PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Reporting and interpretation of results from clinical trials that did not claim a treatment difference; survey of four general medical journals
AUTHORS	Gates, Simon; Ealing, Elizabeth

VERSION 1 - REVIEW

REVIEWER	Michèle B. Nuijten Tilburg University, The Netherlands
REVIEW RETURNED	22-Aug-2018

GENERAL COMMENTS	The authors investigated how non-significant results from RCTs in 4 major medical journals were interpreted. This is an important question, and I think this is an interesting study. However, I do think this manuscript could benefit from rewriting certain paragraphs to increase clarity (and in some cases accuracy). Furthermore, I wonder about the validity and reliability of the coding of the interpretations of non-significant results. Specifically, I miss information about the interrater reliability, information about how these ten categories were determined, and it is not always clear to me how the authors would have classified certain interpretations. Finally, the data of this study are not fully available, and I see no reason why they can't be. I list my remarks below, in "chronological" order. Signed, Michèle Nuijten Introduction • The introduction needs some rewriting in order to get the main point across more clearly, and to make sure that all statements about power and significance are 100% accurate. Specifically: o P. 2, line 35-40: "In a trial … non-significance." This sentence is very long and difficult to follow. I would cut this up in 2-3 sentences. In this same sentence, be careful to give the accurate explanations yourself: if power is 80%, you'd expect a non-significant result in 20% of the cases not only if the treatment effect is as large *or larger* than the expected effect. Also, lower power gives a higher probability of a non-significant result *but not if there is no population effect*, in which case you'd expect a non-significant result in 95% of the cases.

	o Since this is such an important point, maybe consider giving a list of reasons why a non-significant result could arise (truly no effect, true effect smaller than estimated, not enough participants, noisy measurements). Method • The most important measure in this study is the categorization of how the results were interpreted. Because of this, I would really like to see the interrater reliability, and an explanation of the most common reasons why coders disagreed. • How did the classifications in Box 1 arise? Were they created a priori or did they arise naturally during coding? • It is important to code whether authors concluded that there *was* no difference, or that they * didn't find/observe* a difference. Right now it's not clear to me if this distinction was made in coding the results, but this is a very important distinction. Concluding that there *is* no difference is not justified by a non-significant result, whereas concluding that *finding* no difference, is justified. Results • I would like to see a bit more structure in the presentation of the results. A table would be very insightful to see in one glance what type of conclusions were drawn. Of course the Figures would also be in the results section, but I think it would be useful to have the exact percentages listed in a table somewhere. You could for instance add two columns to Box 1 and add the percentages of each type in the results section & in the conclusion. • In the Figures, I would consider using different colors or patterns or paneling, to indicate which of the interpretations were correct and which weren't. This explicit distinction would also be useful in Box 1. The authors could even consider making this an ordinal scale, by labeling interpretations as more or less wrong (if the authors think that that makes sense) Discussion • P. 5, line 21-39: this seems to be an argument against dichotomous decision making, but in the context of clinical trials, a dichotomization does seem necessary at some point. Maybe authors of the RCTs can make clear that scientific results are always probabilistic, and need to be interpreted with caution, but in the end a recommendation needs to be made: should this treatment be implemented or not? Simply stating that dichotomization is too simplistic is not sufficient, I think. There is also a difference between results meant to (further) develop a scientific theory, or results meant to directly inform practice. In the first case, dichotomizations aren't immediately necessary, in the second case I'd argue that they are. • In the discussion, I miss the authors view on *why* these errors seem to occur so often. Is this because authors don't understand statistics? Is it because they want to draw a clear conclusion? It is important to think about this in order to make useful recommendations to solve this problem. I realize that this study cannot answer these questions, but I would be interested in the authors' view on this. • Another limitation of this study that is not mentioned in the limitations section, is that only the abstracts were analyzed. I sympathize with this choice, but it is important to consider the possibility that authors did interpret their results correctly in the main text. If this is the case, it is likely that authors do understand

	how to interpret results, but might be tempted to draw more clear-cut conclusions in their abstract. General • Please publish the (thoroughly documented) data in an online repository (e.g., https://osf.io). There is overwhelming evidence that "data are available upon request" statements are often void, and data disappears over the years. Minor remarks • P. 2, line 42: "trials are regarded as the highest standard of evidence", maybe this is nitpicking, but wouldn't that be a meta-analysis of trials? • P. 3 line 23-27: I have no idea what these types of trials are (non-inferiority, single/multiple armed, etc.). My background is in psychology, and I'm not familiar with these terms. Maybe I'm not the target audience for this paper, but this might be something to keep in mind. • P. 4, line 28: "The majority of …". This does not seem completely correct. I would rewrite to: "the majority of trials that found non-significant results, concluded that…". • P. 4, line 46- p. 5 line 3: I'm not sure what this example illustrates. In both cases, the trials did not show a significant effect/difference, but the authors seem to want to make the point that these results do point into a certain direction (or did I misunderstand?). Please clarify. • P. 5, line 10: "quantified" instead of "qualified"?

