Reviewer's report

Title: Faster title and abstract screening: evaluating Abstrackr a semi-automated online screening program for systematic reviewers

Version: 1  Date: 7 March 2015

Reviewer: Josiah Poon

Reviewer's report:

Major Compulsory Revisions

1) The authors used four datasets to evaluate Abstrackr in the paper. No rationale or justification why these datasets were chosen. If the aim is to evaluate the system, then this evaluation is not systematic. A proper test plan should be created where each dataset has certain characteristics that is different from the others, and how each test can validate what functions/claims of Abstractr. Simply running a system four times (if having similar characteristics) is not a proper evaluation. It seems the current four datasets were used because of its convenience and availability.

2) Although there was a bit of diversity of the systematic reviews chosen (three of the four datasets are all about treatment effectiveness and one diagnostics accuracy), a more in-depth study of intra-comparison of the same type of reviews and inter-comparison of different types of reviews.

Minor Essential Revisions

3) Line 3 of page 7, I meant you meant to refer to Figure 1, NOT figure 3.

4) Although this paper is not meant to be a technical paper, I still reckon it is desirable to briefly describe Abstractr, its goals, its functions, and to give reasons why we have to evaluate this Abstractr, and not system ABC. Is it considered state-of-the-art? Is it “talk of the town”?

5) Detailed discussion on the observations required, e.g. it seems the system worked really well for Rituximab review, best precision and no missing paper, however, according to workload saved, it came last, WHY?

Discretionary Revisions

6) p8, 2nd paragraph, 6 lines from the bottom: ECHO dataset is an imbalanced dataset and, therefore, had the lowest precision, so is Abstractr still a good system? I still think many reviews will start with a high imbalanced situation. I heard that it is a saving of 80%, however, it is not only the percentage but also the actual number of articles to go through, and it still required a user to read more than 3000 papers! What recommendation will you offer? to the user? and to the developer of Abstrackr?
**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests