Potential roles of large language models in production of systematic reviews and meta-analyses

Xufei Luo, Fengxian Chen, Di Zhu, Ling Wang, Zijun Wang, Hui Liu, Meng Lyu, Ye Wang, Qi Wang, Yaolong Chen

Submitted to: Journal of Medical Internet Research
on: February 20, 2024

Table of Contents

Potential roles of large language models in production of systematic reviews and meta-analyses

Xufei Luo[1]; Fengxian Chen[2]; Di Zhu[3]; Ling Wang[4]; Zijun Wang[5]; Hui Liu[5]; Meng Lyu[3]; Ye Wang[3]; Qi Wang[6]; Yaolong Chen[7]

[1]Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou, China. Lanzhou City CN
[2]School of Information Science & Engineering, Lanzhou, China. Lanzhou City CN
[3]School of Public Health, Lanzhou University Lanzhou City CN
[4]School of Public Health, Lanzhou University Lanzhou CN
[5]Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University Lanzhou City CN
[6]Department of Health Research Methods, Evidence and Impact, Faculty of Health Sciences, McMaster University Hamilton CA
[7]Lanzhou University Lanzhou City CN

Corresponding Author:
Yaolong Chen
Lanzhou University
No.199 Donggang West Road, Chengguan District
Lanzhou City
CN

Abstract

Large language models (LLMs) like ChatGPT have become widely applied in the field of medical research. In the process of conducting systematic reviews, similar tools can be employed to expedite various steps, including defining clinical questions, literature search, document screening, information extraction, and language refinement, etc, thereby conserving resources and enhancing efficiency. However, when utilizing LLMs, attention should be given to transparent reporting, distinguishing between genuine and false content, and avoiding academic misconduct. This article reviews the potential roles of LLMs in the creation of systematic reviews and meta-analyses, elucidating their advantages, limitations, and future research directions, aiming to provide insights and guidance for authors involved in systematic reviews and meta-analyses.

(JMIR Preprints 20/02/2024:56780)
DOI: https://doi.org/10.2196/preprints.56780

Original Manuscript

Potential roles of large language models in production of systematic reviews and meta-analyses

Xufei Luo, PhD student[a,b,c,d,e], **Fengxian Chen**, PhD[f], **Di Zhu**, MPH[g], **Ling Wang**, MPH[g], **Zijun Wang**, PhD student[a,b,c,d,e], **Hui Liu**, PhD student[a,b,c,d,e], **Meng Lyu**, MPH[g], **Ye Wang**, MPH[g], **Qi Wang**, PhD[h,i] ,**Yaolong Chen**, MD, PhD[a,b,c,d,e*]

a. Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, Lanzhou, China.
b. World Health Organization Collaboration Center for Guideline Implementation and Knowledge Translation, Lanzhou, China.
c. Institute of Health Data Science, Lanzhou University, Lanzhou, China.
d. Key Laboratory of Evidence Based Medicine and Knowledge Translation of Gansu Province, Lanzhou University, Lanzhou, China.
e. Research Unit of Evidence-Based Evaluation and Guidelines, Chinese Academy of Medical Sciences (2021RU017), School of Basic Medical Sciences, Lanzhou University, Lanzhou, China.
f. School of Information Science & Engineering, Lanzhou, China.
g. School of Public Health, Lanzhou University, Lanzhou, China.
h. Department of Health Research Methods, Evidence and Impact, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada.
i. McMaster Health Forum, McMaster University, Hamilton L8S4L8, Canada.

***Correspondence to**: Professor Yaolong Chen, Evidence-Based Medicine Center, School of Basic Medical Sciences, Lanzhou University, No.199, Donggang West Road, Chengguan District, Lanzhou 730000, China. Email: chevidence@lzu.edu.cn. Phone number: (86)0931-8912639. Fax: 0931-8915023.

Email address:
Xufei Luo: luoxf22@lzu.edu.cn
Fengxian Chen: chenfx@lzu.edu.cn
Di Zhu: zhudiyx@163.com
Ling Wang: wangling_working@163.com
Zijun Wang: bdwzj_0312@163.com
Hui Liu: sxyafxlh@163.com
Meng Lyu: meng.lyu0908@gmail.com
Ye Wang: wye2023@lzu.edu.cn
Qi Wang: wangq87@mcmaster.ca
Yaolong Chen: chevidence@lzu.edu.cn

Word Counts: 2431
Figure: 1
Table: 1
eFigure: 12

Abstract

Large language models (LLMs) like ChatGPT have become widely applied in the field of medical research. In the process of conducting systematic reviews, similar tools can be employed to expedite various steps, including defining clinical questions, literature search, document screening, information extraction, and language refinement, etc, thereby conserving resources and enhancing efficiency. However, when utilizing LLMs, attention should be given to transparent reporting, distinguishing between genuine and false content, and avoiding academic misconduct. This article reviews the potential roles of LLMs in the creation of systematic reviews and meta-analyses, elucidating their advantages, limitations, and future research directions, aiming to provide insights and guidance for authors involved in systematic reviews and meta-analyses.

