Human *vs* machine in medical search strategy development: a comparative evaluation of ChatGPT-4.1

Julia Walz
Library of the Center of Dental Medicine, University Medical Center, Freiburg, Germany

Abstract
This study examined the potential of generative artificial intelligence, specifically ChatGPT-4.1, to support the development of search strategies. Using two Cochrane Review topics as benchmarks, AI-generated MEDLINE strategies were compared to the expert strategies using PRESS (Peer Review of Electronic Search Strategies) criteria. Results show that ChatGPT-4.1 can accurately translate research questions and apply Ovid syntax and Boolean operators, but shows considerable weaknesses in subject heading and text word selection.

Key words: *artificial intelligence; information storage and retrieval; systematic reviews as topic; methods.*

Background

In the context of Evidence-Based Medicine (EBM) effective Information Retrieval (IR) is essential but often challenging, as it requires domain knowledge, advanced search skills, and considerable time – resources that are limited in clinical practice. In consequence, integrating EBM into everyday workflows remains difficult. Recent advances in generative Artificial Intelligence (AI) offer new possibilities to support IR by facilitating the development of search strategies and by mitigating limitations in time and IR expertise.

Purpose

The study aimed to investigate the current potential of generative artificial intelligence – specifically ChatGPT-4.1 – to support the development of search strategies for EBM. To address this, the study tested two hypotheses:

- *H1:* ChatGPT-4.1 is capable of generating high-quality medical search strategies that are comparable to or better than those created by human experts;
- *H2:* the use of ChatGPT-4.1 significantly reduces the time required to develop medical search strategies.

Methods

To assess the potential of generative AI in supporting the development of medical search strategies, two Cochrane Review topics – Example 1 (1), Example 2 (2) – were selected as basis for the intended comparison. Cochrane Reviews were chosen as benchmarks due to their methodological rigor and expert-designed strategies. Both reviews were published after the training cutoff of ChatGPT-4.1 (June 2024 (3)) and had no prior versions, thus minimizing the risk of training-related bias.

Search strategies for MEDLINE (Ovid) were generated with ChatGPT-4.1 (OpenAI) via API for both topics using three different prompting approaches: Zero-Shot (the model answers without being given any example), One-Shot (the model gets one example to guide its answer), and Chain-of-Thought (CoT; in a step-by-step conversational style, the model, breaks a complex task into smaller steps and processes them one after the other before giving the final answer). The prompt structures were standardized across topics; only clinical questions and background details varied. The One-Shot technique included an illustrative Cochrane example (4), while the CoT approach involved stepwise reasoning guidance. For Zero- and

Address for correspondence: Julia Walz, Library of the Center of Dental Medicine, University Medical Center, Hugstetterstr. 55, 79106 Freiburg, Germany. E-mail: julia.walz.zmk@uniklinik-freiburg.de

One-Shot, only the model's initial outputs were analyzed; for CoT, the first complete strategy was used. The AI-generated and the benchmark strategies were evaluated using the PRESS (Peer Review of Electronic Search Strategies) checklist (5), covering six categories: 1. translation of the research question, 2. Boolean and proximity operators, 3. subject headings, 4. text word searching, 5. spelling, syntax and line numbers, and 6. limits and filters (see Data availability statement before the References).

Results

Translation of the research question

ChatGPT-4.1 consistently applied the PICO framework across all prompting approaches, except for the Zero-Shot condition for Example 2. Evaluation against the PRESSed expert strategies indicated that the AI systematically captured the relevant aspects of the research question, demonstrating strong capabilities in text analysis, contextual understanding, and thematic grouping of research topics. Across most conditions, the model used the same aspects as the benchmark strategies for searching. An exception was the Zero-Shot technique, where all identified thematic aspects were included in the search. This resulted in highly precise but overly narrow retrievals, which are generally unsuitable for comprehensive systematic searches. Overall, in structuring the research question and identifying relevant components, the One-Shot and CoT approaches aligned closely with expert strategies.

Boolean and proximity operators. Spelling, syntax, and line numbers

Across all strategies ChatGPT-4.1 applied the Ovid-specific syntax correctly. Nonetheless display issues occurred: in the Zero-Shot outputs of both examples and the CoT strategy for Example 2, truncation symbols were interpreted as Markdown formatting, which resulted in italicized text. This required time-consuming manual correction, otherwise the searches would have been largely ineffective. Correct Boolean logic was applied consistently. However, hallucinated MeSH terms caused misalignment in line numbering, requiring manual adjustment of OR and AND connections. Proximity operators (ADJ) were formally correct but not always semantically optimal: some terms were searched separately rather than being placed in close proximity to other relevant terms which would have yielded better results. Some terms were searched with a proximity operator, although treating them as a phrase would have been more appropriate. Overall, the AI demonstrated strong formal correctness but only limited capability to encode nuanced IR knowledge for adjacency operator selection.

Subject headings

ChatGPT-4.1 demonstrated insufficient understanding of MeSH hierarchies and term definitions: the model applied term explosion arbitrarily, occasionally even introducing false descriptors. The model seems unable to distinguish between relevant and irrelevant terms of the controlled vocabulary leading to unnecessary term inclusions. Hallucinated terms occurred in five of six AI-generated strategies (except in the Zero-Shot strategy in Example 2). Furthermore, AI is unaware of thesaurus term relationships. In CoT 2, the AI used the newer term without referencing the older one. While the CoT approach showed greater creativity and generated more search terms for complex topics, the AI struggled to fully capture multidimensional aspects, which require the use of a range of relevant descriptors. Consequently, AI-generated subject heading usage was qualitatively inferior to the expert strategies, requiring manual verification and correction, which mitigates any potential time savings. Nonetheless, the model reliably suggested central MeSH terms for each PICO aspect, offering modest assistance in initial term identification. With increasing topic complexity, performance deteriorated.

