifteen of the 27 GRADE working group members and the 10 students served as raters and completed the assessment of all 16 outcomes.

Table 1 summarizes the baseline characteristics and previous experiences of the raters. Eight of 15 members in the GRADE working group were involved in developing aspects of the GRADE approach. Thirteen raters (nine from the GRADE working group and four students) were familiar with other grading systems. Six members of the GRADE working group held a PhD degree, another six held a masters degree, one had informal training, and two had no training in health research methods. Three students

Table 1. Baseline characteristics and previous experiences of the raters

Baseline characteristics (number of times)	GRADE working group, $N = 15$ Mean (min–max)	HRM students, $N = 10$ Mean (min–max)
Prepared GRADE summary tables	10.9 (1–>15)	1.6 (0–5)
Used GRADEpro software	10.4 (0–>15)	3.3 (0–15)
Attended GRADE meeting	5.4 (0–>15)	0.2 (0–2)
Attended GRADE workshop	3.3 (1–10)	0.6 (0–2)
Facilitated GRADE workshop	4.5 (0–>15)	0.0 (0–0)
Participated in guideline development	3.4 (0–15)	0.7 (0–5)
Led work in guideline development	3.8 (0–>15)	0.5 (0–2)
Involved in systematic review	13.2 (0–>15)	4.5 (1–10)
Led systematic review	7.4 (0–>15)	2.8 (1–8)

Abbreviations: GRADE, Grading of Recommendations, Assessment, Development, and Evaluation; HRM, Health Research Methodology; Min, minimum score; Max, maximum score.

held a master's degree, whereas seven were in the process of earning a doctoral degree in HRM. Eight raters identified themselves as epidemiologists, 16 as clinicians, and none as biostatisticians. Additionally, the raters included a clinical pharmacist, health informatics specialist, research coordinator, editors, and systematic reviewers.

Table 2 summarizes the results of reliability of assessments of QoE by the ICC for different numbers of raters using the VAS and the GRADE approach. The IRR of a global rating of QoE using a VAS without following the GRADE approach when two individual raters evaluated the body of evidence was 0.31 (95% CI = 0.21–0.42) and 0.27 (95% CI = 0.19–0.37) among the HRM students and members of the GRADE working group, respectively. On the other hand, the IRR of the GRADE approach in assessing the QoE when two individual raters evaluated the body of evidence was significantly higher at 0.66 (95% CI = 0.56–0.75) and 0.72 (95% CI = 0.63–0.79) among the same groups. The calibration exercise significantly improved the IRR for the HRM student group (GRADE-naive group) from 0.11 (95% CI = 0.05–0.19) to 0.66, whereas it did not significantly change the IRR for members of the GRADE working group from 0.62 (95% CI = 0.52–0.71) to 0.72. The IRR further increased for three (0.80 [95% CI = 0.73–0.86] and 0.74 [95% CI = 0.65–0.81]) or four raters (0.84 [95% CI = 0.78–0.89] and 0.79 [95% CI = 0.71–0.85]).

The IRR of the GRADE approach in assessing the QoE when pairs of raters evaluated the body of evidence and reached consensus was 0.51 (95% CI = 0.38–0.63) and 0.47 (95% CI = 0.35–0.58) in the HRM students and members of the GRADE working group, respectively. Appendix D (available on the journal's website at www.jclinepi.com) shows summary of the outcomes evaluated with the number of studies informing each outcome (Table A) and the ratings of the QoE per outcome by individual raters (Tables 1B and 2B) and the ratings of the QoE per outcome by pairs of raters (Tables 3B and 4B). Appendix E (available on the journal's website at www.jclinepi.com) summarizes the variance calculated for each of the facets that contributed to calculating the reliability coefficient.

4. Discussion

We found substantial IRR when two individual raters assessed the QoE of 16 outcomes from four systematic reviews. The IRR of the GRADE approach using a GRADE-naive group (students) improved significantly after the calibration exercises from slight agreement (0.11) to substantial agreement (0.66). IRR using the GRADE approach and raters that were familiar with the approach was already high initially (0.62) and improved only slightly (0.72) after the calibration exercises. IRR was similar among members of the GRADE working group and among students in the HRM program after two calibration exercises. The GRADE approach demonstrated higher reliability than a global rating using a VAS.

