Author's response to reviews

Title: Systems for grading the quality of evidence and the strength of recommendations II: Pilot study of a new system

Authors:

David Atkins (DAtkins@AHRQ.GOV)
Peter A Briss (pxb5@cdc.gov)
Martin Eccles (Martin.Eccles@newcastle.ac.uk)
Signe Flottorp (signe.flottorp@nchs.no)
Gordon H Guyatt (guyatt@mcmaster.ca)
Robin T Harbor (r.harbour@sign.ac.uk)
Suzanne Hill (hillsu@mail.newcastle.edu.au)
Roman Jaeschke (aeschke@mcmaster.ca)
Alessandro Liberati (alesslib@tin.it)
Nicola Magrini (n.magrini@ausl.mo.it)
James Mason (jmason123@orange.net)
Dianne O'Connell (dianneo@nswcc.org.au)
Andrew D Oxman (oxman@online.no)
Bob Phillips (bob.phillips@doctors.org.uk)
Holger Schunemann (hjs@buffalo.edu)
Tessa Tan-Torres Edejer (tantorrest@who.ch)
Gunn E Vist (gunn.vist@nchs.no)
John W Williams (jw.williams@duke.edu)

Version: 3 Date: 20 December 2004

Author's response to reviews: see over
Responses to BMC Editorial and reviewers comments re:
A pilot study of a new system for grading the quality of evidence and the strength of recommendations. 4352636028844837.

We agree to change the titles of this and the associated manuscript (A critical appraisal of systems for grading the quality of evidence and the strength of recommendations. 9976868452873787.

Hence, titles are now: Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches.
&
Systems for grading the quality of evidence and the strength of recommendations II: Pilot study of a new system.

Apologies for the lateness of this response.

Q: Question or comment from referee.
R- our response

Reply to comments from Michael Bigby:

Major compulsory revisions:

QP3: It is impossible to know what was done reading the abstract.
R – The methods section of the abstract has been expanded to include following additional information:
A GRADE evidence profile consists of two tables: a quality assessment and a summary of findings. Twelve evidence profiles were used in this pilot study. Each evidence profile was made based on information available in a systematic review. Seventeen people were given instructions and independently graded the level of evidence and strength of recommendation for each of the 12 evidence profiles. For each example judgements were collected, summarised and discussed in the group with the aim of improving the proposed grading system. Kappas were calculated as a measure of chance-corrected agreement for the quality of evidence for each outcome for each of the twelve evidence profiles.

QP4: The process of developing the system needs to be described in more detail.
R – The Background section has been expanded to include the following additional information:
Based on the critical assessment of existing approaches, the agreement we had reached about the key elements that should be included in an approach for grading the level of evidence and strength of recommendations and our previous experiences we put together a suggestion for a grading system. We then applied the suggested system to a series of examples, and discussed and revised the system based on this experience and consideration of other examples. Examples were selected to challenge our thinking. All the examples used in this pilot study were questions about interventions.

QP5: The procedure for developing “evidence profiles” should be detailed.
R – The methods section has been expanded to include additional information:

Old text:
The evidence summaries included information about study design, study quality, the consistency of results across studies and the directness of the evidence for each main outcome. The summaries of findings included relative and absolute measures of effect for each outcome.

New text:
The quality assessment table was designed such that the quality of each outcome was evaluated separately. For each outcome, the table contained information regarding the number of studies that had reported the outcome, information about the study design (RCTs or observational studies) and the quality of the studies that reported on that outcome (was there any limitations in the design or conduct of these studies). Also included in the quality assessment table was information about the consistency of the results across studies for each outcome and information regarding directness of the study population, outcome measure, intervention and comparison. The summary of findings table was also designed such that each outcome was presented separately. For each outcome information are presented about both the experimental and the control group patients, for dichotomous outcomes the number of events and the total number of participants, and for continuous outcomes the means (standard deviation) and the number of patients were presented. Also included in the summary of findings table is information about the effect, relative effect (95% confidence interval) and absolute effect for each outcome.

