Author's response to reviews

Title: Judging the quality of evidence in reviews of prognostic factor research: Adapting the GRADE's framework

Authors:

Anna Huguet (anna.huguet@dal.ca)
Jill A Hayden (jill.hayden@dal.ca)
Jennifer Stinson (jennifer.stinson@sickkids.ca)
Patrick J McGrath (patrick.mcgrath@iwk.nshealth.ca)
Christine Chambers (Christine.Chambers@iwk.nshealth.ca)
Michelle Tougas (michelle.tougas@iwk.nshealth.ca)
Lori Wozney (lori.wozney@iwk.nshealth.ca)

Version: 2 Date: 25 June 2013

Author's response to reviews: see over
Dear editors and reviewers,

Thank you so much for your feedback. Your comments have been useful to improve our manuscript. We have addressed each of reviewers’ concerns. Below we explain what we have done to address each concern and indicate where changes have been made in the manuscript.

We were requested to format the abstract into separate sections. However, these sections do not coherently fit with our manuscript (i.e., methods). We have not made this modification, but are willing to do so if required. The abstract does fit within the criteria of no more than 350 words. We have made minor editing changes throughout the manuscript as suggested. These have not been highlighted as changes, but are integrated within the manuscript.

Reviewer: Nick Meader

Reviewer’s report:
This an article that seeks to adapt the GRADE approach to assessing the quality of evidence for prognostic studies. This is a good and well thought out attempt to do this.

Major Compulsory Revisions:
1) Discussion of confounding (GRADE framework for prognosis paragraph 1) – I understand the point you make by suggesting confounding should be dealt with as an issue of risk of bias rather than as an additional factor to rate up the quality of evidence. I think this seems sensible. But I think you have to justify much more clearly in your paper how this would be an improvement on the approach GRADE currently takes. My guess is that your modification reflects a critique of the role of confounding in the GRADE system as a whole – rather than just in relation to its application to evaluating prognostic studies.

So I think you need to set out what the rationale GRADE uses for their current approach regarding confounding and then provide a clearer critique of that and why your modification improves the logic of this approach both in the context of prognostic studies but also beyond.

Our manuscript does not intend to evaluate the current GRADE approach that is used for interventions studies and we do not intend to be critical with the role of confounding in the GRADE system when implemented in intervention research. Rather, we intend to apply the GRADE concepts to a new type of literature, prognostic factor research.

The current GRADE approach for interventions upgrades for appropriate potential confounding, and does not account for confounding in the risk of bias assessment. The approach also assumes that the lack of control of confounding likely contributes to an observed effect that is larger than the truth. In our
proposal for prognostic studies, we assess potential confounding in the risk of bias assessment. We have further explained in the manuscript the relevance of this item when assessing risk bias. In prognostic research we feel that it is difficult to judge whether potential confounders are likely to alter the effect size consistently in the same direction (see GRADE framework for prognosis, page 6).

2) Inconsistency (points 1-3) – This is pretty much as reflected in the current GRADE system. But I think some refinement of the criteria may be helpful. Point 1 – I would probably add that the differences between point estimates should be clinically meaningful rather than just based on statistical significance. But some further rationale for either approach would be an improvement.

We completely agree with this comment and have elaborated this need for consideration of clinically meaningful differences in the manuscript. When deciding if there is inconsistency we propose a decision based on the presence of inconsistency suggested by the statistical test for heterogeneity and I² and also based on clinically meaningful differences (see Inconsistency, pages 9-10, Table 3, page 26).

Point 3- I’m wondering whether I²= 75% as a criterion for inconsistency is a little liberal. Granted heterogeneity tends to be much higher in prognostic studies – but I think this just reflects the inherent uncertainty of much of this literature. So I’m not sure that we should use a more liberal threshold just on this basis.

We have modified this criterion to be less liberal by setting an acceptable I² cut-off point of 50%, as outlined by the Cochrane Handbook (Higgins & Green, 2011) (see Inconsistency, page 10).

3) Publication bias – I think your discussion of publication bias is not very clear. This is particularly the case when judging if publication bias is unlikely to be present. You seem to be arguing publication bias should be judged unlikely when adequate adjustment for confounding has been undertaken. I don’t follow how this would suggest publication bias is unlikely. I may have misunderstood your argument but this seems to be referring to a different issue. So I think you need to elaborate on this section and argue it more clearly.

Thank you. We have removed the mention of the need for adequate confounding, and we articulate that publication bias should always be assumed unless (1) there is a large number of studies looking at that particular factor; (2) ideally that some of them are studies conducted in an advanced stage of investigation; that is, there are researchers particularly interested in these factors and studies have been planned a priori and designed specifically to test the hypothesized association. However, since phase of investigation is already considered for GRADEing, we do not recommend downgrading publication bias for only phase of investigation (see Publication bias, page 14, Table 3, page 25).
We also state in the paper that it is very difficult to assess whether publication bias exists because there is not any type of registration for prognostic studies, as there is for intervention trial registries. Until this comes into effect, we can only make the best guess and we have decided to be stringent in order to be more cautious. (see Publication bias, page 14)

4) The article needs a thorough read through and edit.

