MSR107

Developing and Testing AI-Generated PICOS Summaries to Aid in Literature Reviews

Rawal A,[1] Ashworth L,[2] Luedke H,[3] Tiwari S,[3] Thomas C,[4] Murton M[4]

[1]Costello Medical, Boston, MA, USA; [2]Costello Medical, Manchester, UK; [3]Costello Medical, London, UK; [4]Costello Medical, Cambridge, UK

Objective

This research aimed to develop a workflow for artificial intelligence (AI)-generated PICOS summaries of abstracts and integrate this into an in-house literature review platform. The impact of this tool in terms of time savings and accuracy of screening was evaluated across three 'test' literature reviews.

Background

- Ever-increasing volumes of published literature make literature reviews increasingly resource-intensive. Leveraging AI-based tools may improve efficiency and streamline processes, but benefits must be balanced with potential inaccuracies and hallucination risks.[1]

Methods

Prompt Engineering and PICOS Summary Generation

- Various generative AI models and parameters were evaluated to develop a generic 'context prompt' capable of generating PICOS summaries for any abstract. A workflow was established to integrate the output into our bespoke, in-house, literature review platform.
- The selected model was gpt-3.5-turbo with a temperature setting of 0.2. The context prompt stated the requirement for a response to be returned in JavaScript Object Notation (JSON) format (**Figure 1**).

Efficiency and Accuracy Testing

- **Test Reviews 1** and **3** conducted preliminary investigations into efficiency and accuracy, respectively, while **Test Review 2** investigated both aspects in greater depth (**Table 1**).
- To measure efficiency, reviewers recorded their screening rate (abstracts/hour) when PICOS-assisted and PICOS-unassisted. To estimate accuracy, a senior reviewer compared the PICOS summary against the abstract content. Additionally, conflict rates for a PICOS-unassisted/PICOS-unassisted pair and a PICOS-assisted/PICOS-unassisted pair were collected. Qualitative data on user experience were gathered using a questionnaire.

Results

Efficiency and Accuracy

- In **Test Review 1**, abstract review rates were 60/hour for PICOS-unassisted vs 90/hour for PICOS-assisted, representing a 50% increase in efficiency. However, in **Test Review 2**, rates were 190/hour vs 210/hour, representing a smaller increase in efficiency (11%) (**Figure 2**). There was greater variability in screening rates for PICOS-unassisted reviewers.
- Across **Test Reviews 2** and **3**, 92–96% of articles were identified as having a correct or partially correct PICOS summary, with inaccuracies in the intervention and comparator domains being most common (**Figure 3**).
- Additionally, in **Test Review 2**, conflict rates were similar irrespective of whether PICOS summaries were used (18% for PICOS-unassisted vs 20% for PICOS-assisted). Of PICOS-assisted conflicts, only 6% (9/159) were determined as likely being due to the PICOS summaries.