Keywords: large language model, ChatGPT, systematic review, chatbot, meta-analysis

Introduction

A systematic review is the result of a systematic and rigorous evaluation of evidence, and a meta-analysis may or may not be a part of it [1]. Due to its strict methodology and comprehensive summary of evidence, high-quality systematic reviews are considered the highest level of evidence in the hierarchy of evidence [2]. They are positioned at the top of the evidence pyramid[2]. Additionally, high-quality systematic reviews and meta-analyses are often used to support the development of clinical practice guidelines, aid clinical decision-making, and inform healthcare policy formulation [3]. Currently, the methods of systematic reviews and meta-analyses are also applied in various disciplines beyond medicine, such as law [4], management [5], economics [6], and have yielded positive results, contributing to the continuous advancement of these fields [7].

The process of conducting systematic reviews demands a substantial investment in terms of time, resources, human effort, and financial capital [8]. To expedite the development of systematic reviews and meta-analyses, various (semi)automated tools, such as Covidence, have also come into play [9,10]. However, the emergence of large language models (LLMs), particularly Chatbots such as GPT, presents a set of challenges and opportunities in the realm of systematic review and meta-analysis [11]. This article conducts a comprehensive review of relevant literature, aiming to investigate the potential for harnessing LLMs to accelerate the production of systematic review and meta-analysis, while also scrutinizing the potential impacts and delineating the crucial steps involved in this process.

The process and challenges of conducting a systematic review and meta-analysis

The procedures and workflows for conducting systematic reviews and meta-analyses are well-established. Currently, researchers often refer to the Cochrane Handbooks recommended by the Cochrane Library for intervention or diagnostic reviews [12,13]. In addition, some scholars and institutions have also developed detailed guidelines on the steps and methodology for performing systematic reviews and meta-analyses [14-17]. Generally speaking, researchers should take the following steps to produce a high-quality systematic review and meta-analysis: determine the clinical question, register and draft a protocol, set inclusion and exclusion criteria, develop and implement a search strategy, screen literature, extract data from included studies, assess the quality and risk of bias of included studies, analyze and process data, write up the full text, and submit for publication, as illustrated in **Figure 1**. These different steps contain many sub-tasks, therefore

conducting a complete systematic review and meta-analysis requires fairly complex and time-consuming work.

Figure 1 The process of conducting a systematic review and meta-analysis

Although systematic reviews and meta-analyses have been widely applied and play an important role in developing guidelines and informing clinical decision-making, their production process faces many challenges. One of them is the long production time and large resource requirements. Studies suggest that the average estimated time to complete and publish a systematic review is 67.3 weeks, requiring five researchers and costing around $140,000[18-19]. For some time, (semi-)automated tools utilizing natural language processing and machine learning have accelerated systematic review and meta-analysis production to some extent[20], with studies showing such tools can produce a systematic review and meta-analysis within two weeks[21]. However, these tools also have some limitations. First, no single tool can fully accelerate the entire production process of systematic reviews and meta-analyses. Second, these tools cannot process and analyze literature in different languages. Finally, the reliability of results generated by these (semi-)automated tools needs further

validation as they are not yet widely adopted.

Large language models in medical research

Chatbots based on LLMs, such as ChatGPT, Google Gemini, and Claud, have become widely applied in medical research. These chatbots prove valuable in tasks ranging from knowledge retrieval, language refinement, content generation, and medical exam preparation to literature assessment. Research indicates that ChatGPT excels in accuracy, completeness, nuance, and speed when generating responses to clinical inquiries in psychiatry[22]. Moreover, LLMs like ChatGPT play a pivotal role in automating the evaluation of medical literature, facilitating the identification of accurately reported research findings[23]. Despite their significant contributions, these chatbots are not without limitations. Challenges such as the potential for generating misleading content and susceptibility to academic deception necessitate further scholarly discourse on effective mitigation strategies. Standardized reporting practices may contribute to delineating the applications of ChatGPT and mitigating research biases [24].

In the process of conducting systematic reviews and meta-analyses, ChatGPT demonstrates significant application potential and promise. Existing studies [11,25-32] indicate that ChatGPT can play a pivotal role in formulating clinical questions, determining inclusion and exclusion criteria, screening literature, assessing publications, generating meta-analysis code, and assisting in full-text composition, etc. In this context, we will provide a detailed exposition of these capabilities (**Table 1**).

Table 1 The possible functions of chatbots in the creation of systematic reviews and meta-analyses encompass separate stages.

No.	Tasks	Potential roles and application steps of chatbots	References
1	Determine the research topic/question	• Identify previously published systematic reviews and meta-analyses on the same topic. • Assist in determining the rationale for the research question. • Clarify the PICO (Population, Intervention, Comparison, Outcome) question.	[11,33-35]
2	Register and write a research proposal	• Generate preliminary, unverified	[11, 36-37]

