PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review
AUTHORS	Page, Matthew McKenzie, Joanne Higgins, Julian

VERSION 1 – REVIEW

REVIEWER	Gordon Guyatt and Luis E Colunga-Lozano Gordon Guyatt Distinguished Professor, Department of Health Research Methods, Evidence, and Impact McMaster University Canada Luis E Colunga-Lozano MSc Health-related methodology student McMaster University Canada
REVIEW RETURNED	05-Oct-2017

GENERAL COMMENTS	Please elaborate on any 'No' answers in the free text section below. Yes No N/A 1. Is the research question or study objective clearly defined? 2. Is the abstract accurate, balanced and complete? * 3. Is the study design appropriate to answer the research question? 4. Are the methods described sufficiently to allow the study to be repeated? ** 5. Are research ethics (e.g. participant consent, ethics approval) addressed appropriately? 6. Are the outcomes clearly defined? * 7. If statistics are used are they appropriate and described fully? 8. Are the references up-to-date and appropriate? 9. Do the results address the research question or objective? 10. Are they presented clearly? 11. Are the discussion and conclusions justified by the results * 12. Are the study limitations discussed adequately? 13. Is the supplementary reporting complete (e.g. trial registration; funding details; CONSORT, STROBE or PRISMA checklist)? 14. To the best of your knowledge is the paper free from concerns over publication ethics (e.g. plagiarism, redundant publication, undeclared conflicts of interest)? 15. Is the standard of written English acceptable for publication?

	_____ This paper represents a well-thought-out, well-implemented, well-presented summary of instruments available for addressing publication and selective reporting biases. The paper appears to be the first step in a project to develop a new and presumably definitive instrument to address these issues. We do have several substantive concerns with methods, results and discussion that we present below. 1* = The authors may consider adding a background section to the abstract. 2* = The authors fully acknowledge the limitation associated with only one person reviewing and abstracting the all information, as well as the language restriction (English). There authors could have contacted other experts in seeking difficult-to-identify instruments and should perhaps add this to the acknowledgement of limitations. 3. The authors could do a better job of clarifying the issue of assessment at an individual study level and a body of evidence level. This, to us, is crucial. Publication bias can only be assessed at the body of evidence level. Selective reporting can be assessed at both the individual study level and the body of evidence level. For each instrument, it should be clear whether assessments are at the individual study level or the body of evidence level. 4. Related to the prior point, we are most familiar with the GRADE approach. In their Table 4, the authors state that GRADE addresses selective non-reporting at a study and not outcome level. The relevant text from the GRADE writing is as follows: "For example, a systematic review of the effects of testosterone on erection satisfaction in men with low testosterone identified four eligible trials [14] . The largest trial's results were reported only as ''not significant'' and could not, therefore, contribute to the meta-analysis. Data from the three smaller trials suggested a large treatment effect (1.3 standard deviations, 95% confidence interval 0.2, 2.3). The review authors ultimately obtained the complete data from the larger trial: after including the less impressive results of the large trial, the magnitude of the effect was smaller and no longer statistically significant (0.8 standard deviations, 95% confidence interval _ 0.05, 1.63) [15]…" "One should suspect reporting bias if the study report fails to include results for a key outcome that one would expect to see in such a study or if composite outcomes are presented without the individual component outcomes." Note that within the GRADE framework, which rates the quality of a body of evidence, suspicion of selective reporting bias in a number of included studies may lead to rating down of quality of the body of evidence. For instance, in the testosterone example above, had the authors not obtained the missing data, they would have considered rating down the body of evidence for the selective reporting bias suspected in the largest study." The authors were intending to describe assessment of selective reporting on an outcome by outcome basis at both a study and body of evidence level. In retrospect, this is not absolutely explicit that this is on an outcome by outcomes basis, or perhaps that it addresses the body of evidence level, but it still seems to these reviewers that the implication is very strong on both counts (the testosterone example is clearly by outcome across studies and implies inferences on the body of evidence). The first of the criteria currently listed in Table 4 at a study level need to be modified at an outcome level: should be "A particular outcome was clearly measured, but no results were reported." The authors should correct this in their table. Give the mistake here, I wonder if a similar misinterpretation has occurred for other instruments. 5. We may be mistaken, but we suspect the way this paper is put together that the authors are launching an initiative to develop a new instrument. Whether this is true or

	not, given that the authors introduce the idea of a new instrument, some further discussion is warranted: 1) No matter how good the new instrument is, unless reviewers have full access to protocols and trial registries, and the day arrives that IRBs require investigators to enter all proposed trials in trial registries, there will continue to be major problems with confident assessment of publication/reporting bias. 2) Current instruments such as the Cochrane RoB for RCTs and the ROBINS are directed at individual studies. The authors, if we understand, are looking for an instrument that would be applied at the systematic review level. A comment on how this instrument would relate to assessments of selective reporting at an individual study level would be informative. 3) Given that body of evidence level assessments of quality/confidence/certainty of evidence are what the GRADE working group has quite successfully addressed thus far, such an instrument would naturally complement current GRADE guidance. Collaboration with GRADE in the development of such an instrument, or even creating a new GRADE project group to address the issue, might help ensure the highest quality instrument and its rapid uptake in the methods community.

