Artificial intelligence in systematic reviews: An investigation into the impact of eligible studies being excluded by artificial intelligence

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) sit at the top of 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 is increasing over time, the process of conducting SRs is becoming more time-consuming and resource-intensive. Consequently, the use of artificial intelligence (AI) in SRs to increase efficiency by automating key steps, such as study selection, is gaining traction.[2]
- For AI tools to be confidently adopted in clinical and public health contexts, it is crucial to understand whether they are sufficiently accurate and reliable to replace human reviewers.
- As well as the accuracy of AI tools, it is important to consider the potential impact of AI use on the results of the SR. The use of AI in reviews, especially tools with a low accuracy, could result in key evidence being missed, ultimately compromising the conclusions made the by SR.
- One such AI tool is EasySLR, a web-based tool that utilises an AI model to assist in all stages of a review, from search to report.
- **Aim:** To investigate the impact of eligible studies being excluded by a web-based AI tool (EasySLR) on the results of SRs.

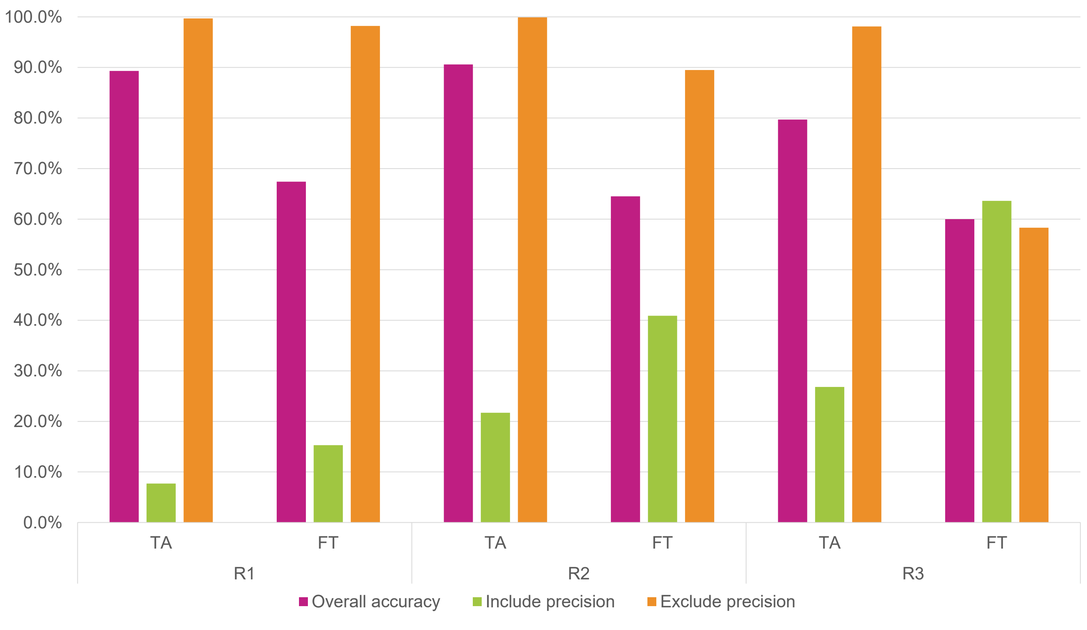
METHODS

- The AI reviewer within EasySLR was assessed by comparing the eligibility decisions made by the AI tool with the decisions made by two independent, human reviewers across three completed SRs of health interventions:
  - **R1**: Clinical effectiveness of multiple sclerosis treatments.
  - **R2**: Clinical effectiveness of adult-onset Still's disease treatments.
  - **R3**: Health-related quality of life and healthcare recourse use in AOSD.
- The overall include and exclude precision of the AI was briefly explored. The potential impact of false exclusions at full text stage was assessed by examining the influence each eligible record had on the results of the review. This impact was assessed by examining the number of AI exclusion studies that were:
  - **High impact**: Primary studies that added a unique element to the review.
  - **Moderate impact**: Primary studies that contributed similar data to other included studies.
  - **Low impact**: Associated records of correctly included studies.
- We also tested EasySLR's protocol optimisation tool, which suggests improvements for a study's PICO (e.g. where clarity is needed). The number and impact of AI-excluded studies were re-assessed using the AI-"improved" protocol.

RESULTS

- Analysis into the precision of the AI demonstrated that across all reviews, there was a higher accuracy at title and abstract screening (80% to 91%) compared to full text stage (60% to 67%). Include precision was consistently lower than exclude precision across stages, suggesting AI was overly inclusive. R3 at full text stage demonstrated the lowest precision. Full details are presented in **Figure 1**.

Figure 1: AI precision across reviews and study selection stages



RESULTS

Before protocol optimisation

- Overall, the number of studies incorrectly excluded by AI was 28: 2 (1% of all AI excludes) for R1, 13 (11%) for R2, and 13 (45%) for R3. 17 of the studies were primary publications and 11 were associated records. These studies were split over the impact groups as shown in **Table 1**.
- Of the 28 publications that were incorrectly excluded by AI:
  - 12 were **high impact** exclusions because they reported on a unique geographical cohort or unique patient subgroup, reported an outcome or timepoint that was not addressed by other included studies, or reported conflicting findings to those of other included studies.
  - 4 were **moderate impact** exclusions because, although they were primary studies, the outcomes and populations reported were similar to other included studies.
  - 12 were **low impact** exclusions because they were associated studies of primary publications which were unlikely to contribute new information.

After protocol optimisation

- Protocol optimisation had a minor impact on the AI performance. After optimisation, there were 13 high, 4 moderate, and 10 low impact studies incorrectly excluded by AI. Whilst the numbers were not dissimilar to those prior to protocol optimisation, the studies were different.
  - 6 correctly included studies were previously incorrectly excluded. This suggests the AI-trained protocol was effective in aiding AI understanding of these studies.
  - 8 incorrectly excluded studies were previously correctly included, suggesting the protocol optimisation negatively impacted the AI's understanding of these studies.
- Common AI errors identified by EasySLR suggest that population, outcomes, and study design are areas of weakness that may explain why studies were incorrectly excluded.

Table 1: Number of incorrect excludes by review and impact group

Impact group	R1	R2	R3
High impact	1	1	10
Moderate impact	1	1	2
Low impact	0	11	1

CONCLUSIONS

- The number of discrepancies between AI and human reviewer decisions during full text screening varies considerably across reviews, as does the impact of these discrepancies. AI accuracy and precision was higher at title and abstract stage across reviews than at full text stage. R3 had the lowest precision of all reviews, for both stages. This is not unexpected and reflects the complexity of the review's outcomes.
- Before protocol optimisation, 12 (43%) of the 28 included studies incorrectly excluded by AI were unique primary studies that, if excluded, would have a high impact on the results of the reviews. After protocol optimisation, this increased to 13 (48%) of 27 studies. To circumvent the AI-human discrepancies in included studies, any differences in decisions should be checked to ensure studies offering unique data are not incorrectly excluded from the review.
- Future investigations should expand on this idea and examine the impact of studies incorrectly excluded by AI at title and abstract stage. It would also be beneficial to investigate the reporting of the studies incorrectly excluded. If AI is to be used routinely in the future, reporting standards may need to change to account for this.

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

Emma.Bishop@york.ac.uk

+44 1904 325503

York Health Economics Consortium

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