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.

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

ARTICLE DETAILS

TITLE (PROVISIONAL)	A plea for routinely presenting prediction intervals in meta-analysis
AUTHORS	IntHout, Joanna; Ioannidis, John; Rovers, Maroeska; Goeman, Jelle

VERSION 1 - REVIEW

REVIEWER	Lee, Wen-Chung College of Public Health, National Taiwan University, Taipei, Taiwan, Institute of Epidemiology and Preventive Medicine, and Research Center for Genes, Environment and Human Health
REVIEW RETURNED	22-Jun-2015

GENERAL COMMENTS	The authors of this paper argue for the prediction intervals be routinely reported to allow more informative inferences in meta-analyses. The paper is very well written. I have two comments as listed below: 1. The calculation in the paper is based on normality assumption. But in fact, the random effects (heterogeneity across different studies) may not be normally distributed. It is true that the random error in a study will be normally distributed as the sample size of the study increases without bound (the famous central limit theorem). But this central limit theorem does not apply to the random effect. Even if the number of studies in a meta-analysis increases without bound, the random effect will not tend to a normal distribution. 2.A prediction interval makes a good prediction for the treatment effect of a future study that is similar to those already been done --- if I understand the paper correctly. But do we need to consider the sample size of that future study? Say, the prediction interval of a small-sized future study is wider than that of a large-sized future study?

REVIEWER	Leeflang, Mariska Academic Medical Center; University of Amsterdam, Dept. Clinical Epidemiology and Biostatistics
REVIEW RETURNED	23-Jun-2015

GENERAL COMMENTS	This article is a plea for reporting 95% prediction intervals as a measure of heterogeneity. Although I agree with most of what is said in the paper, I wondered if I should read this article as a scientific

	paper proving the value of prediction intervals, or as an opinion paper with a plea for prediction intervals, or as a more tutorial-like paper that explains me how to implement the use of prediction intervals in the systematic reviews I am involved in. If it is a scientific paper that should prove the value of prediction intervals, I am not sure the provided evidence has convinced me. MAJOR COMMENTS: 1. I miss a few references to earlier work on this topic. See: Chiolero et al (letter). Meta-analyses: with confidence or prediction intervals? Eur J Epidemiol. 2012 Oct;27(10):823-5. Guddat et al. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. Syst Rev. 2012 Jul 28;1:34. (they introduce the same way of presenting the prediction interval in a forest plot as the authors of the current manuscript do) Graham & Moran. Robust meta-analytic conclusions mandate the provision of prediction intervals in meta-analysis summaries. J Clin Epidemiol. 2012 May;65(5):503-10. (this paper seems to have a lot in common with the current one) 2. On page 5, last paragraph, the authors explain the results of a review on antidepressants. I found this difficult to follow, mainly because of the mentioning of the SE on the log scale. It made me wonder if the other figures were on the log scale as well (which I assume they aren't). Would it be possible to remove the SE here? Or just add a few words, for example that SEs for the odds ratio are always expressed on the log scale? It may also help to provide the OR plus confidence interval directly and then move on to a measure on the log scale. The authors do not explain on what scale the tau-square is reported. Readers not too familiar with tau-square may be left wondering if this is on the log scale as well. 3. In the same paragraph, they define an I-square of 59.9% as being substantial, with a reference to the Cochrane Handbook. However, on page 7 (first paragraph) they cite the same Handbook and state that an I-square between 30% and 60% is 'moderate'. Please be consistent. 4. The authors claim in several places in the manuscript that the PI can be easily calculated (e.g. first sentence on page 6). I would not state it that way. What may seem easy for the authors, may be very difficult for the average non-statistician. 5. At the end of this paragraph, the authors state that the estimated probability that the antidepressants will be ineffective is 13.4%. This went a bit quick... Where does the 13.4% come from? I can find it in the Appendix, but a few words of explanation would have been helpful. 6. Results section. Overall, almost three quarters had a PI that included the null effect. Is it possible to state something about the extent to which this was the case? If a PI just includes the null effect, I would less worried than in case when the PI includes an opposite effect that is just as large as the mean (pooled) effect (e.g. OR of 2 and a PI that includes 0.5). 7. I also wondered if the observation about the PI including the null effect was in line with the observed results in the included studies. In other words, if I would have summarized the results in terms of range of individual study results, would I have come to the same

	conclusions? 8. In the results section , the authors explain that they assessed the impact of low heterogeneity by calculating the 95% PIs under different assumptions. I have a few remarks about this section. a. First, I do not see the value of such a sensitivity analysis on an arbitrary assumption. Please explain the rationale of this exercise, or of such a sensitivity analysis if it was done in a meta-analysis. b. Second, I think this should move to the methods section. c. They only seem to have tested one other assumption, namely that of a I-square of 20%. Please show the other assumptions as well, or just state that you assessed one other assumption. d. Again they claim that this is an analysis that is easy to do, while I wonder if that really is the case for everyone. So that should be removed too. e. Have they followed the explanations in the Appendix here? Please refer to the Appendix then. 9. In the first lines of their conclusion, the authors state that the CI is inadequate for clinical decision making, thus implying that the PI would be adequate for decision making. I wonder if this is really the case. Although I do agree that the PI may be much more informative than the CI, a PI including the null effect or even opposite effects may be limiting for decision making. It may prevent people from making the wrong decision, but at the same time it may be preventing policy makers from making any decision. 10. I may be a bit unfair to stress the assumptions and limitations of the other measures used to express variability in the introduction, and only present some of the limitations of the PI as caveats in the Discussion. An overarching tables presenting the pros and cons of all methods would perhaps have been more helpful. 11. The Discussion section ends with a strong statement about the use of PIs. “Therefore it should be reported as the main tool for clinical prediction making”. Although I do agree with this statement, it does not do right to the provided evidence. Such a statement would actually require a study under decision makers to investigate whether a PI really leads to better decision making than using the CI. I think all the outcome measures (the mean effect, the CI and the PI) provide different and possibly equally important information. MINOR REMARKS 12. In meta-analyses of diagnostic accuracy, prediction regions (as the outcome has two dimensions, sensitivity and specificity) are also recommended. However, there is less stress on the use of a 95% PI; some experts rather prefer 90% PI or 50% PI. Could the authors elaborate on their choice for a 95% PI? Or is that just because all previous papers suggest using a 95% PI? 13. Just to clarify: on page 6, first paragraph, the authors state that the prediction interval can be seen as 95% range of true ORs to be expected in similar studies. So it does not reflect the 95% range of observed ORs?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1, Prof. Wen-Chung Lee

