Training Artificial Intelligence for Literature Reviews: Can an AI Classifier Match a Human Reviewer?

Metcalf T[1], Dodd O[1], Peatman J[1], Kiessling J[1], O'Donovan P[1], Yakob L[1], Heron, L[1]

[1]Adelphi Values PROVE™, Bollington, Cheshire, Great Britain

ADELPHI VALUES | PROVE Expertise in Access and Value Evidence Outcomes

MSR222

Introduction and objectives

Introduction

- With an increasing volume of medical literature being published, artificial intelligence (AI) is becoming recognised as a tool to improve screening efficiency during literature reviews.[1]
- Recent guidance from the National Institute of Health and Care Excellence (NICE) acknowledges the advantages of AI in assisting literature review processes, highlighting progression in acceptance of AI-assisted processes to support reimbursement submissions.
- AI classifiers offer an alternative to AI screening, by performing binary classification of publications in response to a set question.
- AI classifiers are not restricted to a single use setting and can be applied across multiple reviews with potential for iteratively improved accuracy.
- While an abundance of literature exists for AI screening, evidence evaluating AI classifiers and their comparability with human reviewers is limited.

Objectives

- To demonstrate the comparability of four independent AI classifiers with human reviewer decisions in a real-world data set.

Methods

- Four classifiers were independently trained using an online platform to categorise abstracts based on criteria for case reports, elderly populations, paediatric populations, and randomised controlled trials (RCTs).
- Each classifier was trained using ≥1,000 abstracts until either a ≥0.80 F1 score was achieved, or 3,000 abstracts were screened.
- For training datasets, a balance of 'Yes' and 'No' responses were sought across classifiers, to ensure optimal accuracy.
- A total of 2,245 abstracts from a previously completed systematic literature review were classified by both each AI classifier and one human reviewer, to mimic application in a literature review with dual screening.[2]
- Matching decisions were assumed to be accurate, with a senior reviewer making a final decision on conflicts between the human reviewer and the AI classifier.
- Classifier responses were compared with those of a human, with matched responses reported as a decision match percentage.
- The decision match percentage, sensitivity, specificity, and their respective 95% confidence intervals (CIs), were calculated for each classifier.
  - To calculate the 95% CIs for the decision match percentage, sensitivity, and specificity, the binomial test was used with a significance level of 0.05.
- To summarize the decision match percentage across classifiers, the arithmetic mean (95% CI) was calculated, since all classifiers were tested on the same number of records.
- To calculate the pooled mean for the sensitivity and specificity across classifiers, inverse variance weighting (IVW) was used. This method gives more weight to the more "certain" results. Since it is anticipated that the sensitivity will be highly variable due to fewer articles meeting the criteria of each classifier (simulating real-world screening, where only a small proportion of articles are deemed relevant and progress to full-text screening), this method minimizes the variance for the aggregate value.

Figure 1. Approach to AI classifier development

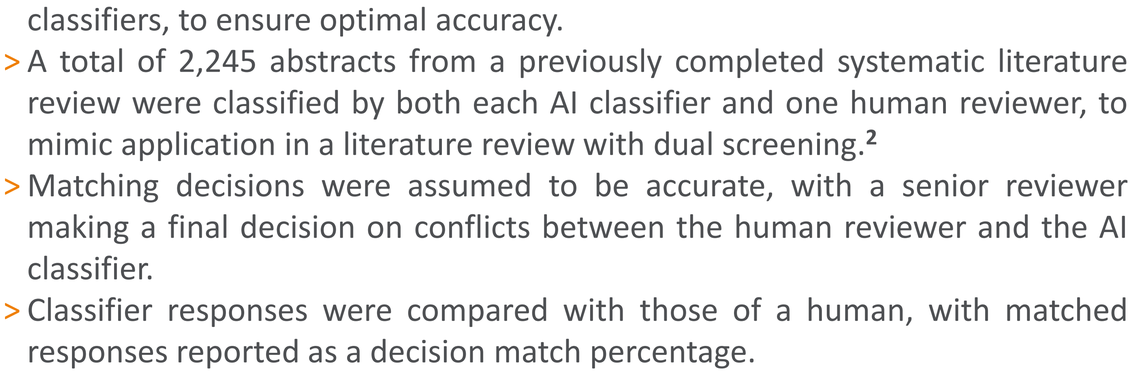


Table 1. Overview of classifier training datasets

Classifier	Number of references used to train classifier		
	Total	% references 'Yes'	% references 'No'
Case reports	1,354	48.7	51.2
Elderly population	2,709	41.1	58.9
Paediatric population	1,854	41.4	58.6
RCTs	1,581	49.5	50.5

RCT: Randomised controlled trial.

