Author's response to reviews

Title: No Short-Cut in Assessing Trial Quality: A Case Study.

Authors:

Karim F Hirji (kfhirji@aol.com)

Version: 2 Date: 10 February 2008

Author's response to reviews: see over
No Short-Cut in Assessing Trial Quality: A Case Study
Karim F Hirji

Cover Letter to the Editors

Dear Editors of Trials:

My paper has been revised in accordance with the helpful comments from the two referees. A point by point response to each referee’s comments is attached to this letter. There are, however, three issues relating to their comments I would like to address here.

1. Length of the Main Paper: My paper makes three somewhat strong claims: (i) Burke et al. BMJ, 1991, is a highly, if not a fatally, flawed study. (ii) The quality evaluations in nine relevant systematic reviews, including the Cochrane Review, generally failed to noticed this defect, and declared it a good to high quality trial. (iii) This failure points to a possible deficiency in the methodology of trial quality evaluation.

I expect that these claims will, if my paper is published, be thoroughly scrutinized by the readers and other concerned parties. For this reason, the initial version had two fairly long additional files laying out in detail the evidence on which I based my claims. The main paper just had a descriptive summary of this evidence.

Both referees were not satisfied with the summary, citing lack of “numerical data,” “convincing evidence” and specific “references.” Dr. Gøtzsche suggested that I bring at least a part of the evidence into the main paper. This has been done in the revised version.

Dr. Gøtzsche also suggested that I shorten the paper. I have thus removed some repetitions and unimportant material from the main paper. But overall, I could not fulfill both of his suggestions at the same time. Bringing important aspects of the evidence into the main body of the paper has increased its length. In the journal publication format, the main paper (including tables and references) is 14 pages long. I believe this still falls within the paper length range for your journal.

2. Two Additional Files: The revised version also has the two Additional Files. Apart from minor editing, and reversed order, these are essentially the same as before. As the sole of author of a potentially controversial paper, I feel obliged to provide the reader the complete evidence on which my strong claims are based. Please also note that the two referees have evaluated my paper in general terms only. A detailed scrutiny of all my specific claims with respect to Burke et al. and the nine systematic reviews will therefore come from the readership.
It is for this reason that I sincerely urge you, should you accept my paper, to also allow the publication (as supplementary material on the journal website) of the two Additional Files in essentially their present form. In this era of electronic publishing, I think this is not too much to ask.

3. Inter-Rater Reliability: Dr. Moher has raised the issue of inter-rater reliability. That I am the sole author of this paper is a limitation: I agree and have noted it now. But also, I do not think a formal assessment of inter-rater reliability is an issue of relevance here. My paper is a case study. It does not propose a new criterion for evaluating trial quality. I use the type of criteria that are normally used in clinical trial context like bias at baseline, biased follow up, biased patterns of missing data; recruitment of ineligible subjects; wrong denominators in data analysis; incorrect computations; etc. to evaluate Burke et al. My point is (i) to show the presence of these problems, (ii) the extent to which they distorted the findings of the trial, (iii) that virtually all of them were overlooked by the quality assessment done for systematic reviews, including three with independent, and one with blinded, reviewers, and (iv) the manner in which these problems have been hidden by the style of presentation.

A case study stands on the merits of the evidence it presents. And I have presented the evidence I have gathered in extensive detail for the referees and others to come to a judgment on this matter. Conclusions of a general nature I draw are tentative and speculative.

It is up to future investigations to determine whether this case is an isolated one or not. How commonly are flawed trials declared good quality trials by check-list assessments in systematic reviews? What is the general level of effort and difficulty of associated with detecting such trials? I feel that it is in such a general context, with multiple investigators and many trials from a variety of fields, that the issue of a formal inter-rater reliability assessment may be pertinent.

I thank you and the referees for reviewing my paper and I sincerely hope that the revised version will meet your approval.

Sincerely,

Karim F. Hirji
Professor, Department of Epidemiology and Biostatistics
Muhimbili University of Health and Allied Sciences
P. O. Box 65015
Dar es Salaam, Tanzania
email: kfhirji@aol.com
Responses to the comments of Dr. Peter Gøtzsche

Thank you for your thoughtful comments. I have tried to make the revised version more focused. It also contains required numerical data. On shortening the paper, please see the letter to the editor. Responses to your specific comments are given below in the order and with the labels used in your report.

