PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Reporting Quality in Abstracts of Meta-Analyses of Depression Screening Tool Accuracy: A Review of Systematic Reviews and Meta-Analyses
AUTHORS	Thombs, Brett; Rice, Danielle; Kloda, Lorie; Shrier, Ian

VERSION 1 - REVIEW

REVIEWER	Eric Youngstrom, Ph.D., & Jacquelynne Genzlinger, B.A. University of North Carolina at Chapel Hill, United States of America
REVIEW RETURNED	04-Jul-2016

GENERAL COMMENTS	This is a well-written paper that examines the quality of reporting of components of abstracts suggested by the PRISMA guidelines, using meta-analyses of the diagnostic accuracy of rating scales for depression published since 2005 as the sample of interest to establish a benchmark for current practices in quality of reporting. Strengths of the study include delineating a thoughtful adaptation of the PRISMA abstract criteria, which initially were developed for reports of clinical trials (a la CONSORT)(Moher, Schulz, & Altman, 2001) to make them more appropriate for tests of diagnostic accuracy (a la STARD)(Bossuyt et al., 2003), an exemplary approach to the formal search strategy, and a clear and balanced description of the limitations of the literature as well as the review itself. The entire paper displays excellent attention to detail. The central findings, that the level of reporting tends to be poor overall, and that it shows a medium to large (r=.45) correlation with the overall quality of reporting within the rest of the paper, are helpful descriptions of the current state of affairs. Though sobering, they are not surprising inasmuch as the PRISMA abstract recommendations are new (2013), and most of the articles included in the review predate the recommendations (as the authors note in the Discussion section). At times the article reads like a critique of the reporting in prior articles, which might feel unfair to the authors to the extent that their publication predated the criteria. Given that the authors also do not control the submission guidelines at journals, and any change on the part of editors is likely to be a slow and piecemeal process, it would be helpful to offer tips and suggestions about how to maximize the reporting of information within the current constraints of abstract lengths. The density required approaches that of classified advertisements in newspapers or online forums, so even suggested abbreviations or template sentences could be helpful. What follows are minor suggestions for refinement: It would be helpful to define acronyms (e.g., PRISMA, DTA, AMSTAR) on first use, lightening the cognitive load for readers.

When introducing the secondary aim of testing the correlation between editorial word limits and PRISMA abstract score, it would be helpful to mention the rationale for expecting a correlation – namely that short word limits may make it difficult to cover all of the required elements. This is included in the Discussion, so it is a simple matter of setting the frame for the hypothesis or research question earlier as well.

Several search terms occurred to us that were not included. Among them were “major depressive episode,” “dysthymia,” “dysthymic disorder” – all of which refer to types of depressive disorder in the DSM nosology. Similarly, we wondered about not using “area under curve,” AUC, ROC, and related terms as ways of finding studies of "diagnostic accuracy," since ROC is a more general framework and AUC is an effect size (Kraemer et al., 1999; Swets, Dawes, & Monahan, 2000; Youngstrom, 2013, 2014). Neither of these is a fatal flaw, but it is worth mentioning as a limitation; it is plausible that studies focused on these permutations might have been omitted. I am aware of at least two examples (focused on other disorders) that have published AUC and diagnostic likelihood ratios (Straus, Glasziou, Richardson, & Haynes, 2011), but not sensitivity and specificity for a single threshold, showing that it is empirically possible for a relevant study to escape the search net that was cast.

The authors do not report any estimate of inter-rater reliability for their coding scheme. The methods appear sound, even so. They can acknowledge this as a limitation, per the PRISMA or MARS guidelines.
It also might be helpful to offer a rationale for splitting item 3 into 3a and 3b. It was not intuitively obvious to me how the referents for 3a and 3b differed conceptually.

Overall, the paper offers a meticulous analysis and description of current reporting practices. It also provides a good prescription, in terms of both promulgating the PRISMA abstract criteria, and also allowing flexibility on the abstract word count in order to create sufficient room for more complete reporting.

