Journal of Clinical Epidemiology 198 (2026) 112390

PROTOCOL

Cochrane evaluation of (semi-)automated review methods: protocol for an adaptive platform study within reviews

Gerald Gartlehner[a,b,*], Susan Banda[a], Max Callaghan[c], Jo-Ana Chase[d], Andreea Dobrescu[a], Angelika Eisele-Metzger[e,f], Ella Flemyng[d], Sean Gardner[d], Ursula Griebler[a], Bartosz Helfer[a], Pawel Jemiolo[g,h,i], Biljana Macura[j], Jan C. Minx[c], Anna Noel-Storr[d], Noosheen Rajabzadeh Tahmasebi[e], Amin Sharifan[a], Joerg J. Meerpohl[e,f,1], James Thomas[k,1]

[a]*Department for Evidence-Based Medicine and Evaluation, Cochrane Austria, University for Continuing Education, Krems, Austria*
[b]*RTI International, Center for Health Economics, Methods & Evidence Synthesis, Research Triangle Park, NC, USA*
[c]*Member of the Leibniz Association, Potsdam Institute for Climate Impact Research, Department for Climate Economics and Policy, Potsdam, Germany*
[d]*Cochrane Central Executive, Cochrane Collaboration, London, United Kingdom*
[e]*Institute for Evidence in Medicine, Medical Center − University of Freiburg/Medical Faculty − University of Freiburg, Germany*
[f]*Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany*
[g]*AGH University of Krakow, Krakow, Poland*
[h]*Faculty of Medicine, Department of Hygiene and Dietetics, Jagiellonian University Medical College, Krakow, Poland*
[i]*Cochrane Poland, Krakow, Poland*
[j]*Stockholm Environment Institute, Stockholm, Sweden*
[k]*EPPI Centre, UCL Social Research Institute, University College London, London, UK*

Accepted 16 June 2026; Published online 19 June 2026

Abstract

Background and Objectives: Artificial intelligence (AI) has the potential to improve the efficiency of evidence synthesis and reduce human error. However, robust methods for evaluating rapidly evolving AI tools within the practical workflows of evidence synthesis remain underdeveloped. This protocol describes a study design for assessing the effectiveness, efficiency, and usability of AI tools in comparison to traditional human-only workflows in the context of Cochrane systematic reviews.

Methods: Members of the Cochrane Evaluation of (Semi-)Automated Review Methods (CESAR) project developed an adaptive platform study-within-a-review design, modeled after clinical platform trials. This design employs a master protocol to concurrently evaluate multiple AI tools (interventions) against a standard human-only process (control) across 3 key review tasks: title and abstract screening, full-text screening, and data extraction. The adaptive framework allows for the addition or removal of AI tools based on interim performance analyses without necessitating a restart of the study. Performance will be assessed using metrics such as accuracy (sensitivity, specificity, precision), efficiency (time on task), response stability, impact of errors, and usability, in alignment with Responsible use of AI in evidence SynthEsis principles.

Results: The study will generate comparative data about the performance and usability of specific AI tools used in a semiautomated or fully automated manner relative to standard human effort. The protocol provides a flexible framework for the assessment of AI tools in evidence synthesis, addressing the limitations of static, one-time evaluations.

Conclusion: This study protocol presents a novel methodological approach to addressing the challenges of evaluating AI tools for evidence syntheses. By validating entire workflows rather than individual technologies, the findings will establish an evidence base for determining the viability of integrating AI into evidence synthesis workflows. The adaptive design of this study is flexible and can be adopted by other investigators, ensuring that the evaluation framework remains relevant as new tools emerge.

Keywords: Study protocol; Evidence synthesis; Artificial intelligence; Workflow validation; Study within reviews; Systematic reviews

Funding: Part of this work was supported by the Wellcome Trust, grant number 323143/Z/24/Z, and part by the Cochrane Collaboration.

[1] These authors contributed equally to this work and share last authorship.

* Corresponding author. Department for Evidence-Based Medicine and Evaluation, Cochrane Austria, University for Continuing Education, Dr. Karl Dorrek Strasse 30, Krems 3500, Austria.
E-mail address: gartlehner@cochrane.at (G. Gartlehner).

https://doi.org/10.1016/j.jclinepi.2026.112390

Plain Language Summary

Doctors and researchers rely on systematic reviews, which are thorough summaries of all available research on a health topic, to guide decisions about patient care. However, creating these reviews is a slow and demanding process, often taking more than a year to finish. AI tools could help speed up this work and reduce human errors, but there are currently no reliable ways to test how well these tools perform in real-world settings. This article describes the design of a study that will rigorously test how well AI tools perform when used in actual systematic review workflows, specifically within Cochrane Reviews. The study will compare AI-assisted methods with the traditional approach, where 2 trained researchers independently complete each step. It will look at 3 main tasks: choosing which studies might be relevant based on their titles and abstracts, reading the full-text publication to confirm which studies should be included, and extracting important information from those studies. A key strength of this study is its flexible design. Instead of testing just 1 AI tool at a single point in time, the study allows researchers to add or remove AI tools as new ones become available, similar to how some modern drug trials are run. This approach helps the study keep up with the fast pace of AI development. Researchers will assess the AI tools based on their accuracy, the time they save, how consistent their results are, and how easy they are to use. The ultimate goal of this study is to give the research community strong evidence about when and how AI can be safely and effectively used in systematic reviews to help summarize medical research.