Our study has several strengths. First, we designed it to replicate a typical situation during the conduct of systematic reviews or development of clinical practice guidelines in which the GRADE approach would be used. Raters assessed

Table 2. Inter-rater reliability for different number of raters using GRADE four categories of QoE or using a global judgment without using GRADE

System used to assess QoE raters groups	Reliability coefficient (95% CI)			
	One rater	Two individual raters	Three individual raters	Four individual raters
GRADE				
GWG	0.57 (0.47–0.67)	0.72 (0.63–0.79)	0.80 (0.73–0.86)	0.84 (0.78–0.89)
HRM	0.49 (0.38–0.54)	0.66 (0.56–0.75)	0.74 (0.65–0.81)	0.79 (0.71–0.85)
VAS without using GRADE				
GWG	0.16 (0.10–0.24)	0.27 (0.19–0.37)	0.36 (0.26–0.46)	0.42 (0.32–0.52)
HRM	0.19 (0.11–0.28)	0.31 (0.21–0.42)	0.41 (0.30–0.52)	0.47 (0.36–0.58)

Abbreviations: GRADE, Grading of Recommendations, Assessment, Development, and Evaluation; QoE, quality of evidence; 95% CI, 95% confidence interval; GWG, members of the GRADE working group; HRM, Health Research Methodology; VAS, visual analog scale.

evidence from several existing systematic reviews covering a range of clinical and public health areas as well as study designs. Second, we recruited two different groups of raters with different levels of expertise in the GRADE approach. All participants were required to complete two calibration exercises. These exercises were similar to the exercises used in any GRADE training workshop. It also represented reality in that assessing research evidence, but not GRADE per se, requires a high skill level similar to that of a specialized clinician (''one would not ask an internist to perform brain surgery despite a general knowledge about the anatomy and function of the brain''). This design allowed us to assess the effectiveness of the required training before raters could reliably use the GRADE approach. Third, recruiting two different groups helped us evaluate the potential effect of intellectual conflict of interest among the GRADE working group on the results. Some may argue that members of the GRADE working group are invested in showing good reliability. The fact that both groups showed substantial agreement provides evidence that this potential bias was not the reason for the observed reliability. Fourth, to further strengthen our results, we blinded investigators and data analysts to the identity of the raters (except for one investigator who held randomization code and was not involved in data analyses). This has likely minimized bias when assessing reliability and potentially minimized any theoretical pressure on raters to perform well. Finally, we used generalizability theory to analyze our results, which allowed us to adjust for the outcomes being nested in systematic reviews.

This study has some limitations. First, selection of well-conducted and well-reported systematic reviews may have affected the generalizability of our results as it may not reflect the reliability when using systematic reviews prepared with lower methodological rigor or when the reporting is suboptimal. However, we believe that for this study it was necessary to separate the variance related to reliability from any potential error variance from the quality of the systematic reviews. Furthermore, little emphasis should generally be placed on systematic reviews of lower methodological rigor so that high or low reliability becomes an irrelevant characteristic. Second, our study did not assess the reliability of the GRADE approach in systematic reviews that do not provide quantitative estimates of effects. Although GRADE can be used to make these judgments, its reliability in the judgment of the QoE of such outcomes remains to be explored. Third, the merit of the GRADE approach is in producing a systematic and transparent assessment of the evidence that may include ''close call'' judgments, that could affect the results of this study, but which we have not been able to take into account in the reliability analysis. Therefore, explanations for ratings remain a critical aspect of evidence assessment.

Although our study indicated substantial IRR of the GRADE approach, raters used the whole spectrum of categories of QoE from very low to high for 10 of the 16 outcomes assessed. This finding emphasizes the need to transparently report the reasons for downgrading or upgrading in the footnotes of every assessment.