References

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. British Medical Journal, 326, 41-44. doi:10.1136/bmj.326.7379.41

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1999). Measuring the potency of risk factors for clinical or policy significance. Psychological Methods, 4, 257.

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parellel-group randomised trials. The Lancet, 357, 1191-1194.

Straus, S. E., Glasziou, P., Richardson, W. S., & Haynes, R. B. (2011). Evidence-based medicine: How to practice and teach EBM (4th ed.). New York, NY: Churchill Livingstone.

	Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. Psychological Science in the Public Interest, 1, 1-26. doi:10.1111/1529-1006.001 Youngstrom, E. A. (2013). Future directions in psychological assessment: Combining Evidence-Based Medicine innovations with psychology's historical strengths to enhance utility. Journal of Clinical Child & Adolescent Psychology, 42, 139-159. doi:10.1080/15374416.2012.736358 Youngstrom, E. A. (2014). A primer on Receiver Operating Characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. Journal of Pediatric Psychology, 39, 204-221. doi:10.1093/jpepsy/jst062

REVIEWER	Ingrid Arevalo-Rodriguez Instituto de Evaluación Tecnológica en Salud- IETS Colombia
REVIEW RETURNED	21-Jul-2016

GENERAL COMMENTS	This is a well written study and also relevant to show faults of reporting in integrative studies. I congratulate the authors for this important effort. I only have a few comments related to this important manuscript: 1. Authors included in this study only DTA systematic reviews with meta-analysis of accuracy figures. This aspect is very clear in methods section. However, in the abstract, summary and other sections they called these reviews only as “meta-analysis”. My main concern is that readers can believe that “meta-analysis” is a specific study design, but we know that it is only the statistical analysis of a systematic review. Is there any way to handle this? 2. Methods/ Assessment of reporting: After modify the PRISMA tool, authors performed a pilot study to check if the modifications works or no? 3. Methods/ Data extraction: Regarding the abstract word limit, what was the source of this information? 4. Results: It can be useful to report the years of publication of included studies, in order to check if there are some trend in the scores obtained.

REVIEWER	Caroline Terwee VU University Medical Center Amsterdam the Netherlands
REVIEW RETURNED	19-Sep-2016

GENERAL COMMENTS	The study is well performed and clearly written. However, I have some doubts about the relevance of the study for a broader audience. This study is actually a satellite project of a previously published study on the methodological quality of meta-analyses of diagnostic accuracy of depression screening tools (reference 14 of the current paper). The authors should acknowledge this more clearly. For example, the search strategy described in this paper was not newly performed for this study, but was already done in the previous study.

	The studies included in this review were all published before the PRISMA for Abstracts tool was developed. Therefore, it was likely that these abstracts do not meet all of the PRISMA criteria. The authors state in the discussion that this study provides baseline results of reporting quality of studies performed before the PRISMA for Abstracts tool and can be used for future comparison with studies published after the PRISMA for Abstracts tool was published. However, I doubt if this study can be considered a good baseline measurement because the number of abstracts was very small (only 21) and all abstracts were from a specific field (screening for depression). To serve as a baseline measure for assessing the effect of the PRISMA for Abstracts tool on the quality of reporting of abstracts, the number and variety of abstracts should be much larger. Some minor comments: • I think that this study can also be considered a systematic review. Therefore, the abstract of this study should also meet the PRISMA for Abstracts tool criteria (it doesn't currently). • In the introduction the authors refer to a previous study that used the PRISMA for Abstracts tool to evaluate the reporting quality of 197 abstracts. It would be helpful to mention that these were abstracts of reviews of trials (not diagnostic papers), otherwise readers may think the current study has already been performed. • Also three studies in dentistry were mentioned: where they also looking at reviews of trials? • If so, it may be helpful to indicate in the introduction that the current study is the first study that looked at the quality of abstract reporting of diagnostic reviews. • I can feel why the authors wanted to split up two items from the PRISMA for Abstracts tool. However, this makes the ratings incomparable to those of other studies. Also, these items are now more heavily weighted because they count twice. Did the authors take that into account when making this decision? • Is any of the authors of this paper also a co-author of any of the 21 included reviews? It would be helpful to include the answer in the conflict of interest statement. The study is well performed and clearly written. However, I have some doubts about the relevance of the study for a broader audience. This study is actually a satellite project of a previously published study on the methodological quality of meta-analyses of diagnostic accuracy of depression screening tools (reference 14 of the current paper). The authors should acknowledge this more clearly. For example, the search strategy described in this paper was not newly performed for this study, but was already done in the previous study. The studies included in this review were all published before the PRISMA for Abstracts tool was developed. Therefore, it was likely that these abstracts do not meet all of the PRISMA criteria. The authors state in the discussion that this study provides baseline results of reporting quality of studies performed before the PRISMA for Abstracts tool and can be used for future comparison with studies published after the PRISMA for Abstracts tool was published. However, I doubt if this study can be considered a good baseline measurement because the number of abstracts was very small (only 21) and all abstracts were from a specific field (screening for depression). To serve as a baseline measure for assessing the effect of the PRISMA for Abstracts tool on the quality of reporting of abstracts, the number and variety of abstracts should be much larger. Some minor comments: • I think that this study can also be considered a systematic review.

