Author's response to reviews

Title: Boolean versus Ranked Querying for Biomedical Systematic Reviews

Authors:

Sarvnaz Karimi (skarimi@unimelb.edu.au)
Stefan Pohl (spohl@csse.unimelb.edu.au)
Falk Scholer (falk.scholer@rmit.edu.au)
Lawrence Cavedon (lcavedon@csse.unimelb.edu.au)
Justin Zobel (jz@csse.unimelb.edu.au)

Version: 2 Date: 17 August 2010

Author's response to reviews: see over
Response to the Reviewers Comments

We thank all the reviewers for the feedback on our work. We have done our best to incorporate all the comments as explained below. The revised or added texts in the paper are highlighted by blue texts, except for the typographical corrections.

Reviewer 1

1. Sections 1 though 3 of the manuscript are far too long – this background information should be presented in a more succinct fashion. The reader has to wade through far too much detail in the end losing the key messages.
Response: The introduction and background sections have been shortened substantially (around 1 page of text removed in total).

2. Findings of the study are presented in Section 1 which should be presented in the results and discussion sections of the manuscript.
Response: Detailed discussion of the findings has been toned down. However, we feel that it is important to give at least a rough indication of the major findings here.

3. Some of the references are dated (e.g., Dickersin et al, 1994 – Problems with Reported Search Strategies) - are these data still relevant? Searching has evolved since 1994 particularly with the addition of Clinical Queries (i.e., methodologic search filters) in Pubmed and Ovid Limits.
Response: The citation of Dickersin et al (1994), together with citations of work by Avenell et al (2001) and McGowan and Sampson (2005) which are provided in the same paragraph, demonstrate that the problem has already existed for quite some time but has not yet been resolved.

4. In the Problems with Reported Search Strategies section of the manuscript the authors indicate “this problem is partly due to the fact that the different resources have different implementations for parsing, indexing, and search documents.” The problem is also likely due to the less than 35% overlap in the journals indexed in Medline and Embase.
Response: We included this information in the revised text.

5. Some of the sections of the manuscript seem to be outside of the scope of the research question, for example, the extent of detail provided in Section 3.
Response: We intentionally include a detailed example of a search strategy to demonstrate the range of different operators available in such search systems, with which some readers may not be familiar. The section is also used to highlight the potential problems that these operators can bring to a search strategy in this context.

6. Sections 4 and 5 of the manuscript are difficult to follow because of too much detail, too many acronyms, figures not explicitly described (e.g., Figure 1 – not clear on how this representation relates to the Boolean search strategy presented), and table data being difficult to interpret.
Response: Again other reviewer asked for more details for Sections 4 and 5. To make the tables more clear, we added some more explanation to their captions. Figure 1 which was in Section 3, is now removed, and we also removed some of the acronyms.

7. The methodology described was difficult to follow and it was not clear what methodology was used to arrive at the data presented in the tables.
Response: All the tables and figures are referenced in the corresponding section that explained the corresponding method. In the revised version, we included more explanation to the methodology section (where it was too brief) and especially to the results (in the form of captions) to make them more clear.

8. A rewrite of the manuscript focusing on the research question and relevant background information, methods, and results would be worthwhile. As indicated earlier the research question is important and the results of this study should be published but in a more succinct, focused manner.

Response: We made the text shorter in the background, and added some more explanation to the methods.

Reviewer 2

1. p. 11/ the creation of systematic reviews typically includes synthesis (T2). This process goes beyond the identification of documents meeting the inclusion criteria. It was not clear if T2 is the last process of the search aspect of systematic reviewing or not.

Response: T2 represents final set of documents that end-up being in the review. Therefore, they have passed the synthesis step. In other words, T0, T1, T2 only represent the “set of documents” rather than “the processes”. We made this clear in the text (after their definition).

2. A related project to the one reported here is the ASSERT project (http://www.nactem.ac.uk/assert/). In this work, terms have been used as metadata. It would be useful to look into this work and compare the approaches. Reference "Supporting systematic reviews using text mining" by S. Ananiadou, B. Rea, R. Procter and J. Thomas (2009) Social Science Computer Review.

3. Other useful references:

Response:

We thank the reviewer for the suggestions of related work. However, these papers focus on applications of clustering and classification algorithms in the process of systematic reviewing, which is not central to our work. Moreover, Reviewer 3 strongly recommended shortening the background sections, and we have therefore not incorporated these additional citations at this time.