The IRR did not improve when pairs of raters discussed their judgments and reached consensus. The lack of better agreement among consensus ratings was initially surprising. However, by forcing consensus and acknowledging that the ''true'' QoE when two raters disagree frequently lies in between, we introduced an additional source of error variance that contributed to the worsening of the reliability coefficients in the consensus ratings.

The only other published study we are aware of that evaluated IRR of a grading system is our study from 2005 [12]. For that study, based on the first iteration of GRADE, guidance for downgrading and upgrading the QoE was very limited. Additionally, we provided a kappa coefficient per outcome but did not provide an overall kappa or weighted kappa that is a more appropriate statistic for such a study. Other unpublished studies that assessed the reliability of grading systems have not clearly separated the variance that is introduced by the level of training, different evidence assessment approaches, or clarity of instructions to complete the exercises.

On the basis of our current findings, we suggest that two individual-trained raters assess the QoE and discuss it. If indeed more than one rater assesses the QoE, these raters should not force consensus after individually completing ratings but describe the reasons for disagreement and ensure that there are no misunderstandings or errors.

Realizing that the current four categories of GRADE limit the ability of averaging the QoE without reaching consensus, it is possible that increasing the number of categories of QoE may further improve the reliability of the GRADE approach and that needs to be evaluated. GRADE is increasingly being used to evaluate observational research and incorporating evidence from decision and economic analysis modeling. We are sometimes unable to clearly define the gradient between these types of evidence within the current four categories. The benefit of having four clear categories includes the ease of communication and conceptualization of the QoE assessments.

In summary, the results of this study support the presumption that using the GRADE approach is more reliable than intuitive judgments when assessing QoE about outcomes of health care interventions. Our findings support the notion that two individual raters without reaching consensus are sufficient to reliably assess the QoE using the GRADE approach.

Acknowledgments

Competing interests: No financial competing interest.

Some authors are involved in the development and dissemination of GRADE, and GRADE's success has a positive influence on their academic career.

Authors' contributions: H.J.S. conceived of the study. R.A.M. and H.J.S. designed the study. J.B., E.A.A., S.D.W., R.C., and G.H.G. contributed to the conception and design. R.A.M., G.N., and M.K. performed the statistical analysis. R.A.M. and H.J.S. drafted the manuscript. All the coauthors critically revised the manuscript and approved the final draft before submission. Y.F.-Y., S.C., M.H.M., G.E.V., T.L.,

G.H.G., V.S., X.S., C.W., P.N.P., E.L., K.T., I.K., H.A., J.J.M., A.C.A., I.F.N., S.G., M.C.E., A.C.-L., R.K., G.N., J.K., A.S., and R.B.-P. contributed as raters in the study.

Authors' information (intellectual conflict of interest): H.J.S. and G.H.G. are co-chairs of the GRADE working group. R.A.M., N.S., J.B., E.A.A., R.C., Y.F.-Y., S.C., M.H.M., G.E.V., T.L., G.H.G., V.S., X.S., C.W., P.N.P., E.L., K.T., I.K., H.A., and J.J.M. are members of the GRADE working group. R.A.M., N.S., A.C.A., I.F.N., S.G., M.-C.E., A.C.-L., R.K., G.N., J.K., A.S., M.K., and R.B.-P. are students in the Health Research Methodology program at McMaster University.

