Title: A critical appraisal of systems for grading the quality of evidence and the strength of recommendations

Version: Date: 8 March 2004

Reviewer: Benjamin Djulbegovic

Reviewer's report:

Thank you for the opportunity to review this paper. Let me say from outset that publication of this paper has been long overdue. The pioneering 1979 work of the Canadian Task Force in the Periodic Health Examination (later followed by the rise of EBM movement) has promoted the concept that not all evidence is created equal and that certain findings are closer to the truth than others. This realization has been followed by the attempt to develop the hierarchy of clinical evidence, which should enable physicians and the patients to know how much confidence they can have in the practice recommendations and evidence that is used to support these recommendations. In fact, as a recent AHRQ systematic review indicated [ref 32], 106 different systems to rate the strength of scientific evidence and more than 51 schemes to devise practice recommendations have been developed to help optimize appropriate care. These systems were all based on 6 main systems, which the GRADE group evaluated in this paper. The group found that all existing systems have important shortcomings. Therefore, their continuing use will not help reduce errors in making of recommendations, will not help improve communication between guidelines developers and guidelines users, and thus will not help people make well-informed decisions. This is important finding that deserves wider dissemination. This realization also prompted the GRADE group to develop a new system.

I agree with the authors and their results. I also think this paper will help clarify a simple-minded thinking about EBM, which lead to equating the study design with the “truth” and in turn with decision-making and practice recommendations. Consequently, I recommend that this paper is published.

I do have two minor (optional) suggestions:
I am a bit puzzled that there is a dissonance between normative problems which plague the existing evidence grading systems and the results shown in the paper (Table 1). All existing systems are grossly inadequate on normative ground because they confuse the issues between the (quality of) evidence and decision-making/practice recommendations. The authors understand this and address this issue in the development of their new system. However, a clear link between the results of this paper and 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. 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”. I also think that the approach taken by the AHRQ [ref 32] to evaluate the existing evidence grading systems should be discussed in more detail since this is the only other work, which addressed the same issue the GRADE group addressed in their paper. Finally, some comments regarding the methodology of judging the level of agreement among multiple observers should be made (see also the comments on the companion paper).

None of this is, however, critical and should not delay publication of this important paper. I leave it up to the authors if they believe the paper would benefit if these clarifications are incorporated.
What next?: Accept after minor essential revisions

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests:

Although I personally know some of the authors, I believe I was able to evaluate this paper in fair and impartial way.