QP7: 11 Greater participation by the investigators is needed.

R – All of the authors contributed to the pilot study as described under Contributions, but only 11 of the 17 completed the questionnaire about the sensibility and understandability of the approach.

QP8: The levels of agreement need to be defined and kappas calculated.

R –
Added to the Methods section:
For each example the kappa agreement was calculated [14] for the 17 graders across the four levels for the quality of evidence across outcomes for each example (number of outcomes per example range from two to seven), across all outcomes (46) and for the judgements about overall quality of the evidence (12).

Added to the Results section:
The kappa statistics for each question are shown in Table 5. The number of outcomes per example range from two to seven and the kappas ranged from 0 to 0.82. In some instances, the agreement among the graders was slightly worse than by chance as indicated by the negative kappa values seen in Table 5. The kappa across the 46 outcomes included in the calculation was 0.395 (SE 0.008). Kappa for agreement beyond chance for the 12 final judgements about the quality of evidence was 0.270 (SE 0.015).

Added to the Discussion section:
Guideline generation includes judgement. Individual, residual judgements will impact on the agreement we measured in this study. Thus, lower kappa values are expected. Further refinement of the GRADE system and additional instructions will improve agreement.

Minor essential revisions: None
Reply to comments from Benjamin Djulbegovic

Q: My critique regarding the paper itself is relatively minor: the authors need to explain how consensus was “measured” and how they defined categories of “high”, “intermediate” etc levels of consensus. I understand that there are some methodological problems how to measure a level of agreement among 17 peoples, but this group should be able to make a comment about it.
R – We have included kappas, as noted above.

Reply to comments from Anja Tuulonen

Major compulsory Revisions:

Q: The authors should define what they mean by good, moderate and poor agreement and use the same criteria in both papers A and B.
R – We have included kappas, as noted above.

Q: Since the two papers of the authors represent a continuum of the same process they could be condensed into one paper, or –alternatively – be published together as 2 papers with the same main title and descriptive subtitles, e.g. marked with A and B.
R – Titles changed according to BioMed Central editors suggestion, please see above.

Q: In general, paper B (“A pilot study…”) is quite laborious to read and it takes quite a while before the reader understands what actually has been done. I The main reason for confusion is the fact that it is difficult to figure out what information was given to the evaluators before their analysis and what were the actual results. I suggest that the tables (and appendices) were be given to the readers in chronological order, i.e. as if he or she would be one of the evaluators: first the data collection form, then the instructions how to grade and finally an example of the evidence profile.
R – The fully prepared evidence profile was given to the evaluators together with the instructions on how to apply the GRADE approach and the data collection form for each of the 12 examples. In the Methods section we wrote:

“Instructions and a form for recording each judgement were included with each example (Appendix). The judges were instructed to apply the approach without second guessing the information presented in the evidence profile or the approach.”

Q: In Instructions, “serious flaws”, “important inconsistency”, “directness” and “associations” are not defined. In “Rules of thumb”, e.g. “90-100% of people likely to do it” etc. Whom does “people” refer to? Clinicians?
R – By ‘people’ we meant those making a decision to either apply or submit oneself to the intervention if they were informed about the evidence. However, the lack of precise definitions is one of the factors for the poor agreement, as noted in the discussion:
“Many of the disagreements were a direct result of a lack of information. We concluded that there is a need for detailed additional information in evidence profiles, and have modified the evidence profiles accordingly. When we have found an empirical basis or compelling
arguments, we have also provided precise definitions. For example, we have agreed on a basis for defining strong and very strong associations. However, in many cases we continue to rely on judgement. We have addressed this by always including the rationale for such judgements in footnotes attached to the evidence profile.”

Methods
Q: Were the 17 evaluators also the authors of the paper? Were 12 of them the same people as in paper A, i.e. were they experts in grading the evidence (please, see similar comment for paper A)?
R – Yes & yes, from the Contributions section:
“All authors except GEV judged the quality of the evidence and strength of recommendation based on information presented in the evidence profiles.”