Comments:

The authors of this paper argue for the prediction intervals be routinely reported to allow more informative inferences in meta-analyses. The paper is very well written. I have two comments as listed below:

1. The calculation in the paper is based on normality assumption. But in fact, the random effects (heterogeneity across different studies) may not be normally distributed. It is true that the random error in a study will be normally distributed as the sample size of the study increases without bound (the famous central limit theorem). But this central limit theorem does not apply to the random effect. Even if the number of studies in a meta-analysis increases without bound, the random effect will not tend to a normal distribution.

We agree with the reviewer that it can be considered a limitation that the calculations in the paper are based on the normality assumption, which may be inadequate in certain situations. However, we believe that it is preferable to show that there is heterogeneity (although possibly inadequate) then to neglect it. If there are sufficient studies available, we recommend to investigate whether the normal assumption is appropriate.
We have added a limitation to the bulleted list of limitations: "Limitations are that the calculations and inferences for the prediction interval are based on the normality assumption, which is difficult to ensure." In addition, we discuss this issue in the discussion: "It is straightforward to calculate a prediction interval if we can assume that the effects are normally distributed and that τ^2 is known and stable across studies. However, one should realize that the prediction interval is dependent on this assumption and on the precisions of the estimated τ^2 and study effect, and will be imprecise if the number of studies in the meta-analysis is small. If the number of studies is large, estimates will be more precise and the normality of the distribution of τ^2 can be empirically evaluated."

2. A prediction interval makes a good prediction for the treatment effect of a future study that is similar to those already been done --- if I understand the paper correctly. But do we need to consider the sample size of that future study? Say, the prediction interval of a small-sized future study is wider than that of a large-sized future study?

A prediction interval estimates where the **true** effects are to be expected for 95% of similar (exchangeable) studies that might be conducted in the future. The effects that one can expect to actually observe will have an even wider range, indeed depending on the sample size of the future study. In this paper we choose to emphasize the range of expected true effects, instead of the range of expected observed effects, as we think that this is in the end more important. We have tried to emphasize more that the text is about prediction intervals for the true instead of the observed effects, by using more often the term "true".

Reviewer: 2, Mariska Leeflang, PhD

Comments:
This article is a plea for reporting 95% prediction intervals as a measure of heterogeneity. Although I agree with most of what is said in the paper, I wondered if I should read this article as a scientific paper proving the value of prediction intervals, or as an opinion paper with a plea for prediction intervals, or as a more tutorial-like paper that explains me how to implement the use of prediction intervals in the systematic reviews I am involved in. If it is a scientific paper that should prove the value of prediction intervals, I am not sure the provided evidence has convinced me.

The paper is an opinion paper with some tutorial aspects indeed. We have emphasized this in the current version of the introduction as follows: “Our objective in the current article is to show the potential advantages of obtaining and reporting the prediction interval routinely in meta-analyses because its clinical meaning is much more straightforward.”

MAJOR COMMENTS:

1. I miss a few references to earlier work on this topic. See:

- Chiolero et al (letter). Meta-analyses: with confidence or prediction intervals? Eur J Epidemiol. 2012 Oct;27(10):823-5.
- Guddat et al. A note on the graphical presentation of prediction intervals in random-effects meta-analyses. Syst Rev. 2012 Jul 28;1:34. (they introduce the same way of presenting the prediction interval in a forest plot as the authors of the current manuscript do)
- Graham & Moran. Robust meta-analytic conclusions mandate the provision of prediction intervals in meta-analysis summaries. J Clin Epidemiol. 2012 May;65(5):503-10. (this paper seems to have a lot in common with the current one)

We thank the reviewer for this comment. In the current version we do refer to Chiolero and Guddat. Further, we discuss the findings of Graham and Moran as follows: “Graham en Moran evaluated prediction intervals in 72 meta-analyses with a dichotomous outcome in critical care published between 2002 and 2010. They found a higher percentage of significant meta-analyses (50/72, 69.4%), compared to 28.5% (572/2009) in our set of meta-analyses with an odds ratio outcome. The difference may be caused by publication bias, the higher number of primary studies in their sample (medium 9 versus 4 in our set), and by their use of the DerSimonian-Laird approach which can result in too many statistically significant findings, whereas we used the HKSJ approach. However, results with respect to the prediction interval were remarkably similar. In 32 (64.0%) of their 50 significant meta-analyses the 95% prediction interval included the null, similar to 65.8% in our dataset. Seven (14.0%) of their 50 meta-analyses suggested a high probability of exact reversal of the efficacy or harm, similar to 12.3% of our meta-analyses where the prediction interval contained the opposite effect, despite the fact that they used a different definition for possible “harm” and that they did not mention whether there was positive between-study heterogeneity in their significant meta-analyses.”