	Therefore, the abstract of this study should also meet the PRISMA for Abstracts tool criteria (it doesn't currently). • In the introduction the authors refer to a previous study that used the PRISMA for Abstracts tool to evaluate the reporting quality of 197 abstracts. It would be helpful to mention that these were abstracts of reviews of trials (not diagnostic papers), otherwise readers may think the current study has already been performed. • Also three studies in dentistry were mentioned: where they also looking at reviews of trials? • If so, it may be helpful to indicate in the introduction that the current study is the first study that looked at the quality of abstract reporting of diagnostic reviews. • I can feel why the authors wanted to split up two items from the PRISMA for Abstracts tool. However, this makes the ratings incomparable to those of other studies. Also, these items are now more heavily weighted because they count twice. Did the authors take that into account when making this decision? • Is any of the authors of this paper also a co-author of any of the 21 included reviews? It would be helpful to include the answer in the conflict of interest statement.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

(1) This is a well-written paper that examines the quality of reporting of components of abstracts suggested by the PRISMA guidelines, using meta-analyses of the diagnostic accuracy of rating scales for depression published since 2005 as the sample of interest to establish a benchmark for current practices in quality of reporting. Strengths of the study include delineating a thoughtful adaptation of the PRISMA abstract criteria, which initially were developed for reports of clinical trials (a la CONSORT)(Moher, Schulz, & Altman, 2001) to make them more appropriate for tests of diagnostic accuracy (a la STARD)(Bossuyt et al., 2003), an exemplary approach to the formal search strategy, and a clear and balanced description of the limitations of the literature as well as the review itself. The entire paper displays excellent attention to detail. The central findings, that the level of reporting tends to be poor overall, and that it shows a medium to large (r=.45) correlation with the overall quality of reporting within the rest of the paper, are helpful descriptions of the current state of affairs. Though sobering, they are not surprising inasmuch as the PRISMA abstract recommendations are new (2013), and most of the articles included in the review predate the recommendations (as the authors note in the Discussion section). At times the article reads like a critique of the reporting in prior articles, which might feel unfair to the authors to the extent that their publication predated the criteria. Given that the authors also do not control the submission guidelines at journals, and any change on the part of editors is likely to be a slow and piecemeal process, it would be helpful to offer tips and suggestions about how to maximize the reporting of information within the current constraints of abstract lengths. The density required approaches that of classified advertisements in newspapers or online forums, so even suggested abbreviations or template sentences could be helpful.