Reviewer 3

Major Compulsory Revisions (which the author must respond to before a decision on publication can be reached)

1. The paper provides a clear exposition of the current approaches to searching for evidence to inform systematic reviews. As this paper seems to seek to span two audiences - one familiar with Boolean searching and one more familiar with non-Boolean approaches - it would be helpful if the same detail and care used to describe the Boolean approach was also applied to non-Boolean. If this can be added then readers will be better informed. I think this can be achieved quite easily by providing more detail on how the ranked query approach works, the software used (Zettair) and how queries are entered. In particular an explanation of how the search elements in Figure 2 are operationalised to achieve a search would be helpful. I can see things are differently presented from a Boolean search but I don’t know how these search elements are put into Zettair
and what Zettair does with them. The authors should tell us more about the Okapi BM25 similarity algorithm and why it was used in preference to any other algorithm.

Response: We added more details to ranked retrieval section of to the revised paper. However, unlike Boolean queries that have very specific method of entering query and running the system, Zettair and other ranked retrieval systems are quite simple. They work similar to Google search. Therefore, there is not much to say about how querying works. We added more information to Figure 2 to make it clear.

2. P9 para 1, and elsewhere in the paper. The use of the term ‘metadata’ is not clear and the meaning of metadata with examples should be defined early in the paper. It has unfortunate confusions for me with the metadata we expect to be added to webpages, and so the term seems unhelpful at present and the authors may wish to consider a term with fewer additional connotations.

Response: We defined “metadata” in its first use in the revised paper. It is also explained in the experimental section again.

3. Page 17, para beginning ‘As a first step’. The authors describe the failure of the documented searches (in MEDLINE?) to retrieve the records of documents in the review. This is important to explore further, if I have understood correctly what the authors are reporting. Systematic reviews that result from numerous searches in a range of databases will often include records which are not indexed in MEDLINE (as MEDLINE is itself not comprehensive). It is also the case that records which are indexed and available to be found in MEDLINE are not retrieved by the MEDLINE searches used for a review. Those records are often identified at the end of the review when checking back and will have been retrieved by searches of other databases with different indexing practices. Do these two explanations help to clarify the situation described in this paragraph? If these explanations don’t help then, this makes the record set being used for the research rather doubtful, as it is not being retrieved by the searches which were recorded as retrieving some of it.

Response: Thanks for the clarification. We should note that we only considered those references (as evaluation baseline) that were indexed in MEDLINE. We added the reviewers point on having references found in other methods to the revised paper.

4. On p21 the authors again note inaccuracies in the original review search strategies. It would be useful for the authors to discuss how far they trust their own data set and to report on whether they found the strategy structure, as well as the terms used in the strategies, to be adequate for the reviews’ purposes. There is now a checklist for assessing search strategy quality (PRESS) which might be helpful to refer to even though it may be too late to assess these reviews against it (http://www.cadth.ca/media/pdf/477\_PRESS-Peer-Review-Electronic-Search-Strategies\_tr\_e.pdf).

Response: Thanks for pointing out the reference. We did not claim on having found the perfect queries, but we recommended using more relaxed Boolean queries and ranked queries together to help to facilitate retrieval of more relevant documents in one search.

5. Relative recall gold standards such as the one compiled for this research are potentially useful but the authors may have encountered some of the issues which can undermine their use – a discussion of these issues would be useful.

Response: Using a gold standard for the evaluation of IR systems is a standard part of the Cranfield paradigm. We discuss some issues relating to the gold standard on page 9 (in particular, the prospect that there are additional relevant documents available that should have been included in a comprehensive systematic review, but were missed). In the interests of brevity, we do not include a more general discussion of philosophical issues surrounding the Cranfield paradigm.

6. P20 The discussion of explosion needs to be improved – We need to know definitely whether the Ovid implementation does or doesn’t explode all terms below the main heading. Then the discussion will have a firm basis.

Response: We did not implement query expansion for our simplified queries to avoid any mis-interpretation
of actual Ovid implementation. In the paper it says “we decided to remove these operators from the queries.” which implies that.

7. Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct):

p.5 There is a focus on the exhaustiveness of systematic review searches, but it would be fair to say that the ideal of exhaustiveness is giving way a little these days to a more pragmatic (resource-limited) approach – so the authors may wish to reflect this. For example, the Cochrane Handbook which guides the production of what are often referred to as the most rigorous and extensive systematic reviews acknowledge that searches can rarely be exhaustive or comprehensive (see Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for studies. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 (updated September 2009). The Cochrane Collaboration, 2009. Available from www.cochrane-handbook.org.)

Response:
We acknowledge that exhaustiveness can never be assured (without reading every document in the collection). This is one of the points that our paper is making. However, the referenced Cochrane handbook states: ”Searches should seek high sensitivity, which may result in relatively low precision” (chapter 6). We therefore do not feel that we are mis-representing the very high recall-focused aims of systematic reviews.

Jesse McGowan’s surname is spelled incorrectly through the paper and references.

Response: Fixed.

