MSR116

Give us the tools and we will finish the job: How accurate is AI for study selection in systematic reviews?

Bishop E [1], Sanderson A [1], Reddish K [1], Carr E [1], Edwards M [1], McCool R [1], Ferrante di Ruffano L [1]

[1] York Health Economics Consortium, University of York, York, YO10 5NQ

INTRODUCTION

- Systematic reviews (SRs) are the highest level in the hierarchy of evidence and are the cornerstone of evidence-based medicine. They provide a comprehensive synthesis of available research and are critical for informing clinical guidelines and policy.[1]
- As the volume of medical literature increases over time, the traditional SR process, particularly the screening of studies for eligibility, has become increasingly time-consuming and resource-intensive.
- Consequently, the development and use of artificial intelligence (AI) and machine learning to conduct elements of the SR process is becoming more widespread, promising increased efficiency by automating key steps.[2]
- However, for AI to be confidently adopted in clinical and public health contexts, it is important to understand whether it is sufficiently accurate and reliable to replace human reviewers.
- One such AI tool is EasySLR, a web-based tool that utilises an AI model to assist in all stages of a review, from deduplication to data extraction. The AI within EasySLR can act as a second reviewer, replacing the need for two human reviewers; or as an assistant, offering suggestions to aid human reviewers.
- **Aim:** To investigate the accuracy of a web-based AI tool (EasySLR) in making eligibility decisions for SRs of healthcare interventions.

METHODS

- The accuracy of EasySLR in making eligibility decisions was retrospectively tested by comparing the decisions made by the AI tool with the decisions made by two independent, human reviewers across three completed SRs:
  - **Review 1 (R1):** Clinical effectiveness of multiple sclerosis treatments.
  - **Review 2 (R2):** Clinical effectiveness of adult-onset Still's disease (ASOD) treatments.
  - **Review 3 (R3):** Health-related quality of life (HRQoL) and healthcare resource use (HRU) in ASOD.
- The following information was uploaded to EasySLR for each review:
  - The PICO.
  - The screening decisions made by human reviewers at title and abstract (TA) and full text (FT) stages.
  - The reasons for exclusion at FT.
- The AI tool reviewed this information and independently suggested whether each record should have been included or excluded at TA and FT stages, as well as providing reasons for exclusion.
- The accuracy of the AI decisions was measured against the human standard. The categories used to measure EasySLR's accuracy are presented in **Table 1**.
- We also tested EasySLR's protocol optimisation tool, which suggests improvements for a study's PICO (e.g. where clarity is needed). The accuracy of EasySLR was re-assessed using the AI-"improved" protocol.

Table 1: Categories used to measure AI accuracy

Category	Description	Calculation
AI sensitivity	The ability of AI to correctly identify an **eligible** study	Correct includes / total human includes x 100
AI specificity	The ability of AI to correctly identify an **ineligible** study	Correct excludes / total human excludes x100
Overall agreement	The number of include / exclude decisions that aligned with human decisions	Correct decisions / total decisions x 100
False negative (FN) rate	The proportion of studies incorrectly **excluded** by AI.	(1 – **sensitivity**) x 100
Fale positive (FP) rate	The proportion of studies incorrectly **included** by AI	(1 – **specificity**) x 100