Q: In the second paragraph on page 5, it is stated “we prepared the evidence profile…” Who were “we”, that is did the same people prepare the evidence tables and grade the evidence? If yes, the possible bias should be discussed.
R – GEV with some help from the other authors, prepared the evidence profiles, GEV was not one of the graders in this pilot study. From the contributions section:
“GEV prepared the first draft of this article, had primary responsibility for preparing the evidence profiles used in the study, and coordinated the study.”

Q: It should be stated in the method section that other modalities than treatment and prevention were excluded in this pilot study.
R – Added to the Background section:
“All of the examples used in this pilot study were questions about interventions.”

It is reported that the number of evaluators was different in different phases (ranging from 17 to 11 which corresponds a 35% reduction). It would be logical that also in qualitative studies, the changing number of observers might have a possible effect on the results.

Q: How and by whom was the questionnaire created (pages 7-8), by consensus? The components and judgements should be defined for readers.
R – Added to the Methods section:
“The questionnaire was adapted from Feinstein [15] and the 16 statements were:”

Q: On page 8, it is referred to four type components, but on pages 6-7 five judgement categories are listed.
R – Thank you, corrected.

Results
Q: After defining what the authors mean by good, moderate, and poor (the same criteria as in paper A), the information of tables 3-6 could probably be condensed into one table.
R – As a measure of agreement, kappas have been calculated, please see above response to Michael Bigby, QP8.
We think that the actual results are much more informative than the calculated kappa agreements because of the nature of the study and the limitations of kappa for so many raters and would prefer the standard results tables.
Q: The number of disagreed cases was 5 for outcomes (10%), two for relative importance (4%), and three for the balance between benefits and harms (6%). Were these cases overlapping or cumulative (altogether 10 cases which equals to 21% disagreement rate)?

R – For four of the disagreed outcomes (3 in question 9 and 1 in question 6) this disagreement did not affect the judgements. The disagreement about the outcome in question 3 was a consequence of the need for additional information before agreement and this spilled over into following judgements about the balance between benefit and harm and the recommendation. The 2 disagreements about relative importance both related to question 10 was also the source of disagreement for the judgements about the balance between benefit and harm and the recommendation. For question 12 there is disagreement within the group about the balance between the benefits and harms and this spills over into the recommendation.

Disagreement Quality in Q3, Q6 and Q9
Disagreement Relative importance Q10
Disagreement Balance between benefit and harm Q3, Q10, Q12
Disagreement Recommendation: Q3, Q10, Q12

Q: The last paragraph of the results on page 12 would read better as a table.
R – We disagree because there are many Tables already. Thus, we have not made this change.

Tables
Q: In general, the title texts are scanty and at their current stage the tables are not self-explanatory. The headings should be more informative, e.g. it should be clearly stated in the tables which data was given for the evaluators before the analysis.

R – Table 1, and 11 other similar tables representing the other 11 examples and the Appendix (and another 11 similar appendix’) were given to the evaluators. All the other Tables are ‘results’

We made several changes to the Tables.

Titles altered:

Table 1, old - Example of an evidence profile used in the pilot study
Table 1, new - Example of an evidence profile given to the evaluators for them to grade in the pilot study

Table 2, old - Judgements made for Example 1 of the pilot study. Should depressed patients in primary care be treated with SSRIs rather than tricyclics?
Table 2, new – Results, summary of the judgements made by the 17 evaluators for Example 1 of the pilot study.

Table 3, old - Outcome quality
Table 3, new – Results, summary of the judgements made by the 17 evaluators of the quality for each of the outcomes presented in the 12 examples in the pilot study

Table 4, old - Overall quality
Table 4, new – Results, summary of the judgements made by the 17 evaluators of the overall quality in the 12 examples in the pilot study

Table 5, old - Kappa agreement
Table 5, new – Results, kappa agreement among the evaluators for each of the 12 examples in the pilot study

Table 6, old - Balance between benefits and harms
Table 6, new – Results, summary of the judgements made by the 17 evaluators about the balance between benefits and harms for each of the 12 examples in the pilot study