We thank Reviewer #1 for her or his enthusiasm and helpful comments. We agree that it is difficult, if not impossible, for authors to draft abstracts that are consistent with PRISMA for Abstracts guidance and that also stay within most journal word count limitations. Thus, even when journals indicate that authors should follow PRISMA guidance, the word count makes this impossible. We do not have any easy solutions, but we do believe that an explicit comment in the next version of the PRISMA for Abstracts statement on word counts that are needed and should be permitted by journals would help. We now state in the conclusions (Page 16, Lines 25-37):

"When PRISMA for Abstracts is updated, it should consider the number of words that may be necessary to comply with recommendations. Journal editors should either provide authors with flexibility in abstract headings and abstract word counts, or match their abstract word limit with that recommendation so that authors can more realistically comply with PRISMA for Abstracts recommendations."

What follows are minor suggestions for refinement:

(2) It would be helpful to define acronyms (e.g., PRISMA, DTA, AMSTAR) on first use, lightening the cognitive load for readers.

We have defined all acronyms on first use within the body of the text as suggested.

(3) When introducing the secondary aim of testing the correlation between editorial word limits and PRISMA abstract score, it would be helpful to mention the rationale for expecting a correlation – namely that short word limits may make it difficult to cover all of the required elements. This is included in the Discussion, so it is a simple matter of setting the frame for the hypothesis or research question earlier as well.

The revised manuscript now states in the introduction (Page 7, Lines 13-23), "Our secondary objective was to determine if the quality of the meta-analysis or the word count permitted by the journal of the meta-analyses were associated with PRISMA for Abstract scores, as the feasibility of adhering to the PRISMA for Abstracts items may be compromised by abstract word count constraints set by journals."

(4) Several search terms occurred to us that were not included. Among them were "major depressive episode," "dysthymia," "dysthymic disorder" – all of which refer to types of depressive disorder in the DSM nosology. Similarly, we wondered about not using "area under curve," AUC, ROC, and related terms as ways of finding studies of "diagnostic accuracy," since ROC is a more general framework and AUC is an effect size (Kraemer et al., 1999; Swets, Dawes, & Monahan, 2000; Youngstrom, 2013, 2014). Neither of these is a fatal flaw, but it is worth mentioning as a limitation; it is plausible that studies focused on these permutations might have been omitted. I am aware of at least two examples (focused on other disorders) that have published AUC and diagnostic likelihood ratios (Straus, Glasziou, Richardson, & Haynes, 2011), but not sensitivity and specificity for a single threshold, showing that it is empirically possible for a relevant study to escape the search net that was cast.

We thank Reviewer #1 for these suggestions. These search terms may be useful for identifying primary studies of the diagnostic accuracy of depression screening tools. However, we do not believe that there are systematic reviews with meta-analyses in this field that have used the AUC as an effect size without using more common metrics (e.g., sensitivity, specificity) that were included in our search terms. The reference provided (Straus et al.) is a textbook on practicing and teaching evidence-based medicine, not an example of a meta-analysis. Similarly, we do not believe that there are meta-analyses of the diagnostic accuracy of depression screening tools to detect major depressive episodes or dysthymia, but that do not mention depression or depressive disorders. It is empirically possible in any systematic review, that a study has been missed, but we do not believe that this is likely here and that these are substantive limitations. Thus, we have not edited in response to this comment. If the editor, however, disagrees, we will include a statement in the limitations section.

(5) The authors do not report any estimate of inter-rater reliability for their coding scheme. The methods appear sound, even so. They can acknowledge this as a limitation, per the PRISMA or MARS guidelines.

We agree with this comment and have noted this in the limitations section (Page 15, Lines 22-29): “First, we did not perform a pilot test of out tool. Adjustments were made to our coding manual during the initial part of our meta-analysis scoring and, as such, we were unable to calculate an interrater agreement statistic for the adapted PRISMA for Abstracts items.”

(6) It also might be helpful to offer a rationale for splitting item 3 into 3a and 3b. It was not intuitively obvious to me how the referents for 3a and 3b differed conceptually.

