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 (jaeschke@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:** 2 **Date:** 24 November 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, and 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
Discretionary 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.
Regarding paper A, please see paper A.

Minor essential revisions: all related to the other article

Discretionary Revisions: all related to the other article

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 18\textsuperscript{th} author deserves to be a co-author?
R – The 18\textsuperscript{th} 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.