Published in final edited form as:

Nat Med. 2023 July ; 29(7): 1593–1594. doi:10.1038/s41591-023-02366-9.

AI-generated text may have a role in evidence-based medicine

Yifan Peng[1,✉], **Justin F. Rousseau**[2,3,✉], **Edward H. Shortliffe**[4,✉], **Chunhua Weng**[4,✉]

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA.

[2]Department of Neurology, Dell Medical School, University of Texas at Austin, Austin, TX, USA.

[3]Department of Population Health, Dell Medical School, University of Texas at Austin, Austin, TX, USA.

[4]Department of Biomedical Informatics, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA.

Abstract

Evidence-based medicine (EBM) requires the retrieval and ranking of relevant evidence by epistemological strength, to identify the most appropriate evidence to inform guidelines and policies, with a preference for robust evidence from randomized clinical trials (RCTs), systematic reviews, and meta-analyses. The explosive growth of the scientific literature and the emergence of new sources of evidence, including social media, case reports, and large-scale observational studies, as well as the free-text nature of this large body of evidence, collectively make it difficult to appraise and select the best available evidence.

Large language models (LLMs), exemplified by ChatGPT[1,2], are already affecting document generation, including creating legal documents, news, and medical writing. ChatGPT can interpret the context of prompts and generate grammatically correct and semantically meaningful answers, compose essays indistinguishable from those written by humans, and author captivating medical research abstracts[3]. In our own experience, ChatGPT can summarize pre-written systematic reviews. However, ChatGPT tends to miss important attributes in the summary, such as failing to refer to short-term or long-term outcomes that often have varying risks. In addition, the content generated by ChatGPT has been reported to contain factual errors, incorrect data, and misrepresentations[4]. ChatGPT was not explicitly trained for EBM and is not intended to be used for this crucial task[5]. Nevertheless, there are opportunities to use LLMs such as ChatGPT in EBM as long as pitfalls can be avoided (Table 1).

EBM is distinguished from other forms of medical or clinical research writing by the requirement for precise criteria for including and excluding studies. The PICO (patient, intervention, comparison, and outcome) framework is by far the most widely adopted process for structuring eligibility criteria and, therefore, formulating the clinical evidence

✉ yip4002@med.cornell.edu; justin.rousseau@austin.utexas.edu; ted@shortliffe.net; cw2384@cumc.columbia.edu.

Competing interests

The authors declare no competing interests

queries for evidence syntheses. Clinical evidence retrieval involves finding RCTs that meet the PICO criteria pertinent to the clinical research question so that those criteria may be aggregated in a systematic review and meta-analysis.

LLMs can extract PICO elements using few-shot learning — a technique to generate models from a limited number of labeled examples — which has already been used for text classification and summarization. Natural language generation capabilities enable LLMs to solve tasks by completing a prompt without needing to fine-tune the parameters, thereby reducing annotation costs and speeding up PICO extraction[6]. LLMs should be compared against traditional language models, as well as with PICO assessments extracted manually by evidence reviewers, to check the correctness and completeness of the PICO output. Use of LLMs should be tested across different biomedical domains, to assess the suitability of using LLMs for different topics.

Evidence synthesis aims to systematically integrate findings from multiple studies to draw evidence-based conclusions that inform healthcare decisions, and it is constantly and rapidly evolving[7]. The number of RCTs proliferates at an unprecedented rate, with new sources of evidence that are increasingly difficult to identify by literature searches. Some have referred to this deluge as an ‘infodemic’, in which it is unfeasible for the average clinician to keep up with the latest knowledge. LLMs could, in the future, be valuable tools for providing conversational answers to complex medical questions. However, answers from LLMs may be obsolete, self-contradictory, or inaccurate.

Obsolescence is already a problem with clinical practice guidelines; for example, a study by the US Agency for Healthcare Research and Quality (AHRQ) showed that half of the 17 guidelines studied became obsolete within 5–6 years[8]. This provides an existing challenge to EBM, which is a continuing process with clinical evidence reviewed regularly to assess whether the evidence synthesis is up-to-date. LLMs cannot currently learn and update their knowledge base continuously. They are trained on data collected at a specific point in time (for example, the current version of ChatGPT was trained on data up to 2021), and so cannot incorporate the latest clinical evidence as it becomes available. If LLMs are to be used in medicine, they will need to be continuously updated with new research as it is published.

