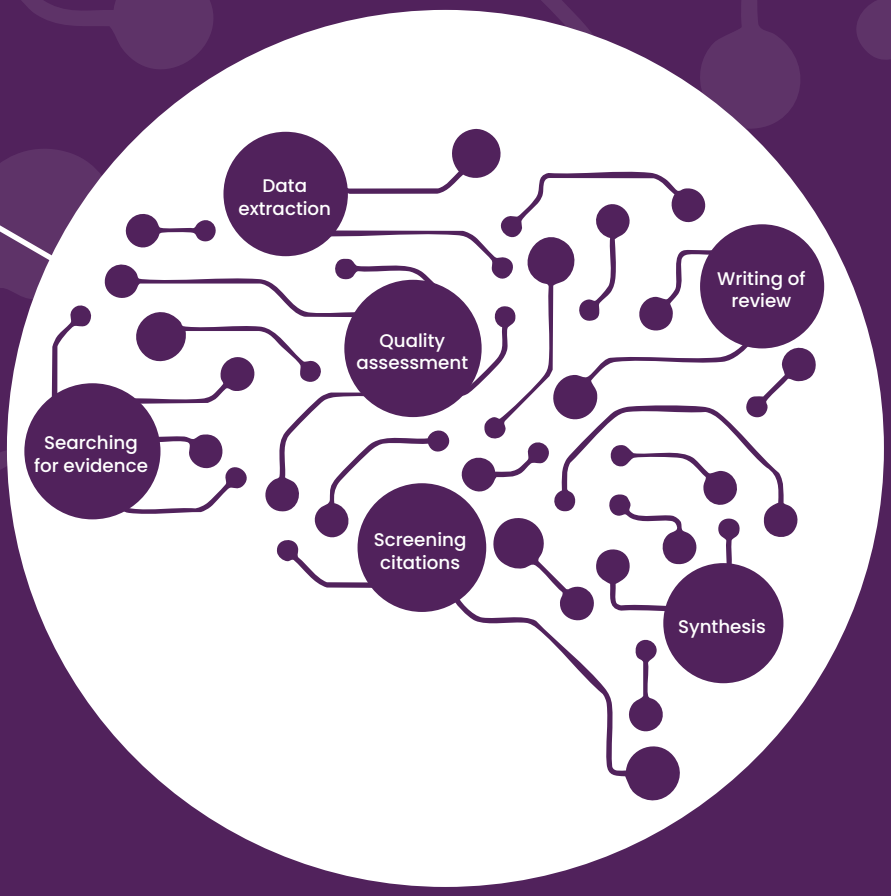
Queen Mary University of London

Health Equity Evidence Centre

A practical guide to using AI tools to assist Evidence Synthesis



Contents

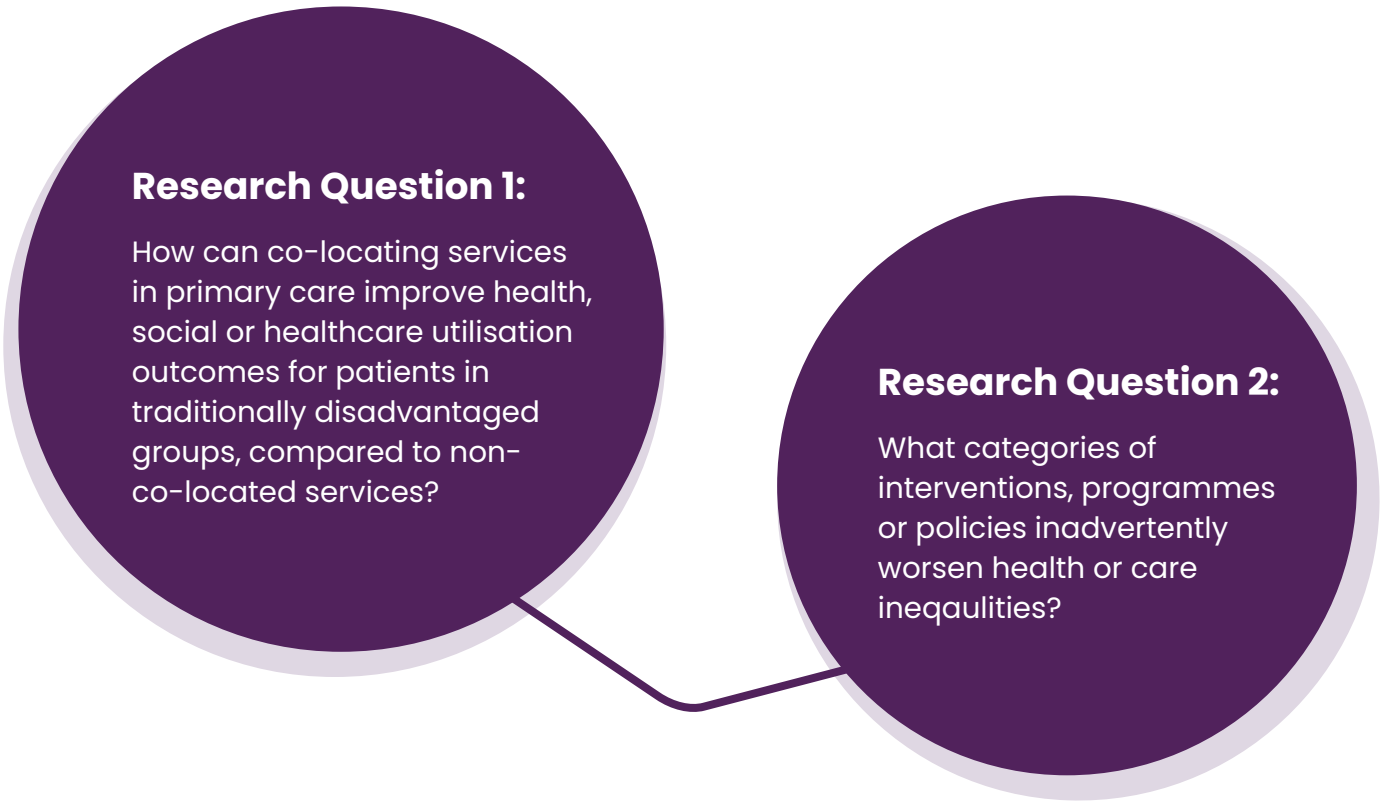
Why create a practical guide?

AI tools are changing how researchers undertake evidence synthesis. There are now tools which can automate every stage of evidence synthesis - from generating search terms to extracting data. With so many tools available and the rapid expansion of generative AI, researchers can find it difficult to know where to start and how to harness of benefits of AI while maintaining rigour and quality.

At the Health Equity Evidence Centre, we have been using AI for priority screening in the creation of living evidence maps using EPPI Reviewer software and have recently finished a research project with the Health Foundation on the use of AI for evidence synthesis. Here we provide tips with practical examples of how to integrate AI into evidence synthesis without compromising quality.

How this practical guide was produced

We have recently finished a programme of research including a scoping review of AI for evidence synthesis, an in-depth review of eight AI tools, a policy workshop and two case studies to compare approaches. Much of the material in this guide is drawn from the case studies where we compared a fully manual, semi-automated, and fully automated approach to evidence synthesis for two complex health and care research questions. Please contact us if you'd like full details of what we did.



Before you start

- Read RAISE guidelines
- Identify which steps in the evidence synthesis pathway would benefit from AI tools
- Match tools to tasks