1. Introduction

Systematic reviews and other forms of evidence synthesis are highly resource intensive [1]. Because tasks such as title and abstract screening, risk-of-bias assessment, data extraction, and rating the certainty of evidence are susceptible to human error, they are typically performed manually by at least 2 independent reviewers. Although this practice enhances the correctness and rigor of review findings, it also substantially increases the time and cost required to produce and update evidence syntheses [2].

Artificial intelligence (AI) offers the potential to improve the efficiency of individual evidence synthesis tasks and reduce the risk of human errors [3]. In this protocol, we define AI as "*advanced technologies that enable machines to do highly complex tasks effectively—which would require intelligence if a person were to perform them*," following the definition provided in Responsible use of AI in evidence SynthEsis (RAISE) recommendations [4].

Recent advances in generative large language models (LLMs) have renewed interest in applying AI to evidence synthesis [5]. These models offer the potential to support more integrated and flexible (semi-)automated review workflows. Unlike earlier natural language processing systems, LLMs can process long-form text without task-specific training and are adaptable to a wide range of review tasks. Notably, studies on literature screening [6−8] and data extraction [9−11] have reported promising results. However, the use of LLMs also introduces risks, including the generation of incorrect responses, fabrication of data or references, perpetuation of biases, and dissemination of misinformation [12]. Furthermore, achieving fully reproducible outputs may be more challenging with LLMs [13].

Evidence synthesis software providers and start-ups have integrated AI to assist review teams with various tasks throughout the review process [14−16]. Comparative assessments within evidence synthesis workflows are essential to determine which tasks AI tools perform worse than, similarly to, or better than human reviewers. Such evaluations help clarify which types of errors are introduced or mitigated when using AI tools for evidence synthesis. Methodologically, RAISE guidance outlines foundational principles for the transparent and scientifically sound evaluation of the effectiveness of AI tools in evidence synthesis [4].

The rapid pace of AI tool development for evidence synthesis necessitates new methodological approaches to assess effectiveness. In clinical research, platform trials gained significant momentum during the coronavirus disease 2019 pandemic, driven by the urgent need to rapidly evaluate multiple potential therapies. Platform trials are adaptive clinical trial designs that use a single master protocol to assess multiple interventions simultaneously against a common control group within a specific disease area [17]. Unlike traditional trials, platform trials allow researchers to add or remove treatment arms based on interim results while the trial is ongoing, making them highly efficient and flexible [17].

Drawing on the principles of platform trials, we ntroduce an innovative methodological approach in this study protocol: an adaptive platform study within a review (SWAR) or adaptive platform SWAR. SWARs are embedded methodological studies conducted within systematic reviews to evaluate alternative methods or processes in real-world review settings [18]. By adapting the platform trial framework to the context of evidence synthesis, the adaptive platform SWAR enables prospective, flexible, and efficient evaluation of multiple AI tools or approaches within a single review process.

2. Objective

The Cochrane Evaluation of (Semi-)Automated Review Methods (CESAR) project aims to compare the performance

What is new?

Key findings

- This protocol introduces an adaptive platform study-within-a-review design, a methodological innovation enabling concurrent, prospective evaluation of multiple AI tools to support review teams for various tasks of the evidence synthesis process.
- The flexible, prospective design accommodates ongoing technological advances and maintains methodological rigor to evaluate the effectiveness and usability of AI tools for evidence synthesis.

What this adds to what is known?

- By adapting the platform trial model from clinical research, this design addresses the limitations of traditional static AI tool evaluations and allows for the addition or removal of AI tools as technology evolves.
- The protocol incorporates Responsible use of AI in evidence SynthEsis principles in its evaluation framework, ensuring that usability, response stability, time on task, and the impact of errors are systematically assessed alongside accuracy.

What is the implication and what should change now?

- Validating AI tools within evidence synthesis workflows is essential for determining real-world effectiveness and whether AI tools can be viably integrated into standard review processes. This adaptive platform approach offers the flexibility needed to rigorously assess AI tools in a rapidly advancing field.

of conventional human-only approaches with semiautomated and fully automated methods for completing individual review tasks within ongoing Cochrane review updates. CESAR will evaluate the concordance between these approaches, quantify the human effort required, assess user experiences with AI tools, examine the stability of AI tool responses, and characterize both the types and impacts of errors associated with each process.

3. Research questions

The research questions are as follows:

1. What is the comparative performance between conventional human-only methods and semiautomated or fully automated approaches for completing individual review tasks in Cochrane review updates?
2. How does the human effort required to complete individual review tasks compare conventional human-only and semiautomated or fully automated approaches?
3. What is the potential impact of errors from semiautomated or fully automated approaches on review findings and conclusions?
4. How consistent are AI tool outputs when repeatedly used for the same review tasks?
5. How do users perceive the usability and overall user experience of different AI tools when applied to specific review tasks within Cochrane review updates?