Current LLMs are also unable to differentiate current evidence from obsolete studies. Future LLMs should be able to identify time-sensitive evidence and perform temporal reasoning to understand how evidence evolves over time, such as recognizing when new guidelines have supplanted old ones.

Much clinical evidence is of high quality, but individual studies may have issues in their planning, conduct, representativeness, analysis, or reporting; reliance on these studies could therefore result in harm[9]. All doctors should be able to critically appraise and synthesize relevant and reliable evidence to answer a specific clinical question, such as assessing whether a study applies to their patient population.

LLMs can answer questions with explicit reasoning steps — a feature that could be useful in supporting clinical reasoning. Despite this evidence of reasoning by LLMs, further research is needed to incorporate knowledge-based reasoning capabilities into LLMs, to enable them

to perform complex, robust, and explainable thinking, and to determine their potential for use in EBM. The medical evidence community should construct benchmarks to evaluate the reasoning abilities of LLMs. Existing general benchmarks, such as arithmetic, symbolic, and table reasoning, could be incorporated into EBM. New metrics to evaluate EBM reasoning quantitatively may be needed to assess whether LLMs are able to reason in a way that is similar to medical professionals.

A central component of clinical reasoning is its knowledge base, which is often stored in complex structured data. We believe the LLMs and well-curated clinical knowledge bases complement each other, and the biomedical informatics and health data science community can contribute large-scale shareable knowledge bases to equip LLMs with computable knowledge and to enable LLMs to perform reliable reasoning.

Patients or health consumers should be part of clinical decision-making, but the technical and complex medical terminology used within systematic reviews and other health literature can be difficult for some people to understand. LLMs could be used to replace technical jargon with plain language, making medical information more accessible to a wider audience. This could increase engagement and empower patients and the public in their own health decision-making.

Lay summaries generated by LLMs may suffer from inadequate accuracy and should not be used in isolation. Some information may be omitted, and recommendations could be ambiguous or confusing. Artificial intelligence (AI)-generated lay summaries should always be reviewed by experts and corrected or clarified where necessary, so that they provide accurate and comprehensible information. More importantly, both human-written summaries from systematic reviews and lay summaries should be made available to the patients to support information provenance. This will help to ensure that the information provided is accurate and complete. If these concerns are addressed, then medical decision-making could be improved by changing how future evidence is generated, evaluated, synthesized, and disseminated.

References

1. Stokel-Walker C Nature 10.1038/d41586-022-04397-7 (2022).
2. Stokel-Walker C et al. Nature 614, 214–216 (2023). [PubMed: 36747115]
3. Gao CA et al. npj Digit. Med 6, 1–5 (2023). [PubMed: 36596833]
4. Eva AM Nature 614, 224–226 (2023). [PubMed: 36737653]
5. Zhavoronkov A Nat. Med 29, 532 (2023). [PubMed: 36750659]
6. Hedderich MA et al. In Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2545–2568 (2021).
7. Elliot J et al. Nature 600, 383–385 (2021). [PubMed: 34912079]
8. Shekelle PG et al. JAMA 286, 1461–1467 (2001). [PubMed: 11572738]
9. Goldstein A et al. J. Am. Med. Inform. Assoc 24, 1192–1203 (2017). [PubMed: 28541552]

Table 1 |

The advantages of and concerns about use of LLMs in EBM

Use of LLM	Definition of use	Advantages of LLM	Concerns
Evidence retrieval	Search, identify, and collect relevant information for clinical claims.	Efficiently processing large volumes of text.	Lack of understanding of the claim; ethical concerns about privacy, accountability, and biases.
Evidence synthesis	Systematically integrate findings from several studies to draw evidence-based conclusions.	Text synthesis and summarization.	Lack of continious learning capability; lack of temporal reasoning.
Clinical reasoning	Collect and analyze patient's status to arrive at medical decisions about a patient's care.	Answering questions with explicit reasoning steps.	Lack of benchmarks for evaluation of reasoning; lack of knowledge-based reasoning; potential to cause medical harm.
Evidence dissemination	Communicate and distribute evidence-based practice to target audiences.	Coherent, easy to understand.	Lack of factual consistency and comprehensiveness.