Abstract: Suggestion implemented.

Full Stops: Corrected where needed. The end of section trailing dots are placed by the journal LaTeX format file.

P2, line 1: Corrected word.

Tools for trial quality: Noted that some trial quality assessment tools have items that have nothing to do with bias. In my view, items relating to external validity are also important and need to be included. This is a broader issue that cannot be fully addressed in this paper. My paper does not critique specific items employed in the assessment tools but the item-based approach as such.

Para 3: ‘We’ changed to ‘I.’

P3, para 2: Two or more independent reviewers: Three of the nine AOM systematic reviews I identified (Rosenfeld et al. 1994; Marcy et al. 2001; Glasziou et al. 2004 (Cochrane Review)) had independent reviewers and in one they were also blinded. Nevertheless, they also did not detect most of the serious problems associated with this study, and ranked it as a good or high quality study. My point was and is that even with independent reviewers, a check list type of assessment is not a guarantee that major trial flaws will be identified.

I agree that it would have been better to have other people involved in my study; I have noted it as a shortcoming in the revised version. (Actually, I tried to get some colleagues interested but was unable to find a collaborator.)

In this version, I have clarified the role of the students, and acknowledged their help. Please see the explanation given below.

Para 3: The search strategy and inclusion criteria for systematic reviews have been stated clearly and its limitations noted in the discussion section. On the role of students: see below.

P4, para 5: There was only one group of students. How and what they did is
clearly stated in the Methods section now, and the results of their work are given later. This was not done in a clear manner before.

The students were participants of a two week course on Principles of Evidence Based Medicine held at University of Oslo in June 2006. They included physicians with years of experience, academic faculty as well as post-graduate students. I was the sole instructor. On the second day, I dealt with the history of clinical trials and procedures for enhancing trial quality with several examples of good and poor quality trials. For the next day, they had to read two papers, Burke et al. (1991) and Halstead et al. (1967). At the start of the next class, they gave their assessments of the two papers on a five point scale as noted in Table 2 of my paper. This was not a check-list based assessment.

Then I spent two hours going through both papers in extensive detail. My detailed review was like that given in Additional File 2. The point was to show that there are times when it is easy to identify a poor quality trial (e.g. Halstead et al.) and there are times where it is a difficult undertaking. All the students, without exception, were astonished that the major flaws I pointed out in Burke et al. had escaped their attention, and there was not a single student who disagreed with my assessment that it was a very poor quality trial.

Thereby, I state with all honesty that these students did not in any way contribute to the critique developed in Additional File 2 and summarized in the main paper. Their contribution to this paper in providing the data for Table 2 is explicitly noted in the Acknowledgments.

(Last September, around the time I submitted my paper, I made a similar, more detailed, presentation at a seminar held at the University of Oslo. Members of the Nordic Cochrane group were invited to this seminar, though I am not sure how many actually attended it. This audience of about 35 people also generally agreed with my assessment.)

P5, 1st para: On the need to quote systematic reviews: The relevant quotes from systematic reviews and data extracted are in Additional File 1. The point then is what is the best place to put them. In the initial submission, a summary was provided in Table 3 and in the last but one paragraph of Traditional Quality Assessments section. I have retained a similar format in this version. Otherwise the main paper will become even longer.

On the length and overall form of the paper: Please see the cover letter to the editor.

Para 2: The role of the students has been clarified above. They did not use a check list but scored two trials on a five point scale having earlier been given a lecture, with
examples, on the history of clinical trials. All the criticisms given in this paper were generated by the sole author.

Para 3: Length of the paper: Please see cover letter to the editor.

Classification of the “major” problems: I agree that this is important, but in this case I was unable to do that as you suggested as many problems were interrelated and did not stem from a single root source.

Consider one example: The sample size calculation in Burke et al. are declared to show that “some 200 subjects” were needed. The trial recruited 232 subjects. On the surface, it seems all was fine. I repeated their calculation and found that with the figures they used the needed sample size, not allowing for missing data and twenty declared main outcome variables, was 240. Due to missing data, the sample size attained was lower by 10% to nearly 50% for key outcome variables. (All that is described in great detail in Additional File 2). How do I classify this: a computational error at the design stage, a case of misrepresentation resulting from an inability to recruit sufficient subjects, an inadvertent enhancement of the false positive rate, a missing data problem, or something else?