Text word searching

Although ChatGPT-4.1 generally identified central text words for each PICO aspect, it also included irrelevant terms and sometimes overly specific terms. For more complex or multidimensional topics, the AI-generated strategies contained fewer synonyms and related terms than the expert strategies, in particular in the Zero- and One-Shot prompt outputs. The dialogical CoT approach performed better but still included fewer terms than the human experts. In general, truncation was applied correctly but nonetheless arbitrarily, occasionally missing essential variants. Therefore, IR-knowledge of the user for proper interpretation is essential. In most cases the AI gave no explanation for its search field se-

lection, which sometimes seemed appropriate (ti,ab,kf) but sometimes insufficient (only ti,ab) for systematic retrieval. Considering common term occurrence patterns, the search words were not optimally combined (OR, ADJ, phrase) by the AI. However, a notable strength of the model was its ability to consistently account for both British and American spellings of the selected terms. Overall, the AI works adequately for identifying candidate text words, but combination, truncation, and field selection quality are insufficient for systematic searches. Manual review is necessary, reducing potential time savings. The model remains useful for initial term identification.

Limits and filters

Similar to its performance with Boolean operators, ChatGPT-4.1 correctly implemented syntax limits and filters, thereby demonstrating formal precision. But in one instance (Example 2, CoT), the AI applied an older Cochrane RCT filter (2008 version) rather than the current 2023 version, falsely claiming that the version used was the most up-to-date. This discrepancy presumably reflects the AI's training cutoff and its exposure to more examples of the earlier filter. Notably, the AI applied the validated filter correctly and completely, surpassing the experts in this specific case, who used the filter with a slight modification. Overall, the applied filters and limits were relevant for the research questions but often their application was unnecessary, occasionally even counterproductive (e.g., language restriction). Regarding limits and filters, AI cannot replace expert judgment: the asserted currency of filters used cannot always be relied upon, and the AI does not evaluate the necessity, relevance, or potential sensitivity loss associated with limits and filters. Consequently, manual review remains essential, therefore limiting any time savings.

Conclusion

H1: ChatGPT-4.1 displays strengths in structuring the research question and in correctly applying Ovid syntax and Boolean operators, thus enabling the generation of functionally correct search strategies. However, significant limitations exist in subject heading and text word selection: while the AI reliably identifies central terms, it struggles with indiscriminate term explosion, hallucinated descriptors, and incomplete coverage of multidimensional topics. As a result, five of six AI-generated strategies retrieved the relevant articles included in the Cochrane Reviews, but strategies were generally formulated too precisely for comprehensive systematic searches. Expert judgment remains essential to identify and correct errors. AI can serve effectively as a generator of ideas or for drafting initial strategies, particularly when context information and sequential prompts (Chain-of-Thought) are used. Therefore, H1 is confirmed for the PRESS categories "translation of the research question" and "spelling, syntax and line numbers", partially confirmed for "boolean and proximity operators" and "limits and filters", but not supported for "subject headings" and "text word searching".

H2: Regarding efficiency, ChatGPT-4.1's strengths in structuring the research question, applying syntax, and using Boolean operators allow for the rapid generation of functionally correct strategies and the suggestion of relevant search terms. However, time savings diminish when expert validation is needed – particularly for proximity operators (to assess appropriate word distances), subject headings and text word searches (to verify completeness, term placement, and linking logic), and filters or limits (which, while syntactically correct, require confirmation of validity and currency). Additional time may also be needed for correcting technical errors, such as missed truncation symbols or shifted line numbering. Thus, H2 is confirmed for "translation of the research question" and "spelling, syntax and line numbers", partially confirmed for "boolean and proximity operators", but not supported for "subject headings", "text word searching", and "limits and filters".

Acknowledgment

This article is based on the author's thesis to obtain a Master degree from the University of Applied Sciences Cologne. Therefore, methodological limitations of this study include the small number of search examples, evaluation by a single reviewer, and the lack of medical expertise for term relevance assessment, restricting the generalizability of findings. Nevertheless, this study highlights the potential of generative AI to assist in search strategy development, with the caveat that expert oversight remains crucial for high-quality systematic retrieval.

Data availability

Further information and data (prompts, AI-outputs, complete PRESS-evaluation tables in German) are available from the author upon request.

Submitted on 12 November 2025.
Accepted on 24 November 2025.

REFERENCES

1. Williams CR, Huffstetler HE, Nyamtema AS, Larkai E, Lyimo M, Kanellopoulou A, et al. Transfusion of blood and blood products for the management of postpartum haemorrhage. Cochrane Database of Systematic Reviews. 2025 Feb 6;2025(2):CD016168. doi:10.1002/14651858.CD016168
2. Wagner C, Ernst M, Cryns N, Oeser A, Messer S, Wender A, et al. Cardiovascular training for fatigue in people with cancer. Cochrane Central Editorial Service, editor. Cochrane Database of Systematic Reviews. 2025 Feb 20;2025(2):CD015517. doi:10.1002/14651858.CD015517
3. OpenAI. Introducing GPT-4.1 in the API [Internet]. 2025 [cited 2025 May 30]. Available from: https://openai.com/index/gpt-4-1/
4. Whittle SL, Johnston RV, McDonald S, Worthley D, Campbell TM, Cyril S, et al. Stem cell injections for osteoarthritis of the knee. Cochrane Central Editorial Service, editor. Cochrane Database of Systematic Reviews [Internet]. 2025 Apr 2 [cited 2025 May 27];2025(4). Available from: http://doi.wiley.com/10.1002/14651858.CD013342.pub2
5. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement. Journal of Clinical Epidemiology. 2016 Jul 1;75:40-6.

This paper is published under a CC BY license