The following sections outline the core protocol elements applied to the adaptive platform SWAR. Task-specific methodological elements are detailed in Appendices A–C.

4. Master protocol

CESAR evaluates (semi-)automated approaches for 3 core systematic review tasks: title and abstract screening, full-text screening, and data extraction. The study duration will be 6–8 months. For each participating review, we will use a prospective, parallel-group design in which AI tools are applied for semiautomated (AI-assisted) and, where technically feasible, fully automated completion of these tasks. Conventional human-only processes on the same reviews, conducted in compliance with the Cochrane Handbook [19], will serve as the shared control group, providing a consistent baseline for comparison (Fig).

4.1. Selection of AI tools

In November 2025, Cochrane invited developers of AI tools to submit proposals via a structured public form [20]. Proposals that met the initial screening for completeness were numerically scored on four criteria: adherence to the RAISE initiative [4]; transparency and interpretability of AI decisions; alignment with Cochrane's mission and values; and compliance with data protection, copyright, and interoperability standards [21]. Independent rankings from members of the joint AI Methods Group (https://methods.cochrane.org/ai/) determined which two AI tools would initially enter the platform study, with others considered for inclusion if interim analysis led to discontinuation. Tool providers will be contractually required to notify Cochrane about any major model updates.

4.2. Selection of reviews

We will select a convenience sample of 10–15 ongoing updates of Cochrane intervention reviews of diverse topics

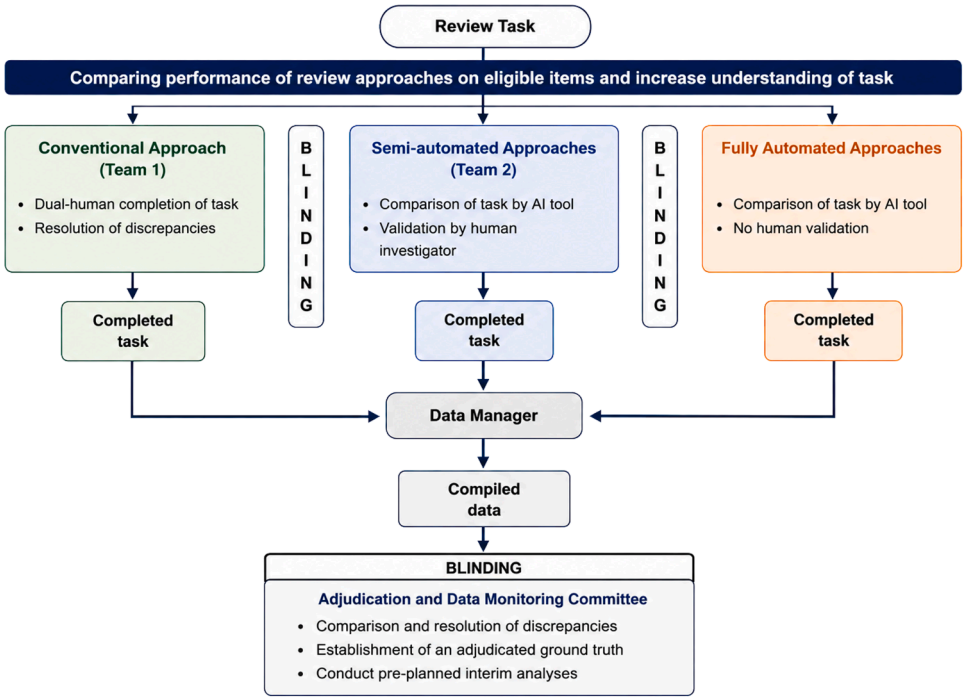

Figure. Study flow diagram comparing conventional, semiautomated, and fully automated approaches to systematic review tasks.

based on the following criteria: (1) the review authors' willingness to participate and the capacity to split into two teams with at least two members per team, (2) alignment of the review timeline with the study schedule, and (3) the expectation that the update will include at least one new study. Reviews requiring major methodological updates (eg, rating the certainty of evidence) are not eligible. Each selected review will be integrated into CESAR after the corresponding review team has completed its literature searches. Because we expect a heterogenous sample of reviews, we will collect major review characteristics to be able to conduct exploratory analyses by review complexity.

4.3. Assembling review teams and preparatory work

For each review, we will establish 2 independent teams (hereafter referred to as Team 1 and Team 2) composed of members from the Cochrane review team conducting the update. While different individuals will comprise these teams, all members must possess review experience and a thorough understanding of the topic and review objectives. Team 1 (Fig) will execute the conventional human-only approach, and Team 2 will implement a semiautomated approach We will use block randomization to assign AI tools. If AI tools allow for fully automated approaches, members of the joint AI Methods Group (https://methods.cochrane.org/ai/) will handle automated tasks.