References

[1] Available at http://www.gradeworkinggroup.org/. Accessed June 1, 2012.
[2] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924−6.
[3] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395−400.
[4] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). J Clin Epidemiol 2011;64:407−15.
[5] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. J Clin Epidemiol 2011;64:1303−10.
[6] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. J Clin Epidemiol 2011;64:1294−302.
[7] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. J Clin Epidemiol 2011;64:1283−93.
[8] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. J Clin Epidemiol 2011;64:1277−82.
[9] Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64:1311−6.
[10] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401−6.
[11] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383−94.
[12] Atkins D, Briss PA, Eccles M, Flottorp S, Guyatt GH, Harbour RT, et al. Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. BMC Health Serv Res 2005;5:25.
[13] West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, et al. Systems to rate the strength of scientific evidence. Evid Rep Technol Assess (Summ) 2002;47:1−11.
[14] Fernandes RM, Bialy LM, Vandermeer B, Tjosvold L, Plint AC, Patel H, et al. Glucocorticoids for acute viral bronchiolitis in infants and young children. Cochrane Database Syst Rev 2010;10:CD004878.
[15] Towheed TE, Maxwell L, Anastassiades TP, Shea B, Houpt J, Robinson V, et al. Glucosamine therapy for treating osteoarthritis. Cochrane Database Syst Rev 2005;2:CD002946.
[16] Rosner S, Hackl-Herrwerth A, Leucht S, Lehert P, Vecchi S, Soyka M. Acamprosate for alcohol dependence. Cochrane Database Syst Rev 2010;9:CD004332.
[17] Ducharme F, Schwartz Z, Hicks G, Kakuma R. Addition of anti-leukotriene agents to inhaled corticosteroids for chronic asthma. Cochrane Database Syst Rev 2004;(2):CD003133.
[18] Chauhan BF, Ducharme FM. Anti-leukotriene agents compared to inhaled corticosteroids in the management of recurrent and/or chronic asthma in adults and children. Cochrane Database Syst Rev 2012;5:CD002314.
[19] Dieleman JM, van Paassen J, van Dijk D, Arbous MS, Kalkman CJ, Vandenbroucke JP, et al. Prophylactic corticosteroids for cardiopulmonary bypass in adults. Cochrane Database Syst Rev 2011;5:CD005566.
[20] Liu BC, Ivers R, Norton R, Boufous S, Blows S, Lo SK. Helmets for preventing injury in motorcycle riders. Cochrane Database Syst Rev 2008;1:CD004333.
[21] Available at http://hrm.mcmaster.ca/. Accessed June 1, 2012.
[22] Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.
[23] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159−74.

Appendix A: Example of GRADE evidence profile table

Self-management for patients with chronic obstructive pulmonary disease

***Patient or population*: patients with chronic obstructive pulmonary disease**
***Settings*: primary care, community, outpatient**
***Intervention*: self-management[a]**
***Comparison*: usual care**

Quality assessment							No. of patients		Effect			
No. of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other	Self-management[a]	Usual care	Relative (95% CI)	Absolute	Quality	Importance
Quality of life (follow-up of 3–12 mo; measured with St George's Respiratory questionnaire; range of scores, 0–100 [worse]; better indicated by lower values)												
7	Randomized trials	No serious risk of bias	No serious inconsistency	No serious indirectness	No serious imprecision	Reporting bias[b]	381	317	—	Mean dyspnea 2.58 lower (5.14–0.02 lower)	⊕⊕⊕○ Moderate	Critical
Dyspnea (follow-up of 3–6 mo; measured with Borg Scale; range of scores, 0–10 [worse]; better indicated by lower values)												
2	Randomized trials	Serious[c]	No serious inconsistency	No serious indirectness	Serious[d]	None	66	78	—	Mean dyspnea 0.53 lower (0.96–0.1 lower)	⊕⊕○○ Low	Critical
Number and severity of exacerbations[e] (better indicated by lower values)												
3	Randomized trials					None	591	—	—[e]	Not pooled[e]		Critical
Respiratory-related hospital admissions (follow-up of 3–12 mo)												
8	Randomized trials	No serious risk of bias	No serious inconsistency	Serious[f]	No serious imprecision	None	95/528 (18%)	10%[g]	OR, 0.64 (0.47–0.89)	Three fewer per 100 (from one to five fewer)	⊕⊕⊕○ Moderate	Critical
								50%[g]		11 fewer per 100 (from 3 to 18 fewer)		
Emergency department visits for lung diseases (follow-up of 6–12 mo; better indicated by lower values)												
4	Randomized trials	No serious risk of bias	No serious inconsistency	No serious indirectness	Serious[d]	None	183	145	—	Mean dyspnea 0.1 higher (0.2 lower to 0.3 higher)	⊕⊕⊕○ Moderate	Important

(*Continued*)