2. On page 5, last paragraph, the authors explain the results of a review on antidepressants. I found this difficult to follow, mainly because of the mentioning of the SE on the log scale. It made me wonder if the other figures were on the log scale as well (which I assume they aren’t). Would it be possible to remove the SE here? Or just add a few words, for example that SEs for the odds ratio are always expressed on the log scale? It may also help to provide the OR plus confidence interval directly and then move on to a measure on the log scale. The authors do not explain on what scale the tau-square is reported. Readers not too familiar with tau-square may be left wondering if this is on the log scale as well.

We agree with the reviewer that the example was rather difficult. We have now selected an example with a continuous outcome, on the effect of topical steroids on the decrease in overall symptom scores. In the appendix, formulas 1 and 2, we have added the explanation of the calculations for outcomes on a binary scale.

3. In the same paragraph, they define an I-square of 59.9% as being substantial, with a reference to the Cochrane Handbook. However, on page 7 (first paragraph) they cite the

same Handbook and state that an I-square between 30% and 60% is 'moderate'. Please be consistent.

We have used a different example now, so this issue is solved (now I^2 in the example is 73.9%).

4. The authors claim in several places in the manuscript that the PI can be easily calculated (e.g. first sentence on page 6). I would not state it that way. What may seem easy for the authors, may be very difficult for the average non-statistician.

We agree, and have removed suggestive wording related to the easiness of the calculations.

5. At the end of this paragraph, the authors state that the estimated probability that the antidepressants will be ineffective is 13.4%. This went a bit quick... Where does the 13.4% come from? I can find it in the Appendix, but a few words of explanation would have been helpful.

The current text states: "The prediction interval contains values below zero, which corresponds to a decrease in symptom scores of at best approximately 1.5 SD after steroid use compared to placebo. But it also contains values above zero which means that the steroids may exhibit no or even a harmful effect (SMD>0) in some settings, with a (95%) worst case increase in SMD of 0.53. Consequently, the effect in a new study may be even the exact opposite to the summary point estimate of the meta-analysis, i.e. an increase of 0.51 instead of a decrease of 0.51 may occur. The estimated probability that the true effect of the steroids will be null or higher in a new study is equal to 13.6%, based on the t-distribution with 6 degrees of freedom (formula 2 appendix)."

6. Results section. Overall, almost three quarters had a PI that included the null effect. Is it possible to state something about the extent to which this was the case? If a PI just includes the null effect, I would less worried than in case when the PI includes an opposite effect that is just as large as the mean (pooled) effect (e.g. OR of 2 and a PI that includes 0.5).

We thank the reviewer for this suggestion, and we have indeed added a paragraph on the number of prediction intervals containing the opposite effect.

7. I also wondered if the observation about the PI including the null effect was in line with the observed results in the included studies. In other words, if I would have summarized the results in terms of range of individual study results, would I have come to the same conclusions?

If we compare the estimated PI with the "naïve" PI, based on the range of observed effects (point estimates and/or confidence intervals) in the included studies, the similarity of the PIs will depend on the number and size of the included studies. The PI aims to estimate the range of true effects in a future, similar study. Therefore, if the number of included studies is small, the naïve PI tends to be too small if it is merely based on the range of the point estimates, but it might be too broad if it is based on the confidence intervals of the most extreme studies, especially if these studies happen to be small. If the number of included studies is large, the naïve PI, whether it is based on point estimates or on confidence intervals, tends to be too broad, because it will reflect observed effects instead of true effects. Observed effects show more variation due to the added variability of the limited sample sizes.

Further, although we agree that the naïve PI might give already an impression of the width and location of the PI, its range cannot easily be summarized with numbers.

8. In the results section, the authors explain that they assessed the impact of low heterogeneity by calculating the 95% PIs under different assumptions. I have a few remarks about this section.
    a. First, I do not see the value of such a sensitivity analysis on an arbitrary assumption. Please explain the rationale of this exercise, or of such a sensitivity analysis if it was done in a meta-analysis.
    b. Second, I think this should move to the methods section.
    c. They only seem to have tested one other assumption, namely that of a I-square of 20%. Please show the other assumptions as well, or just state that you assessed one other assumption.
    d. Again they claim that this is an analysis that is easy to do, while I wonder if that really is the case for everyone. So that should be removed too.
    e. Have they followed the explanations in the Appendix here? Please refer to the Appendix then.

In approximately 50% of meta-analyses the estimated heterogeneity is zero, whereas the true τ is unlikely to ever be exactly 0. We chose to impute an I^2 of 20% for these meta-analyses. We agree that this choice was arbitrary, we could also have used the upper limit of the 95% (or 67%) confidence interval for the τ, or an I^2 of 10%, or anything else. However, our main objective was to exemplify the message that an estimated I^2 of zero can result in a broader range of expected true effects than is suggested by the confidence interval, and that even if heterogeneity is estimated to be zero, it can be present.

With regard to the methods: we have moved the description of the calculation to the methods section: "For significant meta-analyses where the heterogeneity estimate was zero, we assessed the impact of possibly low but non-zero heterogeneity by assuming an I^2 of 20%, calculating prediction intervals using formula 3 (appendix)."

9. In the first lines of their conclusion, the authors state that the CI is inadequate for clinical decision making, thus implying that the PI would be adequate for decision making. I wonder if this is really the case. Although I do agree that the PI may be much more informative than the CI, a PI including the null effect or even opposite effects may be limiting for decision making. It may prevent people from making the wrong decision, but at the same time it may be preventing policy makers from making any decision.