However, I have followed your excellent suggestion to remove the list of the thirty so-called major problems and instead given a detailed description of some of the major problems in the main paper now.

Para 8, bottom: In theory, your proposal to see the impact of excluding data from Burke et al. is very good. But in practice, it faces difficulties. First, as I note in the main paper and detail in Additional File 1, the systematic reviews themselves (including the Cochrane Review) used the data from Burke et al. in a flawed manner, including double counting and using wrong numbers. One then has to compare three things: Results from SR using the Burke data incorrectly, results from SR with corrected data, and results without the Burke study. Also, one has to factor in the different models used (some use the random effects model and some the fixed effects model); and the risk measure used (risk ratio, risk difference or odds ratio). On top of that, the latest systematic review, for which this concern is most critical, is an individual patient data meta-analysis and I would need the raw data for all six included trials. Even this review misrepresents the data for pain and fever as parental diary data when they are derived from researcher home visits. The parental diary data on pain, however, were collected by Burke et al. but are not used in the review. For the remaining studies in the review, all main outcome data are diary data. To correct for all that would be a major undertaking and needs a separate study. I would not like to present only a partially done job here.

Instead, I indicate the importance accorded to this trial in the three most recent
systematic reviews in the field (one of which used individual patient data) in Table SR1 in terms of the number of meta-analyses (for different outcomes) for which its data was used. For the Cochrane Review, I have noted the weight given to Burke et al. in arriving at the overall estimates, and for individual patient data meta-analysis, the percentage of the total sample size.

There are two fundamental issues here: (i) Should the quality of a trial be assessed just in terms of the impact it has on the findings in a field, or is trial quality also a separate entity that needs to be assessed in its own right? If including or excluding a very poor quality trial, or say even a fraudulent trial, has no material impact on the conclusions of a meta-analysis, should one go ahead and include it? (ii) And what if there are a number of other very poor quality AOM studies in these reviews? If deleting one such trial does or does not change the overall estimate, what does that mean?

**P9, para 4:** The noted confounding relates to the fact that age less than three years and otorrhea were exclusion criteria in Burke et al. while the individual patient data meta-analysis used age $\leq 2$ years or not and presence of absence of otorrhea as potential predictive factors. This had been noted in the Additional File 2 previously. In the interest of shortening the main paper, this has now been deleted in the main paper and only retained in the Additional File.

**P11, para 2:** I have now explicitly noted current recommendations for quality assessment, and the importance of concealment of allocation, early in my paper.

I have already provided comments for the Cochrane Review a year ago; however, the authors have chosen not to respond to my comments and, it has taken very long for my comments to appear on the Cochrane website. (At the moment, I am unable to access the full Cochrane Website; my (and all Tanzanian academics’) access to this website was through the WHO/Hinari portal. Because of a password leakage, our access to this website and other journals has been cut-off at this time).

If and when my paper is published, I fully intend to alert the authors of the relevant systematic reviews, so that any necessary remedial action may be taken. But I need to wait until a peer-reviewed journal agrees that my claims have merit.

I have quoted your remark likening quality assessment to a type of “detective work” in this revised version; I hope you will permit me to that.

I agree that not just words but numbers are also important. In the previous version, all the numbers, computations and detailed reasoning were in the additional files. Now I have followed your suggestion and moved some of them to the main paper; but this has increased the length of the main paper.
Responses to the comments of Dr. David Moher

Thank you for your thoughtful comments. My earlier style of presentation did not clearly specify the aims and methods of my paper. Now I clarify that there are only two basic approaches being compared; namely, check-list based evaluation versus in-depth evaluation. My responses to your specific comments are given below in the order you made them. Please also see the cover letter to the editor for other relevant concerns.

Major Revision I: The points you raise about scales and components are now noted in the second paragraph of the Background. The list of 30 ‘flaws’ has been deleted. Instead, I give a detailed description, with references, of some of the main flaws. In the previous version, all these were detailed in the additional files attached to the paper. These files give a complete description and more detailed references for all the problems I found.