P5. Golder researched systematic reviews of adverse effects, so it would be helpful to add this small but important detail.

Response: We had to take whole paragraph out, to make the text shorter.

P6 Although many searchers do access their database via the Ovid interface, many do not – so this statement should be amended.

Response: We had a footnote that explains another popular interface is PubMed. We included that in the text now to make it more visible.

P8 para 1. It would be helpful to briefly mention other reasons why searches are recorded in detail, as these are probably more important than exact reproducibility at some future point in time. Searches are reported to allow the reader to make an assessment of the extensiveness of the search and the likelihood that it would capture relevant records for the review, and hence the likelihood of reducing bias in the review.

Response: We had four reasons listed for reporting the search strategies in the paper. The reviewers comment is already in the paper just after reproducibility.

Page 9 para 1 This paragraph is difficult to follow and this may be partly because of the use of the term “metadata”. Some examples of new metadata fields might clarify the point. However, it would be fair to say that it is generally acknowledged in the systematic review information retrieval community that databases don’t stand still and that searches will be unlikely to be rerun exactly as they were first performed. The reason for reporting searches is not just for rerunning but to allow an assessment of likelihood of bias and recall.

Response: We explained metadata in the begining of the revised paper to make it clear.

Page 9. Para beginning “Within the context”; last sentence. When developing strategies for identifying studies searchers will perform various explorations using different combinations of terms and can NOT out the records they have seen previously, and can show which records are removed by an approach. So there are ways to judge whether altering a query leads to an improvement in the answer set and these are routinely used. The real issue is that these are time consuming.
Response: We appreciate this information. The point we wanted to emphasise is that there is not quick way to judge the results. We added the “time-consuming” factor to the revised version.

Table 6. The changes to the search term limits and combinations have enormous impacts on the precision and sensitivity of the Boolean search: the authors describe this as simplified but it would also help to explain how “expanded” it is as well. For example, replacing “adj n “with “OR” replaces an implicit focused “AND” with an OR, meaning that sensitivity expands enormously and precision drops. Ditto replacing Boolean AND with “OR” in the same table. The Boolean approach has been deconstructed and original intentions reversed.

Response: We intentionally sacrificed precision here to gain more recall. That helped us to have more relevant documents in the pool as shown in Figure 5. Originally Boolean queries were too strict and therefore some of the relevant documents were not found. Then we run ranked retrieval to get the precision back.

P18. Last para – “specificity” is misspelled and the first sentence of the para lacks punctuation.
Response: Fixed.

10. References: I think “Driana Yoshii” may be “Adriana Yoshii”?
Response: Fixed.

Reviewer 4

Minor Essential Revisions

1. In Section 3 Search Mechanisms for Preparing Systematic Reviews, in the first example query, line 12 reads:

   12 (child$ or boy$ or girl$ or schoolchild$ or adolescen$ or teen$ or "young person"$ or "young people"$ or youth$)

   It will be useful to provide an explanation for the brackets for those who are not familiar with complex Boolean queries.

Response: For that particular line of query having brackets or not having them would make no difference in the output because no prioritisation was imposed by brackets. We removed them from this query line not to confuse the readers.

2. In Section 4.1 Experimental Setup: Data and Measurement, in the list of 3 steps in the creation of systematic reviews, the first step is: 1. First, an initial search strategy is employed to retrieve a large pool of potentially relevant items from databases such as medline; Please describe item: is it a document with abstract + title?

Response: Replaced “items” with “citations” to make it clear.

3. In the legend for Table 1 say how Tier 0 statistics for the Drug Dataset were obtained.
Response: We specifically mentioned Misc dataset in that caption because that was the dataset we built ourselves (therefore had no access to the original Boolean query outputs) in contract to the drug dataset which is a dataset provided by other researchers. To clarify this we modified the text in the caption.

4. Other Minor Revisions
   - page 2 bottom paragraph
     researcher – is to use ranked retrieval. Ranked retrieval aims to order a collection of document researcher – is to use ranked retrieval. Ranked retrieval aims to order a collection of documents
searched. Golder et al. argue that the main reason for these problems is the lack of empirical

A number of different evaluation metrics have been proposed for the measurement of search

modeling a patient user who is willing to review an extensive answer set; this reflects the current

Similarly to the retrieval experiments from full medline, trc queries perform better than T

In the legend of Table 4

A, metadata is indexed separately; Method B searches metadata separately, but also includes it

sorting capabilities based on specific fields (e.g., date), we tried to estimate the effort of finding

bag-of-words queries (i.e., the simple queries used in ranked-retrieval) can be improved significantly

We believe this to be a key approach to facilitating higher-performance search for the purpose of

Response: All these typographical mistakes are fixed now.