Dear Miss Wang,

Thank you for sending us your paper. We sent it for external peer review and discussed it at our manuscript committee meeting. We recognise its potential importance and relevance to general medical readers, and would like to offer publication in the BMJ if you are willing to revise as we suggest.

We hope very much that you will be willing and able to revise your paper as explained below in the report from the manuscript meeting, and are looking forward to reading the revised version in due course.

Please remember that the author list and order were finalised upon initial submission, and reviewers and editors judged the paper in light of this information, particularly regarding any competing interests. If authors are later added to a paper this process is subverted. In that case, we reserve the right to rescind any previous decision or return the paper to the review process. Please also remember that we reserve the right to require formation of an authorship group when there are a large number of authors.

When you return your revised manuscript, please note that The BMJ requires an ORCID iD for corresponding authors of all research articles. If you do not have an ORCID iD, registration is free and takes a matter of seconds.

Dr John Fletcher
Associate Editor
The BMJ
jfletcher@bmj.com

*** PLEASE NOTE: This is a two-step process. After clicking on the link, you will be directed to a webpage to confirm. ***

https://mc.manuscriptcentral.com/bmj?URL_MASK=6b6f92a2bf42495ab36ee84ac7b7c0a0

**Report from The BMJ’s manuscript committee meeting**

These comments are an attempt to summarise the discussions at the manuscript meeting. They are not an exact transcript.

Members of the committee were: Joseph Ross (chair), Rafael Perera (statistician), Tim Feeney, Elizabeth Loder, David Ludwig, Tiago Villanueva, Wim Weber

Decision: Put points

Detailed comments from the meeting:
1. As editors we found it interesting to read your manuscript and we think there are enough researchers interested in IPD meta-analysis among our readers for us to offer the potential for publication in the BMJ.

2. A major limitation highlighted by our statisticians and reviewers was the less than optimal approach to appraising the quality of the included reviews. If you could pay close attention to the reviewers’ comments, in particular Prof Riley, and re-run the quality assessments then we would be pleased to look again at your manuscript.

3. At the least, we ask that you add to AMSTAR-2 the critical appraisal questions discussed in Tierney et al referred to by the reviewers.

4. Please revise your paper to respond to all of the comments by the reviewers. Their reports are available at the end of this letter, below.

In your response please provide, point by point, your replies to the comments made by the reviewers and the editors, explaining how you have dealt with them in the paper.

Comments from Reviewers

Reviewer: 1

Recommendation:

Comments:
The stated objectives of the paper are to (i) describe the characteristics of a sample of IPDMA, (ii) assess the methodologic quality of these IPDMA with AMSTAR2; and (iii) investigate factors related to the quality of IPDMA; (iv) suggest how quality might be improved. Considerable work was put into this review; it included 210 IPDAs which were carefully reviewed.

1. The grammar in the manuscript needs considerable revision.

2. IPDMA are often part of a systematic review. According to the guidance provided by Tierney et al. this is a mark of quality of an IPDMA since it ensures many good properties if the systematic review is of high quality. However, there are many more steps in an IPDMA after the conclusion of the systematic review part. To consider only the features relating to the systematic review part in evaluating the quality of the IPDMA seems very limited. The AMSTAR2 guidelines only relate to the systematic review part. Even if the guidance provided by Tierney et al. is not as detailed as that provided by AMSTAR2, some thought should be given to evaluating the parts of the IPDMA that are specific to IPDMA. These include questions 1, 3, 4, 5 (though this may have been taken into account by AMSTAR2), 7a (though this may have been taken into account by AMSTAR 11), 7b, 7c, 8. Moreover, the PRISMA statement for IPDMA may be consulted as well, for details on how these items should be reported. The paper would be strengthened if the IPDMA-specific features were also evaluated for quality. The relevance of this paper is limited given that it has only evaluated the systematic review items.

Minor points

2. Number of included studies in the SR seems like an odd predictor. What is more important is proportion of studies identified that were included in the IPDMA. Please change to this predictor.

3. The paper is supposed to be on the quality of IPDMA, but throughout the paper the authors describe the SRs. Please revise accordingly.

4. In table 1, please give the ranges of the quartiles for the impact factors.

5. In table 1, thre is a number and percent given for funding location. What does that mean?

Sincerely,

Andrea Benedetti

Additional Questions:
The BMJ uses compulsory open peer review. Your name and institution will be included with your comments when they are sent to the authors. If the manuscript is accepted, your review, name and institution will be published alongside the article.

If this manuscript is rejected from The BMJ, it may be transferred to another BMJ journal along with your reviewer comments. If the article is selected for publication in another BMJ journal, depending on the editorial policy of the journal your review may also be published. You will be contacted for your permission before this happens.

For more information, please see our peer review terms and conditions.

Please confirm that you understand and consent to the above terms and conditions:
I consent to the publication of this review

Please enter your name: Andrea Benedetti
Job Title: Associate Professor
Institution: McGill University
Reimbursement for attending a symposium?: No
A fee for speaking?: No
A fee for organising education?: No
Funds for research?: Yes
Funds for a member of staff?: No
Fees for consulting?:
Have you in the past five years been employed by an organisation that may in any way gain or lose financially from the publication of this paper?: No
Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this paper?: No
If you have any competing interests (please see BMJ policy) please declare them here: I have no competing interests.

Reviewer: 2
Recommendation:

Comments:
I have conducted a statistical review of the manuscript "Is the evidence from systematic reviews with individual patient data meta-analysis trustworthy? A cross-sectional study"

Overall, the study provides a good update to previous work in the area relating to characteristics of IPD meta-analyses and is the first study to my knowledge to assess the methodological quality of IPD meta-analyses. The findings seem to highlight some important issues of methodological quality, with the majority of included IPD meta-analyses judged to be critically low quality according to the AMSTAR-2 tool.

