Machine learning outperforms manual screening in literature reviews: a case study with rare disease natural history data

RWD63

Maria Rapoport[1], Antoinette Randet[1], Chris Cooper[2]

[1]Wickenstones Ltd, Oxford, UK; [2]Independent Researcher, London, UK

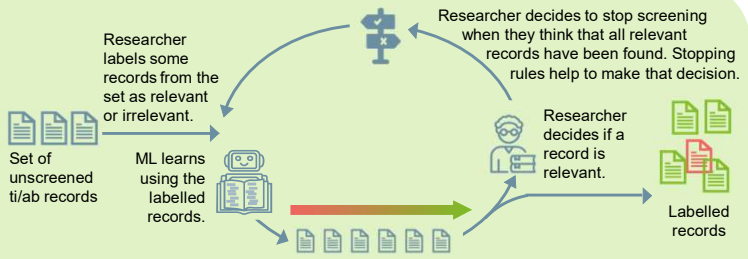
WICKENSTONES
PARTNERS IN BUILDING VALUE

BACKGROUND

Title and abstract screening for systematic reviews can be time consuming. A variety of machine learning applications are now available to assist researchers.

- Machine learning (ML) tools for the title and abstract (ti/ab) screening phase of systematic reviews (SRs) include, among others, generative artificial intelligence and active learning-based tools.
- Active learning tools, such as ASReview,[1] allow reviewers to (Figure 1):
  - Retain control over screening decisions
  - Only screen a subset of the ti/ab records by pre-defining a stopping rule that helps determine when most relevant records likely have been found
- However, concerns remain about:
  - Reproducibility of the screening using ML tools
  - Recall (i.e. proportion of relevant records identified during screening), especially in complex and varied record sets (i.e. sets containing various study types and populations)
  - Choice of optimal stopping rule, as multiple options, ranging from simple (e.g. 10% of the set of consecutive excludes) to complex (e.g. Screen-Apply-Find-Evaluate [SAFE][2] procedure) have been proposed

Figure 1. Screening flow with ti/ab screening tools that use active learning



METHODS

We tested an ML tool and 5 stopping rules for ti/ab screening for an SR of the natural history of a rare genetic disorder to assess the reliability and performance on a variable dataset and with a complex research question.

SET UP

- We conducted an SR to describe the natural history of a rare genetic disorder.
- We searched multiple databases.
- We deduplicated search results using Endnote X8.
- Two reviewers screened the sample. Reviewer 1 used Microsoft Excel; Reviewer 2 used ASReview.

PILOT SCREEN

- Both reviewers assessed 100 random reports to align their interpretation of the inclusion criteria.
- The number of expected includes was calculated based the inclusion rate in the pilot screening after conflicts were resolved.

TI/AB SCREEN

Reviewer 1: Excel

- Screened the set in random order and stopped when all records were reviewed

Reviewer 2: ASReview

- Screened the set in ASReview[1] and stopped when stopping rules were reached
- ASReview was set up for screening using the recommended model settings (TF-IDF, naïve Bayes, maximum, dynamic sampling)[2]
- Decisions from the pilot screen were used to train the algorithm initially

Reviewer 2: ASReview (continued)

- Tested five stopping rules:
  - SAFE procedure, which includes multiple criteria and a second screen of unlabeled records with a more complex model[3]
  - 50 consecutive irrelevant records
  - 2.5% and 10% of the set of unscreened records in consecutively irrelevant records
  - 95% of estimated relevant records found (based on inclusion rate from pilot screen)[4]

RESULTS

With ASReview, 51% of relevant records were found after reviewing only 12% of the set. In Excel, 52% needed to be screened to find the same number of relevant records.