- Upskill in the tools you plan to use

Read RAISE guidelines

RAISE (Responsible use of AI in evidence SynthEsis) are three sets of guidelines being iteratively developed to maintain academic standards for accuracy, quality, and transparency and contain standards for use of AI in evidence synthesis. There are three distinct documents that aid researchers by providing tailored recommendations by role within the evidence synthesis ecosystem, guidance on evaluation and development of AI tools and guidance on selection and usage of such tools.

Key messages from RAISE 1

- Researchers should ultimately remain accountable for their evidence synthesis. AI cannot be credited as an author or used to fabricate data.
- Researchers should be able to justify using AI to automate evidence synthesis and critically evaluate whether it is methodologically sound.
- Researchers should not treat generative AI as knowledge bases.
- Researches should engage in ongoing training and collaboration with other key stakeholders across the entire evidence synthesis ecosystem.

Identify which steps in the evidence synthesis pathway would benefit from AI tools

Consider using AI tools to help with topic familiarisation, title and abstract screening, data extraction, fixing problems with search strategy syntax and identifying additional studies. In particular, AI tools have been shown to save time compared to manual reviews for screening and data extraction. Your research question will influence the appropriate automation level. For example, automating data extraction works well for simple, well-defined questions (like drug effectiveness reviews using randomised trials), but doesn't work so well for complex topics.

