Author’s response to reviews

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

Authors:
Alexandra Bannach-Brown (a.bannach-brown@ed.ac.uk)
Piotr Przybyła (piotr.przybyla@manchester.ac.uk)
James Thomas (james.thomas@ucl.ac.uk)
Andrew Rice (a.rice@imperial.ac.uk)
Sophia Ananiadou (Sophia.Ananiadou@manchester.ac.uk)
Jing Liao (jing.liao@ed.ac.uk)
Malcolm Macleod (malcolm.macleod@ed.ac.uk)

Version: 2 Date: 10 Dec 2018

Author’s response to reviews:

Dear Dr. Florez,

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

We are very grateful to you are your reviewers for your thoughtful and helpful suggestions. We believe we have been able to substantially improve our manuscript as a result. The changes we have made are described in detail below

Dear Dr Nunez-Mir,

Thank you for your helpful comments. We have taken your comments that the manuscript could be split into two and decided to condense the ‘Feature Selection and Classifier’ information and move the additional information to a supplementary file. We have amended the manuscript so it


is more succinct and reads better. We have reflected on the suggestion to split the manuscript into two, but would much prefer to keep the manuscript together because we believe the machine learning methods implementation provides the backbone and context for the error analysis. We have amended the aim to better reflect the goal of the paper and have made further edits to the paper that we believe improves the clarity. Many thanks.

Reviewer 1: Gabriela C Nunez-Mir, Ph.D.

Major Revision

In this study, the authors test the performance of two ML approaches in selecting classifications for inclusion in a systematic literature review. This study is timely, interesting and necessary, as it attempts to resolve two important issues in literature reviews—the massive amount of time and human resources needed to sift through the ever-growing corpora of literature and the unavoidable human bias and error. Although the study itself appears well-planned and nicely executed, the manuscript has a few shortcomings despite having gone through a round of revisions. Addressing these will greatly elevate the manuscript by making it much more attractive to a broader audience of researchers. First, there are instances in the manuscript where the writing feels a little alienating to the reader. As mentioned by previous reviewers, abbreviations for institutions and tools are used throughout the manuscript. These have been spelled out in the new version, but as someone unfamiliar with these institutions, I fail to see the purpose of mentioning them so often, especially since this paper has the potential to be useful to a very broad audience.

• ABB: Thank you for this comment. The primary aim of mentioning these institutions and tools is to highlight the integration of this tool into already established freely-available user-friendly software. We think that integrating software and tools is a key challenge in maximizing the benefits of machine learning tools in research.

In a similar vein, a tool is mentioned towards the end of the manuscript—the Systematic Review Facility tool (SyRF). It is unclear if this tool was used to perform the feature generation and classifying in the study. If this is the case, and if the authors want to encourage the use of this tool, SyRF should be introduced and described much earlier in the manuscript.

• ABB: Thank you for this comment. This SyRF tool was not used to conduct the feature generation. However, in the discussion we mention the future research goal of integrating the automation tools we describe in the paper with existing systematic review platforms such as SyRF. This section of the discussion has been reworded to highlight this.
Second, the manuscript addresses two separate issues—using ML approaches to select citations and using ML to identify and correct human error in selecting citations. I believe both issues are worthwhile research topics; however, it may be too daunting of a task to address both in the same paper. I believe this is the reason why previous reviewers expressed confusion regarding the study's methodology. In the Methods section, the authors separate both goals into two "Steps." At the beginning of Step 1, the authors dedicate a couple of manuscript pages to describe different feature generation and classifier methods. These sections seem out of place in the current format of the manuscript, but they are not without merit.

• ABB: Thank you for your comment. The section on feature generation and classifier methods has been moved to a supplementary methods section with a summarized version retained in the main manuscript. We hope this improves the flow of the manuscript.

In my opinion, this manuscript would benefit from being separated into two manuscripts. The first manuscript would explore different feature generation methods and classifiers, and then test how different combinations of these perform (in the current manuscript, the authors do not clearly explain how they arrived at the feature generation/classifier combinations they used as ML approaches). The second manuscript would explore the issue of using ML approaches to detect and correct human error (this part is underdeveloped in the current manuscript). If the authors choose to keep the manuscript as one, I would suggest paring down the "Feature Generation" and "Classifiers" sections and using the text in these sections to describe the two approaches in the "Approaches" section.

• ABB: Thank you for your suggestion. The application of the error analysis technique is heavily reliant on the implementation of feature generation and classification. Therefore, and after some reflection, we have chosen to keep the manuscript as a single report. We have taken on board your suggestion to reduce the feature generation and classifier selection by moving this information to a supplementary section to improve the readability of the paper.

Specific comments:

• ABB: Thank you for assessing the manuscript closely. We have made changes to the manuscript with regards to all your points below. We have addressed points more in-depth where necessary.

L24-26: Run-on sentence. Please divide in two.

• ABB: Corrected

L39: Switch "assigned" and "the."

• ABB: Corrected
L57: Add comma after "intensive."