I have several methodological concerns and uncertainties relating to the approach taken to analysis and presentation of results of this study. My comments are as follows:

1) Approach to quality assessment
Like the authors, I am unaware of any specific quality tool specifically for assessing the quality of IPD meta-analysis (page 4) and I certainly support the statement within the discussion that there is a need for such a tool.

a) The authors refer to eight questions proposed within guidance from Tierney and colleagues (reference 2) and state that this guidance has: "…neither operational guideline nor assessment criteria. That might hamper the implementation of the criteria for widely adoption and comparison. "(page 3, line 18-19)

I’m not sure exactly what the authors mean here? Certainly, the guidance and questions developed by Tierney et al is not a formal quality assessment tool and does not claim to be, but this does not mean that the guidance should not be adopted?

b) In terms of other literature relating to the conduct of IPD meta-analysis, there is also an extension of the PRISMA guidance for IPD meta-analyses (and a PRISMA-IPD checklist), which was published in 2015 (see Stewart et al reference at the end of this review). The authors do not reference this guidance so may not be aware of it? I note that this guidance is also not designed to be a formal quality assessment tool but I do think that this guidance should be mentioned within this manuscript.

c) The authors use the AMSTAR-2 tool for assessment of the methodological quality of the included SRs. While I agree this is was a reasonable approach given the lack of methodological quality assessment tool specifically designed for IPD-meta analysis, the assessment will concentrate solely on the ‘systematic review’ elements of the included studies, and could miss specific issues relating to IPD meta-analysis. When assessing items of the AMSTAR-2 tool relating to synthesis, such critical items 11 and 15, did the authors take into consideration any of the specific guidance for IPD meta-analyses? Such as the guidance from Tierney et al (or guidance from the PRISMA-IPD statement could have been used?)

d) Furthermore, the AMSTAR-2 tool was published in 2017 and it could be argued that the included IPD meta-analyses published prior to 2017 are at a disadvantage to meet all of the methodological quality items of tool which did not exist at the time they were conducted.

The authors do present the comparative quality categories of IPD meta-analyses published between 2011 to 2017 and 2018 to 2019 within Table 2, but do not really comment on this.

I suggest that this is an important point for the interpretation of overall methodological quality and should be discussed within the findings.

2) Introduction, rationale for the cross-sectional study
The second paragraph of the Introduction eludes to biases which may be present within systematic reviews and meta-analysis but no actual details of what these biases are. Specifically:

a) Page 2, line 55: “However, SR with meta-analysis are not necessarily free from bias (7).”

Does this statement refer to IPD meta-analysis? Or systematic reviews with meta-analysis in general?

b) Page 2, line 56-59: “Empirical evidence has indicated the flaw of SRs... which may in turn threaten the trustworthiness of the evidence derived from the SRs.”

Please briefly summarise the main flaws which have been identified and how these flaws may threaten the trustworthiness of the evidence.

c) Page 3, line 2: “Available evidence has demonstrated that the published SRs with IPD meta-analyses were conducted based on inconsistent standards (12, 13).”

I don’t think that either of these references has demonstrated the use of inconsistent standards in IPD meta-analysis. Neither reference is a review of the general conduct of IPD meta-analyses. Reference 12 is a tutorial paper on how to conduct and report an IPD meta-analysis by Riley and colleagues and Reference 13 is an assessment specifically considering selective inclusion of studies or data in IPD meta-analyses, due to availability of data, publication bias etc. by Ahmed and colleagues.

Please explain what is meant by ‘inconsistent standards’ using appropriate references. For example, I believe that the review of a decade of IPD meta-analysis by Simmonds et al (reference 26 of this manuscript) has details of conduct of IPD meta-analyses which may be relevant.

3) Eligibility criteria

a) IPD meta-analyses of RCTs only were eligible for inclusion yet the AMSTAR-2 tool is also suitable for systematic reviews of non-randomised studies. Please clarify why only IPD meta-analyses of RCTs were eligible for inclusion in the study?

b) IPD meta-analyses published from 2011 onwards were considered to be eligible. Was a start data of 2011 chosen for any reason?

c) Page 4, line 4 “When duplicate IPD meta-analyses were identified, the most updated version was selected with the remaining versions being used as supplementary documents for data extraction.”

Does ‘duplicate IPD meta-analysis’ means the same thing here as an ‘update of a previous review.’ I can see from Table 1 that updates of reviews were included.

I would assume that ‘duplicate IPD meta-analyses’ are co-publications in different journals of exactly the same IPD meta-analysis but that I don’t understand what is meant by the ‘most updated version’ of a duplicate IPD meta-analysis? Please clarify

4) Data extraction and statistical analysis

a) Extraction of the characteristics of the IPD meta-analyses is extremely comprehensive. However, I wonder whether Table 1 may be too big for an author to fully appreciate all of the information within it?

Perhaps (if formatting guidelines allow) this large table could be split into some smaller sub-tables Table 1a, Table 1b etc. by content such as general information (Cochrane Review, Number of authors, location etc.), review topic (treatments, disease area, year of included studies, number of included studies etc.)
and review conduct and reporting (search methods, systematic review methods, risk of bias tools used, statistical analysis etc.)

Alternatively, please see comment 4f below on combining the information in Table 1 and 2

b) Comments on specific items extracted (Appendix 2):

B12 Authorship – I’m not sure that all options would be covered here, for example how would an IPD meta-analysis be categorised with a number of individually named authors (such as those who have conducted the IPD meta-analysis) and then also a collaborative group name (often used to reflect those who have provided IPD)

Related to this for item B7 – would a collaborative group be a single author when counting number of authors?

B13: Does this refer to the funding of the systematic review/ IPD meta-analysis, rather than the funding sources of the RCTs included in the review?)

c) The method used to combine the IPD at the bottom of Table 1 is not included in the Data Extraction form in Appendix 2. It would also be helpful to provide a reference or references for further information on these methods, which may not be well known to all readers.

d) Some characteristics within Table 2 are not included within the data extraction form: The major author is a co-author of the included RCTs, SR accounted for trial characteristics during data analysis, SR accounted for participant characteristics during data analysis, The SR checked the missing data of the RCT.

Also please clarify what a ‘major author’ means in this context.

e) How was the list of independent variables (page 5) to include within ordinal regression decided? For example, based on previous evidence of factors associated with systematic review quality?

It is difficult to interpret the results shown within Figure 3 without understanding of why an association may be present. For example, I am unaware of any reason why methodological quality may vary according to the geographic area of the specifically the corresponding author.

