An AI-Enhanced Targeted Literature Review Workflow to Support Inclusive Research Practices and Population Science Research: A Human-in-the-Loop Approach

SA11

Obaro Evuarherhe,[1] Gemma Carter,[1] Kim Wager,[1] Ruma Bhagat,[2] Nicole Richie,[2] Bruno Jolain[2]

[1]Oxford PharmaGenesis, Oxford, UK; [2]F. Hoffmann-La Roche Ltd, Basel, Switzerland

ISPOR main topic/taxonomy: Study Approaches

ISPOR subtopics: Artificial Intelligence, Machine Learning, Predictive Analytics, Literature Review and Synthesis, Health Disparities and Equity

Scan here to download the poster

INTRODUCTION

Background

- A thorough understanding of population differences in disease outcomes is crucial to performing inclusive research successfully.
  - However, in disease areas with large volumes of evidence, screening titles and abstracts for literature reviews can be so resource and time-intensive that the findings may be out of date by the time the evidence synthesis is completed.[1]
- Large language models, such as GPT-4, offer promising opportunities to accelerate the screening stage and other steps in the literature review process while maintaining high accuracy.[2–4]

Objective

- We sought to develop and test a GPT-4-assisted screening workflow for targeted literature reviews (TLRs) to evaluate its performance across iterative prompt refinements and to demonstrate its feasibility for advancing inclusive research practices and population science research.

METHODS

- The AI-assisted screening process that we employed is illustrated in **Figure 1**.
- Searches for reports on health disparities in inflammatory bowel disease (IBD) were conducted in MEDLINE and Embase on 30 November 2023, yielding 4338 articles for screening.
- For a pilot dataset of 251 articles, a single reviewer manually screened titles and abstracts in Rayyan, a web-based citation screening tool,[5] to generate a ground truth dataset, a reference standard used to evaluate model performance.
- GPT-4 was accessed through its application programming interface, and custom Python code was developed to automate the screening workflow.
- Prompts were designed to identify studies on health disparities in IBD and included a structured list of questions covering predefined topics of interest (e.g. race, ethnicity, sex and social determinants of health). Articles were included if they addressed at least one of these topics.
- Prompts were iteratively refined over five rounds of testing, with sensitivity (the proportion of relevant articles correctly included) prioritized over specificity (the proportion of irrelevant articles correctly excluded) to minimize the risk of excluding relevant articles.
- Following each iteration, screening decisions made by GPT-4 were reviewed manually. Prompts were revised based on observed patterns of misclassification and any apparent misunderstandings of the eligibility criteria.
- In each round, changes were made to one or more individual prompt items.
- The decisions made by GPT-4 were compared with the ground truth dataset to calculate sensitivity, specificity and overall accuracy (the proportion of articles correctly classified).
- An overall accuracy of 80% or higher was considered sufficient to proceed, reflecting a strong balance between sensitivity and specificity for the objectives of the review.

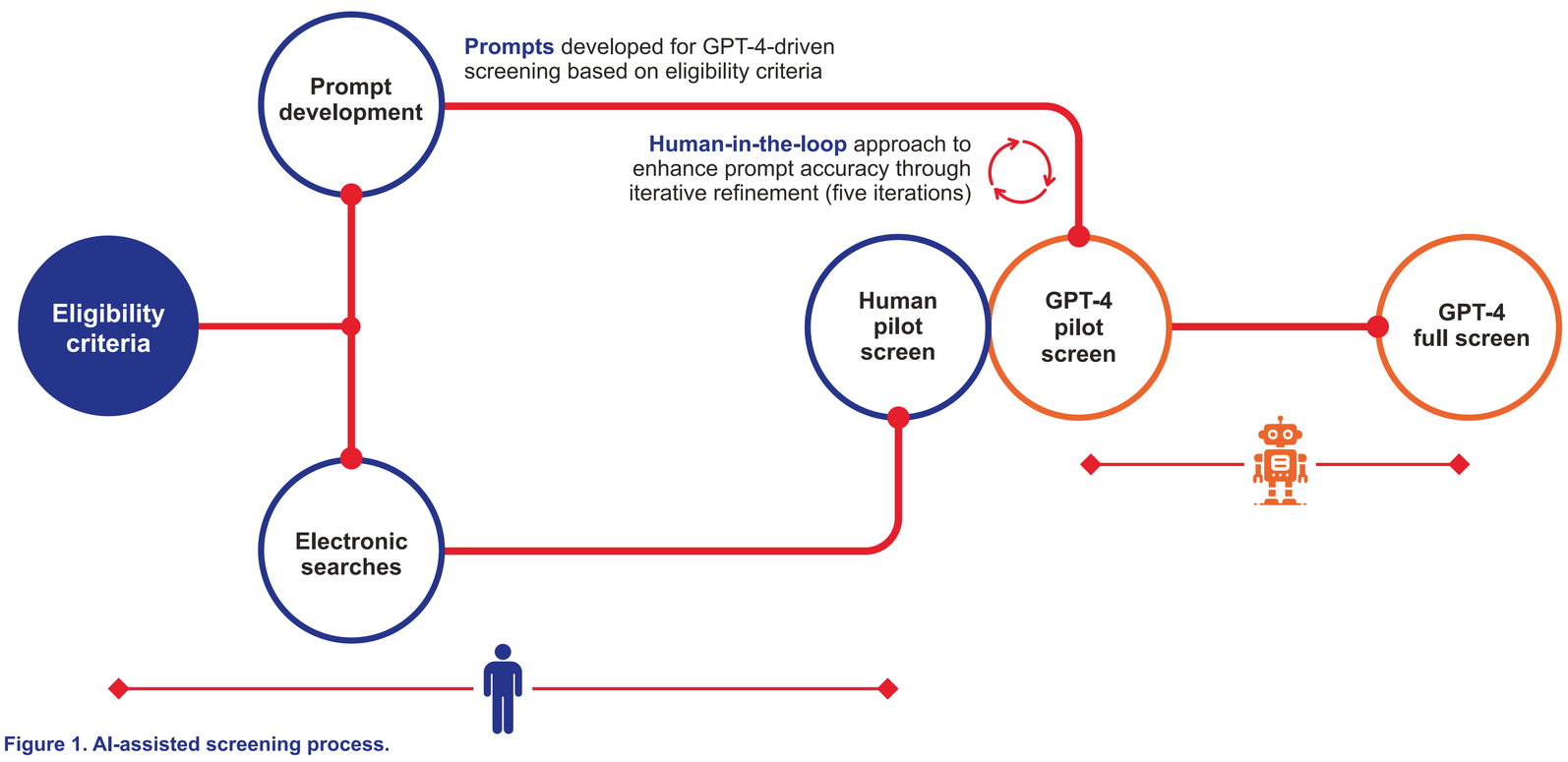
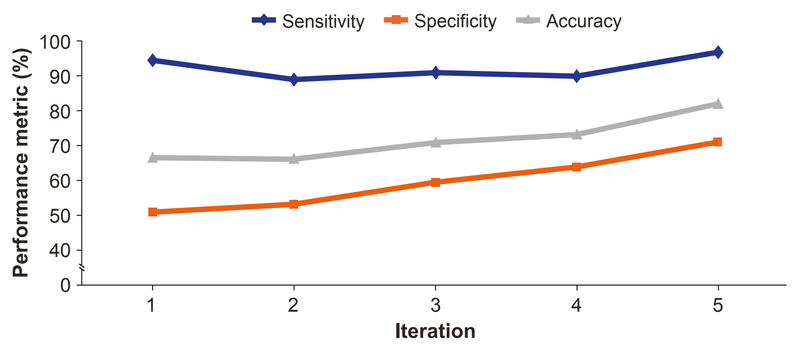