Table 7, old - Recommendation - consensus in the pilot study examples
Table 7, new – Results, summary of the recommendations made the 17 evaluators for each of the 12 examples in the pilot study

Table 8, old - Modified GRADE evidence profile
Table 8, new – Example of a modified GRADE evidence profile.
Table 8 is what Table 1 became when including the improvements made based on the pilot study experience

Table 9, old - Modified quality assessment criteria
Table 9, new – Modified GRADE quality assessment criteria

Q: In Table 1 and Data collection form it reads “Should depressed patients in primary care be treated with SSRI OR tricyclics? Otherwise, the recommendations “Do it, probably do it …” etc. is irrelevant for this question. In Table 7 the phrasing is probably correct “… with SSRIs RATHER THAN tricyclics”? 
R – Corrected, thank you.

Q: The acronyms are defined only in Table 7 but are missing in Table 1. In addition to NNT, shouldn’t it read also NNH in Table 1 (NNT/NNH)?
R – Corrected, thank you.

Q: The scaling is missing of Relative importance in Table 1. Was Relative Importance given beforehand to the evaluators in Table 1? And in addition, the graders evaluated the relative importance (Table 2)?
R – The relative importance was not given in the Table when presented to the evaluators, and should not have been in Table 1, now removed. It was the 17 evaluators that graded the relative importance that is presented with the other results in Table 2.

Q: Table 3 should have subheadings for individual ratings (High, intermediate…). How was the information reported in Table 3 given to the graders? This was not reported earlier.
R – Table 3 reports the results from the 17 evaluators, title of Table 3 is altered as shown above.

Discussion
Q: If the evaluators were experts in EBM grading systems and still agreed poorly, the applicability of the grading system in the hands of less experienced evaluators should be more critically discussed. Even after consensus, the disagreement of recommendation was 25% (Table 6).
R – The aim of the GRADE approach is not to make everyone agree, but to aid the process by ensuring the decisions are made explicit so that sources of disagreement and uncertainty are known. Decisions about recommendations are influenced by values and preferences. Values and preferences differ between those making recommendations. Because there is little
information on values and preferences and how to integrate them in the decision about recommendations, this agreement is expected. Disagreements about quality can be reduced through adherence to the system. This is facilitated by the GRADE profiler software, which we have subsequently developed. Disagreements about quality, as a rule, can be more easily resolved because the reasons for disagreement are made transparent.

Q: It is also interesting that in spite of poor agreement in most dimensions the most graders stated that the approach was clear and understandable, and successfully discriminated between different qualities of evidence.
R – This reflects the discussions of our independent ratings and the points above, as well as our independent ratings.

Q: One conclusion of the paper might be that a similar approach (independent individual evaluations using a systematic grading tool) should be used in preparing the guidelines and the disagreement rates before the consensus should be reported.
R – Added to the Discussion:
“IT may be helpful to guideline panels and others to use this approach before making decisions/recommendations.”

Q: Since the authors are trying define the evidence, the reader expects the same systematic approach in qualitative studies as expected in the clinical studies. However, the problems related to the evaluation of GRADE approach remind me about the problems in evaluating diagnostics in chronic diseases. The GRADE approach is a tool (diagnostic test) to find the evidence (disease) . However, the evidence (chronic disease) does not always follow a dichotomous model, i.e. evidence vs. no evidence (disease vs. no disease). In stead, there is a large grey zone in which we simply do not know whether we have the evidence or not. And still we have to act in every-day life. As in diagnostic tests, whether we have the evidence (disease) depends also on the definition ‘good evidence’ (normal limits).
R – We agree and have therefore used four categories for quality of evidence, rather than “good evidence” versus “no evidence”. The approach we have taken to assessing sensibility is derived directly from criteria for assessing the sensibility of clinical measurements.

Q: Page 14, last paragraph, I suggest the following correction:
“The GRADE system represents our current thinking about how to avoid reduce errors…”
R – Change made, thank you.