Also, arguably, the ‘independent’ variables listed on page 5 cannot necessarily be assumed to be independent of each other or of the outcome of methodological quality.

For example, number of authors and authorship policy are related variables as the authorship policy will determine the number of authors. Whether authors of included studies are authors of the systematic review will also be determined by the authorship policy.

Also, Cochrane Review and updated review are likely to be related variables as Cochrane Reviews are more likely to be updated than non-Cochrane reviews.

Although they are not specific items of the AMSTAR-2 checklist, use of PRISMA flow diagrams and consideration of harms are arguably direct measures of methodological quality which should be considered when performing the AMSTAR-2 assessment, rather than independent variables of methodological quality.

I suggest that the authors reconsider the ‘independent variables’ they are including within this analysis to ensure that only variables which can reasonably assumed to be independent from each other and from the outcome and that rationale for investigating an association between each of these variables and the outcome is clear.
f) Page 5, line 6: “Kruskal-Wallis rank test were adopted to compare the differences in methodological quality across multi-characteristics as the overall rating of the methodological quality was treated as ordinal ranking data.”

I’m not sure I understand what this means, what are ‘multi-characteristics’? Are the p-values presented in Table 2 from the Kruskal-Wallis rank test?

If so, please note that the p-value of this test will only tell you if there is a statistically significant difference in methodological quality category across the characteristic groups, it does not indicate exactly where the difference is. So, for example, the p-value of 0.003 at the top of Table 2 suggests that Cochrane Reviews are different to non-Cochrane reviews in terms of methodological quality, not necessarily that quality is better.

I am not sure that these analyses within Table 2 add anything meaningful over the ordinal regression analyses in Table 3.

I suggest that the information in Tables 1 and 2 could be combined – e.g. remove the p-values from Table 2 and add a total row which would reflect the information in Table 1.

Then distributions of the characteristics and any observed differences in the methodological quality across characteristics can be discussed together, ahead of the formal regression analyses.

5) Methodological quality assessment

a) I consider that there is currently too much focus on this manuscript on characteristics of the IPD meta-analysis and not enough focus on the methodological quality of IPD meta-analyses according to AMSTAR-2, which from my understanding, is the main objective of the cross-sectional study.

I find it quite shocking that the majority of IPD meta-analyses, widely considered to be a ‘gold standard’ of evidence, are of critically low quality according to AMSTAR, and what I really want to know is which areas in particular are these IPD meta-analyses lacking in, and what can be done to improve that.

The authors do state that one of their objectives is to identify aspects where the future IPD meta-analysis might be improved but I do not think there is enough discussion around this objective within the manuscript. While it is interesting to examine characteristics of IPD meta-analyses which may be related to quality, if these characteristics cannot be changed (e.g. year of publication, disease area, number of studies, geographical location of the authors etc.), the implications of these associations are much less important than areas which can be changed relating to the conduct and reporting of the systematic review and IPD meta-analysis.

I would like to see a more detailed discussion of the findings related to the individual items, particularly the critical items within the main results section. Some of this information is currently within the Discussion section, but I suggest the proportions meeting or not meeting each important item should be presented within the results.

Also, this may be a personal preference, but for me, the table in Appendix 7 provides more information than Figure 2 on the AMSTAR-2 tool assessments by item, as the description of the items are listed within the table. Therefore, I suggest that the Table in Appendix 7 should be included within the main results either as well as or instead of Figure 2.

I have made some suggestions within comment 4 on how to condense some of the tables and text relating to characteristics, which should create some space for an expanded discussion on the quality assessment items within the Results text.
b) Page 15, line 28: “Nonetheless, only 56.2% of the sampled SRs assessed risk of bias for included RCTs appropriately.”

Does this statement refer to using a satisfactory technique for RCTs, such as the Cochrane Risk of bias tool? This implies not only that SR reviewers used a satisfactory tool, but also that they used it correctly.

Without going back and checking all of the risk of bias assessments of the studies in all of the IPD meta-analyses (which I would say is outside of the scope of this work), I don’t think it can be concluded that risk of bias was assessed ‘appropriately,’ only that an appropriate tool is used.

Please check the language used when interpreting other items and characteristics that interpretations are made on the systematic review level rather than a study level.

c) Related to the above comment regarding interpretation of findings relating to quality, I am not sure that the wording of the title is appropriate.

Certainly, a review of critically low quality “should not be relied on to provide an accurate and comprehensive summary of the available studies” as stated within the AMSTAR-2 definition of critically low. But to make a judgement on whether the evidence is ‘trustworthy’ I would say would require an additional assessment of the specific areas of quality which are lacking from the review and the likely impact on the results of the meta-analysis (e.g. are results likely underestimated or overestimated? Could results have been very different?). So the impact of the low quality, which could be defined as the trustworthiness, could vary by review characteristics such disease area, and also by which areas of quality are an issue for that review.

I understand the desire to have a title with impact, but I don’t think the current one is appropriate as ‘trustworthiness’ is not directly investigated. I suggest revising the title.

Minor comments

Appendix 3: Please check any copyright restrictions when reproducing the AMSTAR-2 checklist. It would also be helpful if the ‘critical’ items of the checklist could be highlighted (or at least listed within the text) so a reader doesn’t have to go back to the original AMSTAR publication.

Please proof read the manuscript again. Many typographical and grammatical errors are present.

Reference:


Additional Questions:

<em>The BMJ</em> uses compulsory open peer review. Your name and institution will be included with your comments when they are sent to the authors. If the manuscript is accepted, your review, name and institution will be published alongside the article.<strong></strong>

If this manuscript is rejected from <em>The BMJ</em>, it may be transferred to another BMJ journal along with your reviewer comments. If the article is selected for publication in another BMJ journal, depending on the editorial policy of the journal your review may also be published. You will be contacted for your permission before this happens.
Reviewer: 3

Recommendation:

Comments:
Thank you for the opportunity to review this important study, in which systematic reviews with IPD meta-analysis are assessed for methodological quality with the AMSTAR-2 tool, and study characteristics are associated with overall AMSTAR-2 ratings. The authors found that the methodological quality of SR with IPD meta-analysis is unsatisfactory, and that several factors (such as more recent publication, main author part of large included RCT) were associated with methodological quality. Strengths of this study include the comprehensive number of IPDs included and assessed (all IPDs published since 2011), and the important study question.
I have two major comments, and several minor comments.

