MAYO CLINIC

MAYO CLINIC PROCEEDINGS:
DIGITAL HEALTH

COMMENTARY

Future of Evidence Synthesis: Automated, Living, and Interactive Systematic Reviews and Meta-analyses

Irbaz Bin Riaz, MD, MS, MBI, PhD; Syed Arsalan Ahmed Naqvi, MD; Bashar Hasan, MD; and Mohammad Hassan Murad, MD, MPH

Systematic reviews and meta-analyses (SRMAs) are a mandatory step for making clinical, public health and policy decisions; however, they require manual human effort, making them time-consuming and resource intensive requiring 6-18 months to complete.[1] With new studies published every day, SRMAs become outdated very quickly and require frequent updates. In settings such as the coronavirus disease pandemic and cancer care, where new data are rapidly generated, traditional SRMAs become challenging or impractical. A strong rationale for automating SRMA has existed for many years, and researchers and methodologists have envisioned living SRMAs that are continuously updated with new evidence.

Although several previous efforts using conventional machine learning and deep learning techniques have attempted to automate different SRMA steps, the results have been limited in scope and performance.[2] The recent emergence of large language models (LLMs) represents a paradigm shift in artificial intelligence (AI) and provides an opportunity to transform the practice of evidence synthesis. In this study, we address the provocative question: can Chat generative pretrained transformer (ChatGPT) or similar LLMs conduct an SRMA? We describe the current landscape and propose an AI-empowered integrated framework for living interactive SRMAs for efficient and scalable evidence synthesis.

Can ChatGPT (or Other AI Approaches) Conduct an SRMA?

The success and limitations of current models vary on the basis of different steps in an SRMA.

Literature Search. Typically, an information specialist or a medical librarian designs a comprehensive search strategy executed in at least 2 databases to identify all relevant studies. One study[3] found the potential utility of ChatGPT in generating Boolean queries that lead to high search precision (positive predictive value; defined as the ratio of correctly identified positive observations to the total identified positive observations by the model), albeit at the cost of lower recall (true positive rate or sensitivity; defined as correctly identified positive observations by the model to the total number of actual positive observations). However, they relied on a single, oversimplified search using Boolean queries, which does not fully capture the complex search strategies designed by medical librarians or information specialists as in systematic reviews. Another study[4] highlighted the shortcomings of using ChatGPT for literature searches in systematic reviews. First, it missed synonymous terms, resulting in overlooked studies. Second, its search strategies lacked proper structure and included irrelevant search terms. Third, it hallucinated unjustified search deadlines. Fourth, it lacked validated filters for specific study designs, such as randomized controlled trials. This step of SRMA remains in need of sophisticated AI approaches that balance precision and recall while leveraging the expertise of medical librarians.

Selecting Studies for Inclusion. Traditionally, this is done in 2 sequential steps: title and abstract screening and full-text screening. Screening titles and abstracts is where most AI effort and success have been made. Several machine learning approaches, including natural language processing, have

From the Evidence-based Practice Center, Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Healthcare Delivery, Mayo Clinic, Rochester, MN (I.B.R., S.A.A.N., B.H., M.H.M.); Division of Hematology-Oncology, Department of Medicine, Mayo Clinic, Phoenix, AZ (I.B.R., S.A.A.N.); and Division of Public Health, Infectious Diseases and Occupational Medicine, Mayo Clinic, Rochester, MN (B.H., M.H.M.).

been used to automate the process and have shown promising results in reducing the manual effort required in screening new citations. These models have been implemented in various proprietary and open-source platforms.[5–8] Most of the recent LLM efforts have also focused on automating the screening of titles and abstracts.[9] A recent study[10] evaluated the performance of 8 generative LLMs. The study found that the LlaMa2 (7 billion parameters) model outperforms other variants, including the 13-billion parameter LlaMa2 model, and achieves results comparable with or better than other models, such as Bio-SIEVE.[11] The study also demonstrates that calibration methods, including extrapolation from the collection and using seed studies, can improve recall, making the approach more useful for SRMA automation. Likewise, combining instruction fine-tuning with an ensemble of zero-shot models can substantially improve performance and time savings compared with existing approaches.[10]

Limited research has been done on automated full-text screening, and little progress has been reported to date. However, the increasing ability of LLMs to use linear-time sequence modeling and process larger size of text by accepting increasing token sizes holds promise.[12,13] Full-text review also requires considerable computational resources and may not always be necessary, particularly if the title and abstract screening process is highly accurate.

