Author’s response to reviews

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

Authors:

Karim F Hirji (kfhirji@aol.com)

Version: 4 Date: 31 July 2008

Author's response to reviews: see over
July 31, 2008

Dear Professor Altman
Editor-In-Chief, Trials:

With this cover letter and a point-by-point response given below, I enclose a further revised versions of the above named paper and its two Additional Files.

I have moderated the language further and made the minor changes as suggested. Your letter also raised four major queries: (i) the wisdom of guessing; (ii) inclusion of children with bulging ear drums; (iii) missing baseline data for crying; and (iv) allowing for the level of reporting in 1991. I provide detailed responses to these queries below. I have made needed changes relating to the concern expressed in query (iii). Elsewhere, the wording has been changed in a few places to clarify the issue at hand. Otherwise, the new versions of the main paper and the Additional Files remain as they were previously.

I thank you for checking the claims of my paper by reading Burke et al. (1991). The major queries you have raised are precisely the type of specific queries I expected in the initial reviews of my paper. However, none of your very distinguished panel of reviewers and the editorial board earlier raised any such matter. (At the same time, I do acknowledge that their comments were quite beneficial.)

So, it is well that you have raised them. Each requires a detailed explanation. Given that the paper has been shortened, such details now are in the Additional Files. By just reading my main paper and checking it superficially against Burke et al., one may get the impression that its claims are not well founded, or based on pure guess work. I would like to assure you that I document and justify each claim, and the bases of my conclusions are carefully laid out in the Additional Files. As the reader may also have such concerns, the need to read these files is categorically stated in paragraph 2 of the subsection In-Depth Dissection of the main paper.

The world of health research methods has no shortage of controversy, in terms of methodology, substance and interpretation. I agree that some will disagree with key aspects of my paper, and may regard it as dealing with minor issues. The key question, however, are whether I provide adequate documentation to back my main points, and whether the approach I present deserves a hearing. This, I believe, is the case.
Before I give a detailed response to your four queries, let me state that there was one aspect of your comments I could not reconcile. Before making “Some minor points,” you state that you could not check my claim about crying at baseline as the requisite Additional File 2 was a copy of Additional File 1. But, in the 3rd paragraph of your letter, you note the extreme phrase “the biased analytic and presentation style” and indicate that it occurs on “p 13 of the supplement.” What you call “the supplement” here is precisely the file Additional File 2, and this phrase was present in the last version of that file. So it is unclear what the problem you faced was. In any case, please bear with the details I repeat from this file in my point by point response below.

I do hope that with the present revised version both Additional Files are available as they should be.

I know that the job of editors and reviewers is a demanding one and I appreciate all the effort you and the three referees undertook to review my paper. The thoughtful comments and suggestions given so far have helped me considerably in improving it, and I sincerely hope that the revised version will meet your final approval. I also thank you for the idea of commissioning a commentary on my paper.

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
Point-by-Point Response

1. **Moderate Language**: Done as suggested.

2. **On Guessing**: One task of my study, as noted in the Methods Section, was to formulate “plausible explanations and an overall perspective for the distinct problems ...” I think this is a reasonable type of undertaking that appears in many scientific papers. The statement you quote about possible non-occurrence of home visits (and others as well) accord with this objective. The key question is: Is it a pure guess, or is it based on reasonable evidence? I request you to consider the following information on home visits:

2.0: In Burke at al. study, who made the home visits? The physicians, their nurses, some of the authors, or special research assistants? Did this vary from practice to practice? None of this information is specified.

2.1: At the first home visit, historic baseline data was to be collected. The level of missing data for such variables cannot be judged as their denominators are not fully given anywhere. However, from Table III, I noted that for “previous episodes of AOM,” 15 cases (out of 118), including 4 (out of 17) treatment failures, in the placebo group had missing values. The respective numbers for the antibiotic group remain unknown.

2.2: During the home visits, the researcher “... asked about current pain, measured the child’s temperature, weighed the medicine bottles, ...” For the first home visit, there are only 3 missing values for pain, none for analgesic use, but 52 missing values for fever.

