Author's response to reviews

Title: The effect of journal impact factor, reporting conflicts, and reporting funding sources, on standardized effect sizes in back pain trials: a systematic review and meta-regression

Authors:

Robert Froud (r.froud@warwick.ac.uk)
Tom Bjørkli (tomeirikkjeldsen87@yahoo.com)
Philip Bright (philbright@eso.ac.uk)
Dévan Rajendran (devanrajendran@eso.ac.uk)
Rachelle Buchbinder (rachelle.buchbinder@monash.edu)
Martin Underwood (m.underwood@warwick.ac.uk)
David Evans (dwe@backpainclinic.co.uk)
Sandra Eldridge (s.eldridge@qmul.ac.uk)

Version: 3
Date: 2 October 2015

Author's response to reviews: see over
Dear Ms Ma. Luz De Guzman,

RE: Response to reviewer comments – MS:1881226202172370. The effect of journal impact factor, reporting conflicts, and reporting funding sources, on standardized effect sizes in back pain trials: a systematic review and meta-regression

Thank you for your invitation for a revision and for providing reviewer comments. In relation to your editorial requests, we have formatted the manuscript using the latex BMC template review style and added line numbering. I can confirm that a statistician has assessed the statistics; two of the authors are statisticians: RF, and SE, a professor of biostatistics at Queen Mary University of London. In addition, we consulted two statisticians for their views on aspects of our approach: Simon Gates, professor of biostatistics at University of Warwick, and Gary Abel, a senior statistician at University of Cambridge. Professor Gates and Dr Abel are thanked in our acknowledgements section.

In addition to responding to reviewer comments, we have made changes for readability, altered keywords, and corrected a typo in Table 2.

Yours sincerely,

Robert Froud

Response to reviewer comments

Reviewer 1

1. Pg 2, 2nd paragraph; I suggest clarification, or deletion of the sentence reporting that OR for improvement from LBP interventions is 2.4, without explaining what the comparator is, the statement is uninterpretable.

The comparator was 'not improving'. However, on reflection our statement about improvement was confusing as Bekelman's comparison was at aggregate level whereas ours was at an individual level. It is not correct to compare aggregate and individual results and therefore we have removed the sentence as requested.

2. I found the justification for exploring the association of journal IF with effect size somewhat weak, compounded by the fact that the hypothesis does not specify which direction the association is expected. It comes across as something of an academic exercise. Given that this is the primary aim of the study I recommend a stronger case be mounted to underpin the necessity, and utility of the analysis.

We note that there is disagreement between reviewers on the importance of this work. We thank Reviewer 1 for pointing out his concern about the justification and we have aimed to improve the explanation of the case for undertaking this work.

3. Pg 4, 1st paragraph; in reference to the methods of selecting the outcome selected to represent each study, it seems to me that both methods 3 and 4 could result in preferential selection of large or significant effects. The impact of this on the study findings will be dependent on the number of times that these methods were used, compared to the number of times methods 1 or 2 were
used. The decision to extract the largest effect size from studies with more than 2 groups could have a similar influence. I recommend that the authors consider this potential source of bias and its implication for interpretation of the findings, possibly in the Limitations section.

The first paragraph of page four deals with how we identified the primary outcome measure. Methods 3 and 4 describe, in the absence of a primary outcome measure being nominated, or a sample size calculation being based on an outcome, the method by which we selected the first outcome measure identified in the abstract (3) or in the paper (4). The reviewer also questions the way we selected the largest effect size for trials with more than two groups.

This approach has been used in several previous methodological reviews. We have now added a reference for one such review. We consider it unlikely that this approach would result in preferential selection of large effects, because the methods were applied systematically and in order. We had no a priori reason to believe that, if methods 3 or 4 needed to be used, that the outcome measure these identify would be directly associated with differing effect sizes. However, since a posteriori we considered a link to study quality in our discussion, we suppose it could be that lower quality studies are associated with both absence of definition of primary outcome measure, and larger effects (i.e. an unmeasured confounding factor). For this reason, we accept there may be some reasonable criticism on choice of outcome measure in method 3 and 4, and have added a line in the limitations on this point. We have also added in the results how many authors explicitly identified their primary outcome measure. While we recorded this, for the trials in which this was not explicitly identified, we did not record how often primary outcome identification method 2, 3, or 4 needed to be used. This has also been noted under limitations.

Regarding effect sizes, the approach of using the largest comparison was also applied systematically (i.e. the same way for each trial irrespective of its publication impact factor) and there is no reason to believe that the number of trial arms, and hence effect size selection, relates to either quality or to impact factor. We made no changes to our manuscript on this second point.