Appendix A. Continued

Self-management for patients with chronic obstructive pulmonary disease

***Patient or population*: patients with chronic obstructive pulmonary disease**
***Settings*: primary care, community, outpatient**
***Intervention*: self-management[a]**
***Comparison*: usual care**

Quality assessment							No. of patients		Effect			
No. of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other	Self-management[a]	Usual care	Relative (95% CI)	Absolute	Quality	Importance
Doctor and nurse visits (follow-up of 6–12 mo; better indicated by lower values)												
8	Randomized trials	No serious risk of bias	Serious[h]	No serious indirectness	No serious imprecision	None	334	295	—	Mean dyspnea 0.02 higher (1 lower to 1 higher)	⊕⊕⊕○ Moderate	Important

Abbreviations: GRADE, Grading of Recommendations, Assessment, Development, and Evaluation; 95% CI, 95% confidence interval; OR, odds ratio.

[a] Self-management is a term applied to any formalized patient education program aimed at teaching skills needed to carry out medical regimens specific to the disease, guide health behavior change, and provide emotional support for patients to control their disease and live functional lives. Of the 14 studies, there were four in which the education delivery mode consisted of group education; nine that were individual education; and one study that was written education material only. In six studies, the use of an action plan for self-treatment of exacerbations was assessed.

[b] Seven other studies were not pooled, and some showed nonsignificant effects.

[c] No allocation concealment in one study. Incomplete follow-up.

[d] Sparse data.

[e] Different definitions of exacerbations were used, and studies could not be pooled.

[f] Two studies with very severe chronic obstructive pulmonary disease patients weighted heavily in meta-analysis. Therefore, there is some uncertainty with the applicability of effect to all risk groups.

[g] The low- and high-risk values are the two extreme numbers of admissions in the control groups from two studies (8% was rounded to 10% and 51% to 50%).

[h] Unexplained heterogeneity.

Appendix B: Example of GRADE summary of findings table

Self-management for patients with chronic obstructive pulmonary disease

***Patient or population*: patients with chronic obstructive pulmonary disease**
***Settings*: primary care, community, outpatient**
***Intervention*: self-management[a]**
***Comparison*: usual care**

Outcomes	Illustrative comparative risks (95% CI)		Relative effect (95% CI)	No. of participants (studies)	Quality of the evidence (GRADE)	Comments
	Assumed risk usual care	Corresponding risk self-management				
Quality of Life St George's Respiratory Questionnaire. Scale from 0 to 100 (follow-up: 3–12 mo)	The mean quality of life ranged across control groups from 38 to 60 points	The mean quality of life in the intervention groups was 2.58 lower (5.14–0.02 lower)		698 (7)	⊕⊕⊕○ Moderate[b]	Lower score indicates better quality of life. A change of less than four points is not shown to be important in patients
Dyspnea Borg Scale. Scale from 0 to 10 (follow-up: 3–6 mo)	The mean dyspnea ranged across control groups from 1.2 to 4.1 points	The mean dyspnea in the intervention groups was 0.53 lower (0.96–0.1 lower)		144 (2)	⊕⊕○○ Low[c,d]	Lower score indicates improvement
Number and severity of exacerbations[e]	See comment	See comment	Not estimable[e]	591 (3)	See comment	Effect is uncertain
Respiratory-related hospital admissions (follow-up: 3–12 mo)	**Low-risk population[g]**		OR, 0.64 (0.47–0.89)	966 (8)	⊕⊕⊕○ Moderate[f]	
	10 per 100	7 per 100 (5–9)				
	High-risk population[g]					
	50 per 100	39 per 100 (32–47)				
Emergency department visits for lung diseases (follow-up: 6–12 mo)	The mean emergency department visits ranged across control groups from 0.2 to 0.7 visits per person per year	The mean emergency department visits in the intervention groups was 0.1 higher (0.2 lower to 0.3 higher)		328 (4)	⊕⊕⊕○ Moderate[d]	
Doctor and nurse visits (follow-up: 6–12 mo)	The mean doctor and nurse visits ranged across control groups from one to five visits per person per year	The mean doctor and nurse visits in the intervention groups was 0.02 higher (1 lower to 1 higher)		629 (8)	⊕⊕⊕○ Moderate[h]	

Abbreviations: GRADE, Grading of Recommendations, Assessment, Development, and Evaluation; 95% CI, 95% confidence interval; OR, odds ratio.

