Reviewer's report

Title: A pilot study of a new system for grading the quality of evidence and the strength of recommendations

Version: 1 Date: 30 March 2004

Reviewer: Anja Tuulonen

Reviewer's report:

General

The authors should be congratulated for accepting the great challenge to analyze and improve the complex system of grading the evidence and recommendations. First the authors analyze six current systems (hereon referred as study A) and based on the poor agreement in study A, they develop a new grading system with further improvements which they present in the second paper (B).

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.

------------------------------------------------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

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?)

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.

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

The authors should define what they mean by good, moderate, and poor agreement and use the same criteria in both papers A and B.

Results

The analysis is also based on discussions. Either the use of discussions should be described in the method section, or optionally the results based on discussions should be removed from the results into 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 us THOUGHT...”

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.

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.

Discussion

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 directly negative statement with high strength of evidence. One would assume that positive statement leads more often to treatment in spite of poor evidence.
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.

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…”).

---

### Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

“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.

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? Would the previous experience with one system possibly bias the results?

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.

In addition, the readers 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).

---

### Discretionary Revisions (which the author can choose to ignore)

These changes would improve the readability of the paper:

**Abstract**

For better logics, the last sentence of results (“System used…”) should be transferred to the end of Objective.

**Appendices**

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 the criteria were published etc.

Although referred to, tables 2 and 3 are missing in appendix 6.

---

**What next?**: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** No

**Declaration of competing interests:**

None