Author's response to reviews

Title: Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches

Authors:

- David Atkins (DAtkins@AHRQ.GOV)
- Peter Briss (pxb5@cdc.gov)
- Martin Eccles (Martin.Eccles@newcastle.ac.uk)
- Signe Flottorp (signe.flottorp@nchs.no)
- Gordon H Guyatt (guyatt@mcmaster.ca)
- David Henry (mddah@mail.newcastle.edu.au)
- Suzanne Hill (hilsu@mail.newcastle.edu.au)
- Alessandro Liberati (alesslib@tin.it)
- 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: 9 November 2004

Author's response to reviews: see over
Responses to BMC Editorial and reviewers comments re:
A critical appraisal of systems for grading the quality of evidence and the strength of recommendations. 9976868452873787

We agree to change the titles of this and the associated manuscript (A pilot study of a new system for grading the quality of evidence and the strength of recommendations. 4352636028844837).

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: none

Minor essential revisions:

Q: The authors should search the paper and remove all uses of “this” as a noun. In most cases it is entirely unclear to what “this” refers.
R- We have done this.

Q: Suggested that the results section should only include the one first sentence: There was poor agreement among 12 assessors who independently assessed the six systems (Table 1).
R- As noted below the Results section has been reduced to these two sentences:

There was poor agreement among the 12 assessors who independently assessed the six systems. A summary of the assessments of the sensibility of the six approaches to rating levels of evidence and strength of recommendation is shown in Table 1.

Discretionary revisions: none

Reply to comments from Benjamin Djulbegovic

Major compulsory revisions: none

Minor essential revisions: none

Minor (optional) revisions:

Q: “a clear link between the results of this paper and the normative problems of the current systems for rating of strength of scientific evidence was missing in the paper. That is, the
discussion of the paper is really not based on the data (Table 1) but rather on the authors’ understanding about normative defects of the systems they evaluated.”
R- We disagree. We believe the discussion flows directly from our appraisal of the 6 systems. We have added the following to the Methods to clarify the link:

Our independent appraisal of the six systems were summarised and discussed. The discussion focused on differences in the interpretation of the criteria, disagreement about the judgements that we made and sources of these disagreements, the strengths and weaknesses of the six systems, and inferences based on the appraisals and subsequent discussion.

Q: “Perhaps it would be helpful if the authors explained a bit better how they developed their sensibility criteria, which are based on adoption of Feinstein’s criteria of “clinimetrics”.”
R-The criteria is a direct application of the criteria described by Feinstein in the reference.

Q: suggested more discussion of the approach by AHRQ- ref 32, to evaluate systems to rate the strength of scientific evidence.
R- This is not relevant and we have elected not to do this. West et al [32] evaluated systems for rating the quality and strength of scientific evidence of individual articles, systematic reviews, RCTs, observational studies and diagnostic tests separately. They identified a large number of systems to rate the quality of scientific evidence, including several that were considered acceptable. In contrast, our current aim was to evaluate systems for grading the quality of evidence and the strength of recommendations with the aim of identifying a system or the dimensions needed in a system that could be used to grade the quality of evidence across each outcome for all types of studies.

Q: “some comments regarding the methodology of judging level of agreement among multiple observers should be made (see also comments on the companion paper”
R- For this, the critical assessment of grading systems our assessment was largely qualitative and the conclusions reflect a consensus after independent judgements followed by discussion. The extent of disagreement is obvious from the data in the table. Kappas or another measure of agreement would be superfluous. For the pilot study –companion paper of the pilot study- we have now reported and discussed kappas.

Reply to comments from Anja Tuulonen

Major compulsory Revisions:

Q: “The authors should define “prominent” (and “important” as stated in results), i.e. on what ground they selected the 6 systems out of 51 to be analyzed (e.g. the ones with the longest experience, the ones with revisions and improvements?)
R- We have added the following to the text:

These were selected based on the experience and knowledge of the authors through informal discussion.

Q: In Box 1 and Table 1 (“To what extent can the system be used with different audiences?) It is unclear whether the “system” here refers to the end product (i.e. recommendation or
guideline) or to the tool of grading the evidence. The confusion between guideline developers and users, however, is later on discussed with clearness and simplicity of the system.

R- We agree and have refereed to this confusion already in the text:

Some of the criteria were not clear and were interpreted or applied inconsistently.

Q: In Box 1, could the authors give an example what they mean by “How often will information not usually available be necessary?”

R- Examples were already given:

The dimension for which we considered that information would most often be missing was trade-offs; i.e. knowledge of the preferences or utility values of those affected.

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- In paper A we have not used good or moderate agreement but the term ‘poor agreement’ is used in summary statements three times, this is obvious from the data in the table.