The evaluation by students (what you call the second summary approach) did not use a check-list. What the students did and how they did it has now been clarified in the paper. For your information, I give some additional information.

The students were participants of a two week course on Principles of Evidence Based Medicine held at University of Oslo in June 2006. They included physicians with years of experience, academic faculty as well as post-graduate students. I was the sole instructor. On the second day, I dealt with the history of clinical trials and procedures for enhancing trial quality with several examples of good and poor quality trials. For the next day, they had to read two papers, Burke et al. (1991) and Halstead et al. (1967). At the start of the next class, they gave their assessments of the two papers on a five point scale as noted in Table 2 of my paper. This was not a check-list based assessment.

Then I spent two hours going through both papers in extensive detail. My detailed review was like that given in Additional File 2. The point was to show that there are times when it is easy to identify a poor quality trial (e.g. Halstead et al.) and there are times where it is a difficult undertaking. All the students, without exception, were astonished that the major flaws I pointed out in Burke et al. had escaped their attention, and there was not a single student who disagreed with my assessment that it was a very poor quality trial.

The students did not in any way contribute to the critique developed in Additional File 2 and summarized in the main paper. Their contribution to this paper in providing the data for Table 2 is now explicitly noted in the Acknowledgments.
**Major Revision II:** The data were checked by me; this has now been stated explicitly. The fact that I am the sole author and investigator in this study is a shortcoming. I agree, and I have noted that in the conclusion of the revised version.

On the question of inter-rate reliability: Please note that in my detailed review, I do not come up with or propose a new indicator of quality; I deal with issues that are mainly accepted and used, like bias at baseline, biased follow up, biased patterns of missing data; recruitment of ineligible subjects; using wrong denominators in data analysis; incorrect computations; etc.

My point is to show (i) the presence of these problems, (ii) the extent to which they distorted the findings of the trial, (iii) that virtually all of them were missed by the quality assessment done for systematic reviews, including three assessments with independent, and one with blinded, reviewers, and (iv) the manner in which these problems were hidden by the style of presentation.

My paper is a **specific case study** which points to a possible problem in the methodology of quality assessment. It took me three months of careful scrutiny of Burke et al. to develop my critique; and as I note, even the class of students did not uncover the problems I detail.

It is for this reason that in the previous and current versions, I have given my data extraction, reasoning, computations and conclusions in extensive detail in the additional files attached to my paper. A case study needs to be judged on its own merit, and the claims it makes should be checked in terms of the evidence given. I think that is also a function of the peer-review process. To see if the claims I make for this trial valid and reliable, I request you to examine the evidence provided.

How often do systematic reviews declare faulty trials as good quality trials? With what level of difficulty and reliability can one in general identify highly faulty trials using a detailed evaluation? Those are relevant questions that would follow on my case study. In the broader investigations undertaken to study them, the issue of inter-rater reliability assessment in formal terms may be pertinent. I do not think I can address it now.

**Minor Essential Revisions:** I have tried my best to improve the paper in terms of wording, style and format.

**Minor Concerns:**

1: ‘We’ changed to ‘I’.

2: Internal validity has now been defined with a relevant reference given.
3: I agree the reason why reference 24 was not included reference 23 is probably what you state. My reason for stating this fact was to just note that in past, a detailed assessment of 50 acute otitis media trials had been done; but the trial I am critiquing was, for whatever reason, not included in it.

4: I agree that there has not been a single universally accepted quality assessment instrument, and there “is no traditional approach.” I was referring to the approach to quality assessment based on a check list, even though the items in these lists were different. The term “traditional” has been removed in the revised version.

5: The reasons for representing the trial quality as in Table 1 (using Balk et al. as the template) are now given in the Methods section. The simple or few-elements approach refers to what is currently recommended for systematic reviews.

The Balk approach is different from the others you note in that it has a larger list of items. Nonetheless, it is still a check-list based approach. And it too failed to detect the problems found by my detailed review. The Balk template revealed little beyond that uncovered in the Cochrane Review. I do not agree that its results differ substantially from that in the nine systematic reviews.

6: No check-list was used by the students (what you call the second summary approach). It was an overall five-point evaluation based on a general reading of the paper. This has been explained in the revised version.