REVIEWER	Rob Scholten Dutch Cochrane Centre | Julius Center for Health Sciences and Primary Care University Medical Center Utrecht The Netherlands
REVIEW RETURNED	30-Oct-2018

GENERAL COMMENTS	A nicely written report of the way authors of RCTs report their results and formulate their conclusions in case of results that are not statistically significant. Simple methods were used and I have only minor suggestions. Title and Abstract 1. In the methods the authors state that they selected RCTs that had a non-significant treatment effect for the primary outcome. Please, mention also here and in the title. 2. You selected RCTs, so please use the label 'RCTs' instead of 'clinical trials' throughout. Methods 3. What was decided in case of one RCT with multiple primary outcomes of which only one was not-significant? Can you clarify, please. Results 4. Denominator: papers (RCTs) or comparisons: there were 85 reports including 111 comparisons. The authors here and there provide results of both (which to me is very useful, because of the possible dependency of the results regarding comparisons).

	Please, report results with both RCTs and comparisons as denominator throughout. 5. If no CIs or significance tests were reported, what was reported instead? How could the authors conclude about the non-significance of the results? 6. Would it be helpful to make a Table of Box 1 and add the numbers (#/RCTs and #/comparisons) to the 10 categories? Discussion 7. You introduce three 'possibilities'. It might be helpful for the reader if you could present an example of each directly after having introduced those (which you implicitly do later in the text for some of the possibilities). Could also be done in a box. 8. How can we do better? From reference 4, I have learned to state 'A difference in effect of X versus Y could be demonstrated nor refuted'. In Cochrane we are not allowed to use terms like 'not significant', so at least Cochrane tries to educate their authors. This may be considered as at least a small step forwards in educating authors. 9. I started working in clinical epidemiology 30 years ago and I was immediately introduced in the controversy between frequentists and Bayesians. After 30 years, they still quarrel and we still mainly use the frequentist approach (even 100% in this sample). So, we could also blame the Bayesians for haven't done a very good job ;-) You cite Frank Harrell's blog. Maybe you could provide some examples of Bayesian statements for the various (non-significant frequentist) results that you presented before (maybe also in the box of item 7). Box 1 10. This one needs some attention regarding the punctuation. See also item 6: you may wish to add the numbers (#/RCTs and #/comparisons) to the 10 categories.

REVIEWER	Prof. Willem J.J. Assendelft Radboud University Medical Center, Nijmegen, the Netherlands
REVIEW RETURNED	09-Nov-2018

GENERAL COMMENTS	General It is a nice project to study the current reporting on non-significance in 4 major journals. I think that the paper is not suited for a general readership, simply because the underlying knowledge on statistical testing and interpretation still lacks, as illustrated by the results of the project. Now more or less didactical information has to be spread over the Introduction and Discussion, but because of the format of an empirical paper this is too hard to read and understand for the reader (how to deal with uncertainty, advantages of Bayesian statistics, etc.). I would recommend to submit the paper in a more specialized epidemiological or statistical journal, or give it another format, more didactical: explain the basics, and present the results, including examples of how to report properly and how mistakes can be recognized. Some comments in more detail. Abstract

	The special nature of the paper doesn't make it fit for a structured abstract with the standard headings. For example, "Design: retrospective survey" is not appropriate for a review of journal articles, "Participants: Reports of journal articles...", "Intervention: no intervention", etc. I think that most readers will still don't know if the results in the abstract mean that something worrisome is going on or not. Do they know if it is okay if a paper states "there was no treatment benefit", "no significant benefit" or "no significant difference". Provide more guidance in the interpretation of the results, also in the abstract. Methods I miss details on the Methods. Were the assessors blinded for authors and journals?; Was the assessment in duplicate? What if no consensus could be reached?; How many papers were considered, which were excluded on which grounds (a flow diagram similar to QUOROM); Level of initial agreement between assessors in % and Kappa. Results Results are difficult to interpret for the non-initiated reader: what results are 'good' and which are 'bad'? I propose a more didactical approach, where Box 1 is integrated with Figures 1 and 2 (because the words in Figures 1 and 2 are difficult the interpret; a more extensive wording is necessary for the interpretation). In an integrated table (in landscape?) you can also add an example of 'bad' presentation and a recommendation on how it should/could have been reported. Discussion The points 1 - 3 under Strengths and Limitations (presented at the beginning of the paper) are not discussed here in sufficient detail. The discussion goes too far in detail for the statistically uninitiated reader. Restrict the discussion and recommendations to the material itself. Use of different statistics (Bayesian) is another discussion, I think. Now there are two messages: reporting is wrong and the statistics are wrong, while the research question was only on reporting. In addition, the Discussion should use more examples, because the readers will most likely not recognize what is 'wrong' and what the proper way of presenting should be.