1. The AMSTAR-2 developers explicitly point out that their tool was 'not intended to deal with the special requirements of diagnostic test reviews, individual patient data meta-analyses or network meta-analyses, scoping reviews, or realist reviews.' (BMJ 2017;358:j4008)
   Whilst I understand no appropriate tool for IPD-meta-analysis currently exists, it would have been preferable if the authors had created an adaption of AMSTAR-2 for their purposes: including some domains that are important for IPD-MA (e.g. % retention of IPD, appropriate data cleaning and analysis), and potentially omitting some items that are less important for this type of meta-analysis (e.g. up-to-date search may be less feasible, since IPD take longer to complete that traditional meta-analyses).

2. In the methods and results section, there is not enough focus on each respective item of AMSTAR-2, why each item is important (and in particular why it is important for IPD-MA), and to what extend the included studies fulfil each item. Generally, the use of AMSTAR-2 to create overall ratings has been discussed quite critically in the literature, since not each item can be regarded as equally important. The developers of AMSTAR-2 have noted the following: ‘users should consider the potential impact of an inadequate rating for each item’ (BMJ 2017;358:j4008). This is particularly important for the current study (since not all items apply equally for IPD-MA). In addition, for this manuscript a major aim should be to give future study authors and reviewers insight to which items they need to pay particular focus to, to improve quality of IPD-MAs. This aim is currently not fulfilled sufficiently from the information that has been provided in the manuscript.

I would recommend the study authors give an overview of AMSTAR 2 domains and why they are important for IPD study quality in the methods section. This could be in form of a Figure or Table, in the main manuscript (not just in the Appendix). Then, they should give more information on compliance with each item across the studies (instead of overall quality ratings) in the results section. Currently in Figure 2, the AMSTAR-2 items are not labelled so the reader does not know what each item means.

I recognise that this point is partly addressed in the discussion, by discussing key items future IPD-MAs should focus on, but it would be good to see more of this in the methods and results to understand how these recommendations were derived.

Title:
1. ‘Trustworthy’ is a strong word, and arguably, some of the included studies may be ‘trustworthy’ despite not fulfilling all domains of AMSTAR-2. It may be a good idea to think about a more factual title.

Introduction:
2. The authors mention eight questions developed by Tierney et al to assess IPD-MA. Were these questions used to collect study characteristics in the present study? If so, how did the assessed IPD-MA perform against these questions?

Methods:
3. Proportion of obtained data is an important characteristic for which this current study provides valuable information. Did the authors also look at the proportion and number of participants that data were obtained for? For IPD-MAs, it is particularly important to obtain data from the biggest trials, to obtain a high sample size. It would have been good if the authors had somehow formally included this in their quality assessment, since proportion of obtained data is one of the main limitations and source of potential bias in IPD-MAs.

4. Publication bias AMSTAR-2 item – this would not apply to the studies with <10 included trials, since the commonly used tools cannot be applied in this case? What did the authors do in this case?

5. The authors discuss lack of inclusion of non-English studies as a potential source of publication bias. Did the authors downgrade for the related AMSTAR-2 item if non-English studies were not included?
Results:
6. Table 2 reports associations of study characteristics with AMSTAR 2-rated quality. However, it appears some of those characteristics were in fact used to assess quality for AMSTAR-2 items (e.g. ROB)? Thus, the predictor and outcome variable would not be independent from each other by nature, and it would thus be misleading to give such associations. For instance, studies that did not use a risk of bias tool were downgraded in quality rating for this reason – which means there will be an association between risk of bias and methodological quality by nature, as risk of bias (the predictor variable) is used to assess quality (the outcome variable).

7. Generally, (as pointed out above) I would find it more useful if more information was provided on which items the IPDs performed well or not so well in, and potential reasons for this, and less information on study characteristics associated with quality. It would be good to learn more about specific areas IPDs need to improve in (i.e. writing protocols or assessing risk of bias), but it may be less relevant for the research community to learn characteristics such as which country the best IPDs were conducted in.

Discussion
8. Have any studies been done looking how traditional systematic reviews and meta-analyses perform on AMSTAR-2? It would be great to understand how IPD-MA perform in comparison to traditional MA.

9. It would be good to add some more information on limitations to AMSTAR-2 to the discussion section.

10. It would be good if the authors could discuss unavailability of data across studied IPDs and implications in more detail.

11. IPD are described as the 'gold-standard', not because of the high adherence to AMSTAR-2 criteria, but due to the additional advantages these analyses they offer, such as advanced analyses (including individual-level subgroup analyses) and higher data quality. It would have been good to assess some more of these advantages in this study (as additional quality criteria), but at least, the authors should discuss them in more detail in the discussion section.

Figures and Tables
Figure 2 – colour coding is not intuitive – partial yes should not be red (would expect no to be red). As outlined above, items need to be labelled (what they entail, instead of simply item 1 to 15) so the reader understands what each item refers to.

Appendix 7 – this is important and should be included in the main manuscript

Language
When revising the manuscript, authors should check the manuscript carefully for language – there were a few grammatical errors throughout the manuscript, and the argument structure and syntax could be improved in some parts. It may be a good idea to ask a native English speaker to carefully review the manuscript.

Anna Lene Seidler
Research Fellow
NHMRC Clinical Trials Centre, University of Sydney

Additional Questions:
<em>The BMJ</em> uses compulsory open peer review. Your name and institution will be included with your comments when they are sent to the authors. If the manuscript is accepted, your review, name and institution will be published alongside the article. <strong></strong>
If this manuscript is rejected from <em>The BMJ</em>, it may be transferred to another BMJ journal along with your reviewer comments. If the article is selected for publication in another BMJ journal, depending on the editorial policy of the journal your review may also be published. You will be contacted for your permission before this happens.

For more information, please see our <a href="https://www.bmj.com/about-bmj/resources-reviewers" target="_blank">peer review terms and conditions</a>.

