The use of a large language models to create plain language summaries of evidence reviews in healthcare: a feasibility study

Colleen Ovelman¹, Shannon Kugley², Gerald Gartlehner², and Meera Viswanathan²

¹Cochrane
²RTI International

November 8, 2023

Abstract

Introduction: Plain language summaries (PLSs) make complex healthcare evidence accessible to patients and the public. Large language models (LLMs) may assist in generating accurate, readable PLSs. This study explored using the LLM Claude 2 to create PLSs of evidence reviews from the Agency for Healthcare Research and Quality (AHRQ) Effective Health Care Program. Methods: We selected 10 evidence reviews published from 2021-2023, representing a range of methods and topics. We iteratively developed a prompt to guide Claude 2 in creating PLSs which included specifications for plain language, reading level, length, organizational structure, active voice, and inclusive language. PLSs were assessed for adherence to prompt specifications, comprehensiveness, accuracy, readability, and cultural sensitivity. Results: All PLSs met the word count. We judged one PLS as fully comprehensive; 7 mostly comprehensive. We judged 2 PLSs as fully capturing the PICO elements; 5 with minor PICO errors. We judged 3 PLSs as accurately reporting the results; 4 with minor result errors. We judged 3 PLSs as having major result errors for incorrectly reporting total participants. Five PLSs met the target 6th-8th grade reading level. Passive voice use averaged 16%. All PLSs used inclusive language. Conclusions: LLMs show promise for assisting in PLS creation but likely require human input to ensure accuracy, comprehensiveness, and the appropriate nuances of interpretation. Iterative prompt refinement may improve results and address the needs of specific reviews and audiences. As text-only summaries, the AI-generated PLSs could not meet all consumer communication criteria, such as textual design and visual representations. Further testing should explore how to best leverage LLM support in drafting PLS text for complex evidence reviews.

Hosted file

7. Accuracy (dimensions of PICO are described correctly)

- Yes: 2
- Mostly yes (1-3 minor errors): 5
- No (more than 3 minor errors; a...): 3

8. Accuracy (results are correct)

- Yes: 3
- Mostly yes (1-3 minor errors): 4
- No (more than 3 minor errors; a...): 3