Author’s response to reviews

Title: Elaborating on the Assessment of the Risk of Bias in Prognostic Studies in Pain Rehabilitation using QUIPS – Aspects of Interrater Agreement

Authors:

Wilhlemus Grooten (Wim.Grooten@ki.se)
Elena Tseli (Elena.Tseli@ki.se)
Björn Äng (Bjorn.Ang@ki.se)
Katja Boersma (katja.boersma@oru.se)
Britt-Marie Stålnecke (britt-marie.stalnacke@umu.se)
Björn Gerdle (bjorn.gerdle@liu.se)
Paul Enthoven (paul.enthoven@liu.se)

Version: 1 Date: 11 Dec 2018

Author’s response to reviews:

Reviewer reports:

Reviewer #1:

Thanks for the opportunity to review this manuscript. The paper presents a welcome assessment of interrater reliability and critique of the QUIPS tool. Overall the paper is well written and clear. I have comments and suggestions to clarify study methods and guidance recommendations provided, as well as some general suggestions for more consistent language and descriptions for prognosis systematic reviews.

1. The terms 'prognostic factors' and 'predictors' are each used. It is unclear if these are intended to have different meanings (e.g. factors associated (independently?) with outcome, predictive models to identify risk groups, or factors acting as treatment effect modifiers) - since language is still so inconsistent in the prognosis field, it warrants early definition of terms followed by consistent use.

Answer: Thank you for highlighting this point. As indicated by the reviewer, the literature is still inconsistent on the definition of prognostic/predictive factors. As suggested, we have now added
an early definition on this, and we have elected the term prognostic factor throughout the manuscript. Thank you.

2. It appears that the systematic review is a broad review of all prognostic factors/several outcomes. This should be clearly described in the methods, including clear description of how QUIPS was applied, and challenges in this context. Was QUIPS operationalized and assessed separately for each prognostic factor/outcome under consideration? For example, the judgement for assessment of confounding may differ for various prognostic factor/outcome associations; the risk of bias related to prognostic factor measurement will vary considering different prognostic factors. How was this considered? If not, how might this have impacted your results and conclusions? Challenges and recommendations for future reviews?

Answer:

The reviewer is correct in his/her comment on the present wide approach on prognostic factors after MDR. In our study protocol we described that concerning the primary prognostic factors and outcomes, we rated the overall RoB during the QUIPS rating. This has had, of course, an impact on the specificity of RoB judgment in several QUIPS domains (prognostic, outcome, confounders and statistical analyses). We learned that it is not realistic to use several QUIPS forms for each outcome/prognostic factor or for each statistical model. Instead, we used “specific notes” to highlight RoB issues for each specific outcome/prognostic/confounding factor and these were used during our discussions for reaching agreement. These notes were revised and also used during the GRADE process for each specific prognostic factor/outcome and statistical model. Moreover, we have used detailed information on the specific prognostic and outcome factors as well as and included adjustments, as footnotes in the forest plots for the reader to be able to make judgments of potential RoB and confounding by themselves. This, we believe, increased the transparency of the results.

In the manuscript, we have now clarified this aspect by added a new paragraph “QUIPS rating procedure in our method section on page 7

Also, in the Discussion, we have on the reviewers suggestion now added a recommendation for future use of QUIPS in broad reviews (page 17). Thank you.

3. Language related to 'quality' vs 'ROB' seem to be used interchangeably (page 4), but should make distinction for clarity (e.g. page 12 bottom; page 9, line 1). It will likely add clarity to reserve use of 'quality' language for the broader GRADE assessment of the overall quality of
evidence available, with 'risk of bias' used to describe study validity and impact on potential bias due to study methods.

Answer:

Thank you for this point, we acknowledge the value of the suggestion of the reviewer and have now used these terms more consistent as suggested. Changes in the manuscript are made on page 10 and page 18.