4.4. Common pilot phase

Participating teams will conduct a common pilot phase of the tasks for each review. This phase aims to align the teams, ensure consistent interpretation of instructions, and enhance familiarity with the tasks. During this phase, the teams conducting the semiautomated approaches will receive AI tool−specific training. Insights gained from the pilot will be used to refine instructions and improve workflow comparability.

4.5. Review task completion

Teams assigned to different approaches will complete identical tasks under the same instructions but will be blinded to each other's work to prevent performance bias. All completed tasks will be submitted to a data manager, who will compile a standardized, deidentified output matrix containing results from all study arms (conventional, semiautomated, and fully automated). Group identifiers will be removed prior to submission to the Adjudication and Data Monitoring Committee (ADMC).

4.6. Reference standards

Reference standards differ for each review task and are detailed in the respective appendices.

Table 1. Classification of the potential impact of errors (adapted from Gartlehner et al [22])

Potential impact of errors	
Major error	This error substantially compromises data correctness and, if uncorrected, could lead to erroneous conclusions; eg, false exclusion of a large study, grossly incorrect calculations when extracting data, or fabricated data that result in a new or different interpretation of the data.
Moderate error	This error is less severe than a major error, may or may not impact interpretation of existing data, and does not lead to erroneous conclusions; eg, false exclusion of a small study, rounding errors that do not critically affect the data's overall utility, or data extraction errors that do not alter the findings in a meaningful way.
Inconsequential difference	This difference most likely would not impact the interpretation of data or conclusions; eg, differences in the detail level of extracted information on inclusion criteria.

4.7. Data adjudication and interim monitoring

The ADMC serves as an independent body that remains blinded for data adjudication. The committee composition will include a senior topical expert from each review team not involved in the specific tasks under evaluation alongside methodological experts. Within the adaptive framework, the ADMC has two primary mandates: (1) data adjudication and (2) interim monitoring.

4.7.1. Data adjudication

The ADMC will systematically assess and resolve discrepancies between conventional, semiautomated, and fully automated approaches using prespecified adjudication criteria. For data extraction, the ADMC will establish the reference standard using a structured adjudication process. The ADMC will:(1) resolve discrepancies using a consensus-based method and original source documents; (2) determine which approach made an error and assess its severity based on the potential impact on the review's results and conclusions (Table 1); and (3) review a random sample (20%) of concordant items to assess the frequency of jointly incorrect extractions, that is, cases in which both approaches made the same error.

All adjudication decisions, including dissenting opinions and their rationale, will be documented in a standardized adjudication log to ensure transparency and reproducibility.

4.7.2. Interim monitoring

The ADMC will conduct at least one preplanned interim analysis to evaluate whether any AI tool should be discontinued for futility (failing to meet minimum performance thresholds) or harm (introducing systematic errors that could compromise review validity). A formal interim analysis will be conducted for each AI tool after the completion of 5 reviews (33% information fraction). The ADMC will be unblinded for interim analysis.

Each AI-supported approach will be evaluated against performance thresholds considering both point estimates and CIs (Table 2). Specifically, an AI tool will be considered for discontinuation if:the observed point estimate falls below the futility boundary, indicating unambiguous underperformance; or (2) the upper limit of the 95% CI falls below a noninferiority margin, indicating failure to meet acceptable performance even under optimistic assumptions.

4.7.3. Decisions regarding changes in the study design

Decisions regarding modifications to the study design overseen by a Cochrane Oversight Committee, the ADMC, and the principal investigator (G.G.). Specifically, the ADMC will review preliminary analyses to determine whether a tool's performance warrants its removal from the platform study. Should an AI tool be discontinued, the next AI tool on the ranked shortlist will be enrolled in the study for subsequent reviews, ensuring a continuous and efficient evaluation process. All modifications will be documented with a clear rationale and timestamps to maintain a transparent and auditable trail throughout the study.

4.8. Sample size considerations

As this study is exploratory rather than confirmatory in nature, no hypotheses will be tested, and no formal sample size calculations will be performed. This approach allows for flexible, descriptive, and comparative analyses and provides the opportunity to explore multiple aspects of semiautomated and fully automated review processes.

4.9. Unit of evaluation and performance metrics

We will use individual review tasks as the unit of evaluation (rather than entire reviews encompassing multiple review tasks) because this approach provides more detailed insight into the performance of AI tools. Assessing tasks individually allows the use of a consistent baseline for the approaches for each new review task, thereby preventing errors that propagate from one task to the next within a review.

The choice of performance metrics will depend on the specific review task under evaluation and will be presented in more detail in the appendices on the individual tasks. In general, we will follow the RAISE 2 recommendations for performance metrics [24].