		registration information. • Draft an initial research proposal, subject to validation.	
3	Define inclusion an exclusion criterion	• Establish inclusion criteria. • Establish exclusion criteria.	[11, 38]
4	Develop a search strategy and conduct searches	• Develop and optimize search strategies. • Implement retrieval. • Collect grey literature.	[11,25,29, 33, 39-42]
5	Screen the literature	• Remove duplicate records. • Screen literature titles, abstracts, and keywords. • Screen the full text of literature. • Download the full text of literature.	[11,25,27,28,33,34,43-47]
6	Extract the data	• Extract basic information. • Extract patient information. • Extract outcome information. • Extract table information.	[11,25,26,47-50]
7	Assess the risk of bias	• Extract relevant information based on the scale. • Evaluate the risk of bias based on the scale. • Present visual results.	[26,51-53]
8	Analyze the data/meta-analyses	• Extract outcome information. • Generate figures and tables for some results.	[11,26,37,54]
9	Draft the full manuscript	• Search for relevant references. • Polish language and grammar. • Adjust the reference citation format. • Summarize the abstract.	[25,33,55-58]
10	Submit and publish	• Assist in selecting a suitable journal. • Adjust the manuscript format. • Compose a cover letter. • Assist in preparing the submission.	[33,59]

The potential roles of LLMs in systematic review and meta-analysis

Determine the research topic/question

Determining the clinical question represents the initial and paramount step in the process of conducting systematic reviews and meta-analyses. At this juncture, it is crucial to ascertain whether comparable systematic reviews and meta-analyses have already been published and to delineate the scope of the forthcoming review and meta-analysis. Generally, for interventional systematic reviews, the patient, intervention, comparison, outcome (PICO) framework is considered for defining the scope and objectives of the research question [60]. In this context, ChatGPT serves a dual role. On one hand, it expeditiously aids in searching for published systematic reviews and meta-analyses related to the relevant topics (See Figure S1 and S2)[34]. On the other hand, it assists in refining the clinical question that needs to be addressed (See Figure S3), facilitating researchers in promptly determining the feasibility of undertaking the proposed study. However, it is important to be cautious of false literature [35].

Register and write a research proposal

The registration and proposal writing process constitutes a pivotal preparatory phase for the conducting of systematic reviews and meta-analyses. Registration enhances research transparency, fosters collaboration among investigators, and mitigates the redundancy of research endeavors. Drafting a proposal helps in elucidating the research objectives and methods, providing robust support for the smooth execution of the study. For LLMs, generating preliminary registration information and initial proposal content is remarkably convenient and facile (see Figures S4 and S5). For example, ChatGPT can assist researchers in generating the statistical methods for a research proposal [37]. However, considering that LLMs often generate fictitious literature, the content they produce may be inaccurate, thus discernment and validation of the generated content remain essential considerations.

Define inclusion an exclusion criterion

The inclusion and exclusion criteria for systematic review and meta-analyses are instrumental in determining the screening standards for studies. Therefore, strict and detailed inclusion and exclusion criteria contribute to the smooth and high-quality conduct of systematic reviews and meta-analyses. The use of a chatbot based on LLMs can help in establishing the inclusion and exclusion criteria (see Figure S6) [38], however, the inclusion criteria need to be optimized and adjusted according to the specific research objectives, and the exclusion criteria should be based on

the foundation of the inclusion criteria. Therefore, manual adjustments and optimizations are also necessary.

Develop a search strategy and conduct searches

ChatGPT can assist in formulating search strategies, using PubMed as an example [40]. Researchers can simply list their questions using the PICO framework, and a search strategy can be quickly generated (Figure S1 and S2). Based on the generated search strategy, one method is to copy the strategy into the PubMed search box for direct retrieval [40-41]. Another approach involves utilizing the OpenAI application programming interfaces (APIs) to invoke PubMed APIs with the search strategy generated by GPT. This allows for searching the PubMed database, obtaining search results, and applying predetermined inclusion and exclusion criteria. Subsequently, GPT is used to filter the search results, exporting and recording the filtered results in JSON format. This integrated process encompasses search strategy formulation, retrieval, and filtering. However, the direct use of LLMs to generate search strategies and complete the one-stop process of searching and screening may not be mature at present, and poses a significant challenge for generating the PRISMA flowchart. Therefore, we suggest using LLMs to generate search strategies, which are then optimized and modified by librarians and computer experts (specializing in large language models) before manually searching the databases. Additionally, to use search strategies transparently and reproducibly, detailed prompts should be reported [40,42].these search strategies also need validation, refinement, and modification.

Screen the literature

Literature screening is one of the most time-consuming steps in the creation of systematic reviews and meta-analyses. Prior to the advent of ChatGPT, there were already many (semi) automated tools available for literature screening, such as Coevidence, EPPI-Reviewer, DistillerSR, and others [39]. With the emergence of ChatGPT, researchers can now train the model based on pre-defined inclusion criteria. Subsequently, they can utilize ChatGPT to automatically screen records retrieved from databases, obtaining the filtered results . Previous studies suggested that utilizing ChatGPT in the literature selection process for meta-analysis substantially diminishes the workload while preserving a recall rate on par with manual curation [28,44-47].