Data Extraction. Despite notable advancements in information extraction, data extraction poses unique challenges, such as converting portable document formats into natural language processing-readable formats and linking different preceding and succeeding steps and work streams involved in SRMAs. Limited prototypic data extraction techniques and commercial platforms for conducting systematic reviews enable automated extraction of data elements by retrieving candidate sentences.[14] A randomized trial found that computer-assisted data collection in which certain parts of the text are identified, highlighted, and copied for a human to paste, has an error rate of 17%.[15] Even with the use of LLM, the performance for data extraction remains suboptimal at this time.[16–18]

Risk of Bias Assessment. This step is very time-consuming and susceptible to human error and subjectivity. Agreement between humans is often poor.[19] In a study[20] assessing the performance of generative pretrained transformer-4 using the risk of bias in nonrandomized studies of interventions-I tool to appraise nonrandomized studies, the agreement between generative pretrained transformer-4 and humans ranged from 31% to 71% across various bias domains. These findings reflect the need to evaluate and improve LLMs risk of bias assessment-related reasoning abilities.[21] This improvement could be achieved by understanding how LLM makes mistakes and integrating specific training strategies on diverse training data.

Meta-analysis. The statistical pooling of effect estimates across studies is an ideal setting for automation because it is a rule-based quantitative process. However, the decision of which analysis model and statistical method is complex. Initiatives such as the OpenAI advanced data analysis platform[22] offer a promising avenue to perform advanced analysis using natural language prompts. They may bridge the gap between complex statistical methods and the practical needs of systematic reviewers. However, such tools would be standalone analysis tools without an integrated workstream. Therefore, considering the immense potential for automation in this domain, efforts should be directed toward developing user-friendly interfaces that can be easily integrated into existing work streams for SRMAs.

Certainty of Evidence Assessment. This step represents the highest level of synthesis in which we determine the trustworthiness of a body of evidence for each outcome considering factors such as risk of bias, inconsistency, indirectness, imprecision, and publication bias. The Grading of Recommendations Assessment, Development, and Evaluation approach is widely used.[23] This complex and multifaceted task relies heavily on contextual awareness and a subjective assessment of the evidence, making it challenging to automate fully. To our knowledge, there are no available data regarding the performance of LLMs to adjudicate certainty of evidence.

Narration and Interpretation of Results. Expressing treatment effects using plain language and narrative statements is critical to accurately communicating research findings to end users. Large language models' ability to process and generate human-like text coupled with the availability of standard frameworks[24] to describe treatment effects suggests that this step is likely to be automated in the future.

A Framework for Living Interactive Evidence Synthesis. The ability and accuracy of LLMs only represent part of the solution. A framework that harnesses the potential of LLMs and integrates them into an evidence synthesis workflow is needed. Such a framework should include specific roles for humans and AI and allow their interactions. Moreover, SRMAs can have appendices that exceed 100 pages, particularly in the case of network meta-analyses that evaluate multiple interventions. Limited journal space constraints, unidimensional tables, and static figures often omit critical details and result in a poor presentation of findings. Hence, the proposed framework must also have interactive features that tailor the presentation to the end-user's needs and support shared decision-making at the point of care.

Therefore, we propose an AI-supported, linear, integrated, interactive, and collaborative framework—the living interactive evidence synthesis framework—that enables continuous, real-time evidence updating as new studies become available, enabling a "living" synthesis. A prototype of this framework is being developed and piloted at the Mayo Clinic[25] and has been used for several network meta-analyses in the field of oncology.[26,27] The process is categorized into 6 major steps each addressed in a module: evidence search (The Watcher), screening (The Screener), data extraction (The Extractor), data analysis (The Analyzer), and the summary of results (The Tabulator) (Figure).