Q: The authors report that the modified GRADE evidence (Tables 7 and 8) represents considerable improvements to the system. Although most probable, this is, of course, not known since there is no data available using the improved criteria. Based on paper A, I expected that the new criteria would combine the advantages and clinically understandable terms of the 6 different criteria presented in paper A. I was left with feeling that somehow the “more clinically familiar” language of these 6 criteria disappeared with the new criteria. Or maybe these expectations are just biased by own experiences.
R – We think that we have combined the advantages and ended up with an improved approach. However, this remains to be verified. We agree.

Q: Table 8 is not quite clear. How the number of studies is taken into account (study design is in singular format).
R – Study design is either RCT or Observational. When there is evidence from RCTs then the consensus is that there is no need to include observational studies within each outcome. An
exception is a case where the RCTs have sufficient limitations that would make their quality lower than that of observational studies based on the GRADE system. For example, if there is imprecise or sparse data from small or poor quality RCTs so that observational studies provide important additional information, then the overall quality would be based on observational data.

Q: How should one apply “Higher if” in the case where the quality of evidence is already high?
R – in that case, the quality remains ‘High’. Added to Table 9:
“The highest possible score is High (4) and the lowest possible score is Very low (1). Thus, for example, randomised trials with a strong association would not move up a grade.”

Q: How does one get evidence of Reporting bias?
R – One can get strong indications from funnel plots and statistical tests for reporting bias. One can also have strong and substantiated suspicions that one explains in a foot note. It is important to be explicit about it.

Minor Essential Revisions

Q: The spelling mistakes should be corrected.
R – Spell check completed.

Reply to comments from Joseph Watine

Major Compulsory Revisions: None.

Minor Essential Revisions:
Q1: The question posed by the authors is new but they only answer to a part of it. In fact, what the authors have done is “a pilot study of a new system for grading the quality of evidence and the strength of recommendations about the effectiveness of therapeutic and prophylactic interventions”. The system which has been pilot-tested is their preliminary and “oldest” system. Their modified and “newest” system, which they propose at the end of their manuscript (and which is also described in a paper currently “in press” in the BMJ) is not the system which is pilot tested in this study.

Also, in the Background section, mention should be made of the paper “in press” in the BMJ.
R – That is correct. The system that was pilot tested in this manuscript was revised based on the results of the pilot test. Our revised system is presented in the ‘BMJ paper’.

We have added the following changes in the abstract:
The aim of this study was to pilot test and further develop the GRADE approach to grading evidence and recommendations.

This was already clear in the Background section:
“The aims of the pilot study were to test whether the approach is sensible relative to diverse examples of evidence and recommendations, and to agree on necessary changes to the approach, decision rules, and changes in how the evidence profiles used in the pilot study were constructed.”

Added to the Background section:
The revised approach is described elsewhere [15].
Added to Conclusions:
Based on the results of this pilot study we have been able to considerably improve our system for grading the quality of evidence and strength of recommendation [15].

Q2: The Methods section could perhaps be divided into under-sections.
R – The following under-sections have been added: Evidence profiles; Questions and judgements; Sensibility and understandability.

Q: At the very beginning of the Methods section, it is written that “17 people independently judged the quality of evidence”. There are 18 authors for this manuscript. Who is the author who did not participate in the judgement of the quality of evidence? Why does this 18th author deserves to be a co-author?
R – The 18th author is GEV, she had primary responsibility for preparing the 12 evidence profiles and she co-ordinated the study.
The following has been added to the Contributions section:
All authors except GEV judged the quality of the evidence and strength of recommendation based on information presented in the evidence profiles.

Q: How were the evidence profiles made?
R – This information has been added to the text, see QP5 above.

Q: Whom composed these evidence profiles?
R – GEV with some help from the other authors where necessary. See Contributions.