SEARCH RESULTS

- 2,521 reports retrieved
- 1,074 reports deduplicated
- 1,447 reports included in the title and abstract screen

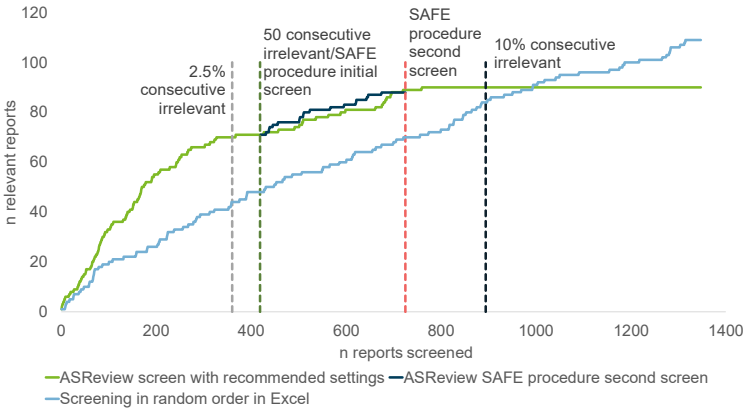
PILOT SCREEN

- Substantial 93% (κ=0.68) agreement between reviewers
- 12% inclusion rate estimated, resulting in 161 expected relevant reports remaining in the unscreened set

TI/AB SCREEN

- Of 1,347 records, Reviewer 1 marked 109 and Reviewer 2 marked 90 records as relevant (Figure 2).
- The agreement between reviewers was lower than in the pilot screen (91%; κ=0.37 [fair agreement]).
- After conflict resolution, 97 reports in addition to the pilot screen were included (overall inclusion rate: 7.5%).
- Screening in ASReview found relevant reports notably faster than screening in Excel.
  - In ASReview, most (51%) relevant reports were found after screening only 12% of the sample.
  - In Excel, 52% of the sample needed to be screened to achieve the same recall.
- The stopping rule of finding 95% of estimated relevant records was not reached.

Figure 2. Recall during ti/ab screening in Excel and ASReview



Note: Stopping rules apply to ASReview screen only.

RECOMMENDATIONS AND CONCLUSION

ML-supported screening can substantially decrease reviewer workload while maintaining high recall, even when applied to complex research topics.

RECOMMENDATIONS FOR TI/AB SCREENING USING ACTIVE LEARNING ML

Tailor the screening experience to your preferences

- Screening in a browser-based application (e.g. ASReview) allows the use of browser add-ons (e.g. multi highlight add on in Google Chrome) that can increase screener comfort and efficiency.

Ensure that the screening criteria are clear and reliably applicable before screening

- Because every reviewer decision affects which reports are shown to the reviewer next, a thorough pilot screening is imperative for reliable screening results.

Choose a stopping rule that fits your project

- Stopping rules balance reviewer workload and recall and range from simple to complex. Different rules will be needed to suit differing projects and objectives.

Follow the evolving best practice

- ML-supported SR methods have become more widely accepted in the last year, with, for example, the National Institute for Care Excellence (NICE) issuing guidance for company submissions.[5]

CONCLUSION

- Use of ASReview can greatly limit screening time during the ti/ab screening stage, as 51% of relevant records were found after screening 12% of the sample (vs screening 52% of the sample to find the same number of relevant records in Excel).
- Despite the variability of the set, ASReview reliable presented relevant records early in the screen.
- ASReview (or other ML applications) may supplement or replace manual ti/ab screening in future.

REFERENCES

1 van de Schoot *et al., 2021. 10.1038/s42256-020-00287-7.*
2 Ferdinands *et al.*, 2023. 10.1186/s13643-023-02257-7.
3 van Haastrecht *et al.*, 2021. 10.3389/frma.2021.685591.
4 Boetje, van de Schoot, 2024. 10.1186/s13643-024-02502-7.
5 Use of AI in Evidence Generation: NICE position statement, 2024. Accessed on 29.10.2024 via https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation--nice-position-statement

ACKNOWLEDGEMENTS

The SR project was supported by Mereo BioPharma, UK.

Presented at: ISPOR EU; Barcelona, Spain; 17–20 November 2024