Limitations and Future Outlook

The emergence of LLMs presents an exciting opportunity to integrate them into the

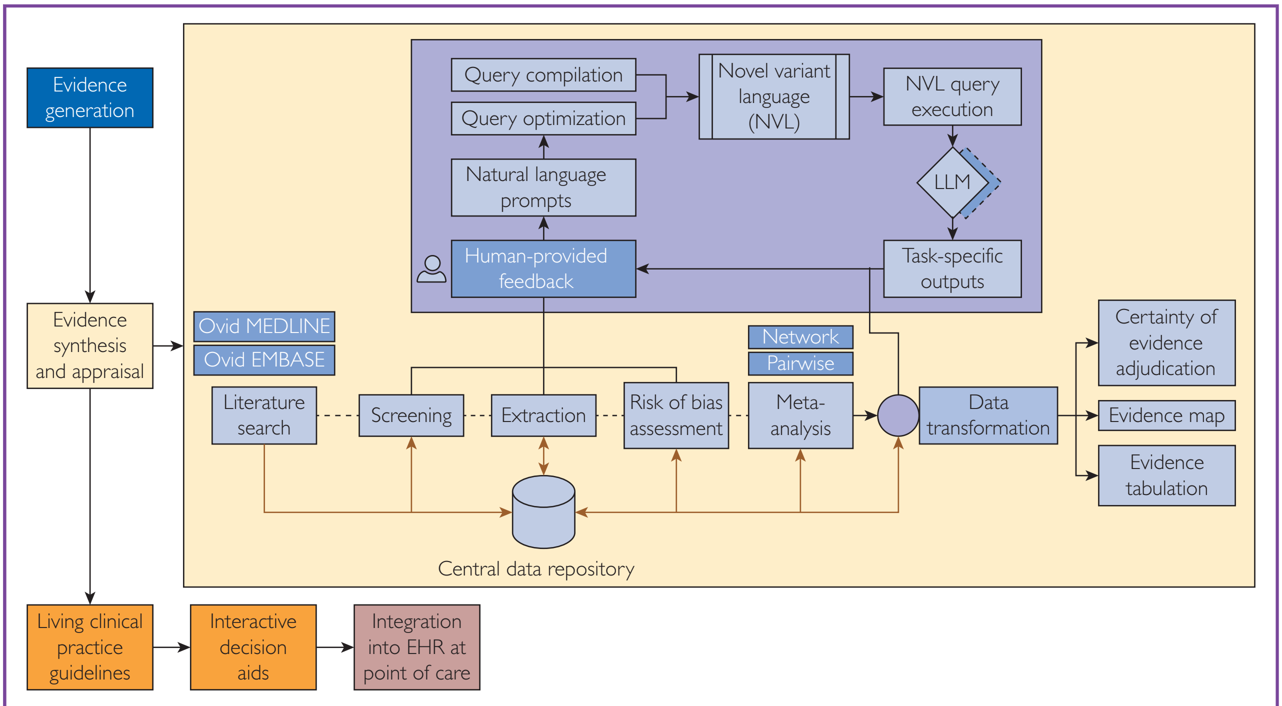

FIGURE. Proposed framework for human-artificial intelligence collaboration: living interactive evidence synthesis framework. EHR, electronic health record; LLM, large language model; NVL, novel variant language.

evidence synthesis workflow. We propose shifting from the conventional approach of relying on LLMs to automate the process to work toward integrating LLM agents in a comprehensive framework.[25] This shift necessitates the development of a human-AI collaboration system that not only leverages the remarkable capabilities of LLMs but also avoids the black-box phenomenon, ensuring transparency, and reproducibility, which are essential to SRMAs. It is also crucial to acknowledge the propensity of LLMs to "hallucinate" or generate outputs that are not grounded in factual information.[28] This phenomenon can lead to the generation of inaccurate or misleading content. Through human validation, iterative refinement, and expert curation, a human-AI system facilitates a collaborative environment where human experts can provide oversight and guidance, thereby reducing the likelihood of hallucinations. Additionally, strategies like truth-based conditional training[29] and corrective retrieval augmented generation[30] can help mitigate the risks of hallucinated outputs.

Emerging data suggest that leveraging a translation mechanism to convert human expertise into novel variant languages—the declarative-declarative language and the declarative-imperative language—for feeding LLMs improves their performance.[31] This approach could be used for any step in the SRMA process. Likewise, using LLM-LLM interactions may be an opportunity to emulate the traditional practice of involving 2 independent reviewers in the real world while leveraging LLM agents' capabilities to enhance efficiency, consistency, and objectivity. Absolute concordance between the agents for screening and quality assessment and agreement matching for data extraction using semantic similarity indices can facilitate a reliable process. Large language models have found exceptional prowess in critique and analysis, a strength that could be exploited through cross-critique and self-critique methods. On discordance or negative matching among LLMs, these critique methods can be used to identify inconsistencies or areas for improvement, ensuring a robust and self-correcting system. In addition, a mixture of expert methods can produce a conglomerate model comprising multiple models with different architectures and expertise. One such example is the integration of transformer architectures like Bidirectional Encoder Representations from Transformers which excel at capturing textual dependencies for tasks like data extraction with recurrent models like Long Short-Term Memory units or Gated Recurrent Units, which may be better suited for sequential tasks like screening updated reports of previously included studies in the context of living systematic reviews.