VERSION 1 – AUTHOR RESPONSE

Reviewers' comments

We thank the reviewers for their detailed and insightful comments, and we have made extensive changes to the manuscript in response to them. Some of the issues that were raised resulted from an earlier submission to a journal with a strict and small word limit, so we have taken the opportunity to expand some points and provide further explanation.

Responses to the Reviewers' specific points are detailed below:

Reviewer: 1

P. 2, line 35-40: “In a trial … non-significance.” This sentence is very long and difficult to follow. I would cut this up in 2-3 sentences. In this same sentence, be careful to give the accurate explanations yourself: if power is 80%, you’d expect a non-significant result in 20% of the cases not only if the treatment effect is as large *or larger* than the expected effect. Also, lower power gives a higher probability of a non-significant result *but not if there is no population effect*, in which case you’d expect a non-significant result in 95% of the cases.

We agree that this sentence is hard to understand, so we have re-written it and split it into several sentences, as suggested, which hopefully make the points more clearly.

We do not fully understand the reviewer’s point about 80% power. If a trial has 80% power for a specified effect size, then there will be a 20% probability of non-significance in a hypothetical long run of trials with that exact true effect size. If the true effect size is larger, power would be higher. We think what we have written is correct, we are happy to make further changes if we have misunderstood the point here.

Since this is such an important point, maybe consider giving a list of reasons why a non-significant result could arise (truly no effect, true effect smaller than estimated, not enough participants, noisy measurements).

This is a helpful addition: we have added some text to include these points.

Method

The most important measure in this study is the categorization of how the results were interpreted. Because of this, I would really like to see the interrater reliability, and an explanation of the most common reasons why coders disagreed.

We did not do this. It was not possible in the time available for the project, which was done during a time-limited student placement. However, although it would be desirable, it does not seem crucial for our main points (and indeed earlier similar studies did not include this). We are not proposing this as a general tool for classifying results, that could be used in other studies, but simply as a useful classification to make the points that we wanted to make in this study.

How did the classifications in Box 1 arise? Were they created a priori or did they arise naturally during coding?

We created the classification before coding by reviewing other papers that were from outside the study's time window, and added extra categories for trials that we found during the study that did not fit any of the predefined categories.

It is important to code whether authors concluded that there *was* no difference, or that they * didn't find/observe* a difference. Right now it's not clear to me if this distinction was made in coding the results, but this is a very important distinction. Concluding that there *is* no difference is not justified by a non-significant result, whereas concluding that *finding* no difference, is justified.

This is a really important point, and the reviewer is completely correct that there is a critical distinction between concluding that that there was no difference and concluding that a difference was not found (but might still exist). The problem is that with almost all of the forms of wording used, it is not clear what the authors believed, and equally unclear how what they said would be interpreted by readers. For example, if a result is described thus: "X did not significantly improve outcome Y," we are not sure whether the authors believed that they had found that there was no difference, or that they meant that the evidence was not sufficient to make a conclusion that X improved Y. Similarly, we do not know how this would be interpreted by readers.

Our aim, therefore, was to record the ways in which results were reported, and not to extrapolate the authors' meaning from what they wrote. A key point is that most forms of words are ambiguous and do not clearly convey accurate information about the meaning of results. Very few studies used a form of words that referred to the evidence for a difference. We have added text to the Discussion to amplify these points, which we agree should be included.

Results

I would like to see a bit more structure in the presentation of the results. A table would be very insightful to see in one glance what type of conclusions were drawn. Of course the Figures would also be in the results section, but I think it would be useful to have the exact percentages listed in a table somewhere. You could for instance add two columns to Box 1 and add the percentages of each type in the results section & in the conclusion.

WE have added the number in each category to the bars in the figure, and also converted Box 1 to a Table with the requested information.

In the Figures, I would consider using different colors or patterns or paneling, to indicate which of the interpretations were correct and which weren't. This explicit distinction would also be useful in Box 1. The authors could even consider making this an ordinal scale, by labeling interpretations as more or less wrong (if the authors think that that makes sense)

We have added text to the Discussion about the limitations of the various interpretations. We are reluctant to label any interpretations as “right” or “wrong” because all can be criticised, and many of them have substantial ambiguity and can be interpreted in various ways.

Discussion

P. 5, line 21-39: this seems to be an argument against dichotomous decision making, but in the context of clinical trials, a dichotomization does seem necessary at some point. Maybe authors of the RCTs can make clear that scientific results are always probabilistic, and need to be interpreted with caution, but in the end a recommendation needs to be made: should this treatment be implemented or not? Simply stating that dichotomization is too simplistic is not sufficient, I think. There is also a difference between results meant to (further) develop a scientific theory, or results meant to directly inform practice. In the first case, dichotomizations aren’t immediately necessary, in the second case I’d argue that they are.