Practical example: Limitations of using Gen AI to develop search strategies

In both of our case studies, we found that the search strategies created by Gen AI (Claude and ChatGPT) missed most studies. In each case, only one of the studies included in the manual reviews (which included a total of 13 and 67 studies for the first and second case studies respectively) was identified using the Gen AI search string. However, when using TERA's automated tools for developing a search strategy which don't use Gen AI and require a higher degree of human oversight, 8 of the 13 included studies were found.

Practical example: Save time by using AI tools with screening

We used two main approaches to automate title-abstract screening:

- **Priority screening:** This machine learning approach is well-validated (1), and is used to rank articles by relevance, allowing you to focus on the most promising studies first. Using EPPI-Reviewer's priority screening, we reduced screening workload by 40% without missing relevant articles, saving 4 hours 40 minutes of work. It still requires manual screening of many articles but maintains a high degree of human oversight..
- **Generative AI:** Using large language models in EPPI-Reviewer, this approach extracts data from title-abstracts and uses this to auto-exclude irrelevant articles. This worked well when abstracts contained clear inclusion criteria information. However, prompts must be carefully crafted to avoid excluding articles that don't meet inclusion criteria in the abstract but might qualify based on full text. As technology develops, full-text screening may become feasible, but current costs make this impractical for large-scale reviews.

Match tools to tasks

Choose the right AI tool for the specific task – there are many tools available which vary in quality and rigour. Many commercially available tools lack independent academic validation, so prioritise those with documented performance in academic settings.

- Tools built specifically for evidence synthesis, particularly screening tools, often outperform general-purpose AI tools.

Practical example: Choosing the right tool

Our scoping review identified **65** AI tools designed to support various stages of evidence synthesis, though evaluations were of often of mixed quality (2). Only a few tools have been thoroughly assessed and validated, particularly for screening the literature.

Some examples include:

- **EPPI-Reviewer:** Extensively validated machine learning for prioritising records during screening while keeping humans in control of final decisions. Works across multiple evidence synthesis stages with additional LLM functionalities for data extraction.
- **ASReview:** Purpose-built for accelerating title and abstract screening using active machine learning.
- **TERA:** Comprehensive suite of automation tools designed to streamline the entire evidence synthesis process.

While general purpose AI tools like ChatGPT and Claude have demonstrated some effectiveness in screening and extraction, their lack of transparency around training data, tendency to hallucinate and high risk of bias warrant significant caution.

Many AI tools also are limited by paywall issues when it comes to accessing publications.

- Tools leveraging traditional machine learning approaches work better on consistently formatted articles that report quantitative, structured evidence, such as randomised trials with standardised eligibility criteria (PICOS format).
- Data extraction tools work well on primary studies, but struggle with systematic reviews, particularly large reviews with heterogeneous primary studies.
- The natural language processing abilities of generative large language models may engage better with qualitative, unstructured evidence – however, they tend to not be grounded in the data provided, often hallucinating outputs even when provided with source materials like full-text articles.

Practical example: Using AI tools for data extraction

EPPI-Reviewer's LLM-based data extraction tool worked well when conducting a systematic review on intervention-generated inequalities. Automated data extraction (using the GPT-4.1 model) performed well for primary studies, as these generally investigate a single research question within a defined population. The relative simplicity and homogeneity of primary study designs reduce ambiguity, allowing automated tools to capture the relevant data with greater accuracy.

However, this tool was less effective when used for an umbrella review on co-location of services. Automated data extraction was less effective for the systematic reviews because they synthesise evidence across multiple heterogeneous studies, encompassing diverse interventions, populations, outcomes, and study designs, and reporting findings in more complex narrative forms. This variability made it difficult for automated tools to consistently identify and categorise the relevant information, resulting in lower accuracy compared to when used on primary studies. For example, one systematic review extraction incorrectly identified "ethnic minorities, people with severe mental illness and elderly" as target populations, but only studies on elderly populations actually examined co-located care as defined by our inclusion criteria. Therefore, significant time was spent manually checking extracted data.

Upskill in the tools you plan to use

It takes time to understand the functionality and limitations of each tool; build in time at the start to understand how each tool works and what it should and shouldn't be used for.

