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: 8 March 2004

Reviewer: Benjamin Djulbegovic

Reviewer's report:

This is a companion paper, which has resulted from the authors’ prior evaluation of the existing systems for grading scientific evidence. The authors, who are internationally known experts in the field, previously determined that the existing systems for grading the quality of evidence and the strength of recommendations are inadequate (see comments related to this paper). They then proceeded to develop a new system with the goal to improve on the shortcomings of the old systems. Faithful to their calling, they then went ahead and critically appraised their own system.

I have been waiting for this moment for a long time. Has the medical community finally gotten a system which can distinguish between a “good” and “bad evidence”, between the “truth” and evidence that we should not believe in? In addition, have we finally got the system to enable optimal decision-making and practice recommendations for our patients and populations? How did authors succeed?

Regarding the paper itself, I don’t have any major criticism to make: this is a well done and written paper, and it should be published. 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 level 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.

However, I can’t help but note that the results of this pilot study are somewhat disappointing. Even among the people who are all so well versed in the guidelines methodology, share the same professional language, committed to this approach of making practice recommendations, the agreement about this new system for grading the quality of evidence and the strength of recommendation was rather poor (see below). I am making this statement not because I don’t value this work and the effort of this remarkable group of individuals. On the contrary, I am a big fan of this project and the authors’ work. I personally believe in rational approach to an uncertain world which surrounds us. However, after re-reading this paper (and other GRADE documents), and despite the fact that I desperately wish this approach to be successful, my reactions continue to be mixed. In the final analysis, this probably should not be surprising since this is an ambitious project and, as the authors note, human judgment and arriving at the “truth” is complex matter. About 80-100 years ago, a group of philosophers of logical positivism movement tried, in a different context, to develop an overarching set of the rules to deduce the “truth”. This brilliant group of people did not succeed. In some distant way, and of course at much less grandeur scale, the approach to develop evidence grading system(and I have had my own attempts too!) reminds of lukewarm attempt to re-invent logical positivism.

So, what does this paper show? Although the authors concluded that the GRADE system was “clear, understandable and sensible”, the fact is that agreement among 17 people was “poor” for 3 elements of the new system and “moderate” and “good” for the remaining elements of the GRADE system. This does not appear to be very good news for the system which is designed to replace 106
existing systems. What is disappointing is that the result is achieved among the people who all share the enthusiasm for this approach and are expert in EBM. Why is this so? And, what one can expect to see when the system begins to be widely adopted and by the people who are less methodologically astute?

Let me try to give my own thoughts why the system did not perform better. First, I cannot agree more with the authors that any new system for grading of scientific evidence and making recommendations should separate assessment of the quality of evidence from decision-making. Evidence is necessary but not sufficient for decision-making/practice recommendations. A five steps approach to assessing the quality of evidence, summarizing benefits and harms, and making recommendations appears to be normatively correct. My only objection is that practice recommendations are left to be formulated intuitively. Theory of decision-making holds that rational decision-making is the one that maximizes the value of consequences: a) based on the decision-maker’s current assets, b) based on the possible consequences (i.e. benefits and harms) and values associated with each consequences of a choice; c) when these consequences are uncertain, their likelihood is evaluated according to the rules of probability theory.

So, I am not surprised that there is a poor agreement about the judgments of benefits and harms and making recommendations. It has been repeatedly demonstrated that people do not manipulate probabilities well in their heads, and the use of formal methods to integrate benefit and harms of competing treatment alternatives or other medical outcomes is likely to be superior to intuitive judgments. Whether one can or should incorporate formal analytic methods in the instrument which is supposed to be used widely is a different matter.

Second, I am also not surprised that there is poor agreement among the assessors related to the quality of evidence (for each outcome and across the critical outcomes). This is because instructions (Appendix) leave so much to judgment and interpretation. For example, what is meant by “serious flaws” or “very serious flaws” in the study quality? What is considered “important consistency”, “minor” or “major uncertainty”, “strong” or “very strong” association etc? Even among the authors who are very familiar with critical appraisal methods, the issues related to internal and external validity of clinical studies etc, better agreement on the quality judgments was not obtained. One has to wonder how will GRADE system fare in the real world.

The good news is that people appear to agree easily on what is important to them, which is also not that surprising particularly in medical decisions related to life and death.

The authors applied their system to 12 different health care interventions. It would be good to know how often their recommendations agree/disagree with practice guidelines developed by various professional organizations. The insights obtained from this comparison may help improve the GRADE system (see also below).

So, we have to come back to the question frequently raised during the last decade: is it possible to develop taxonomy of clinical evidence that is comprehensive enough to capture knowledge obtained from methodologically rigorous clinical trials and knowledge obtained from the full variety of medical evidence, including deterministic science, physician experience and the unique features of individual patients? The new system indeed is attempting to encompass not only observation from randomized controlled trials but also findings from other types of observations and “any other evidence”. Presumably, the goal is to develop such classification of evidence which would combine population-oriented clinical research with all of our observational data, which may range from the chemistry and physics of physiologic processes to ethical issues. At some point in the past, I was hoping that such a system is possible to develop. However, I am increasingly skeptical if the universally acceptable ranking of quality of evidence is indeed possible. One of the problems is that
understanding of nuances of the subject-specific knowledge/content (and context) plays much larger role that we all initially believed to be the case. Today medicine is still dominated by disease-specific experts who, for better or worse, are often the final arbiters of the "truth". Which are those invariant and truth-preserving elements across the entire hierarchy of the scientific spectrum that all experts would agree on, particularly when direct empirical evidence is lacking, will continue to remain the major bone of contention among the guidelines developers. For example, judgments on consistency of evidence will be very difficult to agree upon since consistency across the disciplines or different modes of observation is relatively rare. This is one of the major reasons that the experts from different fields often disagree. Until content and methodological-specific experts find themselves on the common ground, a system of grading of evidence which will be universally acceptable to generalists and specialists alike will be difficult to develop. Perhaps one of the major legacies of EBM movement will be that specialists are now willing to accept a notion that hierarchy of evidence indeed exists. However, that notion so dramatically differs from one expert to another that in practice it often amounts to nothing more than using EBM umbrella to provide a desirable spin on given evidence.

Although I am increasingly skeptical about the feasibility of development of the universally acceptable evidence grading systems, I am wholeheartedly hoping that the GRADE group will succeed. A good initial step is to try to be as specific as possible and develop dictionary of the elements for judging the quality of evidence.

I hope that the authors will forgive me for the above discussion. My intent is not to criticize their system but rather to highlight the issues that are still on the horizon and will not go away. I personally serve on the guidelines panels of three professional organizations which all have somewhat different philosophy to evidence grading and guidelines recommendations: from the one which believes in the importance of the consensus of the experts to formulate recommendations to the one which would like to emulate the GRADE-like system to the third one wishing to develop something in between.

**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 that I was able to make fair and impartial evaluation of this paper.