REVIEWER	Agnes Caille CIC INSERM1415, INSERM UMR1246 France
REVIEW RETURNED	09-Oct-2017

GENERAL COMMENTS	Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review Reviewer report The aim of this paper was to review the existing tools for assessing risk of reporting biases in studies and synthesis of studies. The authors report the content and measurement properties of 18 retrieved tools. The conclusion is that a new, comprehensive tool is needed to cover all aspects of reporting bias. The paper is very well written. The methodology used for the systematic review fulfils the standards for good-quality research except for screening of articles and data collection, which are performed by only one author. Results are complete and clearly reported. Specific concerns 1. One point that needs further explanation is the respective usefulness of statistical methods and other tools to assess the risk of bias due to selective publication. Why did you exclude statistical methods assessing the risk of bias from your work? In which cases should we use a statistical method rather than or in addition to another kind of tool such as those you reviewed? 2. The conclusion lacks some recommendations regarding what tools we should use while waiting for the new tool the authors are currently developing. 3. You selected 42 reports, it would be clearer to find those 42 references in the Table 1. Presently, I retrieve only 28 references in the first column of the Table. 4. I was surprised that the first tool was created in 1998 (Table 1). That seems to me very early. The next one was created in 2010. Perhaps this needs to be commented on. 5. Table 1: you should add that Higgins 2011 Cochrane risk of bias tool for randomized trials is the same as ROB 1.0 in the following tables. 6. Table 2: could-you clarify what is “psychometric or cognitive

	testing" for data sources used to inform tool content? 7. For the tools you reviewed, is there possible to perform an assessment of validity such as for statistical methods? 8. In the Eligibility criteria section, the examples within parentheses should be cited in a similar order, that is, for the second set of examples (e.g., estimate of intervention efficacy or harm, estimate of diagnostic accuracy, association between exposure and outcome) 9. The main limitation of this study is the extraction of data by one person only. It would have been preferable to have duplicate data extraction as is recommended in such reviews to avoid errors as well as subjective interpretation (1). At least, the data extraction form should have been pilot-tested on a sample of papers by at least two authors to obtain an assessment of agreement. 10. In Table 6 and online supplementary Table S5, please explain IPD abbreviation; I guess it stands for Individual Participant Data, but this needs to be reported. 11. Flow diagram Figure 1: - Identification step, records identified … by electronic searches. The "by" is lacking. Could you provide the number of records identified from other sources that have been identified by Google Scholar and by searching the references of included articles? - Screening step, could you add the main reasons for exclusion of records? - Eligibility step, could you clarify the situation for the 7 full-text articles excluded for "No psychometric properties evaluated"? References 1. Higgins JPT, Green SB. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration, 2011. Vol. Version 5.1.0 [updated March 2011]. Available from http://handbook.cochrane.org.

VERSION 1 – AUTHOR RESPONSE

Dear editors,

We would like to thank you and reviewers for their consideration and careful review of our paper entitled "Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review". We think that these comments and suggestions have been useful and we have incorporated them into our manuscript and given explanations of such revisions.

Below we list the reviewer comments followed by our responses.

Reviewer 1
This paper represents a well-thought-out, well-implemented, well-presented summary of instruments available for addressing publication and selective reporting biases. The paper appears to be the first step in a project to develop a new and presumably definitive instrument to address these issues. We do have several substantive concerns with methods, results and discussion that we present below.

1 = The authors may consider adding a background section to the abstract.
Author response: We have added a Background section to the abstract.

2 = The authors fully acknowledge the limitation associated with only one person reviewing and abstracting the all information, as well as the language restriction (English). The authors could have contacted other experts in seeking difficult-to-identify instruments and should perhaps add this to the acknowledgement of limitations.
Author response: We failed to mention that in April 2017, the lead author emailed the list of included tools to 15 individuals with expertise in reporting biases and risk of bias assessment, and asked if they knew about any other tools we had not identified. No missing tools were identified via this method. We have added this information to the "Search methods" section of the manuscript.