4. Description of the Hayden 2013 study should be clarified. This study reported results from various published systematic reviews that had used the QUIPS tool in various ways. I suggest reporting methods (page 5, line 21) as '…kappa values evaluating QUIPS WERE REPORTED TO VARY from…'. On page 10 this should be clarified ('raters' were not selected, they were simply systematic review authors who had used the QUIPS tool, with additional detail from those who responded to a survey). It seems that it would be most relevant in this study to limit comparison of your results to those systematic reviews that were most comparable (i.e. broad prognostic factor reviews that assessed by 6 domains as recommended).

Answer:

As suggested, we have now omitted the first point (on “selected” reviewers). Thank you for pointing out this misunderstanding.

Concerning the issue of the comparison of our results to previous systematic reviews, we believe that it is important to investigate the possible explanations of the differences between our results and especially to the work by Hayden et al (2013), since in their paper they have described and analyzed the developmental process of the tool. We agree with the reviewer that it could be difficult to compare our results directly with the work by Hayden et al (2013), and this has now been indicated in the text on page 15. We believe that this comparison – now more carefully stated – could have scientific value in the sense that it clarifies several pitfalls for future studies on inter-rater agreement.

5. I think that a particularly useful contribution to the literature is a more detailed specific discussion of issues relevant to the chronic pain rehab field. Since the recommendation is that QUIPS should be operationalized for the review question, I would like to see more elaboration and examples of these specific issues; how you resolved and what are your recommendations for full operationalization relevant to this chronic pain rehab field.

Answer:
In table 3, we have added some examples on how we from the chronic pain field have dealt with this.

6. Clarify - were the 5 studies used for pilot testing from the same review? Were they included in the interrater reliability testing?

Answer: These five papers were selected in an early stage of the process, i.e. during the eligibly phase of study selection. These papers were neither included in the review, nor in the interrater reliability testing. This information is now added in the manuscript on page 7. Thank you.


Answer: Thank you for detecting this typo. It should read 15. Changes are made on page 9.

8. Page 7, line 16. 'All ROB domains in all STUDIES (n=43) were separately judged by the raters as having low, MODERATE, or high ROB'. (correct wording)

Answer: Thank you for detecting this typo. Changed as suggested on page 8

9. Be clear and purposeful about use of 'papers' or 'studies'. Were 43 studies assessed, or 43 papers (i.e. these may be different if there are multiple publications of the same study). Since this is common in prognostic factor reviews, it will be good to be consistent with the language used.

Answer:

We have noticed that we haven’t been consequent in this and we have therefore made several changes in the manuscript for a more consistent language. We have used the term “paper” to the included studies in our review.

10. There are a few references to differences in assessment between studies with high and low ROB (e.g. Page 8, line 13). How was this assessed? How were high and low ROB studies defined? Was this before consensus or after? Overall or by domain? This should be described clearly in the methods.
Answer:

Thank you for your point. We have clarified the procedure in the extra paragraph on page 7. In the elaboration part, we also have discussed this issue.

11. Page 8, lines 17-22 describe number of domains rated similarly. It seems that it would be more useful to describe which specific domains tended to be more challenging/less agreement, rather than the number of domains.

Answer:

Thank you for this comment. We have now specified the domains at this point. Moreover, information on which specific domains tended to be more challenging/less agreement is available in table 2 and discussed on page 10. During the revision, we detected a few errors which now have been corrected.

12. The distinction between Round 1 and Round 2 in tables and text of results is of unclear value. If there is an important distinction in the methods between Round 1 and Round 2 (e.g. were the QUIPS items operationalized differently or more fully for Round 2? Were different reviewers involved? Were the studies newer?), any different methods should be clearly described in the methods section, then implications discussed. If not, I suggest dropping these subgroups from results tables.

Answer:

Thank you for your suggestion. As now clarified in the manuscript, between the two rounds we updated our “key-list” and the raters were more confident in their assessment during round 2 than round 1. Interestingly, there were still no large differences between the rounds concerning the kappa or % agreement. However, due to the suggestion of the second reviewer, we found that the proportion of items with high RoB was nearly doubled in the second round in both raters. We have this now added this in the manuscript on Page 11, hence we did not drop the comparison between round 1 and round 2.