In our conclusion, we state that the CI is inadequate for clinical decision making, and that the PI is more informative. We do not say that the PI is adequate, as we realize that the PI is also not perfect. Further we think that a decision should be based on all relevant evidence.

10. I may be a bit unfair to stress the assumptions and limitations of the other measures used to express variability in the introduction, and only present some of the limitations of the PI as caveats in the Discussion. An overarching tables presenting the pros and cons of all methods would perhaps have been more helpful.

We thank the reviewer for the suggestion. We have added a table with the pros and cons of all methods.

11. The Discussion section ends with a strong statement about the use of PIs. “Therefore it should be reported as the main tool for clinical prediction making”. Although I do agree with this statement, it does not do right to the provided evidence. Such a statement would actually require a study under decision makers to investigate whether a PI really leads to better decision making than using the CI. I think all the outcome measures (the mean effect, the CI and the PI) provide different and possibly equally important information.

We agree, and we have changed the sentence as follows: “Therefore it should be routinely reported in addition to the summary effect and its confidence interval, and used as a main tool for interpreting evidence, to enable more informed clinical decision making.” .

MINOR REMARKS

12. In meta-analyses of diagnostic accuracy, prediction regions (as the outcome has two dimensions, sensitivity and specificity) are also recommended. However, there is less stress on the use of a 95% PI; some experts rather prefer 90% PI or 50% PI. Could the authors elaborate on their choice for a 95% PI? Or is that just because all previous papers suggest using a 95% PI?

Indeed, this choice was influenced by the choice of previous papers. Practice will show which percentage will be useful.
We have added the following sentence to the appendix, formula 1 section: “Of course it is possible to estimate prediction intervals with a different coverage, e.g. an 80% prediction interval would be based on $t_{0.20/2,6}$.”

13. Just to clarify: on page 6, first paragraph, the authors state that the prediction interval can be seen as 95% range of true ORs to be expected in similar studies. So it does not reflect the 95% range of observed ORs?

This is correct, the PI we discuss is the expected range of true effects. We have tried to emphasize that more clearly in the text.

VERSION 2 – REVIEW

REVIEWER	Mariska MG Leeflang Academic Medical Center, University of Amsterdam The Netherlands
REVIEW RETURNED	04-Jan-2016

GENERAL COMMENTS	I have peer reviewed an earlier version of this interesting manuscript and almost all of the comments I had back then have been addressed. A few remaining (minor) ones are listed below. In future meta-analyses, I will certainly consider using prediction intervals, thanks to this study. 1. I wondered if the 95% prediction interval always included one or more data points from the included studies. If this is not the case, for example because the data are usually not normally disctibuted, then what is the point of reporting a PI that may only theoretically include the null or opposite effect?

	2. On p4-5, the authors state that the I-square depends on the sample size. Please make explicit whether this s the sample size of the included studies, the sample size of the meta-analysis (i.e. the number of studies included) or both. 3. An I-square of 0% resulted more often in a PI that included the null (in contrast to the corresponding CI) than an I-square between 0 and 30%. I found this fascinating and made me wonder what the difference was between the meta-analyses reporting an I-square of 0% and those reporting an I-square of 0-30% (number of included studies? sample size of included studies?). 4. The percentage of PIs and CIs both excluding the null effect was relatively high in meta-analyses with an I-square of 0-30% (for contiuous data even up to more than 75%). Does this mean that reporting PI may not be necessary in these instances? 5. The authors state that the average number of included studies was 4. I find this very low and wonder how many meta-analyses included only 2 studies. Could this be escribed and perhaps commented on (I mean, what is the use of doing a meta-analysis on 2 studies anyway)? Why didn't the auhtors not set a limit to a minimum of three or four studies? 5. In Table 2, it is not clear where a meta-analysis with I-square of 30 would fall: either in the 0-30 category or the 30-60 category? And in Table W1, the second column heading for the category of 2-6 studies states "I2<30", while the second column heading for the other categories only says "<30".

REVIEWER	Felix Achana University of Warwick, UK I am an author in one of the papers suggested should be cited as an example
REVIEW RETURNED	19-Jan-2016

GENERAL COMMENTS	The authors present an interesting peace of research that highlights the importance of prediction interval as a measure of heterogeneity in random effects meta-analysis. They make a strong case for routine reporting of prediction intervals alongside the usual frequently indices to make the interpretation of random effects meta-analysis more informative especially for a clinical audience. The paper is well written, easy to read and likely to be of interest to clinicians interested in research synthesis, etc. I have only two minor comments (below). Comment 1: Line 38/39 of page 6 and also formula 1 on page 13. The authors assumed a t-distribution with 1 degree of freedom in calculating the 95% prediction intervals. I would suggest a t-distribution with 2 degrees of freedom is more appropriate given that two parameters (i.e. the mean of the random effects and tau) are estimated from the data (see Higgins et al 2006 re-evaluation paper). If as is often the case in tau is assumed to be known exactly rather than estimated, then this should be made clear in the appendix. In any case it would be useful to include a justify of the choice of distribution in the appendix. Comment 2: Lines 21 to 25 of page 10. published examples of the

	statement that prediction intervals that completely lies in the desired direction of effect would increase confidence in the results can be cited e.g. Wilmot et al 2012 (http://www.ncbi.nlm.nih.gov/pubmed/22890825), or similar published example could be cited.