Extract the data

Data extraction involves obtaining information from primary studies and serves as a primary source for systematic reviews and meta-analyses. Generally, when conducting systematic reviews and meta-analyses, we need to extract basic information from the original studies, such as publication date, country of conduct, and the journal of publication. Additionally, characteristics of the population, such as patient samples, age, gender, and outcome data, including event occurrences, mean change values, and total sample size, are also extracted. Currently, tools based on natural language processing and LLMs, such as ChatGPT and Claude, demonstrate high accuracy in extracting information from Portable Document Format (PDF) documents (Figure S7) [47-50]. However, it is important to note that despite the promising capabilities of these tools, manual verification remains a necessary step in the data extraction process when utilizing AI tools[61]. Using large language models to extract data can help avoid random errors; however, caution is still required when extracting data from figures or tables [47-50].

Assess the risk of bias

Assessing the bias of risk involves evaluating the internal validity of studies included in research. For randomized controlled trials, we typically use tools like Risk of Bias (RoB) [62] or RoB 2 tools [63], with an estimated review time of 10-15 minutes per trial. However, automated tools such as RobotReviewer can streamline the extraction and evaluation process in batches [51-53], improving efficiency—though manual verification is still necessary. Additionally, chatbots based on LLMs can aid in risk of bias assessment (see Figure S8), and studies indicate that their accuracy is comparable to human evaluations [23].

Analyze the data/meta-analysis

Data analysis serves as the source of systematic review results, typically encompassing basic information and outcome findings. Meta-analysis may be one outcome, along with potential components like subgroup analysis, sensitivity analysis, meta-regression, and detection of publication bias. Numerous software options are available to facilitate these data analyses, including STATA, RevMan, Rstudio, and others [43]. Currently, it appears that chatbots based on LLMs may not fully execute data analysis independently, but they can extract relevant information. Subsequently,

one can employ corresponding software for comprehensive data analysis. Alternatively, after extracting information with chatbots, the ChatGPT Code Interpreter can assist in analysis and generating graphical results, contingent upon being a ChatGPT Plus subscriber. Moreover, LLM markedly accelerates the data analysis process, empowering researchers to handle larger datasets with greater efficacy [54].

Draft the full manuscript

The complete drafting of systematic reviews and meta-analyses should adhere to the PRISMA reporting guidelines [64]. It is not advisable to use chatbots like ChatGPT for article composition. On one hand, the accuracy and integrity of content generated by GPT require human verification. On the other hand, various research types and journals have different requirements for full-text articles, making it challenging to achieve uniformity in generated content. However, utilizing tools like GPT for language refinement and adjusting content logic can be considered to enhance the quality and readability of the article [33,55]. It is important to declare the use of GPT-related tools in the methods, acknowledgments, or appendices to ensure transparency [24,65].

Submit and publish

Submission and publication represent the final steps in the process of conducting systematic reviews and meta-analyses, aside from subsequent updates. At this stage, the potential role of tools related to LLMs can assist authors in recommending suitable journals (Figure S9). They might also aid in crafting components such as cover letters and highlights [59]. However, it is imperative to emphasize that content generated by these tools requires manual verification to ensure the accuracy of the content, and all authors should be accountable for the content generated by LLMs.

Benefits and drawbacks of using large language models

Systematic reviews and meta-analyses are crucial evidence types that support the development of guidelines [3]. The benefits of employing LLM chatbots in the production of systematic reviews and meta-analyses include increased speed, such as in the stages of evidence searching, data extraction, and assessment of bias risk; they can also enhance accuracy by reducing human errors, for instance, in extracting essential information and pooling data. However, there are also drawbacks, such as the potential for generating hallucinations, the reliability of the models requires human verification and

the entire systematic review process is not replicable. Moreover, when interacting with large language model chatbots, it is important to manage data privacy; when using LLMs to analyze data, especially when it involves patient privacy, ethical approval and management must be properly addressed.

Challenges and solutions

While LLMs can assist in accelerating the production of systematic reviews and meta-analyses in some steps, enhancing accuracy and transparency, and saving resources, they also face several challenges. For instance, LLMs cannot promptly update their versions and information. Currently, ChatGPT 3.5 is trained on data from around 2021. Limitations such as the length of prompts and token constraints, as well as restrictions related to context associations, may potentially impact overall results and user experience [25]. Although LLM-based autonomous agents have made strides in systematic reviews and meta-analyses, LLMs still face numerous challenges due to issues related to personalization, updating knowledge, strategic planning, and complex problem-solving. The development of LLM-driven autonomous agents adept at systematic reviews and meta-analyses warrants exploration [66]. The use of LLMs as centrally controlled intelligent agents encompasses the ability to handle precise literature screening, extract and analyze complex data, and assist in manuscript composition, as demonstrated by proof-of-concept demos like MetaGPT [67]. Moreover, as the use of LLMs continues to grow，ensuring the accuracy of information provided in systematic reviews becomes a significant challenge, particularly if LLMs are indiscriminately overused.

To better facilitate the utilization of tools such as ChatGPT in systematic reviews and meta-analyses, we believe that, first and foremost, authors should understand the scope and scenarios for applying ChatGPT, clearly defining which steps can benefit from these tools. Secondly, for researchers, collaboration with computer scientists and artificial intelligence engineers is crucial to optimize the prompts and develop integrated tools based on LLMs, such as web applications. These tools can assist in seamless transitions between different tasks in the systematic review process. Lastly, for journal editors, collaboration with authors and reviewers is essential to adhere to reporting and ethical principles associated with the use of GPT [24,68]. This collaboration aims to promote

transparency and integrity, and to prevent indiscriminate overuse in the application of LLMs in systematic reviews and meta-analyses.