4. Overall the Results section was wordy and quite difficult to read. I suggest restructuring and simplifying so that the findings, as they relate to the aims and the primary analyses are stated clearly and unambiguously.

We agree with the reviewer. The results contained material that may have been of methodological interest to some, but the hazard was that this distracts from the core aims of the paper, and such has proved to be the case at review. We have now removed material that is unnecessary to the core aims of the paper. This includes the graphical exploration of the relationship between IF and ES (see point 6), comparative log transforms (since regression with REML is known to be adequate), and Spearman's test for monotonic relationship (since we know association is not present from the regression). We have retained information on model fit, which is relevant to point 5.

5. Pg 6, 3rd paragraph; I would guess that the distribution of IFs was not normal, does this influence the performance of the linear regression models?

It is not necessary that outcome or predictor is normally distributed in OLS linear regression - only the distribution of residuals is important. Although it is true to say that non-normal distributions of outcome or predictor variables make non-normal residual distributions more likely. In this case, where the REML estimator is used (GLS rather than OLS), we assess fit using standardized predicted random effects. We discuss the fit of our models and have made no changes in relation to
6. Pg 6, 4th paragraph; I found this very difficult to read. My impression of the Results is that there is no significant association between IF and effect size (as per the a priori hypothesis test). I would suggest that a clear statement of this is made up front, the rest appears to be interpretation of the Results based on some post-hoc mining through the analyses, possibly more at home in the Discussion section?

Yes, we can see how this could confuse - the reviewer is correct to say that there is no association between IF and ES. We were aiming to elaborate on something that we thought might be of interest to readers, but have now removed this material replacing it with a clear and simple statement about there being no significant association. This also helps regarding point 4. Pointing out hallmarks of publication bias may have been a bit tangential and we have removed this material completely rather than adding it to the discussion.

7. Pg 7, 2nd paragraph; if I understand correctly the analysis shows that studies that reported no funding reported an effect size approx. 1 SD higher than studies that did not mention funding. What about compared to studies that did report funding?

Our a priori hypotheses were to compare the categories of reporting none to nothing reported, and reporting some compared to nothing reported. While our view is that it is not correct to make unplanned post hoc comparisons and present these as results, for discussion purposes, making the comparator group 'trials that explicitly reported no funding', to compare this with 'trials that explicitly reported funding' shows strong evidence of a large effect of reporting funding (Beta=-0.89 (95%CI -1.46 to -0.33), P=0.002). We have now included this as a post-hoc comparison in the discussion section.

8. Given the number of analyses, is there the possibility that the single significant result is due to chance?

There is always a risk that results are due to chance (i.e. type I error), and for a single test this risk is set by convention to be dismissed when it is less than or equal to 0.05. The reviewer is concerned about multiple testing and increasing chance of type I error. It is the independence of what is being tested that is the important factor when considering multiple testing and not number of tests. If variables are not sufficiently independent, such as the case where the Bonferroni adjustment is appropriately applied (e.g. in batch testing, where in each test the evaluated variables are exactly the same as those in other tests; for example, if a factory is sampling batches of light bulbs in order to test quality – what is being tested is identical in each case and the null-hypothesis is generic), then it is appropriate to reduce alpha. In this case, what is being tested is independent or at least sufficiently independent (distinctly separate null-hypotheses), that such adjustments are not indicated. We have made no changes to our manuscript on this point. Please see Perneger. BMJ 1998 316(7139): 1236-1238 for more detailed discussion (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1112991/).

9. I miss report of the Results with respect to COI.

These were in tables 2 and 3. We have now highlighted this in the text of the results section as well, setting out the results more clearly as they relate to hypotheses.
10. Pg 7, Main findings paragraph; I am not convinced that the term ‘strong evidence’ is appropriate in this case, for a two main reasons; 1) it is not clear whether this means evidence of a strong (large) effect, or strong (persuasive) evidence of an effect (of any size), 2) the authors did not specify how they were going to classify evidence with respect to its strength. In some ways this may seem like a minor issue but my concern is that is precisely the sort of word/language that is most likely to be picked up by people citing the article, and it is therefore of importance. Same goes for use of the word in the Conclusions section.