REVIEWER	Jason Oke Nuffield Department of Primary Care Health Sciences. University of Oxford. United Kingdom
REVIEW RETURNED	23-Jan-2016

GENERAL COMMENTS	The manuscript is generally well written and the subject matter it coves relevant and provocative. I do however think the manuscript could be improved by addressing the following concerns. **1. Page 3 Strengths and limitations (5th bullet point): In the sentence** *"Further, the interval will be imprecise if the estimates of the summary effect and the τ2 are imprecise".* I think you should write "between study heterogeneity" or similar here in place of greek letter. **2. Page 4, Introduction: first sentence** *"Interventions may have heterogeneous effects across studies because of differences in study populations, interventions, follow-up length, bias, and other factors.".* The phrase "bias and other factors" seems too vague and should be explained properly. **3. Page 5, line 36.** *"A prediction interval always presents the heterogeneity on the same scale as the original outcomes, in contrast to τ, τ^2 or I^2."* I think tau is not supposed to be here? – tau is on the same scale as the outcome as has been stated in the appendices. **Page 5: line 56** "*one can also calculate the probability that the true effect will be harmful"* – minor point but I think that shouldn't say harmful it depends on the context and in the intervention – it may just be a waste of money! **4. Page 6, second paragraph.** *"We derive the SE from the 95% C.I. of the SMD (formula 1*

	appendix), which results in an SE of 0.182." *Appendix says – "it can be approximated by dividing the distance between the limits of the 95% CI of the SMD by 3.92"* I make that (-0.96 - -0.07) = 0.89/3.92 = 0.227 not 0.182? **5. Page 6 paragraph 2** You suggest using the value from the t distribution with k-1 degrees of freedom where k is the number of study estimates. In the Riley paper (Interpretation of random effects BMJ) they suggest in their call out box "calculating a prediction interval" using k-2 degrees of freedom for reasons they don't really explain. However, is there a reason why you have not followed their advice, if so it probably need some explanation or is it an oversight on your part? Also, and this is slightly pedantic – I think you should write $t_{1-0.05,6}$ or $t_{1-\text{alpha},6}$ where alpha= 0.05 not $t_{0.05,6}$ as this is convention and (1-0.05) returns the non-negative value in most software, whereas 0.05 does not. **6. Page 11. Line 12.** The word medium in *"… in their sample (medium 9 versus 4 in our set)"* – is this supposed to be median? **7. Page 12 – Power calculations for a future study.** I agree with the main premise of this section but there are a couple of sentences I do not agree with or perhaps have misconstrued? *"Power may decrease to 5% or less in case of a null effect"* – Surely if the true if effect is zero then the power argument is irrelevant? When the true effect is zero you cannot make a type 2 error, failing to detect a difference as the difference does not exist and hence there can be no power either? "*If the prediction interval shows that 30% of future studies may have a true null or negative effect, the power can never be much larger than 70%."* I think this statement is wrong using the conventional definitions of study power, but it may be a difference in how we think of this problem? I will try to explain how I see it. Let's assume a prediction interval for an intervention (summary effect 0.5) we are interested ranges from -1.35 to 2.36 so that approx. 30% < 0 and positive effects are beneficial. If we plan a new study and assume that the effect in our study will be similar to the ones in the previous MA with the most positive effect estimate (>2) then it will not be difficult to obtain a power exceeding 70% regardless of the within-study variation, the fact that some other estimates are negative is not relevant for our calculation. Similarly, effects sizes in the opposite direction (<-1) could also have power >

	0.7 using the same reasoning. Hence the statement is incorrect. Power calculations based on the effects around zero would undoubtedly be underpowered for any reasonable N but it depends on which estimate you choose. One could calculate the posterior power using the prediction interval as a prior distribution and this could well show that the probability that the power is > 0.7 is low but it will not be zero. **7. Results** My final point is that I think the paper could be strengthened by exploring whether there is any support in the data to back up the main finding that for so many of the prediction intervals in the paper, the intervention effect could be null or in the opposite direction to the summary effect estimate. I would like to know and I think it would be useful for the reader to know in how many of these MA's was there one or more individual studies that reported estimates in the opposite direction to the summary effect. Hence, in part validating the prediction interval normality assumption. I may be wrong, but I think that it is a very real possibility for prediction intervals to overlap regions of no effect but yet there are no actual estimates in this region especially if the summary estimate is small and tau2 is large. In the example given (Figure1) the assumption looks appropriate because one study (Jorissen) has a study effect in the opposite direction (being positive) but I would like to know in how many cases this does happen.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1
Mariska MG Leeflang, Amsterdam, the Netherlands

I have peer reviewed an earlier version of this interesting manuscript and almost all of the comments I had back then have been addressed. A few remaining (minor) ones are listed below. In future meta-analyses, I will certainly consider using prediction intervals, thanks to this study.

1. I wondered if the 95% prediction interval always included one or more data points from the included studies. If this is not the case, for example because the data are usually not normally distributed, then what is the point of reporting a PI that may only theoretically include the null or opposite effect?

Anwer: The prediction interval always included one or more data points from the included studies. However, when a meta-analysis is based on only a few studies, the prediction interval reflects the uncertainty and can be (much) wider than the range covered by the study effects.

2. On p4-5, the authors state that the I-square depends on the sample size. Please make explicit whether this is the sample size of the included studies, the sample size of the meta-analysis (i.e. the number of studies included) or both.

Anwer: We agree with the reviewer that the text is not completely clear. We meant the following: Given the same tau-square, I-square is influenced by the precision of the primary studies: given the same tau-square, more precise studies will result in a larger I-square.
We have added "of the included studies" with a reference to the 2008 paper of Gerta Rucker et. al.