Future perspectives and conclusion

The emergence of LLMs could have a significant impact on the production of systematic reviews and meta-analyses. In this process, the application of chatbots like ChatGPT have the potential to speed up certain steps, such as literature screening, data extraction, and bias risk assessment—processes that typically consume a considerable amount of time. However, it is important to note that if artificial intelligence methods such as GPT are employed in the performing of systematic reviews, disclosure and declaration are essential. This includes specifying the AI tools utilized, their roles, and the areas of application within the review process, etc [24]. In this context, developing a reporting guideline is warranted to guide the application of LLM tools in systematic reviews and meta-analyses. While PRISMA 2020 briefly addresses the use of automation technologies, its coverage is limited to steps such as screening, and lacks comprehensive guidance on the broader spectrum of applications [3564].

Abbreviations

API: Application Programming Interfaces

ChatGPT: Chat Generative Pre-trained Transformer

LLM: Large language model

PDF: Portable Document Format

PICO: Population, Intervention, Comparison, Outcome

PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses

RoB: Risk of Bias

Ethical approval

Not applicable.

Sources of funding
No funding.

Author contributions
X.L.: Conceptualization, Methodology, Data curation, Investigation, Formal analysis, Visualization, Writing-original draft, Writing-review & editing.
F.C., D.Z., and L.W.: Data curation, Investigation, Formal analysis, Visualization, Writing-original draft.
Z.W., H.L., and M.L: Data curation, Investigation, Writing-review & editing.
Y.W., and Q.W.: Formal analysis, Writing-review & editing.
Y.C: Conceptualization, Methodology, Writing-review & editing, Supervision.
All authors read the final manuscript and approved the submission.

Conflicts of interest disclosure
No potential competing interest was reported by the author(s).

Guarantor
Yaolong Chen (corresponding author) takes full responsibility for the work and/or the conduct of the study, has access to the data, and controls the decision to publish.

Data availability statement
No specific data collected for the above manuscript.

Patient and Public Involvement
It was not appropriate or possible to involve patients or the public in the design, or conduct, or reporting, or dissemination plans of our research.

Acknowledgements
We would like to thank ChatGPT 3.5 designed by OpenAI for its assistance in language editing. We take the ultimate responsibility for the content of this publication.

References

1. Jahan N, Naveed S, Zeshan M, et al. How to Conduct a Systematic Review: A Narrative Literature Review. Cureus. 2016 Nov 4;8(11):e864. doi: 10.7759/cureus.864. PMID: 27924252; PMCID: PMC5137994.
2. Wallace SS, Barak G, Truong G, Parker MW. Hierarchy of Evidence Within the Medical Literature. Hosp Pediatr. 2022 Aug 1;12(8):745-750. doi: 10.1542/hpeds.2022-006690. PMID: 35909178.
3. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines; Graham R, Mancher M, Miller Wolman D, et al., editors. Clinical Practice Guidelines We Can Trust. Washington (DC): National Academies Press (US); 2011. Available from: https://www.ncbi.nlm.nih.gov/books/NBK209539/ doi: 10.17226/13058
4. Bystranowski P, Janik B, Próchnicki M, Skórska P. Anchoring effect in legal decision-making: A meta-analysis. Law Hum Behav. 2021 Feb;45(1):1-23. doi: 10.1037/lhb0000438. PMID: 33734746.
5. Geyskens I, Krishnan R, Steenkamp J, et al. A review and evaluation of meta-analysis practices in management research. Journal of Management, 2009, 35(2): 393-419.
6. Bagepally BS, Chaikledkaew U, Chaiyakunapruk N, Attia J, Thakkinstian A. Meta-analysis of economic evaluation studies: data harmonisation and methodological issues. BMC Health Serv Res. 2022 Feb 15;22(1):202. doi: 10.1186/s12913-022-07595-1. PMID: 35168619; PMCID: PMC8845252.
7. Gurevitch J, Koricheva J, Nakagawa S, Stewart G. Meta-analysis and the science of research synthesis. Nature. 2018 Mar 7;555(7695):175-182. doi: 10.1038/nature25753. PMID: 29517004.
8. Tsertsvadze A, Chen YF, Moher D, Sutcliffe P, McCarthy N. How to conduct systematic reviews more expeditiously? Syst Rev. 2015 Nov 12;4:160. doi: 10.1186/s13643-015-0147-7. PMID: 26563648; PMCID: PMC4643500.
9. Scott AM, Forbes C, Clark J, Carter M, Glasziou P, Munn Z. Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. J Clin Epidemiol. 2021 Oct;138:80-94. doi: 10.1016/j.jclinepi.2021.06.030. Epub 2021 Jul 7. PMID: 34242757.
10. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol. 2022 Apr;144:22-42. doi: 10.1016/j.jclinepi.2021.12.005. Epub 2021 Dec 8. PMID: 34896236.
11. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Syst Rev. 2023 Apr 29;12(1):72. doi: 10.1186/s13643-023-02243-z. PMID: 37120563; PMCID: PMC10148473.
12. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.4 (updated August 2023). Cochrane, 2023. Available from www.training.cochrane.org/handbook.
13. Deeks JJ, Bossuyt PM, Leeflang MM, Takwoingi Y (editors). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Version 2.0 (updated July 2023). Cochrane, 2023. Available from https://training.cochrane.org/handbook-diagnostic-test-accuracy/current.
14. Xiao Y, Watson M. Guidance on Conducting a Systematic Literature Review. Journal of Planning Education and Research, 2019; 39(1), 93-112. https://doi.org/10.1177/0739456X17723971
15. Muka T, Glisic M, Milic J, Verhoog S, Bohlius J, Bramer W, Chowdhury R, Franco OH. A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research. Eur J Epidemiol. 2020 Jan;35(1):49-60. doi: 10.1007/s10654-019-00576-5. Epub 2019 Nov 13. PMID: 31720912.
16. Siddaway AP, Wood AM, Hedges LV. How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses. Annu Rev

