Author's response to reviews

Title: Development and Evaluation of a Quality Score for Abstracts

Authors:

Dr Antje Timmer (antje.timmer@klinik.uni-regensburg.de)
Lloyd R Sutherland (lsutherl@acs.ucalgary.ca)
Robert J Hilsden (rhilsden@acs.ucalgary.ca)

Version: 3 Date: 6 Nov 2002

PDF covering letter
Quality score for abstracts

Comments to the Reviewers

Reviewer 1:

Thank for your thoughtful comments on our manuscript.

1. Dichotomy between basic and clinical science, basic science specific items
   We agree that the test performance was worse in basic science, and that this is likely to reflect the biased use of resources. However, this bias could not be avoided, as there were no quality scoring systems available for use in basic science studies. We have discussed the list of items with basic scientists before finalising the instrument, and also have specifically included lab based researchers in the sensibility survey. However, no items specific to basic science research were identified. We give adequateness of the model as an example. However, we feel that this question can only be answered by experts in this field and is likely to be prone to expert bias. It was our aim to develop an instrument that could be used without specific expert knowledge, in order to avoid insider preferences. Certainly, more work needs to be done to develop the instrument for use in basic science, and modifications specific to lab based research might evolve. In the light of the complete lack of work in this field, we could only do a first step. We have modified our conclusion accordingly. Specifically we have indicated that the instrument needs further development and validation for basic science research.

2. Methods for construct validity
   We don’t clearly see our reasoning as circular. We believe that higher quality abstracts are more likely to be accepted for presentation, as selected by experts, than lower quality abstracts. Therefore, we hypothesized that if our instrument accurately measured quality then we should find that abstracts scored highly with the instrument should be more likely to have been accepted for publication than those with lower scores. We are therefore assessing construct validity by comparing our instrument to the expert opinion of the original abstract reviewers.

3. Validity across other specialities
   We did not mean to imply that this instrument would be valid across all other specialities. Rather, as you (and we) noted before, this is a first step. We would like to encourage other researchers to use this measure for abstracts in other fields, but would advise to further test and improve the instrument as necessary. The conclusions have been modified.

Reviewer 2:

We thank you for your comments, and especially, for directing our attention to some more of the work done on quality scoring of RCT’s. However, we feel that most of your criticisms are due to a misunderstanding on the aims and purpose of our instrument. This instrument was not developed to score quality of reports for inclusion into meta-analysis. It is certainly an interesting matter of discussion whether abstract quality in addition to the mere fact of being grey literature should be weighted before inclusion into meta-analysis (I personally do not
think so), and whether information is usually sufficient. However, this was not within the scope and aims of our article.

Rather, we needed an instrument that would describe abstract quality as a possible determinant of abstract acceptance for presentation, as well as subsequent publication. The primary aim of meta-analysis is the unbiased estimation of the effect in question. Subsequently, any measures that would impact on effect size are important to consider; other which don’t are not. In contrast, there are more aspects which might make an abstract worth being presented at a meeting, being read or being followed by a full publication.

1. Combination of methodology and reporting in a single summary score
We agree that the combination of methodology and reporting is debatable. The excellent paper by Jüni was not published when we developed our instrument. As discussed above, it is not directly applicable to our problem, as we were not concerned with meta-analysis, but rather, meant to expand the concept of quality to a wider range of research types and purposes. We were aware of the problem, but decided that both aspects are important when judging whether an abstract is “good”. Also, we feel that within the very limited information available from abstracts, it is not realistic to separate these concepts.

2. Length of the instrument
Possibly, we have been overly optimistic in our assumption that this instrument is quick and easy to use. In fact, “time and effort needed” as well as “clarity and simplicity” scored poorly in the sensibility questionnaire to lab based scientists (Table 2). Interestingly, clinical and health care researchers did not feel so. However, I omitted these attributes in the introduction section.

3. Length of abstracts
We agree that the quality of abstracts is likely to be improved by increasing the length of the abstracts. Interestingly, there don’t seem to be any data supporting this assumption. Please note that AGA abstracts are not confined to 250 words, the length you discussed in your comment recommended for referencing (Moher, 2002). Rather, as pointed out in the methods section, length was limited by space. Depending on the use of additional tables and graphs, up to 350 words were possible. A considerable proportion of the abstracts used less words than would have been possible. We regret we did not do word counts to examine the effect of abstract length.

4. For whom and what purpose was the instrument developed
As pointed out on page 3, we were concerned with the poor inter-observer agreement and the lack of standardization in abstract selection procedures for scientific meetings. We originally set out to examine determinants of abstract selection and subsequent publication. When assessing the literature on possible determinants, we discovered that there was no instrument available to rate the quality of abstracts. For use in research on the fate of abstracts, a single summary score across a sample of very diverse abstracts was needed. We have now used the score successfully for this purpose (Timmer, BMC Med Res Methodol 2001 and 2002) The instrument has not been tested for use in other contexts. We suggest that this instrument will be useful for assessing abstracts for presentation at meetings. We would also like to encourage abstract writers to use this score as a checklist. As discussed above, we do not see it to be very helpful for systematic reviews. These are usually restricted in their scope of topics and designs, so that it would be more helpful to use a more specific score, if at all.
5. Use of a scale rather than components
We are aware of the ongoing debate on this among meta-analysts, and actually take the side that in the context of systematic reviews, the use of components makes more sense. However, this was not about meta-analysis, but about finding a means of judging abstracts without subject expert bias. The use of a summary score rather than components is more helpful if a range of diverse designs and topics is evaluated, as the number of shared components is rare.

6. Use of 1992 to 1995 abstracts
We had to use older abstracts as this was part of a follow up study of abstracts, examining predictors of publication. We share the reviewer’s appreciation for the work by Schultz, and the importance of adequate allocation of concealment. We are now using the instrument on more recent abstracts and will devote specific attention to this topic. However, we are sorry to say that it is not likely that his work has had a major impact on the way gastroenterologists write abstracts for the AGA.

7. Methodology Issues
A large sample (n=1000) was selected for use in our follow up study using computer generated random numbers based on a database containing all abstracts submitted (17 000). This sampling was stratified by research type to increase the proportion of RCT’s. There were 319 RCT’s in this sample. For the test evaluation, a subsample was collected from the study sample, again based on computer generated random numbers (Stata). The abstracts were assessed in a prespecified order.

As reported in the methods section, both raters were blind as to whether abstracts were selected for presentation. The mode of presentation was not known.

24 items were initially generated (see p. 7)

7. Too few controlled trials
As evident from table 1, there were 42 controlled trials in the sample. Due to the stratified sampling, this number exceeds the 5% RCT’s in the sampling base.

8. Weighting
There is weighing in that 3 items relate to randomization (additional design point, randomization at all, method of randomization).

9. Item generation
The list of items was discussed face to face and repeatedly with one researcher each in lab based medicine (J. Wallace), health care research (RJ Hilsden), and clinical trials (LR Sutherland). The final instrument was subjected to any suggestions as to the appropriateness of items included (redundancy vs. comprehensiveness) by sending questionnaires to researchers in the three fields, as reported in our manuscript (Table 2). No further items were suggested to be added or dropped by those surveyed.

We do not think that the introduction of bias is the only measure to decide on the quality of a study report, although we agree that this is the most important aspect in meta-analysis.

10. We moved the “how to calculate” section to the appendix as suggested.