User Experience

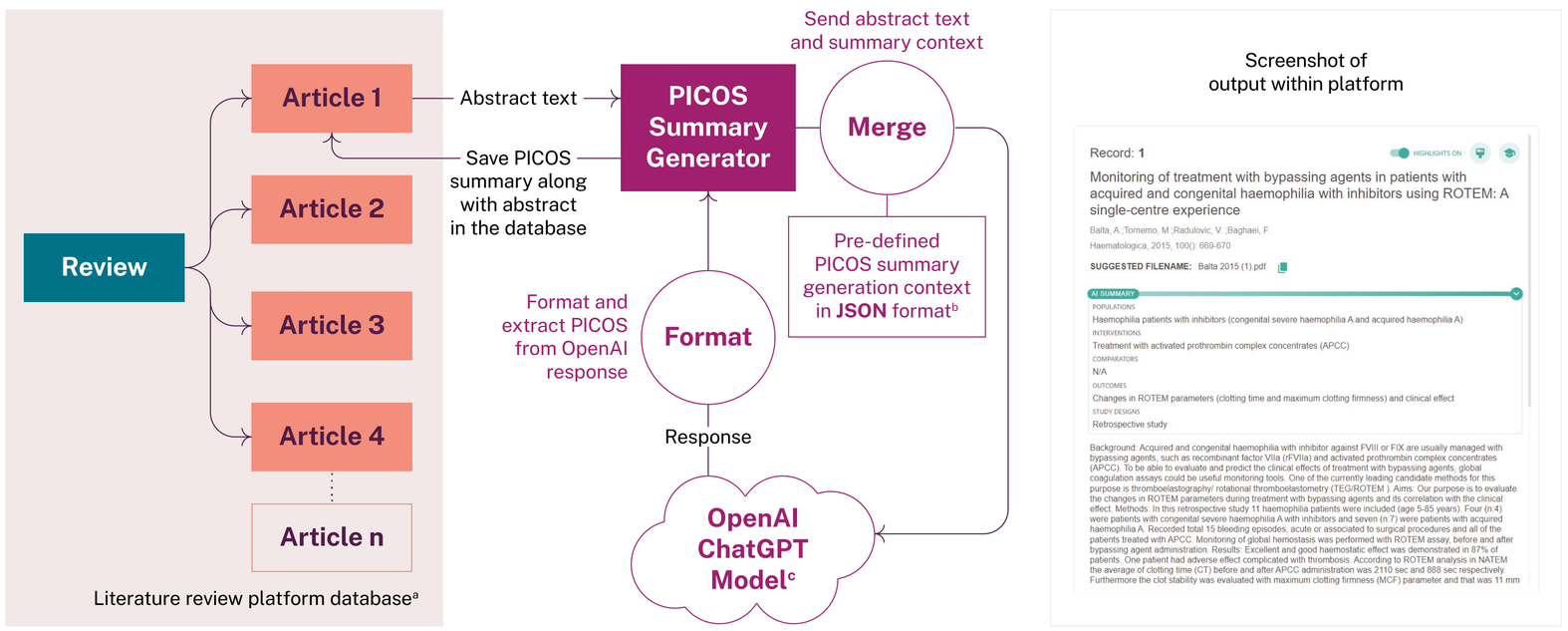
- Reviewer perceptions of the tool were generally positive, with 6/7 reviewers (across **Text Reviews 1** and **2**) enjoying using it and perceiving that it made screening faster. One reviewer did not have confidence in the summaries and therefore felt it slowed their screening rate.

Discussion

- This research demonstrates the potential of AI-generated PICOS summaries to improve review efficiency. However, the magnitude of benefit may vary. Factors such as team experience, topic complexity, and inter-person variability likely influenced outcomes, with **Test Review 2** having both a more experienced reviewer team and a lower complexity rating.
- PICOS summaries were generally accurate, with mistakes in the intervention/comparator domains likely introduced due to a high number of observational study designs in the test reviews. Further testing and refinement is needed to optimise this tool and ensure consistent performance across different contexts and user groups.

FIGURE 1

PICOS summary generation process



A schematic outlining the workflow to generate PICOS summaries for each article and integrate them into the literature review platform. The diagram demonstrates the process for the first article. In the application, this is repeated for n articles sequentially. [a]Article information is contained within a database in the literature review platform. The platform automates record screening and is in its third full release. [b]JSON is a key value format used to translate the PICOS into a format understandable by ChatGPT. [c]The selected model was gpt-3.5-turbo with a temperature setting of 0.2.

TABLE 1

Summary of the three test reviews and efficiency test outcomes

Test review #	Review topic	Review complexity rating[a]	Number of reviewers	Reviewer experience rating[b]	With PICOS summaries (mean abstracts/hour)	Without PICOS summaries (mean abstracts/hour)	Efficiency increase
1	Respiratory disease databases, clinical outcomes	Complex	3	2	90	60	50%
2	Neurological disease, economic outcomes	Simple	4	3	210	190[c]	11%
3	Infectious disease, clinical outcomes	Simple	6	1	Not tested[d]	Not tested[d]	Not tested[d]