<strong>Please confirm that you understand and consent to the above terms and conditions.</strong>

I consent to the publication of this review

Please enter your name: Anna Lene Seidler

Job Title: Research Fellow

Institution: University of Sydney, NHMRC Clinical Trials Centre

Reimbursement for attending a symposium?: No

A fee for speaking?: No

A fee for organising education?: No

Funds for research?: No

Funds for a member of staff?: No

Fees for consulting?: No

Have you in the past five years been employed by an organisation that may in any way gain or lose financially from the publication of this paper?: No

Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this paper?: No

If you have any competing interests <a href="http://www.bmj.com/about-bmj/resources-authors/forms-policies-and-checklists/declaration-competing-interests" target="_new"> (please see BMJ policy) </a>please declare them here:

Reviewer: 4

Recommendation:

Comments:
Synopsis of presented work: Wang et al. aim to assess the quality of Systematic Reviews and Meta-Analyses (SR-MAs) of Individual Patient/Participant Data (IPD SR-MAs) from randomised controlled trials. They first conduct a literature search to identify IPD SR-MAs, then assess their quality using an
The author-created overall score of AMSTAR-2, and finally use statistical methods to identify factors significantly related to (Kruskall-Wallis tests) or predictive of (ordinal logistic regression) their quality metric.

Originality - does the work add enough to what is already in the published literature? If so, what does it add? If not, please cite relevant references.

Originality. Multiple works cited by Wang et al. themselves previously assessed the same question using similar methodology fashion (i.e. identifying IPD SR-MAs using either a specialised database or literature search, and then evaluating 'quality' albeit using different, but arguably more appropriate, tools/criteria to Wang et al.). Wang et al.'s work is therefore not original in its contribution, but potentially incremental.

It is already well-known that IPD SR-MAs vary/lack in quality despite being considered the gold standard of evidence synthesis methodology. For example, Ahmed, Sutton, and Riley, 2012, [https://doi.org/10.1136/bmj.d7762] systematically search for IPD SR_MAs and assess their quality in terms of publication bias, selection bias, and unavailable data and, based on results, 'warn against uncritically viewing any meta-analysis that uses individual participant data as the most reliable'. Simmonds, Stewart, Stewart stated in 2015 ([https://doi.org/10.1016/j.cct.2015.06.012]): '… there remains considerable scope for improving the quality of reporting for both the process of IPD systematic reviews, and the statistical methods employed in them. It is to be hoped that the publication of the PRISMA-IPD guidelines specific to IPD reviews [Tietrney et al., 2015, [https://dx.doi.org/10.1371%2Fjournal.pmed.1001855]] will improve reporting in this area.'

Sufficiency of contribution to research. The potential incremental contribution of Wang et al.'s work is the attempt to identify significant correlates and predictors of IPD SR-MA 'quality' as defined by the authors. The intent is valuable, but the execution has several shortfalls, resulting in a modest contribution to knowledge in this area with predictable research findings. The factors identified as associated or predictive of better methodological quality are those likely to serve as proxies for better research design (e.g. pre-registration), higher funding (e.g. US authors), and greater expertise (e.g. Cochrane review). A major issue with this work is that presented results are unlikely to be reliable due to methodological flaws in both the literature search (resulting in an incomplete data-set for analyses) and in the statistical analyses.

Methodological issues

Search strategy: The authors restrict the literature search such that the earliest year of retrieved records is 2011 without providing a rationale. This choice of date restriction is problematic for the aims of the review, because non-Cochrane IPD SR-MAs have been published since at least 1991, (e.g. Riley, Lambert, and Abu-Zaied, 2010, Figure 1, [https://doi.org/10.1136/bmj.c221]), and Cochrane IPD SR-MAs since at least 2001 (e.g. Scott et al., 2001, [https://doi.org/10.1002/14651858.CD002197]). Wang et al. retrieve 210 IPD SR-MAs, yet previous work *searching the same databases without date restriction* identified 383 in 2005 (full search strategies reported in Riley, Simmonds, and Look, 2007, [https://www.sciencedirect.com/science/article/pii/S0895435606004033#appsec1]). Wang et al's date restriction to 2011 therefore identifies at best 54% of IPD SR-MAs that existed by 2005. This is an odd search result as the frequency of IPD SR-MAs has been shown to increase over time (e.g. (e.g. Riley, Lambert, and Abu-Zaied, 2010, Figure 1, [https://doi.org/10.1136/bmj.c221]).

Interpretation depth: The authors explore a large number of potential quality predictors, but the discussion of results remains largely descriptive. For example, the increase in 'quality' with Year of publication is likely to be a proxy for development of IPD guidelines and increase in familiarity/expertise in IPD methodology. It would have been useful to explore whether IPD SR-MAs that explicitly cite adherence to published IPD guidelines and to explore the implication of the results at greater depth.

Statistical methods:
The conduct, reporting, and presentation of the descriptive and inferential statistics needs to be improved. For example, tests of model assumption-tests as well as tests statistics and associated degrees of freedom need to be reported. There are multiple unaddressed assumption violations which
result in low reliability of a large proportion of presented results (e.g. whole of Table 2 is invalid due to unacceptably low category counts).

Table captions need to clearly state which statistical model/test presented results belong to. Test statistics as well as 95% confidence intervals (95% CIs) should be presented in results tables (this information is currently presented in-text only), rather than just p values (e.g. Table 2).

Table 1 seems to shows descriptive statistics (counts and percentages) using initially defined categories of predictors (e.g. Rank of journal impact factor: Quartile 1, 2, 3, 4), which are non-identical to those used in the statistical analyses (e.g. Kruskall-Wallis test: Rank of journal impact factor dichotomised to Quartile 1 and combined Quartiles 2—4).