13. It is useful to point out where there were particular challenges with application of the QUIPS tool. The manuscript will be further strengthened by describing what your solution was. What was the additional operationalization that you included - did this make a difference in the Round 2 assessments? Table 3 is helpful for further operationalization of the QUIPS items. Did you also further operationalize the tool for your domain judgments? This will be helpful to add to the elaboration specifically for the chronic pain rehab field.
Answer:

Thank you for this comment. This was the main point of this paper. In Table 3, we have now added examples on how we from the chronic pain field have dealt with this and have come with suggestions how to tackle the problems with, for example, multiple prognostic factors and outcomes, uni- and/or multi-variate modelling or the avoidance of double counting of the same methodological flaw in different domains.

14. Page 14, lines 15-20. It is a useful example to describe a potential a cut-point to help inform judgements about risk of bias due to participation and attrition. However, related discussion should include caution that while it may be useful as a guide, reviewers still need to consider this number in the context of other study information (e.g. a study may have less than 33% loss to follow-up, but if this loss was systematic based on a characteristic that could confound the association between the prognostic factor and outcome of interest, then the study still could be at high risk of bias).

Answer:

We agree with the reviewer that we should use recommendations rather than “rules”. Every time the researcher need to make a decision based on previous knowledge they may use the guide/recommendation, but not in the sense of decision tree. We have added a sentence on this in the discussion part, in the implications paragraph on Page 20 Thank you.

Minor editing suggestions:

15. Abstract, line 6. I would not call QUIPS a 'new tool' since it has been around since 2006.

Thank you. We have now rephrased this sentence

16. Abstract, line 10. It seems that goal is to elaborate on the OPERATIONALIZATION of the instrument (or use of the instrument?).

Thank you, we have changed this to “the use” of the instrument

17. Abstract, line 13. We performed a SYSTEMATIC REVIEW and meta-analysis…
Changed as suggested, thank you

18. Page 4, line 13. '…whether a potential prognostic factor is associated with an outcome and secondly to ESTIMATE the strength…'

Changed as suggested, thank you

19. Page 5, line 5. Cochrane Prognosis Methods Group (drop 'The' and 'Collaboration')

Changed as suggested, thank you

20. Page 9, line 11-12. 'Maximal disagreement was not caused by one of the raters judging CONSISTENTLY lower or higher…'.

Changed as suggested, thank you

21. Page 12, line 11. 'Most of the studies included in the present study were classified as having multiple DOMAINS WITH HIGH risk of bias'.

Changed as suggested, thank you

22. Page 13, line 12. 'modification/clarification of the prompting items for each specific REVIEW QUESTION

Changed as suggested, thank you

23. Page 13, line 22. 'discussed before performing ROB ASSESSMENT IN a systematic review'.

Changed as suggested, thank you

Answer minor editing suggestions:

Thank you for your minor editing suggestions. We have made changes accordingly.
Reviewer #2: This paper reports on the experiences of a systematic review team when using the QUIPS tool for assessing risk of bias in a systematic reviews of prognostic factors for long-term outcomes after multidisciplinary rehabilitation in patients with chronic pain. They report inter-observer agreement for risk of bias assessment across all domains and for each of the six risk of bias domains, and make suggestions for improving the process of scoring risk of bias of prognostic factor studies.

(1) Background: The introduction includes a paragraph on meta-analysis (estimating average effect size), which is fine but is not specific to prognosis reviews and not entirely relevant to the objective of this paper. It is also not quite clear if the paper by Page (reference 10) concerns prognosis studies or any type of study design. It would be good to focus the introduction on the importance of assessing quality and risk of bias in prognosis studies specifically. The weaknesses in prognosis research and the importance of assessing risk of bias in these types of studies have been highlighted in many papers (e.g. BMJ series by Moons & Altman; PROGRESS series; Hemingway et al. 2010; Hayden et al, 2013; Kyzas et al 2005; etc), and it would be helpful to highlight issues of specific importance to prognosis studies.

Answer:

Thank you for the suggestions of relevant references and retuning the background so it is better focused to the topic on interest and not too unspecific. We have now made changes in the introduction accordingly. Changes are made on Page 4.

(2) Results / Discussion: I am not quite sure why results and discussion are written as one section - this could be easily separated?

