Author's response to reviews

Title: Can we rely on the best trial? A comparison of individual trials and systematic reviews

Authors:

Paul P Glasziou (paul.glasziou@dphpc.ox.ac.uk)
Sasha Shepperd (Sasha.Shepperd@dphpc.ox.ac.uk)
Jon Brassey (jon@tripdatabase.com)

Version: 3 Date: 25 August 2009

Author's response to reviews: see over
Title: Can we rely on the best trial? A comparison of individual trials and systematic reviews

We have made most of the changes suggested in this second review, and feel the paper has improved further as a result. In addition to a number of changes in the text, we have redone Figure 3 to be a on a log scale (as Dr Schrier suggested), but this has taken us some time as, after some struggle we found that no standard packages could do an adequate job of this Figure, and needed to hand program this in R.

While we have addressed the majority of issues raised, we remain very reluctant to subgroup Figure 2 by sample size rather than “weight”. We had done this subgroup in response to the first review, but feel it would seriously mislead our readers to subgroup by sample size, as suggested in Dr Shrier’s 2nd point. We recognize the apparent “Catch-22” of not being able to find the “largest” trial without doing the systematic review. However, we think that using sample size only would be misleading and it is not the only option (confidence interval width and event numbers being other options). The Discussion now covers this issue much more extensively, and suggested it remains a problem for future research which we do not, and cannot, resolve in this paper.

Our apologies for the delay in doing this 2nd revision. Our detailed replies and changes are on the following pages. We hope the editors are now happy with this version.

Regards

Professor Paul Glasziou
REPLIES TO REVIEWER COMMENTS

Reviewer: Christian Gluud

Reviewer's report:
The manuscript has improved. My suggestions for minor essential revisions are:
1. p. 5: "mega-trials" and systematic reviews [7], but no studies appear to have examined a more representative sample of reviews - ought to become: 'mega-trials' and meta-analyses [7], but no studies appear to have examined a more representative sample of systematic reviews.
RESPONSE: We agree. CHANGED.

2. p.5: aimed to compare its results to that of the meta-analyses. - ought to become: aimed to compare its results to that of the meta-analyses of the main outcome measure in the systematic review.
RESPONSE: We agree. CHANGED – phrase added.

3. p.10: that would change the clinical decision). - ought to become: that would change the clinical decision) and adequately conducted.
RESPONSE: We agree. CHANGED – phrase added.

4. p. 10: and error check - ought to become: an error check.
RESPONSE: We agree. CHANGED.

5. p.10 - a reference seems to be missing next to the DOTS case - and DOTS needs to be explained.
RESPONSE: We have now more fully explained DOTS, but have preferred not to single out specific Cochrane Reviews for discussion, so have omitted the reference.

6. 'Weighted mean difference' ought to become 'mean difference' throughout in text and figures.
RESPONSE: These refer to meta-analyses that used a “weighted mean difference”, whereas the individual trial is a “mean difference” so we have CHANGED the text to read “(weighted) mean difference” to reflect both.

7. Figure 1. "Best Study" ought to become 'Best Trial', which will fit with the figure text and meaning.
RESPONSE: We agree. Wording CHANGED.
Reviewer: Ian Shrier

Reviewer's report:
The authors have revised the manuscript on this very important clinical topic. Although they have addressed many of the concerns I and the other reviewers expressed, there were some important omissions that still need to be covered. In addition, I don't remember Figure 3 and it appears I didn't comment on it before and have made some additional comments

Major Compulsory
1. In my previous comments, I suggested that the authors have to indicate whether they used fixed or random effects summary statistic. In the current version of the analysis section, the authors have still not commented on whether they used the random effects summary statistic or the fixed effects summary statistic. This is important because by definition, the fixed effects summary statistic will be closer to the most precise trial but it is the random effects summary statistic that is more often recommended. So, if they had used fixed effects and redid the analysis with random effects, they would have likely increased the number of discrepancies. The authors may still prefer to believe in the most-precise trial as many researchers and clinicians do, and not all agree that random effects is the more appropriate analysis, but it is important to be transparent about the methods and reasoning.

RESPONSE: We left the most appropriate approach (fixed or random effects, etc) to the meta-analysis to the authors of each meta-analysis, and have not redone their work (and indeed think it would not be appropriate to do so).

CHANGE. We have added a sentence to the Methods section to more clearly explain this.

2. In my previous comment, I suggested that the authors split the old Figure 1 by sample size. The authors have said that they redid the figure and split by "weight" because the number of subjects does not always reflect the weight given. I agree with the authors statement that the number of subjects doesn't reflect the weight given but that is irrelevant to their specific objective. The authors want clinicians to look for the most precise trial without doing a meta-analysis. Therefore, the plot has to show the reader what the result will be if they use the criteria they plan to use - and this can only be the number of subjects because the weight is only available once the meta-analysis is done! In addition, since all the results are presented as RR, then the study with the largest N will almost always be the most precise unless the range of incidence in the comparison groups is extremely large, so that concern is minimized as well. In addition, in seeing this figure, I realize that it would be more transparent to have 3 plots for RR and 3 for WMD using sample sizes I previously suggested (<100, 100-500, 500+) or slight variations depending on the data. Not only will this more appropriately address the authors' objective, it will also provide the reader of the article with insight into how many trials had few numbers of subjects, moderate number of subjects and large numbers of subjects (how ever the authors decide to define few, moderate and large). This will either increase or decrease the credibility of the recommendations, but either way, the change in credibility is important because it is more transparent.