3. The authors could do a better job of clarifying the issue of assessment at an individual study level and a body of evidence level. This, to us, is crucial. Publication bias can only be assessed at the body of evidence level. Selective reporting can be assessed at both the individual study level and the body of evidence level. For each instrument, it should be clear whether assessments are at the individual study level or the body of evidence level.
Author response: In the third paragraph of the Background, we have introduced the distinction between assessment at the study level and body of evidence level (which we call the "synthesis level"). In this paragraph we acknowledge that assessments of risk of bias due to selective publication can be directed at the level of the synthesis, while assessments of the risk of bias due to selective non-reporting can be directed at the level of the individual study and at the level of the synthesis. In Table 1, which lists all included tools, we have indicated in the final column the level of assessment of each tool. We have summarised the frequencies of each level of assessment in the Results text (second paragraph of the sub-section on "General characteristics of included tools").

4. Related to the prior point, we are most familiar with the GRADE approach. In their Table 4, the authors state that GRADE addresses selective non-reporting at a study and not outcome level. The relevant text from the GRADE writing is as follows: "For example, a systematic review of the effects of testosterone on erection satisfaction in men with low testosterone identified four eligible trials [14]. The largest trial's results were reported only as ''not significant'' and could not, therefore, contribute to the meta-analysis. Data from the three smaller trials suggested a large treatment effect (1.3 standard deviations, 95% confidence interval 0.2, 2.3). The review authors ultimately obtained the complete data from the larger trial: after including the less impressive results of the large trial, the magnitude of the effect was smaller and no longer statistically significant (0.8 standard deviations, 95% confidence interval _ 0.05, 1.63) [15]…" "One should suspect reporting bias if the study report fails to include results for a key outcome that one would expect to see in such a study or if composite outcomes are presented without the individual component outcomes." Note that within the GRADE framework, which rates the quality of a body of evidence, suspicion of selective reporting bias in a number of included studies may lead to rating down of quality of the body of evidence. For instance, in the testosterone example above, had the authors not obtained the missing data, they would have considered rating down the body of evidence for the selective reporting bias suspected in the largest study." The authors were intending to describe assessment of selective reporting on an outcome by outcome basis at both a study and body of evidence level. In retrospect, this is not absolutely explicit that this is on an outcome by outcomes basis, or perhaps that it addresses the body of evidence level, but it still seems to these reviewers that the implication is very strong on both counts (the testosterone example is clearly by outcome across studies and implies inferences on the body of evidence). The first of the criteria currently listed in Table 4 at a study level need to be modified at an outcome level: "One or more outcomes of interest were clearly measured, but no results were reported", should be "A particular outcome was clearly measured, but no results were reported." The authors should correct this in their table. Give the mistake here, I wonder if a similar misinterpretation has occurred for other instruments.
Author response: We thank the reviewers for clarifying this aspect of GRADE. We have revised Table 4 as suggested, by clearly indicating that GRADE (and the tools which are based on GRADE) recommends an outcome-level assessment of selective non-reporting. We have also added new rows

to Table 4 to indicate that four tools, including GRADE, guide users to assess the quality/risk of bias in a synthesis due to selective non-reporting in studies. We have revised the Results text under the "Tool content" sub-section accordingly. We have also checked our classifications for all other tools thoroughly and are confident that we have not made any other errors.

5. We may be mistaken, but we suspect the way this paper is put together that the authors are launching an initiative to develop a new instrument. Whether this is true or not, given that the authors introduce the idea of a new instrument, some further discussion is warranted:
1) No matter how good the new instrument is, unless reviewers have full access to protocols and trial registries, and the day arrives that IRBs require investigators to enter all proposed trials in trial registries, there will continue to be major problems with confident assessment of publication/reporting bias.
Author response: The reviewers are correct in their assumption that we are developing a new instrument to assess risk of reporting biases. We have acknowledged this in the Competing Interests section of the manuscript. We agree that without access to detailed trial protocols, assessments of the risk of reporting biases will be challenging. We have noted in the final paragraph of the Discussion that guidance for the new tool will need to emphasise the value of seeking such documents.

2) Current instruments such as the Cochrane RoB for RCTs and the ROBINS-I are directed at individual studies. The authors, if we understand, are looking for an instrument that would be applied at the systematic review level. A comment on how this instrument would relate to assessments of selective reporting at an individual study level would be informative.
Author response: We agree that this is an important issue to resolve when developing the new tool, and we plan to assemble a working group to discuss this. We feel it is pre-emptive to discuss possible solutions in the current paper. To provide further evidence of the need for the new tool, we have noted in the final paragraph of the Discussion that "This tool could guide users to consider risk of bias in a synthesis due to both selective publication and selective non-reporting, given that both practices lead to the same consequence: evidence missing from the synthesis. Such a tool would complement recently developed tools for assessing risk of bias within studies (RoB 2.0 and ROBINS-I), which include a domain for assessing the risk of bias in selection of the reported result, but no mechanism to assess risk of bias due to selective non-reporting."