In our revised manuscript we have provided a statement that describes the difference between Item 3a and Item 3b (Pages 9 and 10, Lines 53-56 and Lines 3-6), “Item 3 was divided into two parts in order to differentiate between characteristics for inclusion in primary studies (i.e., eligible participants. index tests, reference standards and outcomes), and characteristics for inclusion in the systematic review and meta-analyses (e.g., language and publication status of eligible reviews).”

(7) Overall, the paper offers a meticulous analysis and description of current reporting practices. It also provides a good prescription, in terms of both promulgating the PRISMA abstract criteria, and also allowing flexibility on the abstract word count in order to create sufficient room for more complete reporting.

We thank Reviewer #1 for her or his recognition of the importance and strengths of the manuscript.

Reviewer: 2

(1) This is a well written study and also relevant to show faults of reporting in integrative studies. I congratulate the authors for this important effort. I only have a few comments related to this important manuscript:

We thank Reviewer #2 for her or his enthusiasm and helpful comments.

(2) Authors included in this study only DTA systematic reviews with meta-analysis of accuracy figures. This aspect is very clear in methods section. However, in the abstract, summary and other sections they called these reviews only as “meta-analysis”. My main concern is that readers can believe that “meta-analysis” is a specific study design, but we know that it is only the statistical analysis of a systematic review. Is there any way to handle this?

We have edited the wording throughout the manuscript, including the abstract, and now refer to eligible studies as “systematic reviews with meta-analyses”.

(2) Methods/ Assessment of reporting: After modify the PRISMA tool, authors performed a pilot study to check if the modifications works or no?

We did not perform a pilot study of our tool. Instead, any disagreements in coding were discussed among coders, and if any difficulties arose during coding, the item was discussed with all team members and revised for clarity. We have added this limitation to our limitations section (Page 15, Lines 22-29), “Specific limitations should be considered when interpreting the results of our study. First, we did not perform a pilot test of out tool. Adjustments were made to our coding manual during the initial part of our meta-analysis scoring and, as such, we were unable to calculate an interrater agreement statistic for the adapted PRISMA for Abstracts items.”

(3) Methods/ Data extraction: Regarding the abstract word limit, what was the source of this information?

The source of this information was the author information guidelines on each journal website. We have added an appendix (S3 Appendix. Sources of Journal Word Limits) that provides the url for each of these. In our revised manuscript we have referred to this appendix on Page 10, Line 46, where we state “For each meta-analysis publication, one investigator extracted author, year of publication, journal, journal impact factor for 2014, the abstract word limit of the journal where the meta-analysis was published (see S3 Appendix for details), and previously published A Measurement tool to Assess Systematic Reviews (AMSTAR) quality ratings.[14]”

(4) Results: It can be useful to report the years of publication of included studies, in order to check if there are some trend in the scores obtained.

We have added the years of publication of meta-analyses included in our review within our results section in addition to Table 1. Our revised manuscript states (Page 11, Lines 39-46) “Of the 30 articles that underwent full-text review, 9 were excluded because they were not meta-analyses of diagnostic accuracy of depression screening tools (see S3 Appendix), resulting in 21 eligible meta-analyses published between 2007 and 2016 (see Figure 1).[15-35] Characteristics of included systematic reviews with meta-analyses are shown in Table 1.”

Reviewer: 3

The study is well performed and clearly written. However, I have some doubts about the relevance of the study for a broader audience.

We thank Reviewer #3 for her or his enthusiasm and helpful comments.

(1) This study is actually a satellite project of a previously published study on the methodological quality of meta-analyses of diagnostic accuracy of depression screening tools (reference 14 of the current paper). The authors should acknowledge this more clearly. For example, the search strategy described in this paper was not newly performed for this study, but was already done in the previous study.

