Accuracy and efficiency of automated or artificial intelligence tools in systematic literature reviews: A rapid systematic literature review

Hardy, E, Jenkins, A, Ross, J, Lang, S

Mtech Access
Powered by Petauri™

ISPOR poster acceptance code: MSR229
Abstract ID: 144489

Aims and methods

The aim was to assess the comparative accuracy and efficiency of commercially available automated tools versus human reviewers for screening and/or data extraction tasks during a systematic literature review (SLR).

A protocol was submitted to Open Science Framework.[a] Quantitative data for measures of accuracy, sensitivity, or time/workload savings were extracted, grouped by tool, and the range of results reported. A thematic analysis of factors influencing accuracy and efficiency was conducted.

The systematic review identified 43 studies (full list of included studies).[a] The majority of identified studies reported on title/abstract (ti/ab) screening (n=38), with minimal data available for full paper screening (n=5), data extraction (n=3), and risk of bias (n=2) (see Figure 1).

Figure 1: PRISMA flow diagram

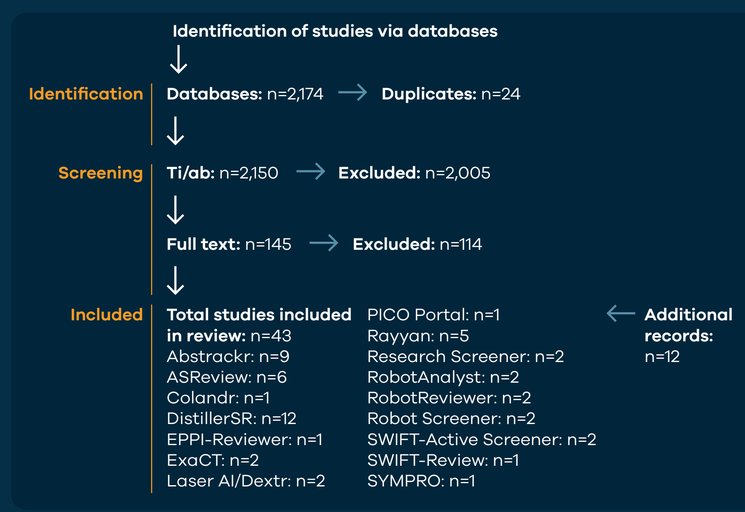

Figure 2: Thematic analysis of reported factors that influenced accuracy or efficiency of screening

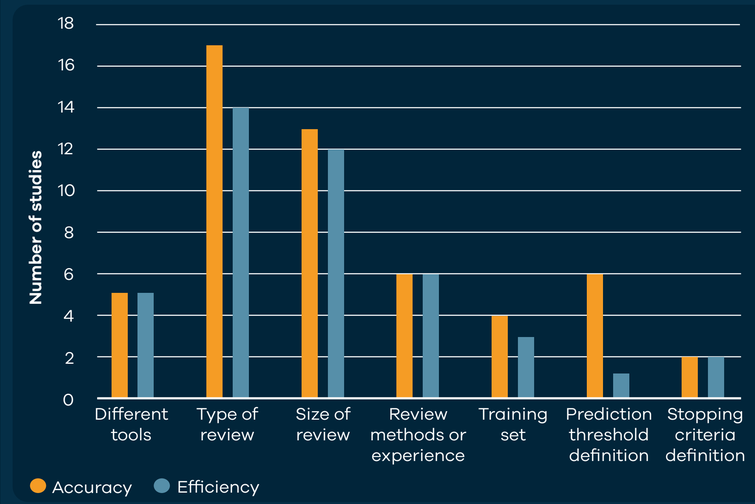

Table 1: Reported accuracy and efficiency of different AI tools for ti/ab screening

AI tool (studies)	Threshold or prediction score	SLR analysed *(citations)*	Training sample	Articles not screened, %	Accuracy	Sensitivity	Workload saved, %	Time saved, %
Abstrackr (n=8)	NR, prediction score 0.395 or 4, 10 to 98% screened, until all included studies were screened	1 to 10 *(500 to 71,284)*	NR or 120 to 296 or 10%	NR		61 to 100†	9 to 67‡	51.5 to 99.3
ASReview (n=6)	NR, stopping after 100 consecutive articles were irrelevant, 95% recall	1 to 14 *(134 to 19,718)*	NR	NR, 75 to 90	80.2	48 to 79§	37 to 92¶††	51 to 77
Colandr (n=1)	NR	1 *(2,797)*	NR	NR		0.65	97	
DistillerSR (n=2)	>0.5 inclusion probability	1 to 8 *(2,472 to 20,100)*	NR or 300 or 10 to 20%	NR	85 to 94	0.14 to 0.78		77.9
DistillerSR (n=2)	>0.8 inclusion probability	2 to 10 *(290 to 5,501)*	10%	NR	73.1 to 98.6‡‡			36 to 91
EPPI-Reviewer (n=1)	NR	9 *(2,605 to 18,281)*	10%	NR				39.9 to 98.8
Laser AI/Dextr (n=1)	None	1 *(4,459)*	100	NA		97.56†	42.69	
PICO Portal (n=1)	Multiple stopping criteria	8 *(4,204 to 14,185)*	NR	50 to 80		95		
Rayyan (n=1)	NR	1 to 2 *(500 to 12,732)*	5 to 50%, or 200	NR	10 to 64	0.78 to 100	0.49††	3 to 46
Research Screener (n=2)	Once certain that all relevant articles identified	1 to 11 *(17,736 to 45,675)*	NR	60 to 96			68 to 96¶	
RobotAnalyst (n=1)	NR	3 *(28,646)*	180 to 197	2 to 3		0 to 100		
Robot Screener (n=3)	NR/>0.8 inclusion probability	2 to 19 *(568 to 23,113)*	NR or 80%	NR	93.1 to 95.9‡‡	27 to 97.1		
SWIFT-Review (n=1)	0.55	1 *(500)*	200	NR		90.9 to 97		
SWIFT-Active Screener (n=2)	NR, 95% accuracy	1 to 26 *(26 to 4,612)*	NR	NR	97.9	0.915 to 1		9.8 to 65.5