The basis for the assumed risk (eg, the median control group risk across studies) is provided in the footnotes. The corresponding risk (and its 95% CI) is based on the assumed risk in the comparison group and the relative effect of the intervention (and its 95% CI).

[a] Self-management is a term applied to any formalized patient education program aimed at teaching skills needed to carry out medical regimens specific to the disease, guide health behavior change, and provide emotional support for patients to control their disease and live functional lives. Of the 14 studies, there were four in which the education delivery mode consisted of group education; nine that were individual education and one study that was written education material only. In six studies, the use of an action plan for self-treatment of exacerbations was assessed.

[b] Seven other studies were not pooled, and some showed nonsignificant effects.

[c] No allocation concealment in one study. Incomplete follow-up.

[d] Sparse data.

[e] Different definitions of exacerbations were used, and studies could not be pooled.

[f] The low- and high-risk values are the two extreme numbers of admissions in the control groups from two studies (8% was rounded to 10% and 51% to 50%).

[g] The low- and high-risk values are the two extreme numbers of admissions in the control groups from two studies (8% was rounded to 10% and 51% to 50%).

[h] Unexplained heterogeneity.

Appendix C: The survey

Dear colleague,

Thank you for volunteering to participate in our study.

The Hamilton Health Sciences/Faculty of Health Sciences Research Ethics Board has evaluated this study, and the requirement for informed consent was waived.

This survey should take about 5 minutes to complete.

You will answer questions about your previous experience and training in the GRADE approach to assessing QoE. For questions that ask about the number of times, please use your best estimate.

1. Please indicate below how many times you have done the following:

Question	Never	1–2 times	3–5 times	6–8 times	9–10 times	11–15 times	>15 times
Prepared GRADE profile or SoF table	☐	☐	☐	☐	☐	☐	☐
Used GRADEpro	☐	☐	☐	☐	☐	☐	☐
Attended a GRADE working group meeting	☐	☐	☐	☐	☐	☐	☐
Attended a GRADE workshop (or educational seminar)	☐	☐	☐	☐	☐	☐	☐
Facilitated a GRADE workshop	☐	☐	☐	☐	☐	☐	☐
Been a guideline panelist	☐	☐	☐	☐	☐	☐	☐
Lead work in a guideline	☐	☐	☐	☐	☐	☐	☐
Been involved in a systematic review	☐	☐	☐	☐	☐	☐	☐
Conducted (lead) a systematic review	☐	☐	☐	☐	☐	☐	☐

2. Please indicate your familiarity with grading systems. Check all that apply.
   - Have used GRADE
   - Been involved in developing GRADE
   - Been one of the leaders in developing GRADE
   - Have used grading systems other than GRADE
   - Been involved in developing grading systems other than GRADE
   - Been one of the leaders in developing grading systems other than GRADE.
3. Please indicate your training in health research methodology
   - Never done formal training
   - Done some formal training but do not have a graduate degree
   - Earned an MSc degree
   - Earned a PhD degree.
4. I am a/an (please check all that apply)
   - Epidemiologist
   - Biostatistician
   - Clinician (please write specialty under other)
   - Others (please specify background under other)

Other (please specify)

Appendix D: Summary of outcomes evaluated and their quality of evidence rating

Table A. Summary of the outcomes evaluated by raters

Outcome (type)	Type	Study design	Number of studies
Return to any drinking	Categorical	RCT	24
Cumulative abstinence duration	Continuous	RCT	19
Return to heavy drinking	Categorical	RCT	6
Dropout because of adverse event	Categorical	RCT	19
Exacerbation requiring systemic steroids	Categorical	RCT	18
Change from baseline quality of life at 6 wk	Continuous	RCT	3
Overall withdrawal	Categorical	RCT	19
Change in proportion of symptom-free days at 6 wk	Continuous	RCT	5
Admission within 7 d	Categorical	RCT	6
Admission within 1 d	Categorical	RCT	10
Length of stay	Continuous	RCT	8
Clinical score inpatients at 3–6 hr	Continuous	RCT	1
Death	Categorical	Observational	4
Head injury	Categorical	Observational	6
Neck injury	Categorical	Observational	12
Facial injury	Categorical	Observational	8

Abbreviation: RCT, randomized controlled trial.