Thinking about using Generative AI? What to consider

- Avoid full automation and don't use AI to synthesise studies or write initial outputs
- Use it strategically
- Optimise prompts
- Consider the environmental impact of generative AI tools

Avoid full automation and don't use AI to synthesise studies or write initial outputs

Generative AI struggles with in-depth synthesis and is prone to hallucinations and inaccuracies. Exercise extreme caution, as outputs can appear comprehensive at first glance while containing significant errors. Instead, use AI to improve readability and correct grammatical errors, but always fact-check the results.

Practical example: Gen AI hallucinations and inaccuracies

- Newer models showed significantly fewer hallucinations when answering our co-location research question. ChatGPT 4.5 performed poorly, with 9 out of 10 citations incorrectly cited (wrong titles, authors, or publication years) and one completely fabricated reference. In contrast, ChatGPT 5's nine citations were all real and correctly cited and had some relevance to the research question, though this may have been unique to our experiment since hallucinations are still highly likely. Claude models (Sonnet 4 and Opus 4.1) consistently provided citation inaccuracies, though they showed fewer outright fabrications compared to the older ChatGPT version.
- Citation inaccuracies were more likely when we prompted the Gen AI tool to meet a specific publication timeframe. This was particularly evident in older models, and demonstrative of their tendency to please users.
- In the semi-automated approach, hallucinations persisted despite providing source material. For example, Claude fabricated additional countries of studies for the second research question during the semi-automated approach, even with access to a detailed data extraction sheet. Achieving moderate-quality analysis of included data required well-formatted extraction tables and iterative prompting.

Use it strategically

Gen AI can help to scope research topics, fix syntax errors in search strategies, identify additional studies and grey literature which may have been missed during the screening phase. If you are not expert at a certain evidence synthesis task yourself, don't allow a generative AI tool to perform it completely.

Optimise prompts

Prompts significantly influence the quality of outputs. Take time to test and compare different prompts, ensuring they are clear, specific, and well structured. This applies even when uploading documents, so be explicit about how the content should be used. For example, when we uploaded the PROSPERO protocol for the co-location research question and simply asked for a report, ChatGPT 4.5 produced a poor output. Instead of using the protocol as a framework for structuring the report, it merely summarised the content despite being prompted otherwise.

Other tips include:

- **Be as descriptive as possible** – when using Gen AI to extract data include detailed study characteristics and population specifications to improve output quality.
- **Request reasoning and methodology** - Ask LLMs to show their rationale and how they made decisions.
- **Trial different approaches** - Test whether taking on roles (like "expert evidence synthesist") makes a difference for your tasks.

Practical example: Impact of prompts on generative LLM outputs

- Identical prompts run on the same day produced vastly different outputs, confirming generative LLMs' non-deterministic behaviour. When we tested our fully automated approach for the co-location research question using ChatGPT 4.5, we ran identical prompts in separate chat windows within minutes of each other, having uploaded the same PROSPERO-registered protocol. Despite identical inputs, the outputs varied dramatically: the first report was succinct and downloadable, citing 8 studies, while the second was more detailed, not downloadable, and structured differently, citing 16 studies with minimal overlap with the first report. Both reports were thematically similar but presented findings in completely different formats.

- Running identical prompts on different days sometimes produced better outputs, though whether this reflects model learning from previous interactions or system updates remains unclear. We observed this again with our fully automated approach for the co-location research question using ChatGPT 5. Despite using the exact same prompt within two days, the later output was distinctly more sophisticated, including sections explaining why co-location matters from an equity perspective and clearly distinguishing co-located care from collaborative care and "just sharing a hallway." These nuanced insights were entirely absent from the earlier attempt.

Does *prompt politeness* matter when using Gen AI chatbots?

Studies suggest being impolite generates poor responses when using chatbots, like ChatGPT or Claude (3). However, being overly polite doesn't guarantee better outcomes. Trial prompts with varying levels of politeness ("please" and "thank you") to see if it affects your outputs.

Consider the environmental impact of generative AI tools

Large language models are both energy and carbon-intensive (4), with studies revealing that a search using ChatGPT consumes significantly more energy than a traditional Google search (5). Therefore, your use of generative AI may need to align with the sustainability commitments set by your research team or institution.

During the evidence synthesis process

- Take ownership
- Don’t skip on critical discussions with colleagues
- Keep researchers in control
- Engage experts throughout

- **Take ownership**

  AI tools exist to support researchers, not to replace them. The design, use and outputs of AI tools remain the responsibility of the humans using them.