Table 2. Stopping Boundaries and Decision Rules for Interim Analyses

Performance metrics	Futility boundaries (point estimate)[a]	Noninferiority margins (upper limit of 95% CI)[a]	Decision rules
Screening			
Sensitivity	<80%[b]	<95%[b]	Stop if either boundary is crossed
Specificity (for full-text screening only)	<50%[c]	<60%[c]	Stop if either boundary is crossed
Data extraction			
Sensitivity	<92%[b]	<97%[b]	Stop if either boundary is crossed
Major error proportion	>3%[d]	>2%[d]	Stop if either boundary is crossed
Usability			
System Usability Scale (score)	<57[e]	<75[e]	Stop if threshold is not met

[a] Futility boundaries are based on the 25th percentile, representing a ''minimum bar''; noninferiority margins are based on the 75th percentile, representing a ''high bar''.
[b] Based on a presentation by Flemyng et al., understanding expectations for evidence synthesis when using AI Survey results (presented at International Collaboration for Automation in Systematic Reviews 2025 meeting, July 2025).
[c] Based on consensus of the Scientific Advisory Board of this study.
[d] Pertains to the lower bound of the CI based on Gartlehner et al [9].
[e] Based on Sauro et al [23].

4.9.1. Accuracy metrics

The assessment of accuracy metrics will vary by review task and is described in more detail for each review task in the appendices.

4.9.2. Stability metrics

We will assess the stability of an AI tool's output if an application programming interface is available. We will conduct 10 independent iterations of fully automated tasks within a single day. For all tasks, our primary focus will be on sensitivity (recall). This approach will provide an initial estimate of variability in sensitivity arising from stochastic model behavior. During these iterations, no human verification will be performed, allowing assessment of the intrinsic stability of the fully automated approach. This process differs from our primary semiautomated approach, in which human reviewers will verify all AI screening decisions.

4.9.3. Efficiency metrics

We will assess time on task, defined as the total time required to complete a review task from initiation to verified output. For conventional approaches, this includes all steps typically performed by trained reviewers. For semiautomated approaches, the same steps are covered, plus time for (1) tool configuration, (2) human verification of tool output and error correction, and (3) any necessary iteration or refinement. We will exclude initial training time for reviewers learning to use AI tools from the primary analysis but report it separately as implementation overhead. Time measurements will be recorded by reviewers using a time-tracking app.

For data extraction, a member of the Cochrane study team will measure the time required to convert the (semi-) automated output into a data format compatible with RevMan [25].

4.9.4. Usability metrics

Usability will be assessed using an electronic survey for each AI tool after completion of each review task. We will apply the System Usability Scale (SUS), a validated 10-item questionnaire that provides scores from 0 to 100 [26]. Scores below 70 indicate products needing scrutiny and improvement, whereas scores above 90 are considered superior [27]. The SUS has been successfully used to evaluate digital tools and systems [23]. Alongside usability, we will assess constructs relevant to the adoption of AI tools, including user satisfaction, trust in automation, and perceived risk. For trust in automation, we will use a subscale from the ''Trust in Automation'' questionnaire [28]. In addition, qualitative feedback will be collected through open-ended questions and analyzed thematically to identify common usability barriers and facilitators. We will also collect information on reviewers' expertise in conducting systematic reviews and using AI tools, as well as their self-assessed topical expertise. The survey will be conducted anonymously, and responses will not be traceable to individual review team members. We expect 20–25 respondents per AI tool for each task. An ethics waiver for the questionnaire was granted by the Ethics Committee of the University for Continuing Education Krems.

4.9.5. Risk assessment and error analysis

In accordance with Paper 2 of the RAISE guidance, we will assess the impact of errors, focusing on 3 outcomes: (1) error severity scoring (which will be performed by the ADMC, see Table 1), (2) proportions of missed critical information, and (3) changes in evidence synthesis results [24].

The definitions of missed critical information differ by task and are detailed for each task in the appendices. Changes in evidence synthesis results will be evaluated

by recalculating effect estimates for the main outcomes in each review, either by excluding missed studies or using erroneous data. This evaluation will address both the cumulative impact of all errors introduced by automating a task and the marginal impact of individual errors using leave-one-out sensitivity analysis [29].

For each pair of meta-analyses, we will assess the agreement between the recalculated and original effect estimates using Lin's concordance correlation coefficient (CCC). We will first convert all dichotomous main outcomes to standardized mean difference to have a common metric for all studies [30]. Lin's CCC measures both covariance and accuracy, with values close to +1 indicating stronger overall agreement and values near 0 or negative indicating poor agreement. We will also investigate the statistical significance of differences between recalculated and actual effect estimates using z-scores, which quantify the difference relative to the combined uncertainty of the 2 estimates.

Finally, we will use methods similar to Nussbaumer-Streit et al [31] to assess whether differences in effect estimates imply a change in review conclusions. Specifically, a changed conclusion will be defined as cases where review authors report (1) the same conclusion with less certainty, (2) a different conclusion (eg, opposite direction), or (3) the inability to draw a conclusion.

4.10. Statistical analysis

The statistical approach for this study will be primarily descriptive, reflecting its exploratory objectives. We will not conduct formal inferential statistics for direct comparisons between individual approaches.

Continuous outcomes, such as time on task, will be summarized using means, SDs, and 95% CIs to characterize the distribution and uncertainty of measurements. For dichotomous outcomes and diagnostic performance metrics (eg, sensitivity, specificity, and precision), we will report proportions alongside 95% Clopper−Pearson CIs. This exploratory design is intended to generate robust descriptive insights that will inform future hypothesis-driven research and provide a foundation for methodological guidance on the responsible integration of AI tools within systematic review workflows.