It is true that eventually, a dichotomisation of trial results is necessary, to make a recommendation that a treatment should or should not be used in clinical practice. However, the analysis of a trial’s primary outcome is not usually the appropriate place to do this. Decisions about clinical practice should be based on all of the relevant information, which includes effects on other outcomes, all clinical benefits and harms, and costs. We have added some text to the discussion to clarify these points.

- In the discussion, I miss the authors view on *why* these errors seem to occur so often. Is this because authors don’t understand statistics? Is it because they want to draw a clear conclusion? It is important to think about this in order to make useful recommendations to solve this problem. I realize that this study cannot answer these questions, but I would be interested in the authors’ view on this.

We have added some possible causes to the discussion. These are speculative, but seem plausible.

- Another limitation of this study that is not mentioned in the limitations section, is that only the abstracts were analyzed. I sympathize with this choice, but it is important to consider the possibility that authors did interpret their results correctly in the main text. If this is the case, it is likely that authors do understand how to interpret results, but might be tempted to draw more clear-cut conclusions in their abstract.

This is a reasonable point. We did initially try to extract information from the main text of papers, but shifted to concentrate on abstracts because extraction from the main text was much more time consuming and difficult because descriptions of results often occurred in several places and in different forms. We have added this as a “limitation” of the study.

General

- Please publish the (thoroughly documented) data in an online repository (e.g., https://osf.io). There is overwhelming evidence that “data are available upon request” statements are often void, and data disappears over the years.

This is an excellent point that we wholeheartedly agree with. The data and explanatory documenthave been put on OSF.

Minor remarks

- P. 2, line 42: “trials are regarded as the highest standard of evidence”, maybe this is nitpicking, but wouldn’t that be a meta-analysis of trials?

Some researchers would strongly agree with this and some strongly disagree! We have changed it to the highest standard of evidence from primary studies.

- P. 3 line 23-27: I have no idea what these types of trials are (non-inferiority, single/multiple armed, etc.). My background is in psychology, and I’m not familiar with these terms. Maybe I’m not the target audience for this paper, but this might be something to keep in mind.

Our belief is that the target audience will be familiar with the terms, but we are happy to insert an explanation if that is helpful (we have not yet done so).

- P. 4, line 28: “The majority of …”. This does not seem completely correct. I would rewrite to: “the majority of trials that found non-significant results, concluded that…”.

We only included trials with non-significant results, so the sentence seems correct.

- P. 4, line 46- p. 5 line 3: I’m not sure what this example illustrates. In both cases, the trials did not show a significant effect/difference, but the authors seem to want to make the point that these results do point into a certain direction (or did I misunderstand?). Please clarify.

We have added more text to expand these examples, which hopefully provides clarity. The two examples make different points. One has a wide confidence interval, so the data are consistent with effects strongly in either direction. In the other, the trial concluded “no improvement” when the 95% confidence interval only just overlapped zero effect. In this case, a different (Bayesian) analysis would conclude that there was a high probability of an effect in one direction, which is completely lost with the standard interpretation.

- P. 5, line 10: “quantified” instead of “qualified”?

No, qualified is correct.

Reviewer: 2

Title and Abstract

1. In the methods the authors state that they selected RCTs that had a non-significant treatment effect for the primary outcome. Please, mention also here and in the title.

We have added this.

2. You selected RCTs, so please use the label ‘RCTs’ instead of ‘clinical trials’ throughout.

We have changed this throughout.

Methods

3. What was decided in case of one RCT with multiple primary outcomes of which only one was not-significant? Can you clarify, please.

Such trials were excluded. “We included multiple-armed trials, and trials with multiple primary comparisons, if no treatment difference was claimed for any of them.” We have tried to make it explicit that if there were several primary comparisons and a treatment difference was claimed for only one of them, the trial was excluded.

Results

4. Denominator: papers (RCTs) or comparisons: there were 85 reports including 111 comparisons. The authors here and there provide results of both (which to me is very useful, because of the possible dependency of the results regarding comparisons). Please, report results with both RCTs and comparisons as denominator throughout.

We have included this information in Table 1.

5. If no CIs or significance tests were reported, what was reported instead? How could the authors conclude about the non-significance of the results?

All of the treatment comparisons except two presented either p-values, or confidence intervals, or both, which were the basis for their claim of non-significance of the treatment effect.

For the comparisons that presented neither confidence intervals nor p-values: one was the fourth co-primary outcome of a trial, for which no events were observed and hence no statistics were presented. The other case was the second co-primary outcome of its trial, which was simply stated as “non-significant”, without any statistics in the abstract, though the p-value was given in the main text of the paper.

6. Would it be helpful to make a Table of Box 1 and add the numbers (#/RCTs and #/comparisons) to the 10 categories?

We have changed this to a table, as suggested.