To address this comment we have added a statement to the first line of the methods section (Page 7, Lines 29-34), “The search strategy used for this study was originally conducted for a study assessing the quality of systematic reviews with meta-analyses of diagnostic test accuracy for depression screening tools.[14]”

(2) The studies included in this review were all published before the PRISMA for Abstracts tool was developed. Therefore, it was likely that these abstracts do not meet all of the PRISMA criteria. The authors state in the discussion that this study provides baseline results of reporting quality of studies performed before the PRISMA for Abstracts tool and can be used for future comparison with studies published after the PRISMA for Abstracts tool was published. However, I doubt if this study can be considered a good baseline measurement because the number of abstracts was very small (only 21) and all abstracts were from a specific field (screening for depression). To serve as a baseline measure for assessing the effect of the PRISMA for Abstracts tool on the quality of reporting of abstracts, the number and variety of abstracts should be much larger.

We agree that it would have been difficult for the authors to have met the PRISMA for Abstracts criteria as only 5/21 included meta-analyses were published a reasonable amount of time after the PRISMA for Abstracts tool was published. We have noted this within our discussion and in our revised manuscript we have deleted the statement regarding our study being able to serve as a baseline. Our revised manuscript now states (Page 15, Lines 3-20), “Our study provides direction for evaluating PRISMA for Abstract adherence in reviews and meta-analyses in the field of DTA. Further, our study

highlights areas where improvement is needed, specifically in systematic reviews with meta-analyses of DTA of depression screening, and will allow future DTA reviews to apply our coding manual, and compare the reporting of abstracts after the PRISMA for Abstracts tool has been more widely endorsed."

(3) Some minor comments: I think that this study can also be considered a systematic review. Therefore, the abstract of this study should also meet the PRISMA for Abstracts tool criteria (it doesn't currently).

We do not believe that this is a systematic review, but rather a cross-sectional study that examines how methods have been applied in a set of systematic reviews with meta-analyses. We do agree, though, that we use systematic review methods, and we have done our best to adhere to PRISMA and PRISMA for Abstracts standards. We were hampered, however, by the word limit for abstracts in BMJ Open, which we describe in our article as a major obstacle to compliance. If the editor is able to allow us to expand the word count, we can revise the abstract.

(4) In the introduction the authors refer to a previous study that used the PRISMA for Abstracts tool to evaluate the reporting quality of 197 abstracts. It would be helpful to mention that these were abstracts of reviews of trials (not diagnostic papers), otherwise readers may think the current study has already been performed.

We have clarified that this study considered only systematic reviews of trials. Our revised manuscript states (Page 5, Lines 46-48), "Only one previous study has used the PRISMA for Abstracts checklist to evaluate the quality and completeness of abstracts for systematic reviews of trials.[7]"

(5) Also three studies in dentistry were mentioned: where they also looking at reviews of trials?

We have also clarified this (Page 6, Lines 10-13): "We identified three studies, all from dentistry literature, that reviewed reporting of abstracts in systematic reviews of trials.[4-6]".

(6) If so, it may be helpful to indicate in the introduction that the current study is the first study that looked at the quality of abstract reporting of diagnostic reviews.

We appreciate this suggestion and have incorporated it within our introduction. We now indicate (Page 6, Lines 46-48), "No published studies have evaluated the completeness of reporting in abstracts of diagnostic test accuracy systematic reviews or meta-analyses."

(7) I can feel why the authors wanted to split up two items from the PRISMA for Abstracts tool. However, this makes the ratings incomparable to those of other studies. Also, these items are now more heavily weighted because they count twice. Did the authors take that into account when making this decision?

We agree that comparing results between studies is beneficial. While our adapted tools divided items 3 and 4, if either item 3a (for example) or 3b were coded as "no" the original item 3 would be coded as "no" as well (both aspects would be required to receive a coding of "yes"). We have added a footnote in the S2 appendix, so that scoring could be easily revised for comparison purposes.

(8) Is any of the authors of this paper also a co-author of any of the 21 included reviews? It would be helpful to include the answer in the conflict of interest statement.

None of the authors of this paper were a co-author for any of the 21 included reviews. As suggested, we have added this statement to the conflicts of interest statement on Page 17, Lines 36-39, which

now reads “The authors of this study did not contribute to any of the included studies that were evaluated.”