In paper B we have calculated kappa for agreement measure, please see responses in B.

Q: The analysis is also based on discussions. Either the use of discussions should be described in the methods section, or optionally the results based on discussions should be removed from the results into the Discussion. E.g. in the first paragraph of the results, only the first sentence contains facts, the rest is discussion, and similarly in page 9 the paragraph “Most of THOUGHT…”

R- We have removed all but two sentences from the Results section to the Discussion section.

Q: The results of Table 1 are unnecessarily repeated in the text. The most relevant points of Table 1 (e.g. “the best” criteria in different dimensions) could be emphasized e.g. using bold font. In this way, paper A could be condensed.

R- As there was poor agreement among the assessors about the six grading approaches, we find it difficult to select “the best” criteria. The text from the results is now part of the discussion. We feel this discussion is important to help with interpretation of the data in the Table.

Q: The last paragraph of the results belongs to the beginning of the paper where it is described how the 6 systems were elected among the 51 systems.

R- The six systems that we appraised were selected prior to identifying and comparing with the systems used by the 51 organisations, as stated in the methods section.

Q: It is not analyzed or discussed in either paper A or B, what is the effect of the positive statement and low grade of recommendation (e.g. 2C in Table 1, appendix 1) versus a direct negative statement with high strength of evidence. One would assume that positive statement leads more often to treatment in spite of poor evidence.

R- This is beyond the scope of these papers.

Q: As stated above (please, see results), much discussion is actually presented as results, which confuses the reader. The second paragraph states important and relevant conclusions that are actually not based on the results of Table 1 but rather discussions with the group.

R- The discussion of the results has been moved from the Results section to the Discussion section.
The conclusions made after discussions in the group of the critical appraisal is not direct discussion of Table 1, but was based on the discussions that the group had following discussion of the critical appraisal exercise and further reflection. Added to the Methods section:

Our independent appraisal of the six systems were summarised and discussed. The discussion focused on differences in the interpretation of the criteria, disagreement about the judgements that we made and sources of these disagreements, the strengths and weaknesses of the six systems, and inferences based on the appraisals and subsequent discussion.

Q: The second paragraph on page 12 could be rephrased. I had to read it several times and still was not sure about the message (“Systematic review…”).
R- The paragraph has been changed to:

- Judgements about the quality of evidence should be based on a systematic review of the relevant research.
- Systematic reviews should not be included in a hierarchy of evidence (i.e. as a level or category of evidence). The availability of a well-done systematic review does not correspond to high quality evidence, since a well-done review might include anything from no studies to poor quality studies with inconsistent results to high quality studies with consistent results.

Minor essential revisions:

Q: “A description of the most recent version of each of these systems was prepared by an individual familiar with system” should state that five authors of the paper prepared the descriptions.
R- Changed to:

A description of the most recent version (as of summer 2000) of each of these systems (Appendix 1 to 6), was prepared by one of the authors familiar with the system, and used in this exercise.
The authors are identified in the Appendix.

Q: Further, it is not clear from the paper who actually were the 12 evaluators. Were these 12 people the authors of the paper, including the five who wrote the descriptions?
R – Added to the Methods section:

all of the authors minus GEV appraised the six systems, three of the authors (DH, SH and DO’C) appraised as a group and reported as one (see contributions).

Q: Would the previous experience with one system possibly bias the results?
R- Almost certainly, however, this bias would apply equally to all of the six systems that were critically appraised in this article as they were all ‘represented’ in the group.
In what was first paragraph of Results but is now first paragraph of Discussion: “Some of us had practical experience using one of the systems or used additional background information related to one or more grading systems.” Added to the text:

And we may have been biased in favour of the system with which we were most familiar.
Under Competing interest it is declared that most of the authors have vested interests in another grading system.

Q: The reader would like to know, whether the evaluators could be regarded as experts in literature grading, in comparison to a less trained evaluator, or a beginner? Obviously, the untrained evaluators would do much worse than the experts.
R- The untrained evaluators would also have less ‘grading system’ bias and might therefore end up with better agreement.

From the Methods section: “The 12 assessors all had experience with at least one system and most had helped to develop one of the six included systems.”

Q: In addition, the reader would need to know, how the evaluators were instructed or trained to use the 6 grading systems. Or were they just given the description provided in appendices 1-6? If yes, in the appendices facts are reported which in fact are under investigation of the study (e.g. strengths and weaknesses, target audiences).
R- Methods section old: “Twelve people independently appraised the six systems.”
New: “These descriptions of the systems were given to the twelve people who independently appraised the six systems.”
Added:
No training was provided and we did not discuss the 12 criteria prior to applying them to the six systems.