It is well-established that the performance of LLMs can be further enhanced through techniques such as reinforcement learning from human feedback[32] and direct preference optimization.[33] These methods can be seamlessly integrated into our proposed framework, allowing for continuous improvement and fine-tuning of open-source LLMs for specific tasks such as screening and data extraction tasks. Methods like reinforcement learning from human feedback and direct preference optimization facilitate a continuous learning process, enabling the LLMs to adapt and improve on the basis of feedback and preferences from human experts. Large language models' decisions can be traced back to human-provided feedback and preferences aligning with the principles of evidence-based medicine.

In summary, integrating LLMs into a human-AI collaborative system that allows continuous optimization of their performance and leveraging the unique strengths of LLMs can revolutionize the SRMA process to create efficient, reproducible, and timely living systematic reviews and meta-analyses.

POTENTIAL COMPETING INTERESTS

The authors report no competing interests.

Abbreviations and Acronyms: **AI,** artificial intelligence; **ChatGPT,** chat generative pretrained transformer; **LLM,** large language model; **SRMA,** systematic reviews and meta-analyses

Correspondence: Address to Irbaz B. Riaz, MD, MS, MBI, PhD, Mayo Clinic Cancer Center, Department of Medicine, Hematology/Oncology Fellowship, Mayo Clinic in Arizona, 5777 East Mayo Boulevard, Phoenix, AZ 85054 (riaz.irbaz@mayo.edu; Twitter: @IrbazRiaz).