Psychol. 2019 Jan 4;70:747-770. doi: 10.1146/annurev-psych-010418-102803. Epub 2018 Aug 8. PMID: 30089228.
17. Tawfik GM, Dila KAS, Mohamed MYF, Tam DNH, Kien ND, Ahmed AM, Huy NT. A step by step guide for conducting a systematic review and meta-analysis with simulation data. Trop Med Health. 2019 Aug 1;47:46. doi: 10.1186/s41182-019-0165-6. PMID: 31388330; PMCID: PMC6670166.
18. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open. 2017 Feb 27;7(2):e012545. doi: 10.1136/bmjopen-2016-012545. PMID: 28242767; PMCID: PMC5337708.
19. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: A call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun. 2019 Aug 25;16:100443. doi: 10.1016/j.conctc.2019.100443.
20. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019 Jul 11;8(1):163. doi: 10.1186/s13643-019-1074-9. PMID: 31296265; PMCID: PMC6621996.
21. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol. 2020 May;121:81-90. doi: 10.1016/j.jclinepi.2020.01.008. Epub 2020 Jan 28. PMID: 32004673.
22. Luykx JJ, Gerritse F, Habets PC, Vinkers CH. The performance of ChatGPT in generating answers to clinical questions in psychiatry: a two-layer assessment. World Psychiatry. 2023 Oct;22(3):479-480. doi: 10.1002/wps.21145. PMID: 37713576; PMCID: PMC10503909.
23. Roberts RH, Ali SR, Hutchings HA, Dobbs TD, Whitaker IS. Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. BMJ Health Care Inform. 2023 Oct;30(1):e100830. doi: 10.1136/bmjhci-2023-100830. PMID: 37827724; PMCID: PMC10583079.
24. Luo X, Estill J, Chen Y. The use of ChatGPT in medical research: do we need a reporting guideline? Int J Surg. 2023 Sep 14. doi: 10.1097/JS9.0000000000000737. Epub ahead of print. PMID: 37707517.
25. 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
26. Mahuli SA, Rai A, Mahuli AV, Kumar A. Application ChatGPT in conducting systematic reviews and meta-analyses. Br Dent J. 2023 Jul;235(2):90-92. doi: 10.1038/s41415-023-6132-y. PMID: 37500847.
27. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open. 2023 Jul 7;13(7):e072254. doi: 10.1136/bmjopen-2023-072254. PMID: 37419641; PMCID: PMC10335470.
28. 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. 2024 Mar 14. doi: 10.1002/jrsm.1715. Epub ahead of print. PMID: 38484744.
29. Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, Shim SR. The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation. JMIR Med Inform 2024;12:e51187. doi: 10.2196/51187.
30. Hossain MM. Using ChatGPT and other forms of generative AI in systematic reviews: Challenges