RESPONSE: We think to suggest the sample size alone is the best guide would certainly be misleading, because: (i) sample size is not the best guide to the weight (confidence
interval width would be much better, or even number of outcomes events), and (ii) while sample size makes some sense *within* a group of studies, it makes little sense across sets of studies where the event rates and effect measures will be very different. Instead, we have (a) changed the term from “largest” to “most precise” trial throughout, and (b) added a more extensive Discussion of this issue and the options. We did consider splitting by confidence interval width or by number of events in the control group (a much better measure of power than N), but again the issue is within a group of studies, not across different meta-analyses. CHANGE: We have tried to explain this problem more clearly in the Results and Discussion (see 2nd last paragraph), and have changed the term to “most precise trial” rather than largest, and Discuss the remaining problem of identifying the most precise trial.

3. I don't remember figure 3 before and am having trouble understanding it. What are the lines going through some of the circles? The authors say that upper right and lower left quadrants are in agreement but this is a scatterplot and they have drawn dotted lines at 0.05. Dichotomizing the p-value this way emphasizes that decision-making should be based on p-values, and this goes against the current advice to avoid p-values and concentrate on confidence intervals. I am okay with including this type of graph but not with the emphasis on p=0.05. I believe that the difference in p-values between the two is more important because that is what would shift decision-making for most readers (one would at least hope that this is the case when confidence intervals are not being used...). So, I would much prefer to see a band around the line of identity - maybe shading an area where the p-values are within 0.05 of each other or 0.1 of each other, or maybe it should be a %difference that is highlighted. I am also thinking that the data might be better presented on a log scale because the difference between 0.02 and 0.05 is not interpreted the same as the difference between 0.50 and 0.53 (both have the same absolute difference of 0.03). In general, plotting on a log scale helps minimize these types of problems (because a doubling of the number is the same all along the scale) but it does depend on the data and the authors should consult with a statistician on the most appropriate graphical method given the data and how it is normally interpreted (we don't usually think of p-value importance in terms of "doubling" per se so maybe log scale is not the best method. For example, there is a p-value from the meta-analysis of approximately 0.4 that corresponds to a p-value from an individual study of 0.1 or less (and one at 0.7 for meta-analysis and 0.07 for individual study) - and these are considered to give the same answers (upper right quadrant). I don't think many clinicians would do so, or think they should do so.
> RESPONSE: We agree that a log plot would be clearer and have done this – thanks.
CHANGE: The revised Figure uses logs on both scales (programming for this was done in R). We have left in the grid line to mark the p=0.05 though, rather than create bands, as we feel this will be more familiar to readers.

4. Point 2 of the conclusion is confusing. This paper is about choosing a strategy for busy clinicians. It is not about how to conduct a systematic review and this particular recommendation lacks the rigour in reasoning required to include it. There is no reason to single out the most precise trial for investigation of heterogeneity. First, their paper is proposing that one doesn't have to do a meta-analysis. If one does the meta-analysis, then
heterogeneity should be explored regardless of study precision. The example that the error in the Cochrane review would have been picked up is a straw-man argument – any check for heterogeneity would have done so. Second, study quality needs to be assessed for all trials, and this is what led to the recent recommendation for Risk of Bias tables in Cochrane Reviews. The authors say small trials are 3x more likely to have bias, but it is the absolute risk of bias that is important. And the statement would need to be qualified because any "large" trial that addresses a population that is different from the patient I have to treat, will by definition give me a biased answer for the question I need to answer. All of these may appear as subtleties but they are in fact essential to the appropriate synthesis of information and decision-making approach. This goes beyond the scope of the authors' study and I strongly recommend that the authors stick to their focused objective.

RESPONSE: We agree. CHANGED. We agree that these suggestions were not central to the paper, and have done two things: (i) we have shifted from Conclusions to become part of the Discussion, and (ii) we have shortened and de-emphasised these suggestions.

Minor Essential Revisions

5. In response to my previous comment that extracting the numbers is only a small part of the process, the authors responded that the people they surveyed said this was about 5%-30% of the total time (I would have said 10% myself). In the current version of the manuscript, the authors say the workload for a best trial would represent a "100 fold reduction in effort". This does not seem to be consistent and the authors need to make it clear that searching for the "largest trial" and appropriately evaluating it is very time-consuming by itself.

RESPONSE: We agree that it could be time consuming, but that this does not happen in practice. CHANGED. We have altered the sentence to emphasize that clinicians attempt to do this, but it is hazardous. New sentence:

"Clinicians therefore often attempt to do a rapid “best trial” review process which will take a few hours or less – over a hundred-fold less effort - but with clear risks of drawing inappropriate conclusions based on a limited search and single trials”.

6. The authors still say "we identified the largest trial...". Although they define the "largest trial" as the trial giving the most weight, I believe this is going to be misinterpreted by readers to actually mean the largest trial is the one to use. I think the authors have to avoid using the "largest trial" words altogether and stick to the "most precise trial" and make sure to explain what this means and why (most readers probably don't realize this).

RESPONSE: We agree. CHANGED. We now use “most precise” throughout (expect one instance where “largest” is explained).

7. Just before the conclusion, the authors again suggest that clinicians only have to evaluate the largest trial (please use "most precise" trial throughout) and don't mention that this requires evaluating all of the studies deficiencies. Their proposed method is only valid if one can do this step and this is a minor but essential point that needs to be stressed.
RESPONSE: We agree. CHANGED. We have added this to the lengthened 2\textsuperscript{nd} last paragraph of the Discussion where we also discuss the problems of finding the most precise trial. This is also stated in the new Figure 1.