Reviewer's report

Title: Judging the quality of evidence in reviews of etiological and prognostic studies: Adapting the GRADE's framework

Version: 1 Date: 22 May 2013

Reviewer: Rob J Scholten

Reviewer's report:

This manuscript is a proposal for assessing the level of evidence regarding outcomes of etiological or prognostic studies by the use of the GRADE framework. The report is well-written and easy to follow. There are no formal underlying methods other than group discussions. Therefore, I have no comments on the methods and I consider this study as a starting point for a GRADEing system (which the authors state in the last sentences of the manuscript) that will be further developed by future discussions either in this journal or in groups of researchers such as the GRADE Working Group or guideline development groups. Some of my comments may require further thoughts and elaboration by the authors, other comments may not require any amendment, but could serve as first start for further discussion.

Major comments

1. Page 6: prognostic studies usually address more than one prognostic factor (prognostic models). If different studies addressed different sets of prognostic factors, can those prognostic factors than be meta-analysed separately (and put in a GRADE Summary of Findings (SoF) Table? The authors may wish to explain this in more detail.

2. Page 6: the rule for upgrading the level of evidence in intervention studies applies to confounding that would have decreased the effect instead of increased (so the effect would be even stronger if the investigators had have adjusted for those confounders). Is this indeed addressed in the risk of bias assessment of etiologic or prognostic studies and could this 'risk of bias' lead to an upgrade? I'm afraid that I don’t understand why such confounding doesn’t have to be addressed separately.

3. Page 8: is the QUIPS tool also suitable for assessing the risk of bias of etiologic studies? Wouldn’t the Newcastle Ottawa Scale (NOS) be more appropriate? The NOS also includes the assessment of case-control studies, which type of studies is very suitable for etiologic questions. The authors seem to ignore case-control studies in this manuscript.

4. Page 10 Inconsistency: downgrading for a low p-value or high I-square. Those statistics can lead to downgrading in its self. However, they can’t be judged in isolation and have to be judged in the realm of the size of the underlying studies. Suppose that there are 10 big studies with quite similar effect estimates with very small CIs (so very low within-study variance). Then I-square will approach 100%
and the p-value for heterogeneity will become very small (which both truly reflect that most of the variance is due to between-study variance). When interpreting those results, the two heterogeneity statistics will probably be ignored, because there are no clinically relevant differences between the 10 effect estimates, and there won’t be any reason for downgrading. The authors may wish to elaborate on that.

5. Page 12 Imprecision (a tough section). The authors refer to reference 21. If possible (and journal space permitting), the authors may wish to include an example. The phrase “… when the confidence interval is not excessively wide …” may require some more guidance (if at all possible).

6. Page 14: upgrading for large effect size. I’m often confused by this rule (also with GRADE for interventions). The size of the OR (or RR) should be judged against the background risk and/or the unit to which it applies. An OR of 1.2 for one extra year may be huge in some cases and a RR of 5 can be trivial if the background risk is 1/1,000,000. There may not be an easy solution for this, but the authors may wish to raise this point.

7. Table 2: the item ‘Imprecision’ is missing.

8. Table 3: in the first column, there are three arrows, instead of two?

9. Tables 4 and 5: it would be very helpful if the authors could fill in this table with real examples of their own.

10. Box 1 might be included in the text. Very clear explanations in the boxes!

Items for further discussion (which may or may not be addressed by the authors at this point in time)

1. Study limitations. I’m not aware of sound empirical studies that addressed the influence of the various risk of bias items on the size of the effect (which also applies to diagnostic test accuracy studies (DTA)). Because the primary studies usually suffer from poor reporting, such empirical studies will often lead to unclear results (like in the DTA domain). Thus, we don’t really know yet, whether there is a relationship and what the strength is of those relationships. So, this item (and the guidance for downgrading) must probably be labeled as work in progress.

2. In addition, there’s also no empirical evidence for the need to leave out ‘low quality’ studies from a meta-analysis. GRADE for interventions would probably include those studies and downgrade for study limitations. Not sure how to handle this and certainly food for future thoughts.

3. The guidance for downgrading by 1 or 2 points (= serious or very serious) is now quite vague (which can’t be avoided in this stage). The authors may wish to emphasize that this is still the case and that further development is necessary (which also applies to GRADE for DTA).

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

**Quality of written English:** Acceptable
**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I'm member of the GRADE Working Group. I have no further competing interests.