Author's response to reviews

Title: Comparison of anticipated and actual control group outcome in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in favour of the novel treatment.

Authors:

Veronica Moroz (v.moroz@bham.ac.uk)
Jayne S Wilson (J.S.Wilson.1@bham.ac.uk)
Pamela Kearns (p.r.kearns@bham.ac.uk)
Keith Wheatley (k.wheatley@bham.ac.uk)

Version: 2 Date: 8 October 2014

Author's response to reviews: see over
7th October 2014

Dear Professor Altman

MS: 2992565741356508
Title: Comparison of anticipated and actual control group outcome in randomised trials in paediatric oncology provides evidence that historically controlled studies are biased in favour of the novel treatment.

Thank you very much for inviting us to resubmit a revised version of this manuscript. We have pleasure in sending you this revised version, which addresses the points raised by your reviewers as detailed below. We hope that you now find this paper suitable for the publication.

Reviewer 1:

1. This paper is brilliant. It is well conceived, soundly executed, and well analysed and reported. Its subject matter -- bias in the clinical trials literature -- and findings are both important. And while the findings come as no enormous surprise, it adds a valuable and convincing piece of evidence to the literature showing historically controlled clinical trials to be unreliable and positively biased for new therapies. It requires no changes in my view. As a discretionary revision, I would suggest just one or two sentences, probably in the introduction rehearsing how sample size calculations are typically carried out. This would make the paper more accessible to those who themselves have never conducted such studies.

Elements of the sample size calculation were added to the background section (page 4).
Reviewer 2:

The paper is a very neat way to address concerns about using historically-controlled studies in paediatric research, and demonstrates that the concerns are well-founded. Overall it has a strong message. There is much to recommend this submission, particularly the discussion and recommendations. My comments below focus on details relating to the methods and presentation of results, which could be improved.

MAJOR COMPULSARY REVISIONS

1. Methods -> Inclusion criteria. States that 'Where publications reported more than one randomised question (RQ), each question was considered separately.' This needs to be justified. A three-arm trial includes two randomised questions, but if you include both you are counting one control arm twice. I am also not convinced it is a good idea for factorial designs, because the control group rate for the two questions overlaps and is thus correlated. It would be reasonable where different randomised questions use different clinical outcomes, but if you do that why not look at the control group rates for secondary outcomes as well as primary?

We respectfully disagree with the reviewer on this point. The main analysis in our paper is the comparison of the actual control group outcome with that anticipated, so this issue does not affect the primary point being made. We considered that two separate clinical decisions – one for each comparison – would be made in factorial and 3-arm trials, and therefore these were included as separate randomised questions. There was only one 3-arm trial and only five randomised questions that were a part of factorial designs (the control arm is not actually the same for both comparisons in a factorial trial, though there is some overlap), so it is unlikely that the overall results would have varied meaningfully had we only included one comparison for each trial (furthermore, how would we have decided which comparison to omit without potentially introducing bias – e.g. leaving out the comparison that fitted less well with our hypothesis?). We have performed sensitivity analyses on the absolute differences excluding randomised questions that formed a 3-arm trial and/or trials with a factorial design and the conclusions were not changed. These justifications and sensitivity analyses have been added to the manuscript (page 5, 6, 9 and 10). A comparison of secondary outcomes rates would not be feasible as the sample size calculations (and therefore the anticipated control group rate) are not usually presented in the publication.

MINOR ESSENTIAL REVISIONS

2. Abstract -> background -> final sentence. Makes sense, but is a bit convoluted. I would suggest something like 'The rationale was that the control group outcome used in an RCT is what would have been used if a historically-controlled study had been done instead.' (and the same with this sentence in the introduction).

The sentence has been rephrased.

3. Methods -> Analysis -> Absolute differences. Need to say which variables are correlations calculated for and why. The 'why' is not explained in the results or discussion either.
The discussion (page 12) has been amended to include the rational behind the correlation calculations. We have also modified the calculations for disease prognosis for the trials with time to event outcomes. Previously disease prognosis was calculated as the average of observed treatment and control outcomes. As primary outcome and their timing varied from trial to trial, we've calculated the survival outcomes for a 3-year time point (assuming an exponential distribution). The scatter plot against the absolute differences was updated with the correlation coefficient.

4. Methods -> Analysis -> Relative differences. 'Rate ratio' usually refers to count variables; 'risk ratio' is surely the appropriate term for binary variables (I grant that an outcome may be good or bad so 'risk' might sound strange).

Rate ratio has been changed to risk ratio throughout.

5. Methods / Results. It might be nice to standardise the anticipated vs. observed outcomes in some way. For example, you could use: (Superiority) The ratio of observed-anticipated to the 'clinically important difference' used in sample size calculation; (NI/equivalence) The ratio of observed-anticipated to NI/equivalence margin (delta) used in sample size calculation. These ratios could be negative, but would have a nice interpretation for how wrong the corresponding historically-controlled study might have been.

Our discussion compares the relative difference presented by Charles and colleagues [reference 32] and ours, both calculated as \[(\text{anticipated-observed})/\text{anticipated}\]. We are unclear how this suggested additional analysis would add to the message of the paper.

DISCRETIONARY REVISIONS

6. Discussion. I would suggest the paragraph beginning 'HC studies compare outcome data...' should be at the start of the discussion, followed by a recap of your main findings.

We consider that the current order works well.

7. Table 2. Is it relevant to include alpha and power? Why?

Alpha and power are part of the sample size calculations and are part of trial design; we feel that adding these helps to describe the trial design better.

8. Figure 5. Caption says 'dashed vertical lines show the median HRs of RRs' - should say 'HRs or RRs'.

Many thanks, this has been corrected.

9. Appendix 2. I am not sure why the correlation between trial size and year of publication are interesting as the methods section is very vague about this (the same for figure 2). Why are the figures in appendix 2 separated from figure 2? Would a Bland-Altman
plot not be more useful than any of these? Such a plot would be worth including in the main body of the paper.

Figure 3 is separate from the appendix as we did not want to select just one figure for the main paper – e.g. selectively choosing the one with the most positive result. We are happy for the plots to be presented in the main document.

The Bland-Altman plot is used for measuring an agreement between two procedures to determine whether the two methods can be used interchangeably. Correlation is not appropriate in these circumstances as the two procedures would have high correlation. We feel that the correlation coefficient is an appropriate measure in our work.

10. There are several minor grammatical errors and I recommend a careful proofread by the authors.

The document has been proofread by a fifth person.