In case of a major model update, we will flag the date and conduct a sensitivity analysis comparing outputs before and after the change on a subset of overlapping records (eg, records of the same review screened before and after the update). If the difference is substantial, we will consider stratifying analyses by model version.

4.11. Data management

A data manager will oversee the data collection process. We will store data electronically in Microsoft Excel datasheets in secure SharePoint folders managed by Cochrane. All investigators will have access to the data. The final prompts, model configuration settings, and human- and model-generated data outputs will be saved and shared upon request as part of dissemination. Likewise, we will make the final study data available upon request. Anyone wishing to withdraw their involvement in the project and delete their data will be able to do so. Anonymized and unidentifiable quotes or responses will only be used in project outputs, which may be public, with written permission.

4.12. Ethical considerations

For the usability survey, an ethics waiver was granted by the Ethics Committee of the University for Continuing Education Krems. For the Cochrane systematic review teams, corresponding authors will provide written consent on behalf of all authors to be part of the study and for Cochrane to use published and unpublished review data. Any data inputs or outputs of the AI tools remain the responsibility and property of the Cochrane review teams, as is standard practice with AI tool use by researchers. We will use AI tools that do not train on uploaded PDFs or user interactions, ensuring that the intellectual property rights of copyrighted study reports and review team data are fully respected.

5. Discussion

Validating AI tools within evidence synthesis workflows is necessary to determine whether these technologies deliver practical value under real-world review conditions rather than only in controlled experimental settings. CESAR addresses this need by introducing an adaptive platform SWAR design that enables the concurrent, prospective evaluation of multiple AI tools across different evidence synthesis tasks. By adapting the logic of platform trials from clinical research, this design offers an important methodological advance over static SWAR evaluations. It allows tools to be added, modified, or removed as technologies evolve while preserving a structured comparative framework. This flexibility is important in a field where AI systems change rapidly and results of conventional evaluation designs risk becoming outdated before they can be implemented.

A major strength of this study design is its orientation as a workflow validation study. Unlike model validation studies, which typically assess performance on preexisting benchmark datasets under tightly controlled conditions, workflow validation embeds the AI tool directly into ongoing review processes [22]. This provides a more meaningful assessment of effectiveness, efficiency, usability, and error consequences in practice. The prospective nature of the design further strengthens the study by reducing the risk of data contamination, which may occur when evaluation datasets overlap with the data used to train LLMs [32]. As developers rarely disclose the contents of training datasets,

the use of ongoing unpublished reviews offers an important safeguard against inflated estimates of performance.

Another strength is the explicit incorporation of RAISE principles into the evaluation framework [4]. Accuracy alone is insufficient for judging whether an AI tool can be responsibly integrated into evidence synthesis. By also evaluating usability, response stability, time on task, and the impact of errors, the study recognizes that implementation decisions depend on a broader set of considerations than accuracy metrics alone.

The study adopted a rigorous approach to defining the reference standard. Due to variability in human inclusiveness when screening titles and abstracts, human decisions alone do not provide an optimal reference standard. Using the final set of included study reports as the reference standard, the study minimized the risk of variability across different reviews. Furthermore, instead of presuming that human-only data extraction is inherently reliable, any discrepancies between human-only and AI-assisted data extraction were adjudicated against the original study reports by blinded assessors. This process reduces benchmark bias resulting from imperfections in the human comparator and enables a more equitable assessment of all approaches.

Several limitations should be acknowledged. First, when using the final set of included study reports as the reference standard for title and abstract screening, it is not possible to reliably distinguish false positives from true negatives in the semiautomated and fully automated approaches. Consequently, metrics such as specificity, precision, and overall accuracy cannot be calculated. Due to time constraints, we limited the number of extracted data items to a maximum of 35, which does not represent the full range of data typically extracted in systematic reviews. Additionally, adjudicating discordant extractions and classifying error type and severity inevitably involve subjective judgment. Although employing 2 independent adjudicators and a third reviewer to resolve disagreements helps reduce arbitrariness, some inconsistency is unavoidable. Furthermore, by adjudicating mostly discordant extractions, concordant but jointly incorrect extractions may go undetected. We will sample 20% of concordant extractions to determine the frequency of jointly incorrect extractions. However, we believe that such cases may be rare because human reviewers and AI tools tend to make errors through different mechanisms: humans are susceptible to fatigue and inattention, while AI errors more commonly arise from ambiguous phrasing, complex sentence structures, or domain-specific terminology.

Overall, CESAR provides a rigorous and forward-looking framework for evaluating AI in evidence synthesis. Its chief contribution is methodological: it moves assessment beyond static benchmarks and toward prospective, adaptive, workflow-based validation that is better aligned with real-world implementation. This study design is readily adoptable by other teams seeking to validate AI tools within evidence synthesis workflows.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this article, the authors used generative AI (Claude 4.5 Sonnet) to assist with text editing; all AI outputs were edited, where necessary, and verified by humans.