Table 1B. Ratings of the QoE by raters of the GRADE working group

	Number of raters that rated			
Outcome	**Very low**	**Low**	**Moderate**	**High**
1	1	6	6	2
2	2	4	8	1
3	2	0	5	8
4	0	3	11	1
5	1	6	3	5
6	1	2	5	7
7	1	6	6	2
8	1	2	6	6
9	1	2	10	2
10	1	3	6	5
11	1	2	7	5
12	1	5	7	2
13	2	10	3	0
14	0	5	9	1
15	14	1	0	0
16	10	3	2	0

Abbreviations: QoE, quality of evidence; GRADE, Grading of Recommendations, Assessment, Development, and Evaluation.

Table 2B. Ratings of the QoE by raters of the students in the HRM program

	Number of raters that rated			
Outcome	**Very low**	**Low**	**Moderate**	**High**
1	0	3	7	0
2	0	4	6	0
3	0	1	5	4
4	0	1	7	2
5	2	2	2	4
6	1	1	2	6
7	2	3	2	3
8	2	0	1	7
9	1	1	6	2
10	0	1	6	3
11	0	0	6	4
12	0	4	3	3
13	2	4	3	1
14	2	2	4	2
15	9	1	0	0
16	7	1	2	0

Abbreviations: QoE, quality of evidence; HRM, Health Research Methodology.

Table 3B. Ratings of the QoE by pairs of raters of the GRADE working group

	Number of raters that rated			
Outcome	**Very low**	**Low**	**Moderate**	**High**
1	0	4	3	0
2	0	4	2	1
3	0	2	2	3

(*Continued*)

Table 3B. Continued

	Number of raters that rated			
Outcome	**Very low**	**Low**	**Moderate**	**High**
4	0	3	3	1
5	1	2	1	3
6	0	0	4	3
7	1	2	4	0
8	1	0	3	3
9	0	2	5	0
10	0	1	5	1
11	0	2	3	2
12	0	3	3	1
13	2	3	2	0
14	0	2	4	1
15	6	1	0	0
16	5	1	0	1

Abbreviations: QoE, quality of evidence; GRADE, Grading of Recommendations, Assessment, Development, and Evaluation.

Table 4B. Ratings of the QoE by pairs of raters of the HRM students

	Number of raters that rated			
Outcome	**Very low**	**Low**	**Moderate**	**High**
1	0	3	2	0
2	0	3	2	0
3	0	1	2	2
4	0	1	3	1
5	0	2	1	2
6	0	2	1	2
7	2	0	2	1
8	1	1	1	2
9	0	0	4	1
10	0	1	2	2
11	0	0	2	3
12	0	2	2	1
13	2	1	2	0
14	1	2	1	1
15	5	0	0	0
16	4	1	0	0

Abbreviations: QoE, quality of evidence; HRM, Health Research Methodology.

Appendix E: Summary of variance estimates

Group	V(O:S)	V(S)	V(R)	V(E)
VAS HRM	0	75	174	156
VAS GWG	0	43	164	70
GRADE HRM (single raters)	0.149	0.247	0.339	0.075
GRADE GWG (single raters)	0.259	0.179	0.314	0.021
GRADE HRM (pairs)	0.326	0.108	0.449	0.396
GRADE GWG (pairs)	0.128	0.171	0.276	0.397

Abbreviations: V(O:S), variance of outcome nested in systematic review; V(S), variance of systematic review; V(R), variance of rater; V(E), error variance; VAS, visual analog scale; HRM, Health Research Methodology; GWG, members of the GRADE working group.