Discussion

6. You introduce three ‘possibilities’. It might be helpful for the reader if you could present an example of each directly after having introduced those (which you implicitly do later in the text for some of the possibilities). Could also be done in a box.

We have substantially re-written the discussion, so hopefully these points have now been addressed.

7. I started working in clinical epidemiology 30 years ago and I was immediately introduced in the controversy between frequentists and Bayesians. After 30 years, they still quarrel and we still mainly use the frequentist approach (even 100% in this sample). So, we could also blame the Bayesians for haven’t done a very good job ;-) You cite Frank Harrell’s blog. Maybe you could provide some examples of Bayesian statements for the various (non-significant frequentist) results that you presented before (maybe also in the box of item 7).

This is a good idea; we have added a reference to the Harrell blog, where examples of Bayesian statements can be found.

Box 1

8. This one needs some attention regarding the punctuation. See also item 6: you may wish to add the numbers (#/RCTs and #/comparisons) to the 10 categories.

We have corrected this in Table 1.

Reviewer: 3

I think that the paper is not suited for a general readership, simply because the underlying knowledge on statistical testing and interpretation still lacks, as illustrated by the results of the project.

Now more or less didactical information has to be spread over the Introduction and Discussion, but because of the format of an empirical paper this is too hard to read and understand for the reader (how to deal with uncertainty, advantages of Bayesian statistics, etc.).

I would recommend to submit the paper in a more specialized epidemiological or statistical journal, or give it another format, more didactical: explain the basics, and present the results, including examples of how to report properly and how mistakes can be recognized.

We disagree with the reviewer on this point. We think that good interpretation of statistical results is a key element of evidence based medicine, because poor interpretation may undo much of the good that is done by rigorous experimental design and conduct. Hence discussion of these issues in publications with general readership is helpful. We think that methodological discussion should not be confined solely to methodological journals, as it may then not be seen by the people who would benefit from reading these studies.

Some comments in more detail.

Abstract

The special nature of the paper doesn't make it fit for a structured abstract with the standard headings. For example, "Design: retrospective survey" is not appropriate for a review of journal articles, "Participants: Reports of journal articles...", "Intervention: no intervention", etc.

I think that most readers will still don't know if the results in the abstract mean that something worrisome is going on or not. Do they know if it is okay if a paper states "there was no treatment benefit", "no significant benefit" or "no significant difference". Provide more guidance in the interpretation of the results, also in the abstract.

This is a good point (also made by another reviewer); we did not provide any guidance as to what are “correct” and “incorrect” interpretations. As noted above, providing guidance is difficult, because it is not easy to come up with forms of words that both describe frequentist results accurately, and convey useful information. We have referred to a recent Datamethods discussion and given some examples from there, and also referred to Frank Harrell’s blog for examples of description of Bayesian results. We have made some changes to the Abstract and the Discussion to try to provide guidance as to what are justified and unjustified ways of describing the results.

Methods

I miss details on the Methods. Were the assessors blinded for authors and journals?; Was the assessment in duplicate? What if no consensus could be reached?; How many papers were considered, which were excluded on which grounds (a flow diagram similar to QUOROM); Level of initial agreement between assessors in % and Kappa.

This was not a systematic review, and we make no claims of comprehensiveness or generalisability. Our aim is simply to use this non-random sample as an illustrative example of current practice.

Results

Results are difficult to interpret for the non-initiated reader: what results are 'good' and which are 'bad'? I propose a more didactical approach, where Box 1 is integrated with Figures 1 and 2 (because the words in Figures 1 and 2 are difficult the interpret; a more extensive wording is necessary for the interpretation). In an integrated table (in landscape?) you can also add an example of 'bad' presentation and a recommendation on how it should/could have been reported.

We have added text to the Introduction and Discussion to try to address these points. Part of our argument is that it is very difficult to find ways of presenting frequentist statistical results that are both accurate and impart useful information to the reader, so it is difficult to recommend any of the methods that were used. We have added text to the discussion to amplify these points.

Discussion

The points 1 - 3 under Strengths and Limitations (presented at the beginning of the paper) are not discussed here in sufficient detail.

The discussion goes too far in detail for the statistically uninitiated reader. Restrict the discussion and recommendations to the material itself. Use of different statistics (Bayesian) is another discussion, I think. Now there are two messages: reporting is wrong and the statistics are wrong, while the research question was only on reporting.

In addition, the Discussion should use more examples, because the readers will most likely not recognize what is 'wrong' and what the proper way of presenting should be.

We agree that it is challenging to present the arguments in a way that is easy to understand, but we have made our best attempt. Our view is that many of the problems of reporting RCT results derive from the use of null hypothesis significance testing, because of the difficulties of understanding what these statistics mean, and translating that into words that convey the results accurately. We think wider use of Bayesian methods would help people to interpret results better, by avoiding the problems related to testing and dichotomisation, and instead making direct probability statements (e.g. “there was 86% probability that the treatment effect was in the beneficial direction”). We therefore think this

is an important argument to include. We have added more examples to the Discussion, which will hopefully make these points more clearly.

