Author's response to reviews

Title: An alternative to the hand searching gold standard: Validating methodological search filters using relative recall

Authors:

Margaret Sampson (msampson@uottawa.ca)
Li Zhang (li.zhang@usask.ca)
Andra Morrison (andram@cadth.ca)
Nicholas J Barrowman (nbarrowman@cheo.on.ca)
Tammy J Clifford (tammymc@cadth.ca)
Robert W Platt (rplatt@po-box.mcgill.ca)
Terry P Klassen (terry.klassen@ualberta.ca)
David Moher (dmoher@uottawa.ca)

Version: 3 Date: 20 June 2006

Author's response to reviews: see over
We have addressed the discretionary comments of the reviewers, and incorporated most of the feedback into revisions in the manuscript. We believe this has strengthened the paper and we thank the reviewers for their thoughtful input. Details of how we addressed each point are below.

Thank you for the opportunity to revise and re-submit this manuscript.

Sincerely,
Margaret Sampson, for the authors

Referee 1
Reviewer: Pamela Royle
Reviewer's report:
1) In Paragraph 5 of the Background the authors state that: ‘The overall time and cost of doing a review is dependent on the size of initial bibliographic retrieval.’. I think this is perhaps a bit strong – as there are a number of other factors that influence the time and cost of doing a review – maybe modify to ‘is one of the key factors ..’ or something similar

   Thank you for this comment. The work cited is also aging, and it would be interesting to see newer material on effort required in systematic reviews. We have revised this to read:
   The overall time and cost of doing a systematic review depends, in part, on the size of initial bibliographic retrieval,[1] thus fine-tuning this initial step in the review process can yield great efficiencies. Cohen et al. demonstrated that even modest improvements in precision could save a week’s effort in a large review.[2]

2) in the Results. paragraph 3 - think there is a typo - should the word be ‘comprehensive’ and not ‘compressive?
   Thank you, this has been corrected.

3) The Results section you state that although all specified RCT or quasi-RCT as an inclusion criteria, a number included other controlled clinical trials. It would be interesting, if you have the data, to include some mention of what these other study designs were and were they confined to one review only? Maybe something in the Discussion would be interesting? Do reviewers misclassify RCTs - or are there other reasons for using non-RCT studies, when Methods specify otherwise.

   We are aware of this, but we have no further data that would enable us to elaborate further.

4) It would be interesting to know what the study designs of the 34 MEDLINE articles missed by HSSS(123) were. (Where they really RCTs or CCTs? If not, then the fact that they were missed is not really a failure of the search filter). If they are RCTs or CCTs, were there any methodological search terms that would have retrieved them?
We are also intrigued, and have identified a number of additional examples of studies that were indexed in MEDLINE but not retrieved by the search strategy used in the review. We intended to analyze and publish these data, but have not yet completed the work.

Referee 2
Reviewer: Carolyn J Green
Reviewer's report:
Discretionary
1. My comments are philosophical rather than technical as I found the paper well thought out, executed and written. Well done.

Thank you.

The following 3 comments have been addressed in a single paragraph:
2. To provide an alternative to the de facto gold standard I expect relative recall to be directly compared. As this wasn't done I suggest it be mentioned as a strategy for further studies. The paper would be improved by outlining how the conclusion could go from 'promising' to 'acceptable' for practice. The definition of “gold Standard” as the “best” available test is practically useful and to go off the gold standard should be done after careful consideration of the consequences. It does appear through this analysis that the consequences could be minimal but I appreciate that this is analysis is useful as a proof of concept and the conclusions appropriately couched.

4. A more efficient standard for developing/validating hedges/filters/search strategies would provide benefits by saving resources and making rigor more widely available to under resourced group. Therefore I think the paper would be more grounded by alluding to these benefits and limitations as part of the discussion.

5. Without a direct comparison I have remaining questions about how much more efficient this is comparatively. E journals make hand searching much more efficient than in the past. In an electronic age perhaps “hand searching” should be renamed “journal issue by issue searching. Still all these issues do not have to be resolved by this one study. Consider extending discussion.

These are excellent suggestions. We have added the following paragraph to the discussion: “While we propose relative recall as an alternative to hand-searching in the formation of a gold standard for search strategy development, our methods are indirect. A useful avenue for further study would be such a direct comparison between a standard based on the included studies of systematic reviews and one derived from hand searching (or, with the growth in availability of electronic full text articles, on-screen searching). The most useful comparison would not only examine the information retrieval characteristics of the two approaches, but would also compare the resources required to assemble the collections. “
3. The method that the authors have devised is indirect. The dataset provided by Cochrane reviews was developed with hand searching as part of the search strategy often and for validating the search strategies used. This makes the systematic review an appropriate denominator for exploring the usefulness of relative recall and therefore useful and innovative.