Table 2 seems to show results of 18 Kruskal-Wallis tests, conducted for each of 18 predictors to assess whether mean ranks of research quality categories (AMSTAR-2 overall score: Critically low, Low, Moderate, High) differ significantly between the predictor categories (e.g. between Cochrane and non-Cochrane reviews). Every single one of the 18 conducted Kruskall Wallis tests presented in Table 2 is unreliable due to insufficient category counts (n < 5). There is an additional problem with vastly uneven counts across predictor categories which affect the variance of the mean ranks and therefore the Kruskall-Wallis H statistic (the basis for the reported p values). For example, for the factor ‘Reported search terms’, Table 2, page 11, the authors seem to compare largely empty categories for IPD SR-MAs without search terms (Critically low: n = 25, Low: n = 1, Moderate: n = 0, High: n = 0) to much greater cell counts for IPD SR-MAS with search terms (Critically low: n = 126, Low: n = 34, Moderate: n = 23, High: n = 1). For Kruskall-Wallis tests, categories with n < 5 should be merged for analyses. In this particular instance, merging would mean that a comparison for the factor ‘Reported search terms’ is not possible, because IPD SR-MAs without search terms would consist of a single category (as counts of all but one category < 5).

It is overall unclear why the authors perform Kruskal-Wallis tests in addition to ordinal logit regression. Both statistics answer the question of which factors is significantly related to quality operationalised as overall AMSTAR-2 score albeit using different methods. The Ordinal Regression approach is preferable as it has greater statistical power than Kruskall-Wallis tests, and provides more detailed information without adding 18 separate statistical comparisons. It is unclear how the predictors were chosen initially and it would be useful to describe their selection process.

The authors use an ordinal logistic regression model with 12 predictors (14 if two dummy-variables for Area of corresponding author are included). This model is in principle appropriate choice as their outcome variable is a derived ‘overall’ AMSTAR-2 score with three ordinal categories (Critically low, Low, and a third category that combines IPD SR-MAs with an AMSTAR-2 overall score of Moderate and High), and their predictors are a mixture of continuous (e.g. year of publication), categorical (Are of corresponding author [Europe, America, Asia, Oceania], and dichotomous (e.g. Cochrane Review [Yes, no]). The operationalisation of both predictors and outcome variables (author-calculated ‘overall’ AMSTAR score) is not appropriate (see ‘Operationalisation of predictors’ and ‘Operationalisation of outcome’). The authors either do not test or fail to report to have tested any assumptions of this model (e.g. multi-collinearity of predictors, proportional odds at each category of predictors. It is very likely that there is strong multi-collinearity between some of the predictors used (e.g. between the factors Cochrane Review [yes, no], Authorship Policy [Collaborative versus Individual], Number of SR authors, PRISMA-like diagram [yes, no]).

Choice of predictors:
Wang et al. use year of publication as a predictor of research quality in an ordinal regression model, but firstly, this variable, which should be continuous, has an artificially restricted range of 8 years (2011 to 2019), and secondly, the authors then dichotomise this artificially restricted range of Year of publication further into two unevenly sized categories (2011—2017 [6 year period] and 2018—2019 [1 year period]) without providing a rationale. Both the date range restriction and the dichotomisation of Year of publication are not advisable (multi-ordinal regression can handle continuous predictors). Wang et al.’s finding, which is that IPD SR-MA quality improves over time (proxy: publication date), is still likely to be qualitatively correct, but the parameter estimates (adjusted odds ratios with associated 95% confidence interval) are unlikely to be accurate (e.g. effects are usually weakened by restricted range analyses). There are similar issues with other predictors (e.g. Rank of journal impact factor is dichotomised into Quartile 1 and Quartiles 2—4).
Choice of outcome/quality assessment tool: The authors reject the use of several tools specifically designed to evaluate the quality of IPD SR-MAs (e.g. Tierney et al., 2015, https://doi.org/10.1371/journal.pmed.1001855; Stewart et al., 2015), stating that they chose AMSTAR-2 as their quality indicator, and then calculate an overall AMSTAR 2 score. AMSTAR 2 is not designed for IPD SR-MAs, and AMSTAR 2 authors explicitly and repeatedly warn against combining AMSTAR 2 items into an overall score. (Shea et al, 2017, doi: http://dx.doi.org/10.1136/bmj.j4008).

Shea et al. (2017) state:

'AMSTAR 2 is not intended to generate an overall score' (p.1, Summary points)
'We strongly recommend that individual item ratings are not combined to create an overall score' (p. 5)
'We stress that responses to AMSTAR 2 items should not be used to derive an overall score.' (p. 6)

• Importance of work to general readers - does this work matter to clinicians, patients, teachers, or policymakers? Is a general journal the right place for it?

It is important that clinicians, patients, teachers, or policymakers *do not blindly trust* that anything labelled IPD SR-MA (or indeed anything else) is reliable and accurate. It is unlikely that the present work itself provides a reliable assessment of IPD SR-MA quality due to numerous methodological flaws. A general journal is the place for this work with major re-analyses and re-writing, but a short communication flagging that, to date, IPD SR-MAs quality is still nowhere near where it should suffice.

• Scientific reliability

Scientific reliability is low due to numerous methodological flaws. Multiple detailed examples of why the reliability of presented results is low are provided in other sections.

• Research Question - clearly defined and appropriately answered?

Clearly defined yes, but inappropriately operationalised and the work is therefore not able to provide reliable answers to the posed research question. Additionally, the Title is inaccurate for the content and should be changed ('Is the evidence from systematic reviews with individual patient data meta-analysis trustworthy? A cross-sectional study'). The work is a systematic review over an extended time-frame (8 years: 2011—2020) with additional regression analyses that defines potential ‘quality-related’ factors and assesses their usefulness as predictors of IPD SR-MA ‘quality’. This work is definitely not a cross-sectional study conducted at single point in time, without any attempts to assess causal relations, or to manipulate variables by selection or otherwise. 'Trustworthiness’ and methodological quality of IPD SR-MAs are non-identical. The authors assess methodological quality of IPD SR-MAs using a tool that is neither recommended nor validated for this purpose (AMSTAR-2 is designed and validated for aggregate data SR-MAs only). There are more appropriate, specialised tools for IPD quality evaluation available, which the authors are aware of and cite, but reject because they feel that these tools ‘were proposed with neither operational guideline nor assessment criteria’ (p. 3, line 18). The authors seem to prefer evaluation tools that condense methodological quality into a single score, which is understandable, but, in my opinion, such an approach is simplistic and impossible. Even AMSTAR-2 does not provide an overall score, so the authors create one, which is explicitly and repeatedly discouraged by AMSTAR-2 authors (because it is simplistic).