2.3: For the second home visit, there are 7 missing values for pain, 17 for analgesic use, and 111 (nearly half the study sample) for fever.

2.4: We know that visit 3 did not occur in a uniform way, since for the consumption of analgesics, the data were reanalyzed to adjust for, “interval between entry to trial and visit [3].” No further details are given but the difference between the groups must have been large enough to warrant this adjustment. Yet, no such adjustment was done for other visit 3 related variables like pain and fever. Were the timings for these variables different?

2.5: For medicine use up to visit 3 (compliance with antibiotic or placebo), there are no missing values, and no adjusted analysis was done. Was there a difference in the manner and timing of collection of these bottles?

A plausible explanation for this pattern of missing data: Medicine bottles can be collected later or can be collected at the in-clinic visit on day 8, pain can be asked
over the phone, but measuring the temperature needs a physical presence of the researcher at the child’s home. In particular, why were fever data missing for about half of the study subjects when the research plan explicitly states that fever was to be measured during the visit? I think it is reasonable to raise the possibility, as I do, that “many home visits did not actually occur .......”

One may consider other possibilities. For example, the missing data on fever may be explained if many children were not at home at the time of the visit (say, was in school). Then how were their levels of pain determined? Was it assumed that their absence indicated they no longer had an ear ache? Then why not also assume they had no fever? etc., etc.

Each possibility raises equally troubling questions. As the paper says that some visits were delayed (and in a biased manner), it is reasonable to postulate that, given the level of missing data for fever and other evidence noted, many visits also did not occur.

Based on such varied evidence, I also raise similar questions and, at times, give a possible explanation that seems to accord with the evidence for other problems I found. I do not make definite statements, and at the end, stress the need for an independent audit to find out what actually happened. Thus note that I do not call it a fatally flawed study but, based on available evidence, a potentially fatally flawed study. Since the data for this trial were recently released for an individual patient data meta-analysis, an independent audit is, I think, a feasible undertaking.

3. On Bulging Ear Drums: Bulging ear drums is not a listed exclusion criterion in the Patients and Methods section of Burke et al. In this paper, several things are not listed where they should be, under scoring the point about not relying just on the Methods section to judge what was planned and done in a study. For example, the first paragraph of the Results and Figure 1 together show that “ear drum not adequately seen” was an exclusion criterion. But that fact is not explicitly stated in the Methods section.

What is the evidence that bulging ear was an exclusion criterion in the study:

3.1: Strong indication for antibiotic, a stated exclusion criterion, is not fully defined anywhere in the paper.

3.2: Investigators held training sessions for participating doctors to explain the methods and aims of the trial. They were shown slides and given sample photos of ear drums to standardize diagnosis, evaluation and terminology.

3.3: A standard form (Fig 1), which included noting whether the ear drums were bulging or not, was devised for the initial examination of the child.
3.4: The Objective in the Abstract reads that this study was to cover “mild” cases of otitis media.

3.5: In the Introduction, we read: “... it is increasingly recommended that use of antibiotics should be selective, being confined to those who are severely affected—for example, those who have bulging ear drums—a physical sign that which predicts bacterial infection.” A reference which claims to show that bulging ear drums are predictive of a bacterial infection is noted. This is the **sole specific indication** for a severe case of AOM noted in the paper. The presence of bacteria constitutes the basic indicator for an antibiotic.

3.6: Only 1 out of the 17 participating practices maintained adequate records about excluded cases. In this practice, 105 of 189 eligible children were excluded. About half (52) of them were for “indication of antibiotic,” and of these, the majority (27) were due to “bulging ear drum.”

3.7: The second paragraph of Discussion reads that children “were selected on the basis that treatment with placebo would pose no ethical problem” and that “excluded children may have been more severely affected ....”