Thank you.

6. There is a good argument for using the true best gold standard to develop a filter that will be widely used. If clinically relevant trials are missed then the consequences of “going off the gold standard” could be consequential. The importance of relevance is therefore worth more explicit discussion. Searches that are informing public health policy and clinical practice warrant rigor and resources and so the emphasis of groups like Cochrane on exhaustive strategies to identify all available primary data is worth reinforcing.

This is a good point. We hope the material we have added regarding the need for a head to head comparison will be helpful here.

7. Hedges and the importance of validating are worth placing in the big picture of what systematic reviews can contribute to the optimization of human health. In my opinion it would ground the discussion. The work of Haynes and colleagues at McMaster on hedges is worth referencing.

We have made some changes to address this but a lengthy discussion is beyond the scope of this paper as most readers who would be interested in this subject likely have an appreciation for the significant contributions the Hedges team and those who have worked on developing and updating the HSSS have made in supporting the optimization of human health. We have changed the leading sentence of the Background from “Search filters or hedges play an important role in evidence-based medicine. They support both focused clinical searches and high recall searches done to build the evidence base for systematic reviews.” To “Search filters or hedges play an important role in evidence-based medicine. For example, work of the HEDGES team has enabled focused clinical searches on PubMed[3] and the original and recently revised highly sensitive search strategy[4] aid in building the evidence base for systematic reviews.”

8. Recall and precision though widely used have been critiqued. The work of Kagolovsky and Moehr are worthy contributions to this discussion.

Thank you, we have added the following paragraph to the discussion: “This paper has focused on recall and precision as the basis for evaluation of search performance. Numerous criticisms of these measures have been made, and they are well reviewed by Kagolovsky and Moehr[5] Never-the-
less, the information retrieval paradigm used in systematic reviews is classically suited to these measures. Retrieval occurs in batch mode, and although preliminary work may be exploratory and interactive, the final published search strategy is not. High recall and high precision are the ideal, with large retrieval sets being the norm. Retrieved documents are classified into a binary relevance scheme as eligible or ineligible for inclusion in the review. Finally, measures are taken to minimize the subjectivity or idiosyncrasy of the relevance assessment: the search result is evaluated against explicit criteria, often by 2 reviewers who must reach consensus, in order that the work could be independently replicated.

9. Generalizability for diagnostic studies is not good given that only 1 systematic review was available. Perhaps a limitations section would be useful in this paper.

Agreed. We have added this sentence to the discussion (p. 11): “Further, almost all reviews examined here studied intervention effectiveness. This technique may not generalize to searches for diagnostic reviews.”

10. The European literature is so rich that I think it is worth mentioning the importance of Embase in avoiding a North American centric perspective.

We have examined “Embase unique” trials in a cohort of Cochrane reviews with similar inclusion criteria and found such trials to be fairly uncommon.[6]

11. The category of 30 studies as an “other” category in table 2 looks like insufficient categorization. Consider labeling as a single groups as described in the text.

We assume you are referring to the originating Cochrane Group in Table 1. These are 20 Cochrane review groups that were represented by 1 or 2 reviews. The entry has been changed to “Review groups with 4 or more included reviews” The footnote indicates that the other groups are: Menstrual Disorders, Skin (3 reviews each), Effective Practice and Organisation of Care, Gynaecological Cancer, Heart, Hepato-Biliary, Hypertension, Lung Cancer (2 reviews each), Acute Respiratory Infections, Colorectal Cancer, Drugs and Alcohol, Ear, Nose and Throat, Fertility Regulation, HIV/AIDS, Metabolic and Endocrine Disorders, Movement Disorders, Oral Health, Pregnancy and Childbirth, Stroke (1 review each).

Other Changes
Some text was inadvertently deleted from the second paragraph of the results section in the previously submitted version. The text here in bold has been inserted

“Reviews were drawn from the 3rd Quarter 2002 issue of The Cochrane Library and most were of recent origin. The vast majority were reviews of treatment interventions.

Please note that one review classified as “other” should have been grouped with Musculoskeletal Injuries, thus MI now has 22 reviews and 29 are classed as “other”.”
We have made a correction on page 10. The sentence “NSSS is narrower again.” Has been corrected to “NBSS is narrower again.”

The Canadian Coordinating Office of Health Technology Assessment has changed its name to Canadian Agency for Drugs and Technologies. We have changed the affiliation of 2 the authors to reflect this change.

Reference List