(Rucker, G., Schwarzer, G., Carpenter, J., & Schumacher, M. (2008). Undue reliance on I2 in assessing heterogeneity may mislead. BMC Medical Research Methodology, 8(1), 79.).

3. An I-square of 0% resulted more often in a PI that included the null (in contrast to the corresponding CI) than an I-square between 0 and 30%. I found this fascinating and made me wonder what the difference was between the meta-analyses reporting an I-square of 0% and those reporting an I-square of 0-30% (number of included studies? sample size of included studies?).

Anwer: Indeed this is fascinating, and we thank the reviewer for bringing this to our attention.
We noted the following differences between the meta-analyses reporting an I-square of 0%, compared to those reporting an I-square >0 but below 30%:

For the meta-analyses with dichotomous outcomes:
- The number of studies per meta-analysis was smaller (5 [3-8] versus 8.5 [6-17], Wilcoxon p<0.001). Consequence: larger t-values in calculations for prediction-interval
- the variance of the typical study was larger (0.40, with interquartile range q1-q3:[0.20-1.06], versus 0.27 [0.15-0.60], Wilcoxon p=0.002).
- (consequently) the imputed tau2 was larger (0.10 [0.05-0.27] versus .01 [.01-.11], Wilcoxon p<0.001
- SE of pooled effect was smaller,0 .12 [0.07 - 0.19], vs. 0.19 [0.13-0.25], p<.001
- The pooled summary effect sizes were very similar.
Together, starting from similar summary effect sizes, and creating prediction intervals with larger t-values and with larger tau, but with smaller SE, the prediction intervals for meta-analyses with I-square =0 were wider and more often contained the null effect.

Meta-analyses with continuous outcomes:
- Number of studies smaller, p<0.001, 3 [2-4] versus 8 [5-12.5]
- the variance of the typical study was similar, p=0.263, 0.06 [.03-.10] versus 0.05 [0.03-0.08]
- consequently imputed tau2 in R0 slightly larger, R0: 0.02 [0.01-0.03] versus 0.01 [0.00-0.01]
- pooled summary effect sizes were very similar, p=0.393
- SE of pooled effect was smaller, p<0.001, 0.05 [0.02-0.08] versus 0.07 [0.05-0.11]
Together, starting from similar summary effect sizes, and creating prediction intervals with larger t-values and with somewhat larger tau, but with smaller SE, the prediction intervals for meta-analyses with I-square =0 were wider and more often contained the null effect.

Overall, the major reasons seem to be the combination of the lower number of studies and the fact that the imputed tau2 is somewhat larger for the meta-analyses with I-square =0. However, we chose not to mention this in the paper, as we think that more research is needed. Moreover, the results of the imputation of the I-square =20% only serve as an example, the fact that I2 is not automatically equal to 0 is the major point.

4. The percentage of PIs and CIs both excluding the null effect was relatively high in meta-analyses with an I-square of 0-30% (for continuous data even up to more than 75%). Does this mean that reporting PI may not be necessary in these instances?

Anwer: This selection of meta-analyses with I-square between 0 and 30% contained also meta-analyses with a very small tau-square, almost equal to zero. In these cases, prediction intervals and confidence intervals will be almost equal. However, we think that prediction intervals are always warranted: Ideally a reference on a meta-analysis not only contains the confidence interval but also routinely the prediction interval in one sentence. Then no further insight in the exact size of tau2 or I2 is needed.

5. The authors state that the average number of included studies was 4. I find this very low and wonder how many meta-analyses included only 2 studies. Could this be described and perhaps commented on (I mean, what is the use of doing a meta-analysis on 2 studies anyway)? Why didn't the authors not set a limit to a minimum of three or four studies?

Anwer: we selected meta-analyses with a minimum of two studies, if those studies were also combined in the original Cochrane review, as we wanted to reflect the current status quo as precise as possible.
Details on these meta-analyses were already described in another paper (IntHout J, Ioannidis JPA, Borm GF, et al. Small studies are more heterogeneous than large ones: a meta-meta-analysis. Journal of Clinical Epidemiology 2015;68(8):860-69), to which I refer in the text.
In that paper it is written: Most meta-analyses were based on a few small studies. Of the 3,263 selected meta-analyses, 1,025 (31%) were based on 2 studies, 1,226 (38%) on 3 to 5 studies, 603 (18%) on 6 to 10 studies, and 409 (13%) on more than 10 studies. The median number of studies per meta-analysis was 4 (Q1-Q3: 2-6). More details are presented in the tables of that paper.

We added the following underlined text in the methods section on the Cochrane database:
“To avoid subjectivity in the selection we used the first meta-analysis with a dichotomous or continuous outcome and based on at least two studies in the Data and Analyses section when these studies were also combined in the original review, as we wanted to reflect the status quo as precise as possible.”

5. In Table 2, it is not clear where a meta-analysis with I-square of 30 would fall: either in the 0-30 category or the 30-60 category? And in Table W1, the second column heading for the category of 2-6 studies states "I2<30", while the second column heading for the other categories only says "<30".

Answer: We agree that this is unclear and inconsistent, and have adapted the column headers.

Reviewer: 2
Felix Achana, University of Warwick, UK

Please state any competing interests or state ‘None declared’:
I am an author in one of the papers suggested should be cited as an example

The authors present an interesting peace of research that highlights the importance of prediction interval as a measure of heterogeneity in random effects meta-analysis. They make a strong case for routine reporting of prediction intervals alongside the usual frequently indices to make the interpretation of random effects meta-analysis more informative especially for a clinical audience. The paper is well written, easy to read and likely to be of interest to clinicians interested in research synthesis, etc. I have only two minor comments (below).

