Author’s response to reviews

Title: Impact of peer review on discussion of study limitations and strength of claims in randomised trial reports: a before and after study

Authors:

Kerem Keserlioglu (k.keserlioglu@amc.nl)

Halil Kilicoglu (halil.kilicoglu@nih.gov)

Gerben ter Riet (g.ter.riet@hva.nl)

Version: 1 Date: 13 Jul 2019

Author’s response to reviews:

Response letter to the reviewers.

Dear Dr. Meerpohl,

Many thanks for your email and giving us the opportunity to make some final corrections and responses to the reviewer comments. Below, we will respond in italics to the points raised by the reviewers.

Reviewer #1: Very interesting and important piece of work which resonates with my own experience as an editor. It is a paper that opens up a new avenue of research in to peer review, and whilst I have some concerns over the reliability of some of the results, the findings are indicative of a real effect.

Answer: We thank the reviewer for these compliments and sharing the belief that there may be a real effect.
I agree with the authors that the main finding of note is that the number of limitation sentences increased. However, I do question the validity of the hedging scores: They capture the right terms, but I have a concern that the hedging tool does not differentiate hedging around limitations and hedging around other appropriate description of the research study. To explore this a bit more, I read through the discussion section to see where hedging was used. I could find uses of 'may', 'could' and 'suggest' as hedging terms to describe features of the study that were listed as strengths. Perhaps hedging is prevalent in responsibly reported discussions, and not necessarily a marker that limitations are being acknowledged? In this regard the findings for the weighted and unweighted analysis are unsurprising, as it is possible that people are using terms such as 'appear', 'may' or 'possibly' throughout the discussion, and not just to highlight a limitation. I also suspect that researchers use different hedging terms as a way of maintaining reader interest in their paper. Under this assumption people end up using different terms interchangeably with the aim of bringing linguistic variety to their manuscript. These terms may well score differently, yet this does not reflect a conscious, preferential selection of terms that peer reviewers will notice and seek to change. In reading this paper I've come to the realisation that I've written different words to express similar probabilities but never thought too much about how different their meanings were from each other ('possibly', 'probably', 'may be', 'likely'). In the context of this paper and the analysis this limitation is important, but in the broader context of editorial and peer review research, I am less worried by how the authors have measured hedging and more interested in where future research efforts could be directed.

Answer: We thank the reviewer for expressing his careful considerations after exploring some potential reasons underlying our findings. We concur with the reviewer that “Perhaps hedging is prevalent in responsibly reported discussions, and not necessarily a marker that limitations are being acknowledged.” To us, the concepts of ‘hedging’ and ‘limitation acknowledgment’ are distinct. We also think (or agree) that the limitation detection results are more readily interpretable than the hedging scores.

The hedge detection software has pretty good measurement characteristics [reference #9], its performance might have been better had we partitioned the text segments within the discussion section even more finely. On the other hand, as reported, we performed sensitivity analyses by excluding the limitations-related sentences, but found that this did not make much of a difference. If the reviewer is right in surmising that hedging may be used somewhat unconsciously and unsystematically, it may in fact explain our “null” findings (a ‘cancelling out’ effect).

We added a paragraph with suggestions for further research. [page 7, line 16 to page 8, line 7]
The only real comments I have really are about the implications for research which could be made more directive. To what extent does the author team agree with these research recommendations that could be added to the article?:

1. A study comparing changes to limitations in the full text and abstract reporting. Summary versions are often where people remove nuance from reporting findings, so does a limitations-rich discussion section reduce spin in the conclusions and in the abstract?

2. A qualitative study based on interviews with editors and researchers to understand better what their behaviour is around limitations reporting and editing.

3. An assessment of how reporting guidelines stress the importance of limitations & whether there is a good understanding of what these are in the context of RCTs (and systematic reviews).

We thank the reviewer for bringing up more suggestions for further research. His first suggestion, as far as we can tell, mixes the issue of limitations with that of hedging and introduces the term ‘spin’. We agree that these phenomena are probably associated. We tried to convince one particular journal to use our software in a low-budget randomized trial to see if the alert of no or very few limitation-related sentences would trigger actions by reviewers and/or editors that would lead to more limitations being mentioned. However, the logistical demands were judged as too high and more resources are needed. We hope that the reviewer agrees that we added this suggestion to the discussion (lines xx-xx). In our small qualitative analysis we focused on the peer review reports which are for some journals readily available, analyzing these on a larger scale could be interesting in understanding how this negotiating process develops and to what results it leads. Adding interviews with editors and researchers, as the reviewer suggest, can help interpreting these qualitative results. (page 7, line 16 to page 8, line 7). As to this reviewer’s last suggestion, we added Appendix 2 in which we list what 7 different reporting guidelines have to say about self-acknowledgment of limitations. We added a few remarks on that in the discussion (page 7, lines 7-8). Please let us know if you consent to an acknowledgment for suggesting the additional work in Appendix 2 and avenues for future work.

I would like to see the research in this area rely less on applying automative techniques on large datasets and instead use judgment made by researchers themselves on a smaller random sample of manuscripts. As noted above I do not have strong faith in the hedging scores, but I do think that the creative approach that the authors took here has generated an interesting and useful aspect of peer review that they should now study further.
We thank the reviewer for his compliment and encouragement. We have added a section on implications for practice were we argue that a flagging up function of software tools may be defensible, but acknowledge the importance of human judgement for the final verdicts. We agree with the reviewer that more research concerning the meaning of hedging scores is needed before one can make reliable judgements based on these. [page 8, lines 1-9]

Reviewer #2: This is a useful study.

My main criticism is that there is no mention of the qualitative analyses of five manuscript-publication pairs until the Discussion (page 6). It should be briefly explained earlier.

We thank the reviewer for pointing this out. In the new version, we inserted this approach in the Methods section and mention the results in the Results section. [page 4, lines 19-22] We should mention that we performed a qualitative analysis on two article pairs only instead of 5 as mentioned earlier. We apologise for the error. We have added the full result of the qualitative analysis in appendix 3. [page 19-21]

Regarding the sensitivity analysis excluding BMJ Open and BMC Medicine articles (pages 4-5), I would have liked the number of editorial team members in the non-excluded journals to be stated there.

These numbers have now been added in the caption to Appendix 1 (page 16, lines 2-3).

Regarding the limitations of the study, not only did it only include open peer review journals (as acknowledged, page 7), but also the journals included (BMJ Open and BMC Medicine) are arguably of a higher than average calibre, and I think it would be good to acknowledge this.

This limitations has been added (page 8, line 12-13).

I am puzzled by the statement that 'Blind peer review may lead to different results, but is uncommon in biomedical journals' (page 7). I don't know the statistics on open peer review, but my perception is that most journals still have blind peer review (unfortunately).

We decided to omit this detail, since there is much confusion on the definition and use of blind review (page 8, lines 13-14).
I would have liked some comment on whether/how the results of the study could potentially be used in practice by academic writers, reviewers, and editors.

We have added a section on implications for practice (page 7, lines 1-9; page 9, lines 8-10)

For consistency, one term should be used instead of 'editorial team' and 'editorial committee' (page 5) and 'editorial board' (pages 4 and several pages later).

Thank you for raising this point. We have now consistently chosen for the term “editorial team”.

Formatting-wise, several paragraphs are too long, particularly on pages 3, 4-5, and 7. Shorter paragraphs (with logical breaks) are beneficial for readers.

We have added several breaks in the sections the reviewer suggested.