Reviewer's report

Title: Development and Evaluation of a Quality Score for Abstracts

Authors:
Antje Timmer (antje.timmer@klinik.uni-regensburg.de)
Lloyd R Sutherland (lsutherl@acs.ucalgary.ca)
Robert J Hillsden (rhillsden@acs.ucalgary.ca)

Version: 2 Date: 8 Sep 2002
Reviewer: Prof David Moher
Level of interest: A paper of considerable general medical or scientific interest
Advice on publication: Reject

8 September 2002

Drs Timmer and colleagues have chosen an important topic to address. Quality, or more accurately the quality of reporting, is of interest to most people involved in health care. There is an established field of knowledge on this topic, some of which is cited by the authors. Similarly, the authors note the paucity of instruments that have been developed to assess the quality of reports of abstracts. From my perspective the development of such an instrument would be valuable for several reasons. As the authors note, not all research presented at scientific meetings goes on to be published. This "gray" literature is often important to include in the conduct of systematic reviews. Currently there is no "satisfactory" way of assessing its quality.

However, I feel the authors have fallen short on several fronts some fundamental while others are less problematic to address.

Fundamentally the authors have combined aspects of methodology and reporting in a single summary score (page 5). Several authors have commented on this problem previously (e.g., Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282:1054-1060). The authors are likely to be significantly criticized for combining aspects of methods and reporting within a single scale.

The instrument is long in terms of the time it takes to complete. Having spent several years trying to persuade individuals to spend 30 seconds completing the Jadad instrument (your reference #13) and an assessment of allocation concealment, I cannot imagine people will take kindly to spending nearly four minutes completing this one. The authors state (page 3) their objective was to develop a "quick and easy..." instrument. I would argue that they have failed to convince me that these objectives have been met.

The authors do not discuss another important possibility for improving the quality of reports of abstracts, namely, increasing their length. Perhaps the commentary by us (Moher D, Schachter HM. Potential approaches to the problem of conducting systematic reviews of new technologies. Canadian Medical Association Journal. 2002; 166: 1674-1675) will be informative.

The authors have not be clear as to "for whom, and what purposes" they developed their scale. For
example, who would use the instrument, a systematic reviewer, journal editors, or scientific conference assessor?

The authors have not stated why they chose to develop a scale rather than focus on specific components, such as has been very successfully developed by Schulz and colleagues (Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; 273:408-412.). In this regard the authors are silent on the controversy of two different approaches that to assess quality, that is, the component versus scale issue. Any revision would need to take these issues into consideration.


There are several issues concerning the methodology that need clarification:

- How was the random sample selected (page 4), such as, computer generation?
- How many abstracts were included in the random sample initially (page 4)?
- How many items were initially generated (page 4)?

I believe the instrument has little internal validity and almost no generalizability for controlled clinical trials (the 5% percent of abstracts is far to low if the total sample was 100 abstracts).

The authors chose to use what I think is a Delphi like system to generate the items and reduce the initial list.

- How many investigators participated in this exercise?
- Was it a face-to-face meeting?
- Were experts in the three content domains (e.g., clinical trials, basic science)?
- Were the quality assessment raters "blind" as to which abstracts were accepted for oral and/or poster presentation?

The Delphi approach is different from the more analytically, and perhaps more precise, psychometric (or clinimetric, although I did not use of Feinstein's sensibility construct) factor analysis approach. The authors are silent as to why they chose one approach over another.

- Did the authors include a measurement/instrument development expert to help develop the instrument?

- The section "How to calculate a summary score" and "examples" should be moved to the Appendix.

The authors are silent on how they developed their scoring system. For example, as best as I can understand it, 4 points are awarded for parallel controlled trials and only 3 points for crossover trials. Why the differential scoring? The authors also assign similar "weights" (e.g., 2 if fully met) for each item. For example, they attribute similar importance to the reporting to whether random allocation is described and whether exact p-values are stated. Given the paucity of data on the importance of reporting exact p-values and the much larger empirical base (e.g., Schulz KF, Chalmers I, Hayes RJ Altman DG. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; 273:408-412 and Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P. Does the quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609-613.) for
the reporting of randomization I find it strange that both items are given the same numerical scores if reported.

Other issues

I find several of the items included in the scale have little to do with quality. And the authors provide little evidence to support the use of them. For example, item 16 asked whether authors report p-values or confidence intervals. Why is this related to quality? I am not aware of any evidence suggesting abstracts that do not report p-values, compared to those that do report this information, are of lower quality, or introduce bias.

Dr. Cho relocated to Stanford/UCSF a couple of years ago (page 13).

**Competing interests:**

None declared.