• ABB: Corrected

L99: The goal as stated in this sentence does not accurately describe the purpose of the study. I don't believe the authors identified the amount of training data needed to use ML algorithms effectively.

• ABB: Thank you for this comment, we agree, this was not achieved in the paper – we have changed the wording to reflect this.

L107: With "systematic error" do you mean human biases/error?

• ABB: Corrected

L113-114: "as part of a preclinical systematic review framework at the classification stage" this phrase is confusing. You may simply say "as part of the citation classification stage in a preclinical systematic review."

• ABB: Corrected

L119: You mean screening of a large body of literature FOR systematic review. The review itself wasn't screened.

• ABB: Corrected

L142: Period after EMBASE.

• ABB: Corrected

L169: This sentence is vague and unclear.

• ABB: Corrected

L312: Five consecutive records seem very low.

• ABB: We chose 5 consecutive instances where the human and machine decision was discrepant as a pragmatic approach in order to reassess the entire dataset. We highlight in the Limitations that this is an initial pilot and further work is needed to expand and develop the error analysis methodology.

L365: How is human error determined? This is an important question the authors need to answer. Once incongruences are found in human and machine decisions, how did the authors confirm that the human had made an error? Was an additional reviewer brought in to inspect the record?
Dear Dr. Stone,

Thank you for your comments and added value which would come from the addition of likelihood ratio. We have added these calculations throughout and updated the conclusions to reflect this. We have added a figure to represent the precision at different levels of inclusion with this likelihood ratio for readers to get a better understanding of applying this algorithm to different reviews. We have made edits to the paper that we believe improves the clarity. Thank you.

Reviewer 2: Jennifer Stone

Minor Revision

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.

• ABB: Corrected
Methods  Line 120: "Because we did could not…" grammar needs correction.

• ABB: Corrected

Line 135: The NRI is not a very useful number, better to use AUC.

• ABB: Both metrics have been used. As below, we have decided to be inclusive and report several measures so readers can compare between tools.

Line 216: "…involving text." remove end bracket.

• ABB: Corrected

Line 254: "Here, the algorithm needs…" no need to repeat this sentence. Already written on line 249.

• ABB: Corrected

Line 291: "cut-off for were determined…” grammar needs correction.

• ABB: Corrected

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

• ABB: There are many metrics on which to measure performance. We have decided to be inclusive and report several performance metrics so readers can easily compare tools from different publications.

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.

• ABB: Thank you for highlighting this. We addressed concerns of low precision below.

Results  Line 345 "…has been reach…” grammar needs correction.

• ABB: Corrected
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 = (sens)/(1-spec) = 0.987/(1-0.86) = 0.987/0.14 = 7.05 Using the 13.2% inclusion prevalence: Convert inclusion prevalence to odds = 13.2% = 0.132/(1-0.13) = 0.132/0.868 = 0.152 Multiple by LR to get post-test odds - 0.15x7.05 = 1.07 Convert to precision estimate - 1.07/(1.07+1) = 52% Now choose a different inclusion prevalence of 5%: Odds 0.05/(1-0.05) = 0.05 x 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.

• ABB: Thank you for highlighting the likelihood ratio calculations. This has been added to a supplementary section and described throughout the paper. We addressed concerns of low precision below.

Discussion/ Conclusions Line 430: Full stop mid-sentence needs correction.

• ABB: Corrected

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

• ABB: Corrected

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.

• ABB: Thank you for this and other comments above regarding the use of the likelihood ratio. We have highlighted throughout the discussion the implications of generalizing the use of this algorithm beyond the types of reviews carried out here and elsewhere within the CAMARADES collaboration. We have added a diagram to the supplementary files where the precision is calculated for inclusion prevalence ranging from 0 -50%. We have added that readers may want to consider the likelihood ratio and precision for projects with different inclusion prevalence rates. This tool has been incredibly useful in the current review and other reviews at CAMARDES, one of which had over 12,000 documents to classify, the second with over 200,000 documents to classify. The amount of human resources saved through the application of tools described here has been crucial for the feasibility of these projects. The documents that are included by the algorithm are checked by human reviewers. Therefore we want to maximise recall (sensitivity), as something missed at this stage will be lost completely. A low-precision error (undeserved inclusion) will simply lower the saved human screening time. The number of studies going through full-text analysis is never going to be larger due to adding machine assisted citation screening - it will be the same, or in case something is missed due to low recall, lower.

*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).

• ABB: Thank you for this suggestion. We currently do not collaborate with any groups that perform reviews of rare diseases but are very open to the suggestion of doing so. We have added a couple of sentences to the discussion regarding the class imbalance issue in training sets and propose a method that may be useful for machine learning in reviews with small inclusion prevalence, perhaps this would be useful for rare diseases.

We believe that our manuscript has been improved substantially following the input of the reviewers, and we would like to thank you again for your time.
Many thanks & kind regards,

Alexandra Bannach-Brown
On behalf of all co-authors.