Figure 1. AI-assisted screening process.

RESULTS

- Using the first iteration of the prompts, GPT-4 achieved a sensitivity of 94.5%, a specificity of 50.6% and an overall accuracy of 66.5% (**Figure 2**) versus the ground truth dataset.
  - During this iteration, five relevant articles were incorrectly excluded.
- After five rounds of prompt refinement, the final iteration achieved a sensitivity of 97.2%, a specificity of 71.0% and an overall accuracy of 82.1% versus the ground truth dataset.
  - At this stage, three eligible articles were incorrectly excluded.
- With this level of performance, the final prompts were used in GPT-4 to screen the remaining 4087 articles without further human involvement.



Iteration	TP	TN	FN	FP	Sensitivity	Specificity	Accuracy
1	86	81	5	79	94.5	50.6	66.5
2	81	85	10	75	89.0	53.1	66.1
3	83	95	8	65	91.2	59.4	70.9
4	82	102	9	58	90.1	63.8	73.3
5	103	103	3	42	97.2	71.0	82.1

Figure 2. Performance metrics across iterations.
Sensitivity = the proportion of relevant articles correctly included by GPT-4: TP / (TP + FN). Specificity = the proportion of irrelevant articles correctly excluded by GPT-4: TN / (TN + FP). Accuracy = the proportion of all articles correctly classified by GPT-4: (TP + TN) / (TP + TN + FP + FN).
FN, false negative (included by humans but excluded by AI); FP, false positive (excluded by humans but included by AI); TN, true negative (excluded by both); TP, true positive (included by both).

DISCUSSION

- These results demonstrate that systematic prompt refinement improves both sensitivity and specificity, reducing the number of missed relevant articles while enhancing overall accuracy.
- Human oversight remained essential to monitor model performance, guide prompt development and ensure that outputs align with review objectives, with human reviewers needing to become experts in prompt development as well as health-related TLRs.
- A key strength of this approach was its ability to screen large volumes of literature more efficiently than human reviewers.
- We prioritized sensitivity over specificity to avoid excluding relevant studies and, as a result, more false positives were included than would have been the case if greater priority had been given to specificity.

Conclusion

- This scalable, robust approach facilitates TLRs that typically have high screening burdens and were previously limited by resource constraints, such as those required for inclusive research practices and population science research to inform clinical trial design and meet regulatory diversity requirements.

Abbreviations

FN, false negative; FP, false positive; IBD, inflammatory bowel disease; TLR, targeted literature review; TN, true negative; TP, true positive.

Funding statement

The population science research TLR was funded by F. Hoffmann-La Roche Ltd.

Author disclosures

Obaro Evuarherhe, Gemma Carter and Kim Wager are employees of Oxford PharmaGenesis, and Gemma Carter is a shareholder of Oxford PharmaGenesis. Ruma Bhagat, Nicole Richie and Bruno Jolain are employees and shareholders of F. Hoffmann-La Roche Ltd.

References

1. Borah R *et al. BMJ Open* 2017;7:e012545.
2. Delgado-Chaves FM *et al. Proc Natl Acad Sci U S A* 2025;122:e2411962122.
3. Khraisha Q *et al. Res Synth Methods* 2024;15:616–26.
4. Scherbakov D *et al. J Am Med Inform Assoc* 2025;32:1071–86.
5. Ouzzani M *et al. Syst Rev* 2016;5:210.

Presented at International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Europe 2025, 9–12 November 2025, Glasgow, UK