ORCID

Irbaz Bin Riaz: https://orcid.org/0000-0003-4249-0311

REFERENCES

1. Beller EM, Chen JK, Wang UL, Glasziou PP. Are systematic reviews up-to-date at the time of publication? *Syst Rev*. 2013;2:36. https://doi.org/10.1186/2046-4053-2-36.
2. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8(1):163. https://doi.org/10.1186/s13643-019-1074-9.
3. Wang S, Scells H, Koopman B, Potthast M, Zuccon G. Generating natural language queries for more effective systematic review screening prioritisation. In: *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 2023:73-83. https://doi.org/10.1145/3624918.3625322.
4. Guimarães NS, Joviano-Santos JV, Reis MG, Chaves RRM; Observatory of Epidemiology, Nutrition Health Research (OPENS). Development of search strategies for systematic reviews in health using ChatGPT: a critical analysis. *J Transl Med*. 2024;22(1):1. https://doi.org/10.1186/s12967-023-04371-5.
5. Chai KEK, Lines RLJ, Gucciardi DF, Ng L. Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. *Syst Rev*. 2021;10(1):93. https://doi.org/10.1186/s13643-021-01635-3.
6. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center. In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. Association for Computing Machinery; 2012:819-824. https://doi.org/10.1145/2110363.2110464.
7. van de Schoot R, de Bruin J, Schram R, et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nat Mach Intell*. 2021;3(2):125-133. https://doi.org/10.1038/s42256-020-00287-7.
8. Li D, Wang Z, Wang L, et al. A text-mining framework for supporting systematic reviews. *Am J Inf Manag*. 2016;1(1):1-9.
9. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. *J Med Internet Res*. 2024;26:e48996. https://doi.org/10.2196/48996.
10. Wang S, Scells H, Zhuang S, Potthast M, Koopman B, Zuccon G. Zero-shot generative large language models for systematic review screening automation. Preprint. Posted online January 12, 2024. arXiv 2401.06320. https://doi.org/10.48550/arXiv.2401.06320.
11. Robinson A, Thorne W, Wu BP, et al. Bio-sieve: exploring instruction tuning large language models for systematic review automation. Preprint. Posted online August 12, 2023. arXiv 2308.06610. https://doi.org/10.48550/arXiv.2308.06610.
12. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res Synth Methods*. Published online March 14, 2024. https://doi.org/10.1002/jrsm.1715.
13. Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. Preprint. Posted online December 1, 2023. arXiv 2312.00752. https://doi.org/10.48550/arXiv.2312.00752.
14. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. *Syst Rev*. 2016;5(1):210. https://doi.org/10.1186/s13643-016-0384-4.
15. Li T, Saldanha IJ, Jap J, et al. A randomized trial provided new evidence on the accuracy and efficiency of traditional vs. electronically annotated abstraction approaches in systematic reviews. *J Clin Epidemiol*. 2019;115:77-89. https://doi.org/10.1016/j.jclinepi.2019.07.005.
16. Alshami A, Elsayed M, Ali E, Eltoukhy AEE, Zayed T. Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. *Systems*. 2023;11(7):351. https://doi.org/10.3390/systems11070351.
17. Kartchner D, Ramalingam S, Al-Hussaini I, Kronick O, Mitchell C. Zero-shot information extraction for clinical meta-analysis using large language models. In: *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*; 2023:396-405. https://doi.org/10.18653/v1/2023.bionlp-1.37.
18. Sun Z, Zhang R, Doi SA, et al. How good are large language models for automated data extraction from randomized trials? Preprint. Posted online February 21, 2024. medRxiv. https://doi.org/10.1101/2024.02.20.24303083.
19. Könsgen N, Barcot O, Heß S, et al. Inter-review agreement of risk-of-bias judgments varied in Cochrane reviews. *J Clin Epidemiol*. 2020;120:25-32. https://doi.org/10.1016/j.jclinepi.2019.12.016.
20. Hasan B, Saadi S, Rajjoub NS, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. *BMJ Evid Based Med*. Published online February 21, 2024. https://doi.org/10.1136/bmjebm-2023-112597.
21. Pitre T, Jassal T, Talukdar JR, Shahab M, Ling M, Zeraatkar D. ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: a methods study. *Preprint*. Posted online November 22, 2023. https://doi.org/10.1101/2023.11.19.23298727. medRxiv.
22. Data Analysis with ChatGPT. https://help.openai.com/en/articles/8437071-advanced-data-analysis-chatgpt-enterprise-version. Accessed March 28, 2024.
23. Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol*. 2017;87:4-13. https://doi.org/10.1016/j.jclinepi.2017.05.006.
24. Murad MH, Fiordalisi C, Pillay J, et al. Making narrative statements to describe treatment effects. *J Gen Intern Med*. 2021;36(1):196-199. https://doi.org/10.1007/s11606-020-06330-y.
25. Riaz IB, Naqvi SAA, He H, et al. The living interactive evidence synthesis framework for living systematic reviews and meta-analyses. https://living-evidence.com. Accessed July 9, 2024.
26. Riaz IB, He H, Ryu AJ, et al. A living, interactive systematic review and network meta-analysis of first-line treatment of metastatic renal cell carcinoma. *Eur Urol*. 2021;80(6):712-723. https://doi.org/10.1016/j.eururo.2021.03.016.
27. Riaz IB, Naqvi SAA, He H, et al. First-line systemic treatment options for metastatic castration-sensitive prostate cancer: a living systematic review and network meta-analysis. *JAMA Oncol*. 2023;9(5):635-645. https://doi.org/10.1001/jamaoncol.2022.7762.
28. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. Preprint. Posted online November 9, 2023. arXiv 2311.05232. https://doi.org/10.48550/arXiv.2311.05232.
29. Yu T, Zhang S, Feng Y. Truth-aware context selection: mitigating the hallucinations of large language models being misled by untruthful contexts. Preprint. Posted online March 12, 2024. arXiv 2403.07556. https://doi.org/10.48550/arXiv.2403.07556.
30. Yan S-Q, Gu J-C, Zhu Y, Ling Z-H. Corrective retrieval augmented generation. Preprint. Posted online January 29, 2024. arXiv 2401.15884. https://doi.org/10.48550/arXiv.2401.15884.
31. Sharma A, Li X, Guan H, et al. Automatic data transformation using large language model—an experimental study on building energy data. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE; 2023:1824-1834.
32. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *Adv Neural Inform Process Syst*. 2022;35:27730-27744.
33. Rafailov R, Sharma A, Mitchell E, Manning CD, Ermon S, Finn C. Direct preference optimization: your language model is secretly a reward model. Preprint. Posted online May 29, 2023. *Adv Neural Inform Process Syst*. 2024;36. https://doi.org/10.48550/arXiv.2305.18290. arXiv 2305.18290.