Figure 1: TA accuracy before and after protocol optimisation

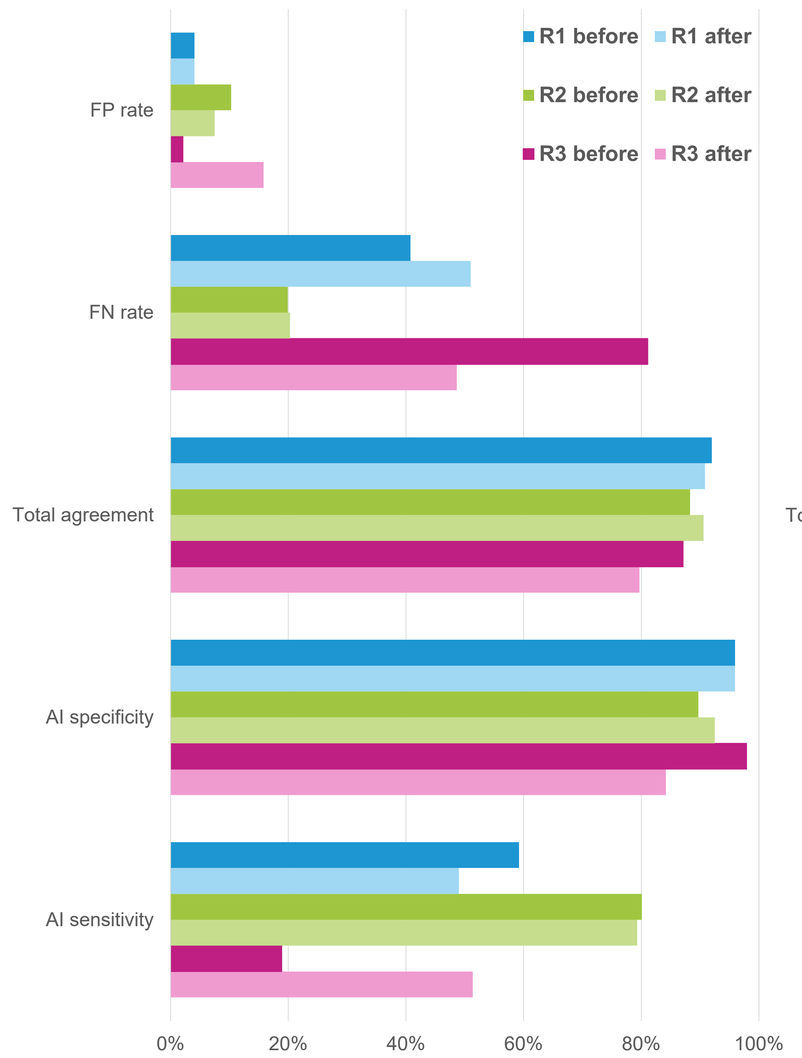
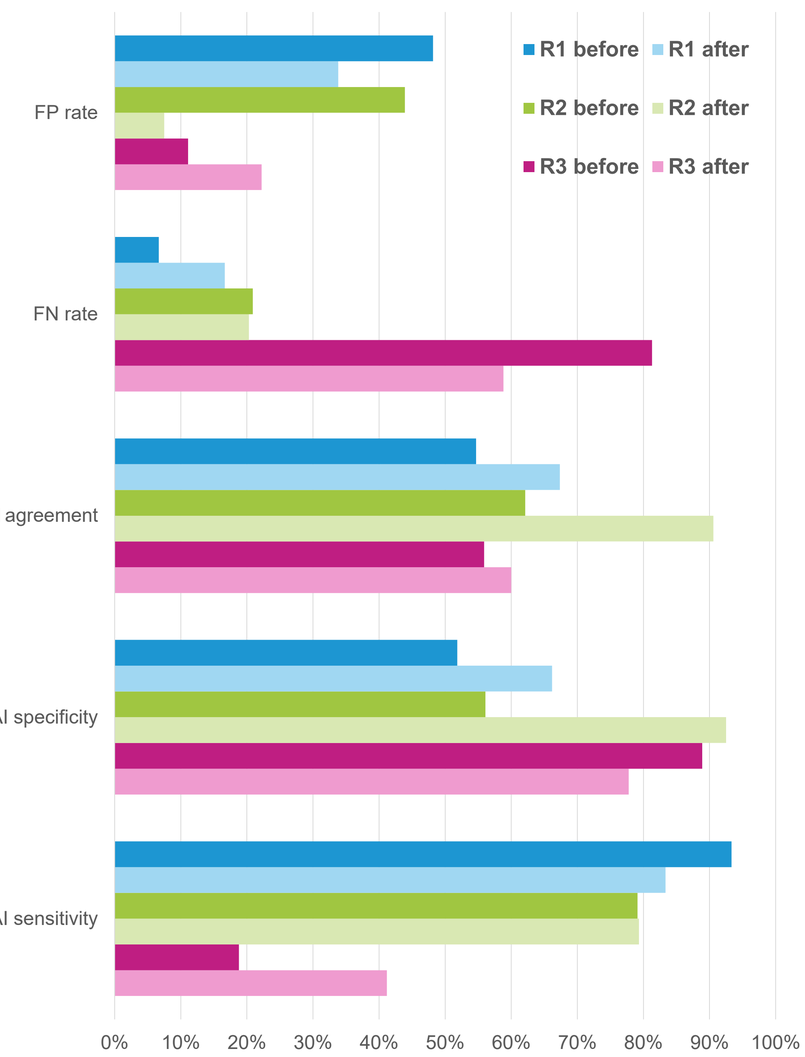


