Artificial intelligence-assisted literature screening: a pilot study

Kim Wager,[a] David Gothard,[a] Andrew Liew,[b] Eleanor J Raynsford[c]

[a]Oxford PharmaGenesis, Oxford, UK; [b]Oxford PharmaGenesis, Melbourne, VIC, Australia; [c]Ipsen, London, UK

Presenter: Olivier Morteau, Ipsen US

For further information, please send your question(s) to **Eleanor J Raynsford (eleanor.raynsford@ipsen.com)**.

To download the poster, please scan the Quick Response (QR) code.

Objective

- Manually screening medical literature for articles of interest is a substantial resource burden. Here we report the development and assessment of an artificial intelligence (AI)-assisted literature screening tool (**Panel 1**).

Research design and methods

- We evaluated a proprietary AI model for:
  - filtering literature search results for relevant articles, and
  - selecting the most relevant articles based on prespecified weighted prompts (study design, population, intervention etc.) (**Panel 1**).
- The model was developed and optimized using one manually screened data set (A; N = 82 articles) and tested on a second data set (B; N = 74 articles) (**Panel 1**).
- We compared AI with manual decisions, based on precision (avoiding false positives), recall (avoiding false negatives), and accuracy (proportion of matching decisions) (**Panel 1**).
- We also evaluated the model's ability to extract and summarize information (20 articles per data set), based on accuracy and likelihood of misinterpretation (Likert scale, 1 [most likely] to 4 [least likely]) (**Panels 1 and 4**).

Results

- The model filtered for relevant articles with 78% (n = 64/82) and 73% (n = 54/74) accuracy in data sets A and B, respectively (inclusion precision, 78% [n = 46/59] and 75% [n = 39/52]; inclusion recall, 90% [n = 46/51] and 85% [n = 39/46]) (**Panel 2**).
  - The model missed only five and seven relevant articles in data sets A and B, respectively.
- However, the model performed suboptimally when selecting the most relevant articles (inclusion recall, 63% [n = 15/24] and 33% [n = 7/21]) (**Panel 3**).
- AI-generated summaries were of high quality; in both data sets, mean accuracy was 96%, and mean likelihood of misinterpretation was 3.6 (4 = lowest likelihood) (**Panel 4**).

Conclusions

- We developed an AI model capable of extracting information and summarizing published articles.
- Although the model could accurately filter relevant articles, further work is needed to improve the model's ability to judge the level of relevance.
- At this stage, AI could be used to help to lower the workload burden by reducing the number of articles that require manual screening.

Panel 1: Objective and methods

What we did

Compared **AI-assisted** literature screening with **manual** literature screening

How we did it

Tests: Two pre-existing literature searches — **82** articles Data set A; **74** articles Data set B

Assessments — How well does AI do with: filtering for relevant articles; shortlisting and selecting articles; summarizing information

Outcomes: Precision[a]; Recall[b]; Accuracy[c]; Accuracy[d]; Quality[e]

[a]Ratio of 'true positives' to 'true positives plus false positives'. [b]Ratio of 'true positives' to 'true positives plus false negatives'. [c]Proportion of matching AI and manual decisions. [d]Proportion of correct responses to eight questions. [e]Likelihood of misinterpretation based on language and ambiguity.

Panel 2: Article filtering

Data set A

Number of relevant articles

Manual decision \ AI decision	No	Yes
No	18	13
Yes	5	46

AI decision to include

	Yes	No
Precision	0.78	0.78
Recall	0.90	0.58
Accuracy	78%	

AI accurately predicted manual screening decisions for **78%** of articles but missed **5** potentially relevant articles and included **13** false positives

Data set B

Number of relevant articles

Manual decision \ AI decision	No	Yes
No	15	13
Yes	7	39

AI decision to include

	Yes	No
Precision	0.75	0.68
Recall	0.85	0.54
Accuracy	73%	

AI accurately predicted manual screening decisions for **73%** of articles but missed **7** potentially relevant articles and included **13** false positives

Panel 3: Article shortlisting

Data set A

Number of relevant articles

Manual decision \ AI decision	No	Yes
No	49	9
Yes	9	15

AI decision to include

	Yes	No
Precision	0.63	0.84
Recall	0.63	0.84
Accuracy	78%	

AI-assisted shortlisting was suboptimal with only **63%** of the topmost relevant articles included

*Note: AI-assisted 'article selection' was also suboptimal with only **50%** of manually-selected articles included (data not shown)*

Data set B

Number of relevant articles

Manual decision \ AI decision	No	Yes
No	39	14
Yes	14	7

AI decision to include

	Yes	No
Precision	0.33	0.74
Recall	0.33	0.74
Accuracy	62%	

AI-assisted shortlisting was suboptimal with only **33%** of the topmost relevant articles included

*Note: AI-assisted 'article selection' was also suboptimal with only **30%** of manually-selected articles included (data not shown)*

Panel 4: Article summarization

Analysis of a subset of 20 articles per data set

Summarization example

AI focuses on the article abstract, similar to manual screening

AI generates answers to eight questions covering:

1. Article type
2. Trial or study name
3. Treatment type
4. Main conclusion
5. Treatment groups
6. Total population size
7. Country
8. Subgroups

Factually correct[a]: ✓ ✓ ✗ ✓ ✓ ✓ ✓ ✗

Likelihood of misinterpretation[b] *Likert scale assessment of language and ambiguity*

1. Unclear language and some ambiguous answers (most likely)
2. Unclear language and no ambiguous answers
3. Clear language and some ambiguous answers
4. Clear language and no ambiguous answers (least likely)

[a]Where questions were partially answered or unanswered because there was no relevant information to extract; these were considered to have been handled correctly by the AI model.
[b]Based on reviewer judgement of all eight answers collectively.

		Data set A	Data set B
Accuracy	Mean proportion of correct answers	**96%**	**96%**
	Range	63–100%	63–100%
Quality	Mean Likert score (max. score 4)	**3.6**	**3.6**
	Range	3–4	3–4

In general, summaries were accurate with low likelihood of misinterpretation

What we found

AI was able to identify articles of potential interest, but further work is needed to align AI decisions with manual decisions. AI was able to extract information and generate clear summaries, and could therefore be used to assist rather than replace manual screening.

Abbreviations AI, artificial intelligence.

Author contributions All authors provided substantial contributions to study conception/design or acquisition/analysis/interpretation of data; drafting of the publication or reviewing it critically for important intellectual content; and gave their final approval of the publication.

Disclosures KW, DG, AL: employees of Oxford PharmaGenesis. EJR: employee of Ipsen.

Medical writing support The authors thank Tamzin Gristwood, PhD, of Oxford PharmaGenesis, Oxford, UK, for providing medical writing support, which was industry sponsored in accordance with Good Publication Practice 2022 guidelines (GPP 2022).

Disclaimer Our research utilized a commercially available large-language model, but does not evaluate the model's function and performance; rather it evaluates our adapted research application.

Presented at the 20th Annual Meeting of the International Society for Medical Publication Professionals (ISMPP) | Washington, DC, USA | April 29–May 1, 2024
Previously presented at the European Meeting of the International Society for Medical Publication Professionals (ISMPP) | London, UK | January 23–24, 2024

This study was sponsored by Ipsen