and opportunities. J Med Imaging Radiat Sci. 2024 Mar;55(1):11-12. doi: 10.1016/j.jmir.2023.11.005. Epub 2023 Nov 30. PMID: 38040497.
31. Giunti G, Doherty CP. Cocreating an Automated mHealth Apps Systematic Review Process With Generative AI: Design Science Research Approach. JMIR Med Educ. 2024 Feb 12;10:e48949. doi: 10.2196/48949. PMID: 38345839; PMCID: PMC10897815.
32. Nashwan AJ, Jaradat JH. Streamlining Systematic Reviews: Harnessing Large Language Models for Quality Assessment and Risk-of-Bias Evaluation. Cureus. 2023 Aug 6;15(8):e43023. doi: 10.7759/cureus.43023. PMID: 37674957; PMCID: PMC10478591.
33. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. Am J Cancer Res. 2023 Apr 15;13(4):1148-1154. PMID: 37168339; PMCID: PMC10164801.
34. Issaiy M, Ghanaati H, Kolahi S, Shakiba M, Jalali AH, Zarei D, Kazemian S, Avanaki MA, Firouznia K. Methodological insights into ChatGPT's screening performance in systematic reviews. BMC Med Res Methodol. 2024 Mar 27;24(1):78. doi: 10.1186/s12874-024-02203-8. PMID: 38539117; PMCID: PMC10976661.
35. Branum C, Schiavenato M. Can ChatGPT Accurately Answer a PICOT Question? Assessing AI Response to a Clinical Question. Nurse Educ. 2023 Sep-Oct 01;48(5):231-233. doi: 10.1097/NNE.0000000000001436. Epub 2023 Apr 28. PMID: 37130197.
36. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. J Glob Health. 2023 Feb 17;13:01003. doi: 10.7189/jogh.13.01003. PMID: 36798998; PMCID: PMC9936200.
37. Richard Evans, Antonio Pozzi. Using ChatGPT to Develop the Statistical Analysis Plan for a Randomized Controlled Trial: A Case Report, 17 October 2023, PREPRINT (Version 1) available at Research Square. https://doi.org/10.21203/rs.3.rs-3433956/v1.
38. Hutson M. How AI is being used to accelerate clinical trials. Nature. 2024 Mar;627(8003):S2-S5. doi: 10.1038/d41586-024-00753-x. PMID: 38480968.
39. Van der Mierden S, Tsaioun K, Bleich A, Leenaars CHC. Software tools for literature screening in systematic reviews in biomedical research. ALTEX. 2019;36(3):508-517. doi: 10.14573/altex.1902131. Epub 2019 May 16. PMID: 31113000.
40. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good Boolean query for systematic review literature search? Available at https://arxiv.org/abs/2302.03495. Published online February 9, 2023. Accessed May 21, 2024.
41. Alaniz L, Vu C, Pfaff MJ. The Utility of Artificial Intelligence for Systematic Reviews and Boolean Query Formulation and Translation. Plast Reconstr Surg Glob Open. 2023 Oct 30;11(10):e5339. doi: 10.1097/GOX.0000000000005339. PMID: 37908326; PMCID: PMC10615538.
42. 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. Published 2024 Jan 2. doi:10.1186/s12967-023-04371-5. PMID: 38167166 PMCID: PMC10759630.
43. Tantry TP, Karanth H, Shetty PK, Kadam D. Self-learning software tools for data analysis in meta-analysis. Korean J Anesthesiol. 2021 Oct;74(5):459-461. doi: 10.4097/kja.21080. Epub 2021 Mar 8. PMID: 33677944; PMCID: PMC8497909.
44. Xiangming Cai, Yuanming Geng, Yiming Du, Bart Westerman, Duolao Wang, Chiyuan Ma, Juan J Garcia Vallejo. Utilizing ChatGPT to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. medRxiv 2023.09.06.23295072; doi: https://doi.org/10.1101/2023.09.06.23295072.
45. Eugene Syriani, Istvan David, Gauransh Kumar. Assessing the Ability of ChatGPT to Screen

Articles for Systematic Reviews. arXiv:2307.06464. https://doi.org/10.48550/arXiv.2307.06464.
46. Kohandel Gargari O, Mahmoudi MH, Hajisafarali M, Samiee R. Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. BMJ Evid Based Med. 2024 Jan 19;29(1):69-70. doi: 10.1136/bmjebm-2023-112678. PMID: 37989538; PMCID: PMC10850650.
47. 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 Jan 12;26:e48996. doi: 10.2196/48996. PMID: 38214966; PMCID: PMC10818236.
48. Polak MP, Morgan D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat Commun. 2024 Feb 21;15(1):1569. doi: 10.1038/s41467-024-45914-8. PMID: 38383556; PMCID: PMC10882009.
49. Mahmoudi, Hesam and Chang, Doris and Lee, Hannah and Ghaffarzadegan, Navid and Jalali, Mohammad S., A Critical Assessment of Large Language Models for Systematic Reviews: Utilizing ChatGPT for Complex Data Extraction (April 17, 2024). Available at SSRN: http://dx.doi.org/10.2139/ssrn.4797024.
50. Zhuanlan Sun, Ruilin Zhang, Suhail A. Doi, Luis Furuya-Kanamori, Tianqi Yu, Lifeng Lin, Chang Xu. How good are large language models for automated data extraction from randomized trials? medRxiv 2024.02.20.24303083. https://doi.org/10.1101/2024.02.20.24303083.
51. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. J Am Med Inform Assoc. 2016 Jan;23(1):193-201. doi: 10.1093/jamia/ocv044. Epub 2015 Jun 22. PMID: 26104742; PMCID: PMC4713900.
52. Lai H, MM; Ge L, Sun M, Pan B, Huang J, Hou L, Yang Q, Liu J, Liu J, Ye Z, Xia D, Zhao W, Wang X, Liu M, Talukdar JR, Tian J, Yang K, Estill J. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA Network Open. 2024;7(5):e2412687. doi:10.1001/jamanetworkopen.2024.12687.
53. Tyler Pitre, Tanvir Jassal, Jhalok Ronjan Talukdar, Mahnoor Shahab, Michael Ling, Dena Zeraatkar. ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: A methods study. medRxiv 2023.11.19.23298727. https://doi.org/10.1101/2023.11.19.23298727
54. Zeeshan Rasheed, Muhammad Waseem, Aakash Ahmad, Kai-Kristian Kemell, Wang Xiaofeng, Anh Nguyen Duc, Pekka Abrahamsson. Can Large Language Models Serve as Data Analysts? A Multi-Agent Assisted Approach for Qualitative Data Analysis. arXiv:2402.01386. https://doi.org/10.48550/arXiv.2402.01386.
55. Kim SG. Using ChatGPT for language editing in scientific articles. Maxillofac Plast Reconstr Surg. 2023 Mar 8;45(1):13. doi: 10.1186/s40902-023-00381-x. PMID: 36882591; PMCID: PMC9992464.
56. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, Pearson AT. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med. 2023 Apr 26;6(1):75. doi: 10.1038/s41746-023-00819-6. PMID: 37100871; PMCID: PMC10133283.
57. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol. 2023 Oct;19(5):598-604. doi: 10.1016/j.jpurol.2023.05.018. Epub 2023 Jun 2. PMID: 37328321.
58. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of Large Language Model Performance and Reliability for Citations and References in Scholarly Writing: Cross-Disciplinary Study.J Med Internet Res 2024;26:e52935.doi: 10.2196/52935.PMID: 38578685.PMCID: 11031695.
59. Nuzula I F, Amri M M. Will ChatGPT bring a New Paradigm to HR World? A Critical Opinion Article. Journal of Management Studies and Development, 2023, 2(02): 142-161.