Q: What is the utility of the whole grading system if the systematic reviews, on which the evidence profiles are based, and therefore on which all the grading processes are based, are not properly done? Which system, if any, did the authors use to judge the methodological rigor of the 12 systematic reviews that they used to compose the 12 evidence profiles? As the authors probably know, many such systems have been proposed and published, e.g. that of McAlister FA, Clark HD, van Walraven C, Straua SE, Lawson FM, Moher D, Mulrow CD [The medical review article revisited: has the science improved? Ann Intern Med 1999 Dec 21;131(12):947-51]. If such a system has been used, this should be said in the Methods section, and probably discussed in the Discussion section. The reference of the system should also be quoted.
R – The purpose of this paper was to pilot test our system for grading the quality of evidence and strength of recommendation. The assessment of systematic reviews is beyond the scope of this paper.

Added to the Methods section:
“For the purpose of testing our grading approach in this pilot study we made the assumption that the systematic reviews that we used were all well conducted. The examples we used and presented here were selected to test our new approach, not with an intention of making actual recommendations for a specific setting based on up-to-date systematic reviews.

Already in the Discussion section:
“Much of the information we found lacking was missing in these original systematic reviews, particularly information about harms and side effects. It was outside of the scope of this study to systematically collect this information. However, systematic reviews of evidence of harms, as well as benefits, are essential for guidelines development panels."

Q4: No keywords are provided
R – The keywords that you suggested to the companion paper are used:
Evidence-based health care, levels of evidence, practice guidelines, strength of
recommendations, systematic reviews.

Q: In Table 7, summary of findings: “absolute” (not “absolute”).
R – Thank you, corrected.

Q: In Table 8, the word “moderate” is used in the table, whereas in the legend, it is the word
“intermediate” which is used.
R – Thank you, corrected.

Q5: Couldn’t the authors discuss the fact that methodological quality of a systematic review
that is used to compose an evidence-profile might be another item to be included in the
evidence profile?
R – One of the conclusions in paper A was that systematic reviews should not be included in a
hierarchy of evidence. As noted above, critical appraisal of systematic reviews upon which
evidence profiles are based is beyond the scope of this paper.

Q: The issue of patients’ choice is perhaps not discussed as one might have hoped.
R – This paper reports on a pilot study of an approach for grading the quality of evidence and
strength of recommendation. A discussion of patients’ choice is beyond the scope of this paper.

Q: In the Table of the appendix, page 30, we do not know what the differences are between
“serious flaws” and “very serious flaws” or between “some uncertainty” and major
uncertainty”.
R – We have not developed clear guidance regarding judgements such as what constitutes a
“serious limitation” or a “very serious limitation”. These are often difficult judgements. A
strength of our approach is that these judgements are made explicitly. When there is empirical
evidence, we do provide guidance.

Q: Is the quality of the evidence of a randomized controlled trial with serious flaws the same
as that of an observational study with a strong, consistent and direct association and no
plausible confounders?
R – Yes.

Q: When reference is made to the first manuscript which is being submitted to BioMed
Central, it would be worth mentioning that it includes not only questions about the
effectiveness and harm, but also about diagnosis and prognosis, whereas these latter questions
are not really the subject of this second manuscript.
R – Added to the Background section (as noted above):
All the examples used in this pilot study were questions about interventions.

Q: It should be made clear that the system developed in the second manuscript only applies to
judgements about the effectiveness of therapeutic or prophylactic interventions, and not to
diagnostic interventions, neither to economic, etiological, or to prognostic studies.
R – Discussion section, we have listed ongoing developments, including considerations of
public health and health policy interventions, costs, and diagnostic tests.
Q6: As already suggested above, a more accurate title would be: ‘A pilot study of a new system for grading the quality of evidence and the strength of recommendations about the effectiveness of therapeutic and prophylactic interventions’. This should be made clearer in the abstract too.
R – Title of both manuscripts are changed according to BioMed Central editors suggestion, see above.

Q: Also in the Methods section of the abstract, the first sentence could perhaps be rewritten something like this: “Twelve evidence profiles were prepared based on the results of 12 systematic reviews” (the word “example” is already used in the following sentence):
R – The Methods section of the abstract has been changed, see above.

Q: Page 11, first sentence of the “Sensibility and understandability” section: there is a spelling mistake (“rates” instead of “raters”).
R – Thank you, change made.