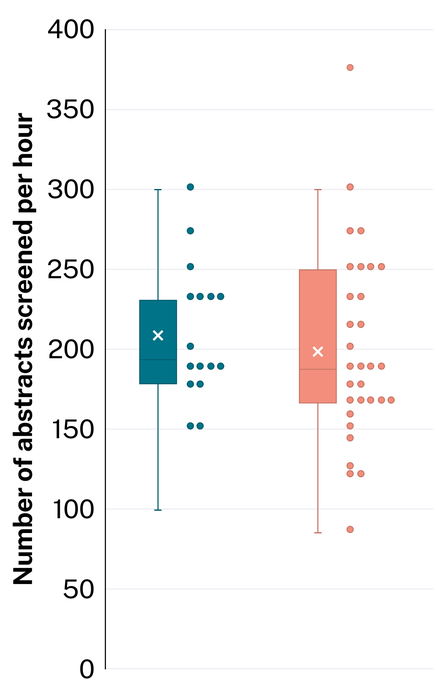
[a]Review complexity (simple/medium/complex) was assigned based on factors such as topic of the review and complexity of the eligibility criteria. [b]Experience rating was assigned based on the number of previous reviews having been conducted by each reviewer, with a higher number representing a more experienced reviewer team. [c]Mean abstracts screened/hour calculated following exclusion of one outlier. [d]Test Review 3 tested accuracy only.

FIGURE 2

Comparison of review rates for PICOS-assisted and PICOS-unassisted reviewers

Test Review 1
Mean abstract review rates were **60/hour** vs **90/hour** for PICOS-unassisted vs PICOS-assisted screening[a]

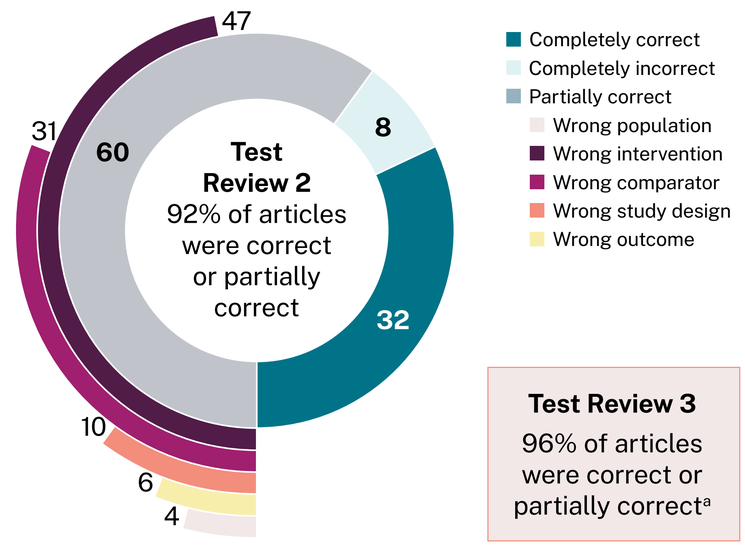
Test Review 2
Mean abstract review rates were **190/hour**[b] vs **210/hour** for PICOS-unassisted vs PICOS-assisted screening



[a]Granular recording of screening rates for individual reviewers was not conducted for Test Review 1. [b]Mean abstracts screened/hour calculated following exclusion of one outlier.

FIGURE 3

Accuracy of PICOS summaries



Test Review 3
96% of articles were correct or partially correct[a]

The numbers on the figure for Test Review 2 represent the number of PICOS summaries. [a]Granular recording of which elements of the PICOS summary were incorrect was not conducted for Test Review 3.

Conclusion

AI-generated PICOS summaries have potential to improve efficiency in the abstract review stage of literature reviews without compromising accuracy. Greater efficiency gains may be possible on reviews that are more complex while efficiency gains are likely to be smaller for reviews which are less complex and/or have a very experienced reviewer team. Further research is warranted.

Abbreviations: AI: artificial intelligence; **JSON:** JavaScript Object Notation; **PICOS:** Population, Intervention, Comparator, Outcomes, Study design.

References: [1]Sallam M. InHealthcare 2023;11(6):887.

Acknowledgements: The authors thank Jon Green, Costello Medical, for graphic design assistance. The authors thank the following systematic reviewers for their contributions to the test reviews: Hannah Borda, Julianna Catania, Shona Cross, Corrine Gregory, Lee Hughes, Emily Kaiser, Ambar Khan, Max Lee, Jose Medrano, Libby Sadler, Kylie Scott.

Presented at ISPOR Europe 2024 | Barcelona, Spain | 17–20 November