Answer:

Yes, we have now separated this according to the reviewer’s suggestion.

(3) Results: inter-rater agreement - If I read this paragraph correctly, the results show that in 25/43 papers at least 4 domains were scored the same by the two observers. Describing agreement in a cumulative way might be more informative than describing agreement in 1, 2, 3, 4, 5, or 6 domains.

Results: Expected agreement is dependent on the distribution of the marginal totals (distribution of low, moderate, or high risk of bias) and consequently kappa is highly sensitive to this. Table 1, which presents the full 3x3 table shows that risk of bias was not evenly distributed and
numbers are low for some categories. This may be more prominent for the analyses of agreement per bias domain, where numbers are smaller. This needs to be clear in the results: perhaps add a table to present information regarding the marginal totals (e.g. for each domain, the proportion (n/N) assessed as high/moderate/low/risk of bias for each observer).

Answer:

We agree that the cumulative way is informative and we have added this into the results section. Moreover, we thank the reviewer for the suggestion to calculate the proportion n/N. This highlighted the difference between the two rounds (a nearly doubling of the proportion with high RoB). For that reason we have added (%) in Table 1.

(4) Discussion: I am wondering if some of the findings for agreement, especially by risk domain, are influenced by the dependence of kappa on the distribution of margin total (see previous comment), and this might partly explain the discrepancy noted by the authors compared to other studies investigating the reliability of the QUIPS tool. On page 12, the authors do discuss the fact that agreement seemed higher in reliability studies where a larger proportion of papers included in the analysis were of better quality - which might indicate a more balanced distribution. Although there has been some debate in the literature to what extent this dependence of kappa on margin totals (or: prevalence) is a problem and how this should be dealt with, I think this needs to be acknowledged in the discussion.

Answer:

Thank you for starting this discussion on the influence of prevalence distribution on the kappa coefficient. We felt that we were not clear enough in the point that we wanted to make: that we felt it easier to read and assess the quality in low RoB studies compared to assess the quality in poorly written papers and we have now added one sentence on this in the discussion. Moreover, we think the reviewer is right that the differences in kappa between our study and other studies that we found could also be due to differences in the prevalence distribution over the 3x3 tables. Sim and Wright (2015) showed that both prevalence and bias (the extent to which the raters disagree on the proportion of positive/negative cases) has an influence on the Kappa values and in their paper they gave examples how to proceed to adjust for this bias using a PABAK (Prevalence-adjusted bias-adjusted kappa). However, they also mention the critics of this procedure by Hoehler (2000) stating that PABAK generates a value of kappa that does not relate to the real situation. In contrast to other statistical tests like Chi2, to our knowledge, there is no problem with a small number of observations or no observations in a cell when using kappa. These two arguments made us make the decision not to adjust the kappa values for prevalence. We have made changes in the discussion on this point on Page 18. Thank you.

References to this answer


(5) Discussion: The authors do not appear to report on the strengths or weaknesses of their study, e.g. was the number of papers large enough for the analysis? Are the results generalisable to other reviews conducted in other clinical areas, number of observers, …?

Answer:

Thank you for commenting this. We felt that we wanted to restrict the paper in length, but we agree with the reviewer that it is important to mention these aspects. We have now added a new paragraph in the discussion section on Page 19 in which several aspects are discussed.

(6) Conclusions: The authors indicate that they put forward suggestions for improving the tool. I am not sure if they have. They elaborate on the process of using QUIPS, and how this can be improved (e.g. by providing more detailed guidance for scoring each signalling items, and how to weigh the different items when assessing risk of bias; by making sure observers are sufficiently knowledgeable; by embedding discussions regarding interpretation of items). This makes perfect sense and is important (tends to be important when assessing risk of bias for other types of studies as well), but do they have recommendations for amending the tool itself? And if so, what would be the next steps for improving and re-testing the tool?

Answer:

Yes, the reviewer is right in this comment of our conclusions and we have now reformulated it, using some of his/her suggestions. Also some of these suggestions are included in the paragraph on suggestions for further research in the discussion section on Page 19. Thank you.

(7) Typo: page 6, line 18: "included presented".

Answer:

Thank you for finding this typo