Comment 1:
Line 38/39 of page 6 and also formula 1 on page 13. The authors assumed a t-distribution with 1 degree of freedom in calculating the 95% prediction intervals. I would suggest a t-distribution with 2 degrees of freedom is more appropriate given that two parameters (i.e. the mean of the random effects and tau) are estimated from the data (see Higgins et al 2006 re-evaluation paper). If as is often the case in tau is assumed to be known exactly rather than estimated, then this should be made clear in the appendix. In any case it would be useful to include a justify of the choice of distribution in the appendix.

Answer:

In the current literature there is disagreement on the number of degrees of freedom that must be used in the prediction interval of a meta-analysis. Indeed, Higgins (2009) and also Riley (2011) present their formula with degrees of freedom equal to the number of studies minus 2 (without clear rationale). Viechtbauer presents in his R package metafor prediction intervals with degrees of freedom equal to the number of studies minus 1 (=k-1), mainly to be consistent with meta-regression. The true number of degrees of freedom that needs to be subtracted will be somewhere between 1 and 2, however, the number is not (yet) precisely known.
For the many meta-analyses in our sample that were based on only a few studies, the prediction intervals with k-2 df would be very wide. In order to increase credibility and acceptance, we chose to use k-1 df. In addition, we used the HKSJ approach, which also multiplies the standard error of the pooled effect with a factor >1.
We added the statement “Note that the interval is calculated under the assumption that the value of τ2 is known (and not estimated).” to the appendix formula 1. We hope that in the near future there will be more clarity with respect to the correct number of degrees of freedom.

Comment 2:
Lines 21 to 25 of page 10. published examples of the statement that prediction intervals that completely lies in the desired direction of effect would increase confidence in the results can be cited e.g. Wilmot et al 2012 (http://www.ncbi.nlm.nih.gov/pubmed/22890825), or similar published example could be cited.

We thank the reviewer for the comment. We have added the following text to section 2 on prediction intervals:
“For example, in a meta-analysis on sedentary time in adults and the association with diabetes, cardiovascular disease and death, confidence intervals were thought to represent insufficiently the different study populations. Therefore also prediction intervals were reported”

Reviewer: 3

Reviewer Name
Jason Oke, Nuffield Department of Primary Care Health Sciences. University of Oxford. United Kingdom

The manuscript is generally well written and the subject matter it coves relevant and provocative. I do however think the manuscript could be improved by addressing the following concerns.

1. Page 3 Strengths and limitations (5th bullet point): In the sentence
“Further, the interval will be imprecise if the estimates of the summary effect and the τ2 are imprecise”.
I think you should write “between study heterogeneity” or similar here in place of greek letter.

Answer: we agree and adapted the text accordingly.

2. Page 4, Introduction: first sentence
“Interventions may have heterogeneous effects across studies because of differences in study populations, interventions, follow-up length, bias, and other factors.”.
The phrase “bias and other factors” seems too vague and should be explained properly.

Answer: In hindsight, this phrase is indeed vague. However, in our opinion the aim of this paper is to give insight in prediction intervals and not to provide an exhausting list of reasons for heterogeneity. We have changed the sentence as follows: Interventions may have heterogeneous effects across

studies because of differences in study populations, interventions, follow-up length, bias, andor other factors like publication bias.

3. Page 5, line 36.
“A prediction interval always presents the heterogeneity on the same scale as the original outcomes, in contrast to τ, τ^2 or I^2.”
I think tau is not supposed to be here? – tau is on the same scale as the outcome as has been stated in the appendices.

Answer: For binary data, tau is also not on the same scale as the original outcome, if the original outcome is presented by a ratio (e.g. odds ratio). We adapted the text as follows:
in contrast to τ (e.g. in case of odds ratios), $\tau 2$ or I2

Page 5: line 56 “one can also calculate the probability that the true effect will be harmful” – minor point but I think that shouldn’t say harmful it depends on the context and in the intervention – it may just be a waste of money!

Answer: We agree that indeed the true effect is not always harmful in a medical sense if it is on the other side of the null. That is why we also specified “be harmful (on the other side of the null)”. For readability we chose to keep the text as it is.

4. Page 6, second paragraph.
“We derive the SE from the 95% C.I. of the SMD (formula 1 appendix), which results in an SE of 0.182.”
Appendix says – “it can be approximated by dividing the distance between the limits of the 95% CI of the SMD by 3.92”
I make that (-0.96 - -0.07) = 0.89/3.92 = 0.227 not 0.182?

Answer: Thank you for your attention. Indeed we made a mistake in the calculations. The text is corrected, and a correct Figure 1 (with updated prediction interval) is included.

5. Page 6 paragraph 2
You suggest using the value from the t distribution with k-1 degrees of freedom where k is the number of study estimates. In the Riley paper (Interpretation of random effects BMJ) they suggest in their call out box “calculating a prediction interval” using k-2 degrees of freedom for reasons they don’t really explain. However, is there a reason why you have not followed their advice, if so it probably need some explanation or is it an oversight on your part?
Also, and this is slightly pedantic – I think you should write t1 – 0.05, 6 or t1- alpha, 6 where alpha= 0.05 not
t0.05, 6 as this is convention and (1-0.05) returns the non-negative value in most software, whereas 0.05 does not.