Results

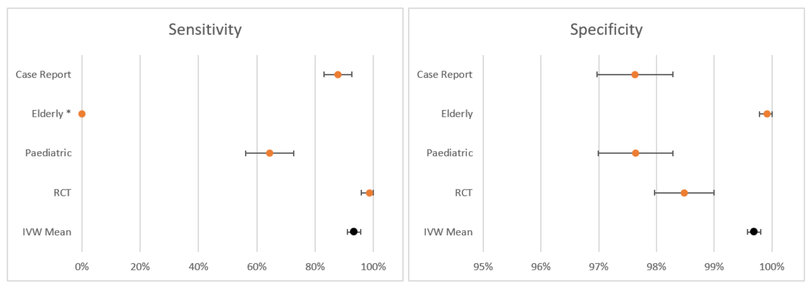
- For all classifiers, the mean (95% CI) decision match percentage ranged from 94.5% (93.6%–95.5%) to 99.6% (99.4%–99.9%) across the four classifiers (Table 2).
  - The arithmetic mean (95% CI) for the decision match percentage was 96.9% (96.5%–97.2%).
- The IVW mean (95% CI) for the sensitivity was 93.4% (91.1%–95.7%). The sensitivity was variable across classifiers, likely due to the small number of references that met the criteria for each classifier.
- All classifiers demonstrated high specificity, significantly greater than 97.6% (Figure 2), with an IVW mean (95% CI) of 99.7% (99.6%–99.8%).

Table 2. Decision match rate (95% CI) for each AI classifier

	Matches (%)	95% CI
Case reports	2,136 (95.1%)	94.3%–96.0%
Elderly population	2,237 (99.6%)	99.4%–99.9%
Paediatric population	2,122 (94.5%)	93.6%–95.5%
RCTs	2,205 (98.2%)	97.7%–98.8%

CI: confidence interval; RCT: randomised controlled trial.

Figure 2. Sensitivity and specificity and AI classifiers, with 95% CIs



* Only three references were found to meet the criteria of the "Elderly" classifier and was not included in the IVW mean sensitivity.
CI: confidence interval; IVW: inverse variance weighted; RCT: randomised controlled trial.

Discussion and conclusions

Discussion

- Our approach to training classifiers resulted in comparable screening decisions between the AI classifiers and a human reviewer, evidenced by a high percentage match rate across all classifiers.
  - The results of this work provide validation of processes taken to develop the classifiers. The approach taken to train the AI classifiers was efficient and resulted in classifiers which closely emulated the screening decisions of a human reviewer, demonstrating their appropriateness for application in a real-world data literature review.
- The references in the "test" dataset were relatively unbalanced, with only 0.13% to 8.06% of the references meeting the criteria for each classifier at final assessment.
  - As such, the sensitivity varied widely between classifiers, ranging from 64.3% to 98.6% for three of the classifiers, excluding the "Elderly" classifier.
  - For the "Elderly" classifier, only three references were deemed relevant at final assessment, which explains the high variance for this value.
  - As a lower sensitivity results in overly inclusive screening decisions, the classifiers were unlikely to exclude records relevant to their criteria, or the PICOTS criteria they represent in practice.
- The specificity was consistently high between classifiers, with an IVW mean of 99.7% (99.6% to 99.8%) which compares favourably to the 88.7% reported in published literature.[3] As such, the classifiers closely emulated the decisions of a single human reviewer on references which did not meet the criteria of the classifiers.
- The amount of training data required to train each AI classifier is dependent on the complexity of the criteria. Classifications which are easily answered by a human (i.e., "is this reference a case report?") are easier to train than more complex classifications (e.g., "does this reference report on a paediatric population?", where "paediatric" can include a variety of age ranges or associated terms).

Conclusion

The approach taken to train the AI classifiers was effective and the classifiers are appropriate to support dual abstract screening in literature reviews.

- Our approach to AI classifier development has been validated, demonstrating comparable results to human reviewers.
- AI classifiers tended towards being over inclusive, meaning they are highly unlikely to exclude any relevant articles.
- AI classifier training processes vary in complexity, dependent on the criteria.

References

1. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic reviews. 2019;8:1-10.; **2.** Higgins J, Lasserson T, Chandler J, Tovey D, Churchill R. Standards for the conduct and reporting of new Cochrane Intervention Reviews, reporting of protocols and the planning, conduct and reporting of updates. Methodological Expectations of Cochrane Intervention Reviews (MECIR). 2018; **3.** al. Se. Diagnostic performance of artificial intelligence tools for article screening during literature review: A systematic review. 2024.