60. Eriksen MB, Frandsen TF. The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. J Med Libr Assoc. 2018 Oct;106(4):420-431. doi: 10.5195/jmla.2018.345. Epub 2018 Oct 1. PMID: 30271283; PMCID: PMC6148624.
61.
62. Brandon Roberts. I tested how well ChatGPT can pull data out of messy PDFs. Available from: https://source.opennews.org/articles/testing-pdf-data-extraction-chatgpt/.
63. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JA; Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011 Oct 18;343:d5928. doi: 10.1136/bmj.d5928. PMID: 22008217; PMCID: PMC3196245.
64. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng HY, Corbett MS, Eldridge SM, Emberson JR, Hernán MA, Hopewell S, Hróbjartsson A, Junqueira DR, Jüni P, Kirkham JJ, Lasserson T, Li T, McAleenan A, Reeves BC, Shepperd S, Shrier I, Stewart LA, Tilling K, White IR, Whiting PF, Higgins JPT. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019 Aug 28;366:l4898. doi: 10.1136/bmj.l4898. PMID: 31462531.
65.
66. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021 Mar 29;372:n71. doi: 10.1136/bmj.n71. PMID: 33782057; PMCID: PMC8005924.
67.
68. Gaggioli A. Ethics: disclose use of AI in scientific manuscripts. Nature. 2023 Feb;614(7948):413. doi: 10.1038/d41586-023-00381-x. PMID: 36788370.
69. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, Zhao WX, Wei Z, Wen J. A survey on large language model based autonomous agents. Front Comput Sci. 2024;18, 186345. https://doi.org/10.1007/s11704-024-40231-1.
70. Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, Jürgen Schmidhuber. Metagpt: Meta programming for multi-agent collaborative framework. arXiv preprint arXiv:2308.00352, 2023.
71. Flanagin A, Pirracchio R, Khera R, Berkwits M, Hswen Y, Bibbins-Domingo K. Reporting Use of AI in Research and Scholarly Publication-JAMA Network Guidance. JAMA. 2024 Apr 2;331(13):1096-1098. doi: 10.1001/jama.2024.3471. PMID: 38451540.
72.

Supplementary Files

Figures

The process of conducting a systematic review and meta-analysis.

Multimedia Appendixes

Using ChatGPT 4 to assist in generating PubMed search strategies for assessing systematic reviews.
URL: http://asset.jmir.pub/assets/a70d49a9846eacc1d989e9d8f7a1cd9f.png

The results obtained after searching the PubMed based on the search strategy generated by GPT.
URL: http://asset.jmir.pub/assets/4dfd298eba0294752a39a52c2bf10280.png

Using ChatGPT 4 to assist in optimizing the clinical question for conducting a systematic review and meta-analysis.
URL: http://asset.jmir.pub/assets/93ce07aefb9625bd0baec08bf529289d.png

Utilizing ChatGPT 4 to generate PROSPERO registration information.
URL: http://asset.jmir.pub/assets/123d6153d6d7ae917662fc368ce41536.png

The exercise for osteoarthritis: systematic review and meta-analysis, generated by Claude 3 based on the provided prompts.
URL: http://asset.jmir.pub/assets/1e15faf93bf081f2dfcf5c1e9a517223.png

The inclusion and exclusion criteria for a systematic review and meta-analysis on exercise therapy for osteoarthritis based on GPT-4.
URL: http://asset.jmir.pub/assets/0afe9bc31a361147ef619b7e1f3a7e70.png

Using Claude 3 for Data Extraction: An Example with Three RCTs.
URL: http://asset.jmir.pub/assets/15fdb8bded94e5dcf3187be067c1153f.png

Using Claude 3 for Risk of bias assessment: An Example with Two RCTs.
URL: http://asset.jmir.pub/assets/0d4c465723ac69e27ec0c3af3735dbcf.png

Using GPT-4 to assist in selecting submitted journals.
URL: http://asset.jmir.pub/assets/e7ff7b65180f72f1ea0126c9f8339551.png