This is the first time we learn that inclusion of children was done on the basis of a set **ethical** criterion relating to severity of AOM. (One systematic review (Rosenfeld 2003) has positively singled out Burke et al. for this reason.) This refers to a context in which doctors routinely prescribed antibiotics to children with AOM and where it was deemed unethical to withhold antibiotics to most children with the condition.

3.8: In paragraph 4 of Discussion, once again the reference to “severely affected” cases arises in connection with children with bulging ear drum.

**Plausible inference:** This was a trial with a declared objective, for ethical reasons, of including only mild cases of AOM, where the sole sign of severe disease and indicator for antibiotic given consistently at several places is bulging ear drums, where the initial assessment form included noting this condition, where training sessions for doctors were held to promote uniformity in diagnosis and selection, and where data from the sole practice with good record keeping show that bulging ear drum was the most common reason for excluding children.

Under these circumstances, I think it is reasonable to say that bulging ear drums was a specific exclusion criterion in the trial, and the inclusion of 27 (12%) with bulging ear drums was a violation of the ethical standard and scientific criteria for this study. (Other serious concerns related to bulging ear drums are in Additional File 2.)

4. Crying at Baseline: Figure 2 shows that about 57% placebo and 34% of antibiotic group **are known to be crying at baseline**. At two places (legend of Fig
2 and first para of Short Term Outcome), the authors misrepresent these data as difference in outcome rather than a baseline difference (and a highly significant one at that). Two systematic reviews have spotted this error, and the latest Cochrane review ascribes it to a “failure of randomization.”

Your query concerns the status of the remaining 43% placebo and 66% antibiotic group cases. Were the data for them missing, or are these children who were not crying at baseline? Please note that I do NOT say that all of them had missing data on crying at baseline. I use the phrase “up to” to indicate that the missingness for the two groups could be as high as 43% and 66%. In the discussion on crying given later, and in the Additional File 2, I clearly state these groups include both, that is, those who were not crying and those whose crying status was not known.

I agree that using the phrase “up to” and referring only to the missingness aspect of the data is confusing and unclear. I have changed the wording in this paragraph to clarify this issue.

Also note that these curves represent the duration of the “initial episode” of crying (and pain), and not those who were crying (or in pain) at some point in time. Thus they ARE meant to be survival curves.

There are two reasons why I say that these groups (the 43% and 66%) include both those who were not crying at baseline and those whose crying status was not known. If they included only those who were not crying at baseline, the shape of the survival curve would have been slightly different. Like for pain, they would start at 100%, and instead of a sharp drop at time zero, there would have been a sharp drop at the next observation point (four hours after the visit – from the first entry in the parental diary). (See Additional File 2 why, for pain, my interpretation is not the same).

The second reason relates to the fact that for several variables, there was a lack of harmony between the at clinic collected and at home collected data. Thus, while fever data are available for the home visits, the fever data for the baseline clinical visit and day 8 clinical visit are 100% missing (or not reported). There are no pain or crying data for day 8 clinical visit while both outcomes appear in parental diaries and pain is reported for the home visits. There is no reason to make an exception for crying and it is plausible that baseline crying status was also missing for some subjects (with completed parental diaries) and as for baseline fever data, the level of missingness may be high. (For more details on this issue and the computation errors for crying, see Additional File 2).

5. Allowing for the Reporting Level in 1991: On a personal note, I would like to say that for teaching my classes I have often used your series of papers in the BMJ in the 1980s that promoted good scientific and statistical practice as good ethical
practice. Your papers monitoring the quality of medical trials over three decades and other related matters have been a source of inspiration to me. I also note there have been many, many well designed, implemented and reported trials since the 1950s.

For a condition that causes much misery among children, has a high societal cost of treatment, is the leading reason for children’s visit to a doctor and antibiotic usage, and where the issue of poor quality studies became prominent in the 1980s, I find it difficult to make any allowance for a study of this quality to appear in one of the top medical journals in the world, and that in 1991. I note that it does acknowledge assistance from a leading medical statistician. Hence I have not made any changes to my paper in relation to this query.

6. **Minor Points**: Corrections done.

Thank you for your thoughtful comments.