To answer their research question, authors could amend their search strategy to that used by Riley, Simmonds, and Look, 2007, https://www.sciencedirect.com/science/article/pii/S0895435606004033#appseca1, which should result in a data-set that substantially larger than 383 IPD SR-MAs. Quality assessment could then be conducted using a tools specifically designed to evaluate the quality of IPD SR-MAs (e.g. Tierney et al., 2015, https://doi.org/10.1371/journal.pmed.1001855; Stewart et al., 2015). I would strongly discourage attempt to derive a single quality score other than perhaps completeness judged against most up-to-date guidelines. I think the authors need to acknowledge that ultimately, methodological rigour does not guarantee trustworthiness of results, which comes back to the quality of the raw data.

• Overall design of study - adequate ?

No.
Participants studied - adequately described and their conditions defined? Unit of analyses is IPD SR-MA, descriptives (Table 1) are adequate.

Methods - adequately described? Complies with relevant reporting standard - Eg CONSORT for randomised trials? Ethical?
No. Numerous methodological issues. No ethical issues other than considerable wasted effort on behalf of the authors.

Results - answer the research question? Credible? Well presented?
Analyses cannot answer the research question(s) credibly due to numerous methodological flaws. Presentation needs to be improved.

Interpretation and conclusions - warranted by and sufficiently derived from/focused on the data? Message clear?
Not warranted based on results. Interpretation would benefit from greater depth (mostly descriptive). For example, it seems that multiple identified predictors of methodological quality are proxies for amount of funding, expertise, and progress in guideline development. It would be good to explore these possibilities at least theoretically, ideally quantitatively. The message is clear (IPD SR-MA quality is still unacceptably low, the ‘usual suspects’ (factors) are associated with quality improvement [planning/pre-registration, greater expertise, etc.]).

References - up to date and relevant? Any glaring omissions?
Bibliographic references are adequate.

Abstract/summary/key messages/What this paper adds - reflect accurately what the paper says?
Apart from previously discussed issues: Yes.

Overall English language issues: Frequent minor errors (e.g. agreement between subject--verb, singular noun instead of plural, incorrect preposition such as ‘on’ instead of ‘of’, understandable but incorrect phrasing, such as ‘may hazard the validity of results’. Page 3, line 2 [threaten] etc.). I documented the first few minor errors only. The whole article would benefit from proof-reading by a native or native-equivalent English speaker. This is simply a language issue, no reflection on content.

Examples of minor typos and/or style issues
-----------------------------------------------------------
Page 1
Line 50: At start of sentence, use words instead of numbers (i.e. replace ‘210’ with two-hundred and ten’).
Line 53: Subject--verb agreement. Only one *SRs* (plural) should be ‘SR’ (singular) *were* (plural) should be ‘was’ of high quality
Line 57: Object--verb agreement. ...assessed the risk of bias of included *study* (singular) should be ‘studies’ (plural), subject-verb agreement ‘and *justifying*’ should be ‘justified’

Page 2
Line 3: Improve clarity of phrasing. Suggest to start sentence with ‘Factors independently associated with methodological quality of SRS were...’
Line 27: *meta-analysis* (singular) *on* randomized controlled trials should be meta-analyses* (plural) *of* randomized controlled trials
Line 38: *meta-analysis* (singular) should be *meta-analyses* (plural)

Additional Questions:
<em><strong>The BMJ</strong></em> uses compulsory open peer review. Your name and institution will be included with your comments when they are sent to the authors. If the manuscript is accepted, your review, name and institution will be published alongside the article.<em></strong>
If this manuscript is rejected from *The BMJ*, it may be transferred to another BMJ journal along with your reviewer comments. If the article is selected for publication in another BMJ journal, depending on the editorial policy of the journal your review may also be published. You will be contacted for your permission before this happens.

For more information, please see our <a href="https://www.bmj.com/about-bmj/resources-reviewers" target="_blank">peer review terms and conditions</a>.

**Please confirm that you understand and consent to the above terms and conditions.**

I consent to the publication of this review

Please enter your name: Kristina Charlotte Dietz

Job Title: Post-graduate Research Associate

Institution: University of Kent

Reimbursement for attending a symposium?: No

A fee for speaking?: No

A fee for organising education?: No

Funds for research?: Yes

Funds for a member of staff?: No

Fees for consulting?: No

Have you in the past five years been employed by an organisation that may in any way gain or lose financially from the publication of this paper?: Yes

Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this paper?: No

If you have any competing interests <a href="http://www.bmj.com/about-bmj/resources-authors/forms-policies-and-checklists/declaration-competing-interests" target="_new">please see BMJ policy</a> please declare them here: In the past, I received an NIHR training fellowship in IPD methodology (capacity building) and I contribute to IPD meta-analyses. I have been trained by and worked for core personnel of the Cochrane IPD group.

Reviewer: 5

Recommendation:

Comments:
This study has focused on examining the methodological quality of IPD meta-analysis which is a much needed research topic. Although this study could have great impact, there are major limitations which could affect validity of the results of this analysis:

1. Searches are missing a wording for capturing IPD meta-analyses 'individual participant data'. Meta-analysis of pharmacological/non-pharmacological interventions reporting IPD data are likely to use wording 'individual participant data' instead of 'individual patient data'. One example is an IPD meta-analysis published by our group which has not been captured in this study:
https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2545074

2. I agree that SRs with IPD have to be assessed against common methodological quality criteria as all SRs (as described in AMSTAR-2). However, space in most peer-reviewed journals is limited. For this reason, describing the process and steps to account for IPD-specific bias is usually the focus in most IPD publications. Authors of IPD meta-analysis are left with considerably less space compared to common SRs to describe their approach towards common methodological quality items captured by AMSTAR-2. Thus, the findings of this study may just reflect suboptimal reporting of common methodological quality items rather than low methodological quality in IPDs. One way to account for this and establish the validity of the conclusions of this study (e.g. do the findings actually reflect low methodological quality or low reporting standards), would be to request additional information by the authors of IPD meta-analysis in relation to AMSTAR-2 items.