Studies were not included in this table if they did not have equivalent outcomes to those listed. We focused on DistillerSR reporting on >0.5 and >0.8 inclusion probability, as these were most frequently reported. Outcome definitions where reported are listed below.

† – 2% of correctly identified studies; ‡ – Proportion of citations predicted as irrelevant out of the total number of citations to be screened; § – Relevant records identified after screening 10% of total records; ¶ – Work saved over random sampling; †† – Work saved over random sampling at 95% recall; ‡‡ – Comparison between AI and human reviewers and used to calculate IRR based on Cohen's kappa statistics.

Table 2: Reported accuracy and efficiency of different AI tools for data extraction

AI tool (task)	SLR analysed *(citations)*	Training sample	Accuracy	Sensitivity	Workload saved, %	Time saved, %
ExaCT n=1 study (Data extraction)	75 trials	NR	48†			44.6
Laser AI/Dextr n=1 study (Data extraction)	NR *(52)*	NR		0.918		
RobotReviewer n=2 studies (Risk of bias)	6,610 trials 10 SLR *(28,646)*	NR	71‡ to 90.6	0.34		

† – Relevant solutions identified; ‡ – Multi-task model jointly modelling all domains and incorporating information about sentence relevance as features.

Results

Accuracy and efficiency of ti/ab screening

Most studies reported on DistillerSR, Abstrackr, and ASReview (Table 1). The results highlight how these tools are implemented with a variety of thresholds or classifiers to determine whether a study will be included or not. The size of training samples, method of application of the tool, comparator definition, and outcome definitions were variable and often not reported. The tools offer clear time savings but the accuracy of the tools was variable.

Accuracy and efficiency of data extraction support

Three published tools were identified for data extraction: ExaCT, Laser AI, and RobotReviewer (Table 2).

Accuracy and sensitivity varied widely. Accuracy varied according to the type of data extracted (e.g. outcomes or baseline characteristics or study designs) or the risk of bias domain being extracted (selective reporting was the hardest to get accurate).

Factors influencing accuracy and efficiency of screening

Tables 1 and 2 indicated considerable heterogeneity between studies. Seven major themes were identified, which influenced accuracy or efficiency (Figure 2):

1. Different tools (six studies reported direct comparisons between tools)
- In four studies, similar results were generally observed across tools: Abstrackr, DistillerSR, EPPI-Reviewer, Rayyan, RobotAnalyst, Robot Screener, and SWIFT-Review (1-4)
- Rayyan (versus Abstrackr and Colandr) and Abstrackr (versus DistillerSR and RobotAnalyst) were identified as the most sensitive and were highly rated by users in two different studies (5, 6)

2. Review methods and application of tool (was the AI tool used standalone, or to assist a human? Was a second reviewer used? [with or without AI]; semi-automation methods [AI assisting a human] usually provided the preferred balance of accuracy and efficiency)

3. Type of review (accuracy and efficiency were generally impacted by the complexity of inclusion criteria)

4. Size of review (efficiency benefits were more consistently observed with increasing size)

5. Training set (larger training sets generally improved accuracy; type of data and reviewer experience informing the training set can also alter results)

6/7. Prediction threshold/stopping criteria (higher thresholds were more accurate but less efficient)

Conclusion

AI tools offer clear screening efficiency benefits, however the impact on accuracy is less consistent. Limited tools support data extraction, further research is required to support any recommendations. Reviewers contemplating integration of AI tools into their workflow should first consider the themes highlighted here to identify published tools reflecting their own intended use.

The practical application and understanding of AI tools in SLR are hampered by complex digital methodologies, which are not always transparent or understood by the user. Clearer methodological reporting is essential for inter-study comparisons and to provide individual research groups and external health technology assessment (HTA) bodies with confidence in the outputs from AI tools.

Mtech Access
Powered by Petauri™

ISPOR Europe 2024,
17–20 November 2024,
Barcelona, Spain

(Stand No. 1406)

Scan to see video walkthrough and references

a Open Science Framework

Abbreviations

AI, artificial intelligence
HTA, health technology assessment
IRR, inter-rater reliability
NA, not applicable
NR, not reported
PICO, population, intervention, comparator(s), outcome(s)
PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SLR, systematic literature review
ti/ab, title/abstract