Thank you. We have reviewed the manuscript in detail and have edited.

Discretionary Revisions:
5) Defining serious and very serious limitations (i.e. downgrading one or two levels) for inconsistency, imprecision, indirectness. You provide some guidance for risk of bias but not for these domains – further discussion of serious or very serious limitations. I accept that this does require judgement depending on the particular situation – but I also think it is possible to develop some general rules of thumb. Some examples from the GRADE working group articles would probably be helpful.

We agree with this point. However, at this stage, it is very difficult to establish rules. We opt to leave it up to the best judgment of each reviewer, and we have included a note that this is an aspect that needs to be further developed in the future (see Conclusions, page 17).

6) Imprecision – I think you would also want to take into account number of studies as well as total number of participants particularly if you were wanting to explore and quantify heterogeneity which I would think is almost always the case in relation to prognostic studies

We have added a new piece in the manuscript in which we acknowledge that when meta-analysis is not possible, it is very difficult to assess imprecision. When this happens, and we find that there is imprecision within the studies, we also state that the number of studies and number of participants should also be taken into account because there is likely to be more imprecision with a fewer number of studies and/or participants (see Imprecision, pages 12-13, Table 3 page 25).

7) Have you contacted the GRADE working group and received feedback from them? I would have thought this would be an important aspect of refining this approach.

We did contact the GRADE working group, we received very general comments at the very beginning, but we did not go any further to pursue this development together. It was a need for us to adapt the GRADE to this type of research for completion of our own systematic review and we moved ahead. Their further work on this initial attempt, if they are interested, may be very useful. For this
reason we have included in our existing discussion that their input would be very valuable.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**
I declare that I have no competing interests
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.

We have made the introduction clearer to emphasize that this manuscript is focused on explaining how to adapt the GRADE to assess the overall quality of evidence of single prognostic factors (and models) (see Introduction, page 3).

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.

As discussed as an answer to the second response, point 1 raised by Reviewer 1 (Nick Meader), it is unclear how confounders influence the effect. They may inflate or minimize the effect. So, this is the reason why we decided not to assess this factor separately. When assessing risk of bias, we are not evaluating how this will influence the strength of the effect, instead we are evaluating the methodological quality of the studies, and we know that when studies do not control for confounders, the risk of bias increases and this limits the validity of results (see GRADE framework for prognosis, page 6).

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.

For simplicity and to add clarity, we have now focused this manuscript on prognostic research rather than both prognostic and etiologic research (these changes were made throughout the entire manuscript).

4. Page 10 Inconsistency: downgrading for a low p-value or high I-square. Those statistics can lead to downgrading in itself. 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.

This concern is similar to the concern raised by Reviewer 1 (Nick Meader). We completely agree that we should not base our decision on p value and I², we should also consider clinically relevant differences. This elaboration has been added to the manuscript (see Inconsistency, pages 9-10).

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).

As for reference 21, we have worded the sentence in a different way to try to avoid confusion. As for the second request, it is very hard for us to quantify when a confidence interval is excessively wide, we have articulated in the manuscript that to decide what is excessively wide, the authors need to use their best judgment (see Imprecision, paragraph 2, page 12).

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.

We completely agree with this comment. To evaluate whether to upgrade the overall quality for large effect, we should go beyond the size of the effect and understand what an OR of 2.0, for example, means in each individual systematic review. We are proposing an arbitrary well-establish cut-off. We have
acknowledged this in the manuscript and recommend judgment considering the review question (see Moderate or large effect size, pages 14-15, Table 2-5, pages 24-27).

Considering this, we have decided to be more liberal and propose a cut-off of 2.5 for OR (or RR), which is considered a moderate effect size instead of a cut-off of 4.25, which is considered a large effect size (see Moderate or large effect size, page 15, paragraph 2).

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

We have added this (see Table 2, page 25).

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

We have taken care of this (see Table 3, page 26).

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

We have presented examples from the qualitative synthesis in our current review in Table 5 (see Table 5, page 28).

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

We considered moving this content to the body of the manuscript. However, we feel that the description included in the box is not the intent of our manuscript and reading it takes away from this focus.

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.

This is a good point. We have added this issue as a limitation, and emphasize that further research on it is necessary (see Conclusions, page 17).

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.

This reviewer may be right, especially when we consider his comment that we do not know how much study limitations affect the size of the effect. We propose conducting subgroup analyses to explore the potential impact of studies with ‘high risk’ for specific bias domains and overall (see Study Limitations, page 9).

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).

We have added this as a future direction to work on. We are in a very early stage of development and more thinking on the influence of the factors on the overall quality needs to be done, before being able to approach these more specific issues (see Conclusions, page 16).

NEW CHANGE THAT WE HAVE ADDED IN OUR MANUSCRIPT

When performing a narrative systematic review, we have decided that when inconsistency is not applicable because there is only one study, we will downgrade the quality of evidence anyway because it is an indicator that the literature is not well established in that area. We have observed that the overall quality of evidence for factors that are explored for only one study are (we think, inappropriately) consistently higher than factors explored by more than one study (see Inconsistency, page 11).