3. It is not clear how the potential factors affecting the quality of IPDs have been selected. Has a priori systematic procedure applied for selecting those factors? Other factors could be important too. For example, the introduction of PRISMA-IPD checklist/diagram could account for the higher quality of the IPDs in recent years.

4. It would be helpful in terms of improving the reporting standard of this study if the authors could provide a table in the appendix which would list the descriptive characteristics, iAMSTAR-2 item ratings and potential factors affecting the methodological quality for each included IPD SR. This will facilitate future updates.

Additional Questions:
<em>The BMJ</em> uses compulsory open peer review. Your name and institution will be included with your comments when they are sent to the authors. If the manuscript is accepted, your review, name and institution will be published alongside the article.

If this manuscript is rejected from <em>The BMJ</em>, it may be transferred to another BMJ journal along with your reviewer comments. If the article is selected for publication in another BMJ journal, depending on the editorial policy of the journal your review may also be published. You will be contacted for your permission before this happens.

For more information, please see our peer review terms and conditions.

Please confirm that you understand and consent to the above terms and conditions.

I consent to the publication of this review

Please enter your name: Maria Panagioti
Reviewer: 6

Recommendation:

Comments:
Thank you for the opportunity to review this interesting paper submitted to the BMJ. It is clear the authors have undertaken a large and important review, on a topic that is rapidly growing in important. It is concerning, though perhaps not surprising, to see that many published IPD meta-analysis projects are of sub-optimal standards. For example, rarely examining risk of bias or examining publication bias issues. Some of the issues identified have been discussed in previous papers, [1-6] which is why I am not surprised of the problems. The paper also covers some similar ground to Nevitt et al. review of prevalence of IPD meta-analyses and whether they obtain all the data desired.[7] However, this review still adds useful information on the current state of play, and raising awareness that improvements are needed is always important.

Major comments:
- Individual participant data is a broader and more appropriate term than individual patient data; the former is more widely adopted nowadays, so I suggest to use it.
- The AMSTAR-2 tool was developed with conventional systematic reviews of aggregate data in mind. IPD meta-analyses have much broader issues that should be evaluated in terms of quality. For example, the process of obtaining, cleaning and harmonising IPD should be critically appraised, but this is not part of the AMSTAR tool. I'm surprised that the critical appraisal questions within Tierney et al. was not used to supplement AMSTAR,[1]
- Also – and crucially – IPD meta-analyses can have main objectives that were not addressed in original trial publications, such as examining effect modifiers (treatment-covariate interactions). Key variables may not be available in all trials in a field, and so we don’t always need to obtain or include IPD from all
trials in a field to address the research question for the IPD meta-analysis. I wonder if this was taken into account when deciding whether an IPD meta-analysis obtained or searched for all relevant trials; that is, was the research question accounted for when making some classifications of the quality of IPD meta-analyses? For example, the question 'Performed any adjustments/sensitivity analyses to account for missing IPD?' is perhaps not fair when examining interactions, as the relevant reported information about the interactions is rarely available in a publication, and so what else could the authors actually have done? The IPD available goes well beyond that possible without IPD, and it seems unfair to criticise IPD authors for not having all the IPD in this instance or for not performing sensitivity analyses.

- A related point: "To the best of our knowledge, there is no specific tool for assessing the methodological quality of IPD meta-analysis" – this is because each IPD meta-analysis could have very different questions, and so having one overall tool is difficult as it is a broad area. Rather, we need different (perhaps overlapping) tools for appraising IPD MAs that have examined treatment effects from those examining interactions, or developing prediction models, etc.

- The overall rating were classified into four levels: high, moderate, low or critically low quality. I cannot see how each level was defined? This must be based on some threshold of the AMSTAR score in total?
- Quality is not an ordinal measure originally (see previous comment), and yet it is examined as such. For example, "Potential factors that might associate with the methodological quality of SRs were explored with Multi-ordinal regression analysis. The overall ranking of the methodological quality was used as the dependent variable with moderate and high quality being combined into one group to account for the fact that only one SR was judged as high quality” – it is more powerful to analyse on a continuous scale where possible. I wonder if the associations identified would hold if the authors rather considered quality as a continuous variable (i.e. total AMSTAR score) and so used linear regression to examine associations.

- The authors refer to PRISMA in places but surprisingly not to PRISMA-IPD or adherence to this.[8]
- Why was publication year dichotomised at 2017? Again, analysing as continuous (and potential consideration of non-linear trends) is more powerful.
- Percentages are routinely given without 95% confidence intervals. I think Cis are needed to help summarise the uncertainty in the estimates reported.
- "Thus, it is important for evidence users to evaluate the publications bias of SRs with IPD meta-analysis when considering adopting the evidence. Future researchers are suggested to put efforts in reducing the risk of publication bias through comprehensive literature search as well as addressing its potential impact to the review results.” – the authors might include reference to Ahmed et al, where these specific issues are discussed in detail.[4]

I hope these comments are useful to the authors and editors moving forwards, and help the authors to add to their comprehensive review and hard work.

Best wishes, Richard Riley

Reference List

Additional Questions:

The BMJ uses compulsory open peer review. Your name and institution will be included with your comments when they are sent to the authors. If the manuscript is accepted, your review, name and institution will be published alongside the article.

If this manuscript is rejected from The BMJ, it may be transferred to another BMJ journal along with your reviewer comments. If the article is selected for publication in another BMJ journal, depending on the editorial policy of the journal your review may also be published. You will be contacted for your permission before this happens.

For more information, please see our peer review terms and conditions.

Please confirm that you understand and consent to the above terms and conditions.

I consent to the publication of this review.

Please enter your name: Richard Riley

Job Title: Prof of Biostatistics

Institution: Keele University

Reimbursement for attending a symposium?: No

A fee for speaking?: No

A fee for organising education?: No

Funds for research?: Yes

Funds for a member of staff?: No

Fees for consulting?: Yes

Have you in the past five years been employed by an organisation that may in any way gain or lose financially from the publication of this paper?: No

Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this paper?: No

If you have any competing interests please declare them here: I have given
IPD meta-analysis training courses at Keele, Roche, Leeds and Barts. I am publishing a book on IPD meta-analysis with Wiley in 2021.