Reviewer's report

Title: Machine learning algorithms for systematic reviews: reducing workload in a preclinical review of animal studies and reducing human screening error

Version: 1 Date: 11 Nov 2018

Reviewer: Jennifer Stone

Reviewer's report:

Overall comments

This is an interesting and useful article for systematic review researchers interested in fast-tracking the systematic review process through use of semi-automated methods or correcting their article screening process. The article does contain small grammatical errors which require closer inspection and revision prior to publication. These have been outlined below. The authors have addressed previous reviewer comments reasonably well.

The main criticism of this article is the results are overstated given the poor precision estimate and greater transparency around what these results mean and how they can be interpreted is required so users of these ML approaches can know exactly how the algorithms perform with their data. This is further addressed in specific comments below.

Title and abstract

- Authors may update their results and conclusions in light of the comments regarding the precision estimate detailed below.

Introduction

- Line 108: "Sources of systematic error…” sentence needs rephrasing.

Methods

- Line 120: "Because we did could not…” grammar needs correction.
- Line 135: The NRI is not a very useful number, better to use AUC.
- Line 216: "…involving text.)" remove end bracket.
- Line 254: "Here, the algorithm needs…” no need to repeat this sentence. Already written on line 249.
- Line 291: "cut-off for were determined..." grammar needs correction.

- Table 1: Accuracy and WSS@95% are not useful measures.

- Table 1: Authors need to make sure readers are very clear on what the precision measure means as this is a very important measure practically, more important than sensitivity and specificity. It is the post-test probability of true inclusion. Sensitivity and specificity are constant whereas precision changes due to change in inclusion prevalence, which is very relevant for systematic review researchers, especially if use of these algorithms are later extended beyond broad and shallow SRs to focused and detailed SRs.

Results

- Line 345 "...has been reach..." grammar needs correction.

- Table 2: The precision estimate is 50% and is then bumped up to 55.9%. These are both low estimates and completely dependent on the inclusion prevalence.

This means if a reference is included there is a 50% probability it is a correct inclusion. For every 1 reference correctly included there is 1 reference incorrectly included. Even with 13.2% inclusion prevalence, half the references included will be irrelevant. This means half of the included references will incorrectly go on to full-text screening/ data abstraction.

If the inclusion prevalence drops, which is likely to be the case for many systematic reviews, the precision estimate will drop even further. For example, if the inclusion prevalence drops to 5%, the precision estimate will be 27% which means for every reference correctly included you will have ~2 false inclusions. That is 63% of the included references incorrectly going on to full-text screen!

Is this then really a useful ML algorithm if half of the included references go on to full text screen? If you have 10,000 references with a 13.2% inclusion prevalence you will be including 26% of all the references in your search so a human has to go through 2600 references at full-text screen/ data abstraction. This is a dismally low precision estimate.

- Likelihood Ratios (LR) can be used to directly compute the post-test probability at different levels of inclusion prevalence and would illustrate this well for readers. E.g.,

\[
LR = \frac{\text{sens}}{1 - \text{spec}} = \frac{0.987}{1 - 0.86} = 0.987/0.14 = 7.05
\]

Using the 13.2% inclusion prevalence:

Convert inclusion prevalence to odds = 13.2% = \(\frac{0.132}{1 - 0.13} = 0.132/0.868 = 0.152\)

Multiply by LR to get post-test odds - 0.15 x 7.05 = 1.07

Convert to precision estimate - \(\frac{1.07}{1.07+1} = 52\%\)

Now choose a different inclusion prevalence of 5%:

Odds \(\frac{0.05}{1 - 0.05} = 0.05 \times 7.05 = 0.371\)
Convert to precision - \(0.371/1.371 = 27\%\)

This means at a 5\% inclusion prevalence the precision estimate drops to 27\%. As the prevalence drops there is a marked decrease in precision.

Authors may wish to display different prevalence estimates on precision in a graph to explain how this impacts systematic review researchers classifying references using these ML algorithms.

Discussion/ Conclusions

- Line 430: Full stop mid-sentence needs correction.

- Line 466: "…discrepancies might be a pinpoint…" grammar needs correction.

- Authors need to correct the way they interpret ML outcomes with specific attention to the precision estimate. The authors need to discuss that the characteristics of ML algorithms are still poor and adjust their recommendations in light of this evidence. It should be discussed that the precision estimate is not really a good result in terms of time restraints for systematic reviewers (as has been outlined in the introduction of the paper as the rationale) and should be made clearer that these algorithms may only be used to correct human error or as a semi-automated method at this stage.

- Authors may wish to look into approaches for screening rare diseases using sequential testing and apply this to ML in a subsequent paper - applying approach 2 after approach 1 for example may increase specificity (but may not retain the sensitivity you require).

Level of interest
Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

Quality of written English
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published
**Declaration of competing interests**

Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

No competing interests to declare

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal

Were you mentored through this peer review?

No