Author’s response to reviews

Title: Software Tools to Support Title and Abstract Screening for Systematic Reviews in Healthcare: An Evaluation

Authors:

Hannah Harrison (hh504@medschl.cam.ac.uk)
Simon Griffin (ProfGP@medschl.cam.ac.uk)
Isla Kuhn (ilk21@cam.ac.uk)
Juliet Usher-Smith (jau20@medschl.cam.ac.uk)

Version: 1 Date: 15 Nov 2019

Author’s response to reviews:

Please see attachment "Response Letter" in which we have colour coded our response.

Alternatively, the response is below:

Reviewer 1 (Ba’Pham, MSc, PhD): Comments to the author
The authors A) conducted a reasonably comprehensive search to identify software tool candidates that support titles and abstracts screening; B) assessed the candidates for eligible tools; C) designed, implemented and carried out a feature analysis of the eligible tools; and D) piloted a set of promising tools to elicit the users' experience. The methods and results are reasonably well described and overall, the manuscript is well written. I however have a few specific questions.
Overall question: Did the authors come up with the items A-D that I could delineate above as part of a study protocol from the beginning?

We planned the four stages of the study at the outset (searching, screening, feature analysis and user survey), we have added a sentence to the beginning of the methods section outlining the overall structure:
“The study had four stages to identify and evaluate the suitability of currently available software tools to support T&Ab screening. The stages are: a search for relevant tools, screening for suitability, a feature analysis and a user survey.”
There is a flow diagram (figure1) showing the number of tools at each stage at beginning of results section.

Item B: Were the selection criteria developed by one author or all three authors?
The selection criteria were developed by three of the authors, the text has been amended to reflect this: “A list of five criteria for inclusion was developed by three of the authors (HH, JUS and SG);”
Item C: Why did only one of the three authors involve in the design of the feature selection?
One author designed the feature selection that was then reviewed by a group of researchers (see response below).
One researcher developed the assessment criteria; this was then reviewed by a second author. The text has been amended to better reflect this process.
“A single evaluator (HH) devised assessment criteria - which can be found in the supplementary materials (S1). These were reviewed by a second researcher prior to the collection of information and feature scoring for each tool.”

What was the process to consult with medical researchers involved in systematic reviews? How many of the researchers did the authors consulted?
As part of the discussion group, which developed the feature weighting (see paragraph below), 7 medical researchers were consulted on a list of potential features and invited to add additional features to the list. Once a consensus had been reached, the group then voted on the relative importance of the different features – the results of this vote was used to develop the weighting (in combination with the opinions of the two interviewed researchers). The text has been amended to clarify this:
“…, as well as consulting medical researchers involved in systematic reviews. Five researchers participated in a discussion group during which a list of potential features was presented; the researchers were invited to revise, remove and add features to this list until a consensus was reached.”

Was the assessment of the eligible tools with respect to each of the selected features conducted by one author?
Yes, one author (HH) carried out the assessment of the tools, using the list of selected features. This is indicated in the text:
“A single evaluator (HH) devised assessment criteria - which can be found in the supplementary materials (S1) - compiled the information about each tool and scored each feature for each tool.”

Did the authors verify to see if the ranking of the tools might depend on alternative coding of the feature scores (e.g., 1-3 instead of 0-2) and weighting scores (e.g., 1-5 instead of 0-4)?
We considered carefully how to code the different feature scores. If you change one (i.e. change from 0-2 to 1-3) but leave the other as 0-4 then you change the outcome (the ranking), as you are giving some points where none were awarded previously. However, this would result in giving “points” for non-implementation of a feature, which would be inflated if it were of high importance to the researchers. If you change both (1-3, and 1-5) then the ranking is not altered, alternatively, if you normalise both scales to 0-1 the ranking is not altered. We specifically chose to award 0 points for “non-implementation” and for features that were considered “irrelevant” by the majority of researchers.

Additionally, we realised we have not made it explicit that we drew the weighting scheme from a previous study. We have now clarified this in the text (methods):
“A similar weighting has previously been developed for the assessment of systematic review support tools within software engineering [8].”
Item D: How did the authors identify the sample of the six researchers? Why were these six researchers selected? Did the authors take steps to ensure a sample of researchers with different experience in using tools that automate titles/abstracts screening?

We initially used a snowballing approach to recruit survey participants, then we purposefully selected a range of different experience levels and career stages. Of the eight researchers we approached six agreed to take part in the study. We have amended the text (methods, user survey) to give more detail of this process.