VERSION 2 – REVIEW

REVIEWER	Michèle Nuijten Tilburg University, The Netherlands
REVIEW RETURNED	04-Feb-2019

GENERAL COMMENTS	I thank the authors for their detailed replies to my previous comments. In my first review I had a question about the interpretation of 80% power, but after seeing the authors' explanation, I think I was confused, my apologies. I also thank the authors for posting their data online. Even though the authors have made some important improvements, I still see two major issues with this manuscript. First, I think that the conclusion that erroneous interpretations of non-significant results are widespread, is not supported by the data. Second, I think the proposed solutions are too one-sided and not always clear. There are several reasons why I think the data do not support the conclusions. The first reason is that the authors are reluctant to sort their categories in right and wrong interpretations. This lack of categorization makes it hard to conclude something about the prevalence of wrong interpretations. In the manuscript and in their reply, the authors indicate it was hard to classify results, because it was often unclear what the authors meant in their interpretation of the results. I think this is problematic: If it can't be determined whether a claim is actually erroneous, the research question can't be answered. Since the whole paper seems to argue that authors do not report non-significant results correctly (this conclusion is quite strongly stated, and repeated throughout the manuscript), I think the authors have to make explicit which categories were correct and which were incorrect interpretations. If this can't be done, a different conclusion needs to be drawn. The second reason why I think the conclusions are not supported by the data, is illustrated by the following statement from the manuscript: "It is likely that most readers would interpret significance in a way similar to the common English meaning of the word, and would take away from a "non-significant" result the impression that there was no important difference between the interventions." The authors often seem to be worried that *readers* may not interpret conclusions correctly, even though the reported conclusion may not be factually erroneous. For instance, the

	question of how readers interpret the word “significance” is a different question entirely. There is a crucial difference between stating “there was no difference” (wrong) and “there was no significant difference” (factually right). The latter case may or may not be interpreted correctly by the reader, but it is too strongly stated that the authors drew the wrong conclusion. My second issue concerns the proposed solutions. Right now, the main conclusion I draw from the presented findings is that we need to improve language for describing results (as the authors also mention in their manuscript). However, I do have some questions about the examples they provide. From page 9: Example 1: “We were unable to find evidence against the hypothesis that A=B (p=0.4) with the current sample size. More data will be needed. As the statistical analysis plan specified a frequentist approach, the study did not provide evidence of similarity of A and B.” Are the authors suggesting that when a non-significant effect is found, one should recommend collecting more data? This seems strange: at which point do we abandon an apparent fruitless research line? Example 2: “Assuming the study’s experimental design and sampling scheme, the probability is 0.4 that another study would yield a test statistic for comparing two means that is more impressive that what we observed in our study, if treatment B had exactly the same true mean SBP as treatment A.” Even though this example is factually correct, it seems very difficult to interpret. It doesn’t read as an interpretation, but more as the definition of a p-value, and I’m not sure how useful it would be to keep repeating that throughout a paper (but that might be a personal preference). Minor additional remark: I’d switch the word “impressive” for “extreme”. Finally, the authors present Bayesian statistics as a solution, but they seem uncritical towards potential limitations of this framework. The authors argue that frequentist stats are hard to understand, but is there evidence that people are generally better in interpreting Bayesian stats? This is an important point to discuss. Signed, Michèle Nuijten Minor remarks • P.5, line 13: typo in “treatments were simila.,”

REVIEWER	Rob Scholten Cochrane Netherlands & Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, The Netherlands
REVIEW RETURNED	27-Jan-2019

GENERAL COMMENTS	I like the revised version very much, which to me is very readable and easy to follow. I will certainly direct medical students (who can't be blamed for misinterpretation of non-significant results, because in many instances, they will receive incorrect examples from their supervisors in the clinic) to this paper. I have only some very minor comments, of which the authors should feel free to ignore those. Title 1. Add 'non-significant' to 'results' in the title? Introduction 2. I may be wrong, but the quote of reference 3 ("absence of evidence is not evidence of absence" [of effect]) doesn't seem to underpin incorrect interpretation of non-significant results. I thought that this applied to situations in which no studies were identified ("no evidence") for a particular clinical question, which was then interpreted as "the intervention is not effective". I may be wrong, though. Results 3. Third paragraph: "The main alternative approach was to re-state the lack of a statistically significant difference (13/111; 11.7%) or statistically significant benefit (14/111; 12.6%).". Add 'lack of' to 'statistically significant benefit'? Discussion 4. Statistical methods (line 28): I'm not sure whether I understand the phrase '… that many trials seek to determine clinical practice,'. Rephrase in 'to inform clin practice'? Some typos 5. There are some typos in the document ('simila.'", 'was not benefit', 'Weshould', 'giving, the probability', 'Despite many years of warnings, inappropriate interpretation of RCT results are widespread'(, but the copy-editor will pick these up, I assume.