CRediT authorship contribution statement

Gerald Gartlehner: Writing − review & editing, Writing − original draft, Supervision, Project administration, Methodology, Conceptualization. **Susan Banda:** Writing − review & editing, Writing − original draft, Methodology. **Max Callaghan:** Writing − review & editing, Writing − original draft, Methodology. **Jo-Ana Chase:** Writing − review & editing, Writing − original draft, Methodology. **Andreea Dobrescu:** Writing − review & editing, Writing − original draft. **Angelika Eisele-Metzger:** Writing − review & editing, Methodology. **Ella Flemyng:** Writing − review & editing, Resources, Conceptualization. **Sean Gardner:** Writing − review & editing, Writing − original draft, Methodology, Conceptualization. **Ursula Griebler:** Writing − review & editing, Writing − original draft, Conceptualization. **Bartosz Helfer:** Writing − review & editing. **Pawel Jemiolo:** Writing − review & editing, Writing − original draft. **Biljana Macura:** Writing − review & editing. **Jan C. Minx:** Writing − review & editing. **Anna Noel-Storr:** Writing − review & editing, Methodology. **Noosheen Rajabzadeh Tahmasebi:** Writing − review & editing, Writing − original draft, Methodology. **Amin Sharifan:** Writing − review & editing, Writing − original draft, Methodology. **Joerg J. Meerpohl:** Writing − review & editing, Methodology, Conceptualization. **James Thomas:** Writing − review & editing, Methodology, Conceptualization.

Declaration of competing interest

G.G., who is employed at a 25% full-time equivalent by RTI International, declares an institutional conflict of interest. RTI International maintains an institutional partnership with Nested Knowledge, including an equity investment made in 2024. G.G. has no personal financial or nonfinancial ties to Nested Knowledge, receives no equity, bonuses, or remuneration from the company, and does not use Nested Knowledge in his RTI-affiliated work. G.G. was not involved in ranking the AI tools, which determined which AI tools entered the study.

G.G., M.C., A.E.M., E.F., P.J., B.M., J.J.M., J.M., A.N.S., and J.T. are co-convenors of the joint Methods Group between Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence.

E.F., A.N.S., J.A.C., and S.G. are employed by the Cochrane Collaboration.

A.D. and B.H. are engaged under a consulting agreement with Cochrane to serve as members of the Adjudication and Data Monitoring Committee.

The other authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank Sandra Hummel from Cochrane Austria for administrative support and formatting.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2026.112390.

Data availability

No data was used for the research described in the article.

References

[1] Nussbaumer-Streit B, Ellen M, Klerings I, Sfetcu R, Riva N, Mahmić-Kaknjo M, et al. Resource use during systematic review production varies widely: a scoping review. J Clin Epidemiol 2021;139:287−96. https://doi.org/10.1016/j.jclinepi.2021.05.019.

[2] 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;7(2):e012545. https://doi.org/10.1136/bmjopen-2016-012545.

[3] Scotti KL, Young S, Gainey MA, Lan H. Artificial intelligence and automation in evidence synthesis: an investigation of methods employed in cochrane, Campbell collaboration, and environmental evidence reviews. Cochrane Evid Synth Methods 2025;3(5):e70046. https://doi.org/10.1002/cesm.70046.

[4] Thomas J, Flemyng E, Noel-Storr A, Moy W, Marshall I, Hajji R, et al. Responsible use of AI in evidence SynthEsis (RAISE) 1: recommendations for practice. Available at: https://osf.io/fwaud/overview. Accessed February 15, 2026.

[5] Wang Z, Cao L, Danek B, Jin Q, Lu Z, Sun J. Accelerating clinical evidence synthesis with large language models. NPJ Digital Med 2025;8(1):509. https://doi.org/10.1038/s41746-025-01840-7.

[6] Guo E, Gupta M, Deng J, Park Y-J, 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.

[7] Oami T, Okada Y, Nakada T-A. Performance of a large Language model in screening citations. JAMA Netw Open 2024;7(7):e2420496. https://doi.org/10.1001/jamanetworkopen.2024.20496.

[8] Tran V-T, Gartlehner G, Yaacoub S, Boutron I, Schwingshackl L, Stadelmaier J, et al. Sensitivity and specificity of using GPT-3.5 turbo models for title and abstract screening in systematic reviews and meta-analyses. Ann Intern Med 2024;177(6):791−9. https://doi.org/10.7326/M23-3389.

[9] Gartlehner G, Kugley S, Crotty K, Viswanathan M, Dobrescu A, Nussbaumer-Streit B, et al. Artificial intelligence-assisted data extraction with a large Language model: a Study within reviews. Ann Intern Med 2025;178:1763−71. https://doi.org/10.7326/annals-25-00739.

[10] Khan MA, Ayub U, Naqvi SAA, Khakwani KZR, Sipra ZBR, Raina A, et al. Collaborative large language models for automated data extraction in living systematic reviews. J Am Med Inform Assoc 2025;32(4):638−47. https://doi.org/10.1093/jamia/ocae325.