In the current literature there is disagreement on the number of degrees of freedom that must be used in the prediction interval of a meta-analysis. Indeed, Higgins (2009) and also Riley (2011) present their formula with degrees of freedom equal to the number of studies minus 2 (without clear rationale). Viechtbauer presents in his R package metafor prediction intervals with degrees of freedom equal to the number of studies minus 1 (=k-1), mainly to be consistent with meta-regression. The true number of degrees of freedom that needs to be subtracted will be somewhere between 1 and 2, however, the number is not (yet) precisely known.
For the many meta-analyses in our sample that were based on only a few studies, the prediction intervals with k-2 df would be very wide. In order to increase credibility and acceptance, we chose to use k-1 df. We added the statement “Note that the interval is calculated under the assumption that the

value of τ2 is known (and not estimated).” to the appendix formula 1. We hope that in the near future there will be more clarity with respect to the correct number of degrees of freedom.

With respect to the subscript of t, we agree and adapted the text accordingly.

6. Page 11. Line 12.
The word medium in “… in their sample (medium 9 versus 4 in our set)” – is this supposed to be median?

Answer: thank you, we adapted the text accordingly

7. Page 12 – Power calculations for a future study.
I agree with the main premise of this section but there are a couple of sentences I do not agree with or perhaps have misconstrued?
“Power may decrease to 5% or less in case of a null effect” – Surely if the true if effect is zero then the power argument is irrelevant? When the true effect is zero you cannot make a type 2 error, failing to detect a difference as the difference does not exist and hence there can be no power either?
“If the prediction interval shows that 30% of future studies may have a true null or negative effect, the power can never be much larger than 70%.”
I think this statement is wrong using the conventional definitions of study power, but it may be a difference in how we think of this problem? I will try to explain how I see it.
Let’s assume a prediction interval for an intervention (summary effect 0.5) we are interested ranges from -1.35 to 2.36 so that approx. 30% < 0 and positive effects are beneficial. If we plan a new study and assume that the effect in our study will be similar to the ones in the previous MA with the most positive effect estimate (>2) then it will not be difficult to obtain a power exceeding 70% regardless of the within-study variation, the fact that some other estimates are negative is not relevant for our calculation. Similarly, effects sizes in the opposite direction (<-1) could also have power > 0.7 using the same reasoning. Hence the statement is incorrect. Power calculations based on the effects around zero would undoubtedly be underpowered for any reasonable N but it depends on which estimate you choose. One could calculate the posterior power using the prediction interval as a prior distribution and this could well show that the probability that the power is > 0.7 is low but it will not be zero.

Answer: We used the word power to denote the probability of a significant result in a new trial This may indeed be confusing as it differs from the conventional definition of study power. Therefore we replaced it by “probability of a statistically significant result in a new study” (in full) and “probability of a significant study” in short. The updated text is:
“Meta-analysis results can also be used for power calculations for a new study. However, the expected true effect in a new study is not necessarily equal to the point estimate of the meta-analysis: it can be any of the values in the prediction interval. In case of heterogeneity the probability of a statistically significant result in a new study may differ substantially from an apparent power of 80% based on the point estimate. The latter will be overly optimistic because the power function is asymmetric. If the true study effect is larger than the point estimate the real power probability of a significant of the study will be higher, up to a maximum of 100%, but if the effect is smaller the power probability may decrease substantially, even to 5% or less in case of a null effect. Consequently the expected power probability of a significant new study in case of heterogeneity will be lower than 80% (formula 4 appendix). For example, if the prediction interval shows that 30% of future studies may have a true null or negative effect, the power probability of a significant new study can never be much larger than 70%. The sample size should be increased to compensate for this loss in power, see also Roloff et al.21

8. Results

My final point is that I think the paper could be strengthened by exploring whether there is any support in the data to back up the main finding that for so many of the prediction intervals in the paper, the intervention effect could be null or in the opposite direction to the summary effect estimate. I would like to know and I think it would be useful for the reader to know in how many of these MA's was there one or more individual studies that reported estimates in the opposite direction to the summary effect. Hence, in part validating the prediction interval normality assumption. I may be wrong, but I think that it is a very real possibility for prediction intervals to overlap regions of no effect but yet there are no actual estimates in this region especially if the summary estimate is small and tau2 is large. In the example given (Figure1) the assumption looks appropriate because one study (Jorissen) has a study effect in the opposite direction (being positive) but I would like to know in how many cases this does happen.

Answer: Thank you for your thoughtful advice. We have added the following text to the results section: "Of the 347 meta-analyses with a prediction interval that contained the null or opposite effect, 199 (57.3%) had also at least one study with an opposite effect. This happened more often in meta-analyses with more than six studies (181/235, 77.0%) than in those based on at most six studies (18/102, 17.6%). Especially in meta-analyses with few studies and substantial heterogeneity, the prediction interval was wider than the range of study outcomes. The opposite (i.e. a smaller prediction interval) occurred in meta-analyses based on many studies and with low estimated heterogeneity. Results for meta-analyses with dichotomous and continuous outcomes were not notably different."

VERSION 3 - REVIEW

REVIEWER	Mariska MG Leeflang Academic Medical Center, Amsterdam The Netherlands
REVIEW RETURNED	14-Mar-2016

GENERAL COMMENTS	I have no further comments.

REVIEWER	Felix Achana University of Warwick
REVIEW RETURNED	09-Apr-2016

GENERAL COMMENTS	The authors have addressed or clarified all issues I raised in my initial review of this manuscript. I have no further comments.

REVIEWER	Jason Oke University of Oxford, United Kingdom
REVIEW RETURNED	22-Mar-2016

GENERAL COMMENTS	No further comment. Thank you for addressing my questions and concerns.