“Potential survey participants were recruited using a snowballing approach, from which we were able to select a range of experience levels and career stages. Eight researchers were approached and six agreed to take part in the user survey.”

However, as shown by Table 2, all our survey respondents had some experience with these types of tools, this potential for bias is acknowledged in the limitations section:

“Additionally, all of the respondents had some previous experience with either Rayyan or Covidence and this may have resulted in a bias in favour of these tools.”

I generally agreed with the discussion, with a few questions. Some considerations regarding the performance of the tools with respect to sensitivity and precision are needed, especially if the authors wanted to recommend some tools over the others. The authors need to confine the way they make sense of the results from the perspective that their study is confined to usability of these tools - data pertaining to the performance characteristics of the tools are also important in selecting a tool for use in practice.

The conclusion and recommendation would need to be from the perspective that this study is about accessibility and usability of the tools (not taking into account how the tools perform the task with selecting eligible abstracts with high sensitivity and high precision).

Thank you for raising this point, we agree that the study would be improved by making our aim of assessing usability rather than the accuracy/performance of the tools clearer. We have amended several sections of the paper’s conclusion to reflect this:

Abstract (background): “In this study, we identified and evaluated the usability of software tools that support T&Ab screening for systematic reviews within healthcare research.”

Abstract (results): “Their usability scored highly across a range of metrics, with all surveyed researchers (n=6) stating that they would be likely (or very likely) to use these tools in the future.”

Abstract (conclusions): “Based on this study, we would recommend Covidence and Rayyan to systematic reviewers looking for suitable and easy to use tools to support T&Ab screening within healthcare research. These two tools consistently demonstrated good alignment with user requirements.”
Background (3rd paragraph): This scoping review aims to identify, describe and evaluate the usability of the available software tools that support the T&Ab screening process for healthcare research to enable researchers to select the most appropriate for their work.

Discussion (key findings): “The six highest scoring tools were trialled by a group of six healthcare researchers with experience of systematic reviews. Out of all the software tools considered, Covidence and Rayyan emerged as the most suitable tools to support T&Ab screening for systematic reviews in both the feature analysis and the user survey.”

Discussion (strengths and limitations, 3rd paragraph): “Collecting both quantitative and qualitative data in the user survey improved our understanding of the usability of the screening tools.”
Discussion (strengths and limitations, 4th paragraph): “There are several limitations of this study, which should be considered when interpreting its findings. This study only considers the features offered by each tool and the user experience they provide. There has been no assessment of any other performance measures, such as the accuracy, sensitivity or reliability of these screening tools. Readers who are considering using the tools discussed in this study should consider these properties in addition to considering the user experience.”

Conclusions: “The results of this study suggest that Covidence and Rayyan provide the best user experience for systematic reviewers carrying out T&Ab screening.”

Reviewer 2 (Brian Hutton): Comments to the author
In this article the authors have performed an assessment of available software tools for title and abstract screening in systematic reviews, a clearly vital component to the systematic review process. The authors have looked at a feature assessment (to assess availability of key components considered important for tools of this nature) as well as a small user survey across tools. I have provided thoughts below regarding the manuscript.

General comments:
The authors have presented a generally clear and readable document summarizing their approach and findings. I think spelling out the target audience for this work (it would seem to be focused not toward systematic review producers on a grander scale but perhaps junior researchers new to SR and perhaps graduate students gaining exposure to knowledge synthesis?) would be beneficial as I do not believe this will be of interest to all researchers who perform reviews.

We are pleased that the reviewer considers our manuscript to a clear and readable document. With regards to their suggestion to spell out the target audience for the work, we have added a line to the introduction highlighting the readers who we think would be most interested in this study.

“This work may be of particular interest to researchers who are new to systematic reviews, looking to change their approach to screening or those in the position of selecting an appropriate tool for a collaboration.”
I feel that in not considering some of the more well known tools and some of the significant gains that can be acquired in adopting platforms that support the entire review process and which may contain additional highly relevant capabilities (advanced data collection, artificial intelligence capabilities, data security and other elements), the study is perhaps not as relevant and pertinent as it could be for a wider audience. Addressing this as a limitation of the study in the discussion section would also be worthwhile, perhaps this could be a means of focusing to whom the article is geared.