- **Don’t skip on critical discussions with colleagues**

  Discussing research questions, methodology, eligibility criteria, and analysis with colleagues is crucial to improving quality.

- **Keep researchers in control**

  Ensure there is a human-in-the-loop at every stage to maintain human oversight. The optimal approach is semi-automated: automate repetitive tasks while keeping researcher control over critical judgments. Structure automation so you can verify each stage. Tools that learn from human inputs (like priority screening) work best by combining AI efficiency with human expertise. Avoid processes where you can’t explore the decision making of tools, especially when using Gen AI (e.g. ChatGPT or Claude), where decision criteria are in a “black box”.

- **Engage experts throughout**

  Discuss critical decisions with topic experts, librarians and evidence reviewers throughout to quality assure outputs.

- State where and how you've used AI in published outputs

Once you've finished

- **State where and how you've used AI in published outputs**

  Document the date, time, model, and key parameters (like machine learning settings) for all AI tools. This disclosure ensures transparency, enables reproducibility by allowing others to replicate your work, and maintains methodological rigor by treating AI tools like any other research instrument requiring technical specifications.

Practical example: What does good disclosure look like?

Any AI tool used to automate evidence synthesis should be clearly described in the Methods section of your report. Although there is no standardised format for disclosing the use of generative AI in this context, examples from primary studies provide useful starting points (see below). In addition, some publishers, such as Elsevier, have issued clear guidance on how to disclose the use of generative AI in scientific writing (6).

Example from Visokay et al. (7):

"Generative AI Disclosure Statement

We utilized multiple Generative AI tools (OpenAI's GPT-4 [including through GitHub Copilot]; Microsoft's Copilot [based on the GPT-4 architecture]; and Anthropic's Claude 3.5/3.7 Sonnet) in the production of this manuscript, in the following ways:

- *Producing computer code for data cleaning and analysis.*
- *Locating relevant research articles in the literature.*
- *Brainstorming ideas and outlining the structure of the paper.*
- *Proposing sentences to include in the manuscript.*
- *Iteratively improving the concision and clarity of the writing.*

We have carefully reviewed all aspects of the manuscript for accuracy and coherence. All scientific insights, analysis and interpretation of data and scientific conclusions are made solely by the authors. All errors are our own. This disclosure is adapted from Professor Tyler Ransom."

In line with RAISE guidelines, this can be improved by:

- Explicitly linking each Generative AI tool to the evidence synthesis stage it supported (e.g., screening, data extraction, analysis).
- Providing details on how the tool was used or if customisations were applied, including the exact prompts crafted, along with the dates when prompts were run.
- Disclosing any financial interests or affiliations related to the AI tools used.

Acknowledgements

An overview and evaluation of the use of automated tools in rapid evidence synthesis project is supported by the Health Foundation, an independent charitable organisation working to build a healthier UK (ref no: FR-0006738).

Disclaimer: This practical guide was developed independently by the Health Equity Evidence Centre (HEEC), hosted by Queen Mary University of London. The guide reflect the authors' own insights and analysis. HEEC currently use EPPI-Reviewer software under a paid licence to produce living evidence maps which are available online.

References

1. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev. 2020 Apr 2;9(1):73.
2. Harasgama S, Pearce H, Appel C, Loftus L, Painter H, Kuhn I, et al. Artificial intelligence tools for automating evidence synthesis: A scoping review (Preprint). Journal of Medical Internet Research; 2025.
3. Yin Z, Wang H, Horio K, Kawahara D, Sekine S. Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. arXiv; 2024.
4. Rillig MC, Ågerstrand M, Bi M, Gould KA, Sauerland U. Risks and Benefits of Large Language Models for the Environment. Environ Sci Technol. 2023 Mar 7;57(9):3464–6.
5. Ji Z, Jiang M. A systematic review of electricity demand for large language models: evaluations, challenges, and solutions. Renew Sustain Energy Rev. 2026 Jan 1;225:116159.
6. Elsevier. Guide for Authors - Journal of Biotechnology [Internet]. 2025. Available from: https://www.sciencedirect.com/journal/journal-of-biotechnology/publish/guide-for-authors
7. Visokay A, Hoffman K, Salerno S, McCormick TH, Johfre S. How to measure obesity in public health research? Problems with using BMI for population inference. medRxiv; 2025. p. 2025.04.01.25325037.