3) Given that body of evidence level assessments of quality/confidence/certainty of evidence are what the GRADE working group has quite successfully addressed thus far, such an instrument would naturally complement current GRADE guidance. Collaboration with GRADE in the development of such an instrument, or even creating a new GRADE project group to address the issue, might help ensure the highest quality instrument and its rapid uptake in the methods community.
Author response: We agree that the new instrument would complement current GRADE guidance (particularly the "Publication bias" domain), and welcome the idea of collaborating with the GRADE Working Group to ensure the highest quality instrument is developed, and to ensure its rapid uptake in the methods community.

Reviewer 2
The aim of this paper was to review the existing tools for assessing risk of reporting biases in studies and synthesis of studies. The authors report the content and measurement properties of 18 retrieved tools. The conclusion is that a new, comprehensive tool is needed to cover all aspects of reporting bias. The paper is very well written. The methodology used for the systematic review fulfils the standards for good-quality research except for screening of articles and data collection, which are performed by only one author. Results are complete and clearly reported.

1. One point that needs further explanation is the respective usefulness of statistical methods and other tools to assess the risk of bias due to selective publication. Why did you exclude statistical methods assessing the risk of bias from your work? In which cases should we use a statistical method rather than or in addition to another kind of tool such as those you reviewed?
Author response: We excluded from our review articles that cover statistical methods only, as these have already been summarised in previous systematic reviews (e.g. https://www.ncbi.nlm.nih.gov/pubmed/27502970 and https://www.ncbi.nlm.nih.gov/pubmed/25363575). We have clarified this in the Methods sub-section on "Eligibility criteria". We also note in this section that "Multi-dimensional tools with a statistical component were also eligible (e.g. those that require users to respond to a set of questions about the comprehensiveness of the search, as well as to perform statistical tests for funnel plot asymmetry)." We do not disregard statistical methods entirely, but rather believe that if they are applied, other factors need to be considered to reach a judgement about the risk of bias due to selective publication. We have clarified this in the Discussion (first paragraph of the "Explanations and implications" sub-section).

2. The conclusion lacks some recommendations regarding what tools we should use while waiting for the new tool the authors are currently developing.
Author response: We have refrained from recommending particular tools over others largely because there was little overlap in the focus of each tool. That is, we identified tools designed specifically for use in reviews of randomized trials, reviews of non-randomized studies of interventions, reviews of prognosis studies, reviews of laboratory animal experiments, and a particular type of analysis in reviews (e.g. network meta-analysis). The tools also varied in regard to the types of reporting biases they assess, and the level of assessment (e.g. synthesis of studies versus result in a study). Given they focus on different issues, we do not think it makes sense to include a recommendation in the conclusion regarding which tools systematic reviewers should use while waiting for the new tool to be developed.

3. You selected 42 reports, it would be clearer to find those 42 references in the Table 1. Presently, I retrieve only 28 references in the first column of the Table.
Author response: The 42 reports pertain to either one of the included tools, a study on the measurement properties of a tool, or an article that describes an included tool and presents data on measurement properties of that tool. There are only 28 references in Table 1 because Table 1 includes references to the tools only (not references to studies evaluating the measurement properties of tools; references to such studies are included in Table 6).

4. I was surprised that the first tool was created in 1998 (Table 1). That seems to me very early. The next one was created in 2010. Perhaps this needs to be commented on.
Author response: The earliest tool included in our review is the Downs-Black tool. This is a 27-item checklist which covers multiple sources of bias, and one of the items assesses risk of bias in selection of the reported result. We have stated this in the Results text (first paragraph of the sub-section on "General characteristics of included tools").

5. Table 1: you should add that Higgins 2011 Cochrane risk of bias tool for randomized trials is the same as ROB 1.0 in the following tables.
Author response: We have added "RoB 1.0" in parentheses in Table 1 as suggested (we have also changed "Higgins 2011 to "Higgins 2008" as 2008 was the year in which this tool was first published).

6. Table 2: could-you clarify what is "psychometric or cognitive testing" for data sources used to inform tool content?
Author response: We have added the following footnote to Table 2: "Psychometric testing includes any evaluation of the measurement properties (e.g. construct validity, inter-rater reliability, test-retest

reliability) of a draft version of the tool. Cognitive testing includes use of qualitative methods (e.g. interview) to explore whether assessors who are using the tool for the first time were interpreting the tool and guidance as intended."