Notwithstanding hypothesis testing, the way statistical evidence is interpreted is set by convention: 0.10>p>0.05=weak evidence; 0.05>p>0.01=some evidence; 0.01>p>0.001=strong evidence; 0.001>p>0=very strong evidence. It relates only to chance (P-value) and not to effect size. To clarify we have added additional text for the interpretation of effect size, which is separate to strength of evidence. Please see statistics textbooks for further discussion; for example, p50 of Chance, B. Rossman, A. Investigating Statistical Concepts, Applications, and Methods. 2006 Thompson Higher Education. For interpretation of effect sizes, please see Cohen's original work on power and effect sizes (Cohen, C. Statistical power analysis for the behavioral sciences.1988 Lawrence Erlbaum Associates or Cohen, C. A power primer. 1992 Psychological bulletin (18)

11. Pg 8; there is considerable space devoted to study quality. This was not an aim of the study, nor was it explicitly analysed. I suggest cutting or dramatically reducing the volume of this material.

We agree with the reviewer and we have now drastically reduced the material relating to quality, stating it only as a possible explanation.

12. Pg 9, as per study quality, I find the material about efficacy and effectiveness trials peripheral to the study and think it will be confusing for the reader. I recommend reducing or deleting this section.

We have now reduced this material.

13. I wonder about the reliability of the information about conflicts of interest. Some journals demand a COI statement, others do not – presumably it is more likely that a COI will be reported if the journal demands completion of a form? Secondly, COIs may include those that relate to the study directly eg. an investigator being paid as a consultant by the company whose device/method is being tested, or unrelated eg. an investigator receiving an untied research fellowship from a government agency. Some COI forms generate a COI statement which includes all these together. This being the case it is not clear what ‘statement of a COI’ actually means in the absence of categorising whether the COI is relevant to the study being reported.

The reviewer is correct to identify that COIs declared may or may not be relevant to the study and that some journals prompt for completion whereas some do not. We did not assess the relevance of COIs to the study. We have now noted this under limitations.

14. Pg 11, Conclusions paragraph; while the recommendation that readers consider internal validity when reading a study is certainly good advice, it is not clear how it comes to be a conclusion from this study.
We agree that we did not demonstrate an association with quality. Our study shows that unfunded trials tend to report larger absolute effect sizes and it is important for readers to be aware of this. Without understanding the mechanism behind this association, it is difficult to make solid recommendations. Giving additional scrutiny to internal validity may indeed be good advice, as the reviewer states, but since this is an association that we did not demonstrate we have now softened the language considerably in relation to this suggestion in the main conclusion and removed it completely from the abstract.

15. Pg 2, 1st paragraph; is ‘contrasting paradox’ a tautology?

Yes. We have removed the word 'contrasting'.

Reviewer 2

1. Methods (Page 3, paragraph 2): It would be useful in this section to give more details on the type of trials considered eligible for this review. Perhaps, list the comparisons that were suitable for this review. It was only in the discussion that the authors mentioned two types of trials: trials with sham/placebo comparator and trials with non-sham comparator. Give examples.

Thanks – we had missed a line about the included trials, since these were essentially 'all' nsLBP trials that were not excluded by the criteria. We have now added this and reordered our methods here so that PICO now comes before the database description, which comes before details of reviewers. We do not think that specific examples are necessary, as this would add to the word count of a paper that is still quite long after the reductions in content. Pragmatic and explanatory trials are only mentioned in our discussion because of a post hoc analysis which relates to our hypothesis that this might be relevant because of the differences in effect size when comparing to sham/placebo, or an active intervention.

2. (Page 3, paragraph 2) Give the rationale for excluding non-inferiority trials.

When trials were considered non-inferiority trials? I’m not convinced that in this area researchers make clear distinction between superiority and non-inferiority trials, particularly when comparing different types of active exercise.

There are few non-inferiority trials in studies of LBP. However, effect sizes in non-inferiority trials, as a sub-population of trials, are likely to be smaller and incomparable with effect sizes in superiority trials, since the comparisons in inferiority trials are hypothesised to have non-different effects. We have now added a brief definition.

3. (Page 8, paragraph 3) Can evidence from this review be used to support the following sentence “However, LBP research trials are more commonly funded by government and charitable organizations rather than by industry”. Did the authors collect data regarding the type of funder? As more than half (54%) of the trials reported some funding, it would be useful to know who are main funders in this area.

This is not based on data collected in the study but on the experience of the study authors. The reviewer is correct to raise the point and on reflection we have changed the language used to reflect this. We did not collect data on type of funder.
4. Have the authors thought to conduct a subgroup analysis looking at pharmacological interventions versus conservative interventions, assuming that trials investigating drugs are more likely to be funded by pharmaceutical companies.

Thanks for the suggestion. We will consider this for the future. We would prefer not to present a further post hoc comparison in this paper.