Figure 2: FT accuracy before and after protocol optimisation



RESULTS

Before protocol optimisation

- AI-human agreement was consistently higher at TA stage (87% to 92%; **Figure 1**) than FT stage (55% to 62%; **Figure 2**).
- AI-human agreement on FT exclusion reasons ranged from 30% to 40% across reviews. The most commonly misassigned exclusion reasons were "patient population" and "study design".
- The AI's sensitivity at both TA and FT was considerably poorer in R3 than R1 and R2 (**Figure 1 & 2**). Consequently, the FN rate for R3 was much higher: the AI incorrectly excluded more eligible studies in R3 than the other two reviews.
- Specificity was consistently higher at TA stage (89.7% to 97.9%; **Figure 1**) than FT stage (51.8% to 88.9%; **Figure 2**). However, the range of values is considerably wider at FT than TA. The FP rate is, consequently, lower at FT than TA.

After protocol optimisation

- Protocol optimisation affected the results for each review differently.
- At TA stage **(Figure 1),** R1 and R2 decreased sensitivity and increased specificity, having a lower ability of correctly identifying eligible records (-1% to -10%) and a higher ability of correctly identifying ineligible records (+0.1% to +3%). Conversely, R3 demonstrated increased sensitivity (+32%) and decreased specificity (-14%).
- Similar results were found at FT for R1 and R3; the AI's sensitivity for R2 at FT did not change after protocol optimisation **(Figure 2)**.
- Protocol optimisation increased AI-human agreement for all reviews at the FT stage, but it decreased agreement for R1 and R3 at TA stage. Therefore, AI-human agreement was still consistently higher at TA stage than FT stage for all reviews.
- The AI-human agreement on FT exclusion reasons increased from 36% to 51% after protocol optimisation. The most commonly misassigned exclusion reason was "patient population"; this was especially apparent in R2, which investigated subgroup populations.

CONCLUSIONS

EasySLR's AI reviewer is a promising study selection tool for clinical SRs, though should be used alongside human reviewers to achieve acceptable screening accuracy. For R3, before protocol optimisation, the AI excluded many eligible studies. This is an expected finding due to the complex nature of HRQoL and HRU reviews and outcomes.

Protocol optimisation resulted in minor accuracy changes in the clinical reviews (R1 and R2) at TA stage and an increase in the number of ineligible studies correctly identified at FT stage. The tool had positive impact on the AI performance in the HRQoL and HRU review (R3) and enabled more eligible studies to be correctly identified. The protocol optimiser tool is still under development at the time of conducting this research, and it is worth noting that the tool hasn't been trained on HRQoL and HRU reviews.

Future evaluations should prospectively assess AI tools to understand how many eligible studies would be missed, and the impact of missing those studies when using AI to make screening decisions alongside a human. Training AI tools on larger and more diverse data sets would also be advantageous in increasing AI accuracy.

REFERENCES AND ACKNOWLEDGEMENTS

1. Bangdiwala, S. I. Int J Inj Contr Saf Promot 2024;31(3):347–349. **2.** Blaizot A. et al. Res Syn Meth 2022;13(3): 353-362.

Acknowledgements: We would like to thank Easy SLR for providing us with free access to their software.

CONTACT US

Rachael.McCool@york.ac.uk

+44 1904 324893

www.yhec.co.uk

Providing Consultancy & Research in Health Economics

UNIVERSITY of York

INVESTORS IN PEOPLE® We invest in people Gold

YHEC

York Health Economics Consortium