We have mentioned in the text why we have not been able to consider the tools DistillerSR, EROS, SluRp and SLR-tool:

“For two tools, DistillerSR or EROS, it was not possible to obtain a free trial, so these were not investigated further. Additionally, SluRp and SLR-Tool were excluded as they required the setup of an SQL server.”

Other than this, we believe our search to have captured most of the available tools (see the supplementary file providing a full list of the software tools). However, some tools – which people using systematic reviews may be familiar with – were not captured by our search (such as RevMan) or were excluded by our inclusion criteria (such as JBI SUMARI as it does not support full text screening). We have added a line to our search results, explicitly stating why this type of software was not included:

“The most common reason for excluding a tool from the study was that it did not have T&Ab screening functionality, for example the tool JBI SUMARI only supports full text screening and RevMan5 does not provide any screening functionality beyond what is offered by a reference manager.”

Additionally, several of the tools we considered were platforms from which users could manage the whole systematic review process. We have added a line to the search results section clarifying this variability:

“The included tools ranged from those providing a basic system exclusively for T&Ab screening (for example Abstrackr) to platforms able to offer support for several stages of the systematic review process (for example EPPI-reviewer).”

This information is also shown in Fig.2 (traffic light diagram), where the section on “screening support” (see table 1 for key to the codes used) indicates other stages of the systematic review process supported by each tool. Additionally, although we are not evaluating the additional capabilities mentioned in the comments above (e.g. AI capabilities) – several of these features are also indicated in the traffic light diagram (e.g T7-F8: indicates machine learning (AI) ranking capabilities and T8-F1: indicates website security). The collection of information portrayed by this figure could be used to help guide decision-making processes for researchers with a range of different needs.

We agree that considering a review platform tool (e.g. Covidence/EPPI-reviewer) alongside a single purpose tool (e.g. Abstrackr) is not a perfect comparison, however, as the screening
process can dominate the time spent on a systematic review we feel that considering this element in isolation is valuable in and of itself. To reflect this in the paper, we have added a line to the limitations section to indicate that this study does not consider the wider benefits/difficulties of adopting a single platform for the whole process, as we have instead chosen to isolate the T&Ab screening stage.

“The tools identified by the search displayed considerable heterogeneity, which makes drawing comparisons between them more difficult. Additionally, in this study, we have chosen to consider the T&Ab screening stage in isolation. While this means we were able to compare a wide range of tools that offer that function, our finding do not consider the potential advantages or disadvantages of using a simple tool just for T&Ab screening or a platform that supports multiple stages of the systematic review process.

The small nature of the user survey and the close ties of respondents in terms of setting also do not carry great weight.

Given the size of the user sample, we have taken several steps to try to limit the bias. This included:
- Recruiting researchers with a range of experience levels/roles (Table 2)
- Using a predesigned spreadsheet for survey responses, which survey participants were required to fill in independently of each other.

We have acknowledged the bias introduced by the close ties of the participants in the discussion section:
“The six respondents were all drawn from the same research community and all but one work in the same department.”

We have also added a line to this section specifically addressing the small sample:
“The number of participants in the user survey was small (n=6), therefore caution is required when interpreting the findings.”

However, given the lack of research in this area to date, we consider the survey findings to be informative to researchers in similar environments.

I am unfamiliar with the approach they’ve used to consider weights in the analysis, and thus perhaps other reviewers will be able to comment on concerns they might have (if any).

We hope this has been addressed in our response to reviewer 1 (see above).

Overall I did not consider the study to present details that will be of high interest to a wide range of readers.

We thank the reviewer for the time spent considering this paper. While we acknowledge his points regarding the various limitations of this work we believe that this study is of value to healthcare researchers. This is the first review of this kind within healthcare research, and it was designed so that the needs of these researchers were considered at every stage. Whilst the findings may not be applicable in different research areas or in settings with different resources,
many healthcare researchers - often at early career stages - carry out systematic reviews and we feel that the findings are relevant to them.

Minor comments:

- in the abstract, I believe the authors mean to mention December 2018 rather than December 2019.

Now corrected

- I would perhaps spell out the DESMET abbreviation within the methods section so this is more clear. We have added this to the methods section:
  “As part of the DESMET method (a methodology for evaluating software engineering methods and tools), ..”

- in Table 2, adding headers to columns 1 and 2 would be worthwhile, perhaps 'Respondent characteristic' and 'Categories' or something along those lines. Thank you for the suggestion, we have added headers to the columns in table 2.