Discretionary Revisions:

Q: Abstract, for better logics, the last sentence of Results (“Systems used… “) should be transferred to the end of Objective.
R- As noted above, this is a result.

Q: Because the descriptions of the 6 systems have no common structure, their quality varies and they may not open up easily to readers unfamiliar with the “Alphabet soup” of the EBM evaluations. Since these 6 systems represent the most prominent ones, it would very useful for the readers if the authors created a summary table of their major features, such as how many levels of recommendation are reported, how they are reported (A-D, A-I, I-IV etc) when they are published etc.
R- We agree that this is a problem and this was an important motivation for the work reported here and subsequent work [33, 34]. However, the aim of this paper is to highlight the limitations of these systems, not to assist with their use which we consider problematic.

Q: Although referred to, tables 2 and 3 are missing in appendix 6.
R- The references to these tables are removed. Thank you for bringing this to our attention.

Reply to comments from Joseph Watine

Major Compulsory Revisions: none

Minor Essential Revisions:

Q1: As far as I can understand, the ensuing chapters of this long “story” are in preparation. I think that the authors should tell their readers more about their plans for the ensuing chapters.
Old Conclusions: “Based on discussions of the strengths and limitations of current approaches to grading levels of evidence and the strength of recommendations, we agreed to develop an approach that addresses the major limitations that we identified [33].”

New Conclusion: “Based on discussions of the strengths and limitations of current approaches to grading levels of evidence and the strength of recommendations, we agreed to develop an approach that addresses the major limitations that we identified. The approach that the GRADE Working Group developed based on the discussions following the critical appraisal has been pilot tested [33]. Based on the pilot testing and discussions following the pilot, the GRADE Working Group has further developed the GRADE system to its present format [34].”

And more in the pilot study article.

Q: Perhaps they could also tell the readers who does what in their team, and how their team is organized. This might help readers to understand, e.g. why there are 14 co-authors for the first manuscript, and 18 for the second one.

R- Added to the Conclusion section:

The GRADE Working Group has continued to grow as an informal collaboration that meets one or two times per year. The group has web pages http://www.gradeworkinggroup.org and a discussion list.

Q2: Box 1: instead of points, couldn’t numbers (1 to 12) be used?

R- We have changed the points to numbers.

Q: Appendix 1: in the second page, reference is made of a table 2 which I cannot find anywhere in Appendix 1.

R- Thank you, the reference to table 2 is removed.

Q: Appendix 2: the last sentence of the second page does not make sense to me: shouldn’t the word “avoidance” be replaced by something else?

R- No, that is what they mean.

Q: Appendix 3: the meaning of the some abbreviations in the table or in its legend should be indicated, e.g.- SR, RCT, CPG, ARR, Rx.

R- Abbreviations added to the Table legend:

Abbreviations:
SR – Systematic review
RCT – Randomised Controlled Clinical Trial
CPG- Clinical Prediction Guide
ARR – Absolute Risk Reduction
Rx – Prescription

Q: Appendix 5: in the second sentence of the “Target audience” section (page 3), there is a spelling mistake (“pprofessional”). In the Figure 1, I do not understand why there is a Figure called “4a” whereas there are no “4b”.

R- Thank you, we have corrected.

Q: Appendix 6: in the first sentence of the “Strength of recommendations” section (page 2), mention is made of a Table 3 that I cannot find anywhere in Appendix 6.

R- Thank you, the reference to Table 2 and 3 have been removed.
Q4: Yes it does, except for the fact that no keywords are provided. The following ones could perhaps be used: evidence-based health care; levels of evidence; practice guidelines; strength of recommendations; systematic reviews.
R- Thank you.

Q5: In the conclusion, one point is missing, i.e. this appraisal work is a preparatory work. This preparatory work has made it possible for the authors to write the two ensuing papers (i.e. the second manuscript submitted to BioMed Central and the paper which is “in press” in the BMJ). Unless I am mistaken, it will also likely enable them to write other papers, regarding e.g. prognostic or etiological studies, or diagnostic interventions.
R- Please see Q1, we have added a sentence describing other work that is ongoing.

Q6: In the last sentence of the Methods section of the abstract, the number of systems which were compared with the six prominent systems should be indicated (instead of “Other”).
R- Old: “Other systems were compared with these six approaches.”
New: “Systems used by 51 organisations were compared with these six approaches”.

Q7: Method section: -end of first sentence: the word “method” could be deleted;
R- done.

Q7 -end of second sentence: “…to assess the sensibility of the different approaches” could be replaced by “…to assess their sensibilities”.
R- We would like to keep this sentence as is.

Page 8, second paragraph, second sentence: shouldn’t “Almost all agreement…” be replaced by “Almost all agreed…”?
R- Yes, thank you.