7. For the tools you reviewed, is it possible to perform an assessment of validity such as for statistical methods?
Author response: Currently there are no formal methods available to assess the validity of quality/risk of bias assessment tools, so we were unable to assess the tools included in our review in this way.

8. In the Eligibility criteria section, the examples within parentheses should be cited in a similar order, that is, for the second set of examples (e.g., estimate of intervention efficacy or harm, estimate of diagnostic accuracy, association between exposure and outcome)
Author response: We have corrected the second set of examples as suggested.

9. The main limitation of this study is the extraction of data by one person only. It would have been preferable to have duplicate data extraction as is recommended in such reviews to avoid errors as well as subjective interpretation. At least, the data extraction form should have been pilot-tested on a sample of papers by at least two authors to obtain an assessment of agreement.
Author response: We agree it would have been preferable to have data collected by two authors independently. We have clearly indicated in the Discussion and Strengths and Limitations section that failure to do so is a limitation of our review, and readers will see from the abstract that "One author screened all titles, abstracts and full text articles, and collected data on tool characteristics".

10. In Table 6 and online supplementary Table S5, please explain IPD abbreviation; I guess it stands for Individual Participant Data, but this needs to be reported.
Author response: We have removed the "IPD" abbreviation and stated "individual participant data" in Table 6 and Table S5.

11. Flow diagram Figure 1:
- Identification step, records identified … by electronic searches. The "by" is lacking. Could you provide the number of records identified from other sources that have been identified by Google Scholar and by searching the references of included articles?
Author response: We have inserted "by" as suggested. We considered Google Scholar as an electronic search, so included the number of records screened from this database (n=300) in the total number of records identified by electronic searches (n=5,538). Therefore, this means that the box indicating "Records identified from other sources (n=16)" is referring only to records identified by screening the references of included articles. We have specified this in footnotes to Figure 1 (see Figure legend).

- Screening step, could you add the main reasons for exclusion of records?
Author response: We did not record the reasons for excluding titles and abstracts. Doing so for the 4,605 records that were excluded at this stage would have been very time-consuming, and is not a practice recommended in guidance for systematic reviews (e.g. Cochrane Handbook, PRISMA Statement).

- Eligibility step, could you clarify the situation for the 7 full-text articles excluded for "No psychometric properties evaluated"?
Author response: In Figure 1 and Table S2 (Table of Excluded Studies), we have changed the reason, "No psychometric properties evaluated" to, "Evaluation of use of tool in practice, but no measurement properties assessed". In such articles, the investigators evaluated how often a particular tool was used by systematic reviewers, or the ways in which reviewers completed the tool

(to see if the tool was applied appropriately or not), but no formal evaluation of measurement properties (e.g. inter-rater reliability) was performed.

VERSION 2 – REVIEW

REVIEWER	Agnes Caille Centre d'Investigation Clinique – INSERM CIC 1415 UMR INSERM 1246 - SPHERE CHRU de Tours - Hôpital Bretonneau - Bâtiment Tertiaire - 2ème étage 2 boulevard Tonnellé - 37044 TOURS Cedex 9
REVIEW RETURNED	23-Nov-2017

GENERAL COMMENTS	I feel my comments have been well addressed and I have no further recommendations. I have just one remaining question, which is not intended to modify the manuscript but just for my better understanding: Could you clarify what you mean by "vested interests of investigators" p27 ? Could you provide an example ?

VERSION 2 – AUTHOR RESPONSE

Dear editors,

We would like to thank you and reviewers for their consideration and careful review of our paper entitled “Tools for assessing risk of reporting biases in studies and syntheses of studies: a systematic review”. We think that these comments and suggestions have been useful and we have incorporated them into our manuscript and given explanations of such revisions.

Below we list the reviewer comments in normal text followed by our responses in bold text.

Reviewer 2
I feel my comments have been well addressed and I have no further recommendations.
I have just one remaining question, which is not intended to modify the manuscript but just for my better understanding: Could you clarify what you mean by “vested interests of investigators” p27? Could you provide an example?
AUTHOR RESPONSE: By “vested interests of investigators”, we are referring to any conflicts of interest that may influence investigators to not report unfavourable results. In some fields, the study investigators may have a vested interest that influences them to not disseminate studies that suggest an experimental intervention is ineffective or harmful (e.g. if the study is sponsored by a company who developed the intervention). For clarity, we have revised “vested interests of investigators” to “conflicts of interest that may influence investigators to not disseminate studies with unfavourable results”.