REVIEWER	Willem J.J. Assendelft Radboud University Medical Center, the Netherlands
REVIEW RETURNED	24-Jan-2019

GENERAL COMMENTS	In general, the changes made the paper more accesible for a general readership. As I previously commented the special nature of the paper doesn't make it fit for a structured abstract with the standard headings. For example, "Design: retrospective survey" is not appropriate for a review of journal articles, "Participants: Reports of journal articles...", "Intervention: no intervention", etc I think the methods could / should have been better designed and executed. Previously I commented: "Were the assessors blinded

	for authors and journals?; Was the assessment in duplicate? What if no consensus could be reached?; How many papers were considered, which were excluded on which grounds (a flow diagram similar to QUOROM); Level of initial agreement between assessors in % and Kappa." And for me as a researcher the answer from the authors ("This was not a systematic review, and we make no claims of comprehensiveness or generalisability. Our aim is simply to use this non-random sample as an illustrative example of current practice.") is unsatisfactory. And I think these limitations are only addressed partially, and need more explanation under Limitations.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 3

Reviewer Name: Willem J.J. Assendelft

As I previously commented the special nature of the paper doesn't make it fit for a structured abstract with the standard headings. For example, "Design: retrospective survey" is not appropriate for a review of journal articles, "Participants: Reports of journal articles...", "Intervention: no intervention", etc

We agree that the specified headings do not fit entirely comfortably with this study, but we have tried to make the abstract comprehensible. We have rewritten it modelled on another recent methodological publication in BMJ Open (Wong JLC, et al. BMJ Open 2018;8), which we hope will be more acceptable.

I think the methods could / should have been better designed and executed. Previously I commented: "Were the assessors blinded for authors and journals?; Was the assessment in duplicate? What if no consensus could be reached?; How many papers were considered, which were excluded on which grounds (a flow diagram similar to QUOROM); Level of initial agreement between assessors in % and Kappa." And for me as a researcher the answer from the authors ("This was not a systematic review, and we make no claims of comprehensiveness or generalisability. Our aim is simply to use this non-random sample as an illustrative example of current practice.") is unsatisfactory. And I think these limitations are only addressed partially, and need more explanation under Limitations.

We have addressed all of these points, in previous revisions and new changes to the text, and addition of a flow diagram. We have also added text to the limitations section of the Discussion.

Reviewer: 2

I like the revised version very much, which to me is very readable and easy to follow. I will certainly direct medical students (who can't be blamed for misinterpretation of non-significant results, because in many instances, they will receive incorrect examples from their supervisors in the clinic) to this paper. I have only some very minor comments, of which the authors should feel free to ignore those.

Thank you for the positive comments.

1. Add 'non-significant' to 'results' in the title?

We prefer not to use the term "non-significant" where it is possible to avoid it.

2. I may be wrong, but the quote of reference 3 ("absence of evidence is not evidence of absence" [of effect]) doesn't seem to underpin incorrect interpretation of non-significant results. I thought that this applied to situations in which no studies were identified ("no evidence") for a particular clinical question, which was then interpreted as "the intervention is not effective". I may be wrong, though.

We agree that the quotation is open to several interpretations, so we would not want to over-use it. However, the situation that Altman and Gardner were referring to in their original paper was the same situation as we have studied here; studies that had non-significant results concluding that there was no difference.

3. Third paragraph: "The main alternative approach was to re-state the lack of a statistically significant difference (13/111; 11.7%) or statistically significant benefit (14/111; 12.6%).". Add 'lack of' to 'statistically significant benefit'?

This is clearer so we have added it – thank you for the suggestion.

4. Statistical methods (line 28): I'm not sure whether I understand the phrase '… that many trials seek to determine clinical practice,'. Rephrase in 'to inform clin practice'?

We have made this change, and re-worded the text around this sentence. Hopefully it is now clearer.

5. There are some typos in the document ('simila."', 'was not benefit', 'Weshould', 'giving, the probability', 'Despite many years of warnings, inappropriate interpretation of RCT results are widespread'(, but the copy-editor will pick these up, I assume.

We have corrected all of the typos we could find – thank you for identifying these.

Reviewer: 1

In my first review I had a question about the interpretation of 80% power, but after seeing the authors' explanation, I think I was confused, my apologies. I also thank the authors for posting their data online.

Thank you for clarifying and we are happy that our explanation was helpful.

First, I think that the conclusion that erroneous interpretations of non-significant results are widespread, is not supported by the data. Second, I think the proposed solutions are too one-sided and not always clear.

There are several reasons why I think the data do not support the conclusions.

The first reason is that the authors are reluctant to sort their categories in right and wrong interpretations. This lack of categorization makes it hard to conclude something about the prevalence of wrong interpretations.

In the manuscript and in their reply, the authors indicate it was hard to classify results, because it was often unclear what the authors meant in their interpretation of the results. I think this is problematic: If it can't be determined whether a claim is actually erroneous, the research question can't be answered. Since the whole paper seems to argue that authors do not report non-significant results correctly (this conclusion is quite strongly stated, and repeated throughout the manuscript), I think the authors have to make explicit which categories were correct and which were incorrect interpretations. If this can't be done, a different conclusion needs to be drawn.

Thank you for these comments, which are really valuable in helping us to make our messages clearer. We tried, in the text that we added to the Discussion, to elaborate the reasons why our view is that the commonest ways of describing results are misleading. The main issue is the incorrect interpretation of non-significance as no improvement or no difference, which was used by over half of the papers and comparisons. We have altered this section to be more explicit about which descriptions we regard as inadequate, and why. Other ways of describing the results are also potentially open to misinterpretations, and we have tried to explain why in the Discussion.

The second reason why I think the conclusions are not supported by the data, is illustrated by the following statement from the manuscript:

"It is likely that most readers would interpret significance in a way similar to the common English meaning of the word, and would take away from a "non-significant" result the impression that there was no important difference between the interventions."

The authors often seem to be worried that *readers* may not interpret conclusions correctly, even though the reported conclusion may not be factually erroneous. For instance, the question of how readers interpret the word "significance" is a different question entirely. There is a crucial difference between stating "there was no difference" (wrong) and "there was no significant difference" (factually right). The latter case may or may not be interpreted correctly by the reader, but it is too strongly stated that the authors drew the wrong conclusion.

This point is really about the use of the term "significant" and we have re-written the relevant sections to try to explain more clearly why we feel this is problematic. We agree that it is technically correct (if one is operating within a frequentist null-hypothesis testing framework, which is itself problematic), but our view is that making statements that are likely to be misinterpreted is not a good way to draw conclusions.

My second issue concerns the proposed solutions.

Right now, the main conclusion I draw from the presented findings is that we need to improve language for describing results (as the authors also mention in their manuscript). However, I do have some questions about the examples they provide.

From page 9:

Example 1: "We were unable to find evidence against the hypothesis that A=B (p=0.4) with the current sample size. More data will be needed. As the statistical analysis plan specified a frequentist approach, the study did not provide evidence of similarity of A and B."

Are the authors suggesting that when a non-significant effect is found, one should recommend collecting more data? This seems strange: at which point do we abandon an apparent fruitless research line?

Here we are quoting Frank Harrell; this is his attempt at a valid wording for interpreting non-significance frequentist results (the suggestion of collecting more data in response to a non-significant result actually comes directly from R.A.Fisher). We have moved these suggestions from the text to a table, to make it clearer that they are quotations of someone else's suggestions.

Example 2: "Assuming the study's experimental design and sampling scheme, the probability is 0.4 that another study would yield a test statistic for comparing two means that is more impressive that what we observed in our study, if treatment B had exactly the same true mean SBP as treatment A."

Even though this example is factually correct, it seems very difficult to interpret. It doesn't read as an interpretation, but more as the definition of a p-value, and I'm not sure how useful it would be to keep repeating that throughout a paper (but that might be a personal preference). Minor additional remark: I'd switch the word "impressive" for "extreme".

Again, this is a direct quotation from Frank Harrell. It is another suggestion for accurate wording of what a "non-significant" result actually means. We agree that such statement are hard to interpret, and we have added some text to this effect. A large part of the general problem with frequentist statistics is that their results are difficult to communicate in a way that is both accurate and understandable.

Finally, the authors present Bayesian statistics as a solution, but they seem uncritical towards potential limitations of this framework. The authors argue that frequentist stats are hard to understand, but is there evidence that people are generally better in interpreting Bayesian stats? This is an important point to discuss.

This is just our own preferred solution, which seems to us (and others) to have some merit. It seems generally agreed that people find Bayesian results easier to interpret, because (a) Bayesian results deal directly with quantities that people can understand and interpret, such as probabilities that one treatment is better than the other and (b) some of the common misinterpretations of frequentist results (particularly confidence intervals) are to interpret them in a Bayesian way (e.g. a range of values with 95% probability of including the true value). However, we are not aware of any studies that have set out to compare understanding of Bayesian and frequentist results. We have changed the text to try to make these points more clearly.

VERSION 3 - REVIEW

REVIEWER	Michèle Nuijten Tilburg University, The Netherlands
REVIEW RETURNED	17-Jul-2019

GENERAL COMMENTS	I thank the authors for incorporating my comments. I look forward to seeing this paper in print.

REVIEWER	Assendelft Radboud University Medical Center, Nijmegen, The Netherlands
REVIEW RETURNED	13-Jun-2019

GENERAL COMMENTS	My comments were properly addressed.