[11] Mitchell E, Are EB, Colijn C, Earn DJD. Using artificial intelligence tools to automate data extraction for living evidence syntheses. PLoS One 2025;20(4):e0320151. https://doi.org/10.1371/journal.pone.0320151.

[12] Zhang G, Jin Q, Jered McInerney D, Chen Y, Wang F, Cole CL, et al. Leveraging generative AI for clinical evidence synthesis needs to ensure trustworthiness. J Biomed Inform 2024;153:104640. https://doi.org/10.1016/j.jbi.2024.104640.

[13] Lieberum J-L, Toews M, Metzendorf M-I, Heilmeyer F, Siemens W, Haverkamp C, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. J Clin Epidemiol 2025;181:111746. https://doi.org/10.1016/j.jclinepi.2025.111746.

[14] Hamel C, Kelly SE, Thavorn K, Rice DB, Wells GA, Hutton B. An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening − impact on reviewer-relevant outcomes. BMC Med Res Methodol 2020;20(1):256. https://doi.org/10.1186/s12874-020-01129-1.

[15] 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;9(1):73. https://doi.org/10.1186/s13643-020-01324-7.

[16] Yao X, Low A, Sivajohanathan D, Vella E, Wang P, Kaur G, et al. Evaluation of artificial intelligence-based tool Covidence in literature screening for guideline updates: a prospective study. Intell Med 2025. https://doi.org/10.1016/j.imed.2025.12.006.

[17] Angus DC, Alexander BM, Berry S, Buxton M, Lewis R, Paoloni M, et al. Adaptive platform trials: definition, design, conduct and reporting considerations. Nat Rev Drug Discov 2019;18(10):797−807. https://doi.org/10.1038/s41573-019-0034-3.

[18] Devane D, Burke NN, Treweek S, Clarke M, Thomas J, Booth A, et al. Study within a review (SWAR). J Evidence-Based Med 2022;15(4):328−32. https://doi.org/10.1111/jebm.12505.

[19] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane handbook for systematic reviews of interventions version 6.4. 2023. Available at: www.training.cochrane.org/handbook. Accessed February 15, 2026.

[20] Cochrane. Call for proposals: AI tools to transform evidence synthesis. Available at: https://www.cochrane.org/about-us/news/call-proposals-ai-tools-transform-evidence-synthesis. Accessed February 20, 2026.

[21] Cochrane. How did Cochrane select AI tools to evaluate in our platform study?. Available at: https://www.cochrane.org/about-us/news/how-did-cochrane-select-ai-tools-evaluate-our-platform-study. Accessed April 9, 2026.

[22] Gartlehner G, Kahwati L, Nussbaumer-Streit B, Crotty K, Hilscher R, Kugley S, et al. From promise to practice: challenges and pitfalls in the evaluation of large language models for data extraction in evidence synthesis. BMJ Evid Based Med 2025;30(6):385−9. https://doi.org/10.1136/bmjebm-2024-113199.

[23] Sauro J, Lewis JR. Quantifying the user experience: Practical statistics for user research. Boston, MA: Morgan Kaufmann; 2016.

[24] Flemyng E, Thomas J, Noel-Storr A. Responsible use of AI in evidence SynthEsis (RAISE) 2: building and evaluating AI evidence synthesis tools. Open Sci Framework. Available at: https://osf.io/fwaud/. Accessed February 15, 2026.

[25] Review Manager (RevMan) [Computer program]. Version 11.4.0. 2026. The Cochrane Collaboration. Available at https://revman.cochrane.org. Accessed July 10, 2026.

[26] Brooke J. SUS-A quick and dirty usability scale. In: Jordan PW, Thomas B, McClelland I, Weerdmeester B, editors. Usability evaluation in industry. London: Taylor & Francis; 1996:189−94.

[27] Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System usability Scale. Int J Human−Computer Interaction 2008;24(6): 574−94. https://doi.org/10.1080/10447310802205776.

[28] Theoretical considerations and development of a questionnaire to measure trust in automation. In: Körber M, editor. Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018). Cham: Springer International Publishing; 2019. https://doi.org/10.1007/978-3-319-96074-6_2.

[29] Viechtbauer W, Cheung MWL. Outlier and influence diagnostics for meta-analysis. Res Synth Methods 2010;1(2):112−25. https://doi.org/10.1002/jrsm.11.

[30] Gartlehner G, Dobrescu A, Swinson Evans T, Thaler K, Nussbaumer B, Sommer I, et al. Comparison of effects as evidence evolves from single trials to high-quality bodies of evidence. research white paper. (Prepared by the RTI-UNC-evidence-based practice center under contract no. 290-2012-0008-I). Rockville, MD: Agency for Healthcare Research and Quality; 2015.

[31] Nussbaumer-Streit B, Klerings I, Wagner G, Heise TL, Dobrescu AI, Armijo-Olivo S, et al. Abbreviated literature searches were viable alternatives to comprehensive searches: a meta-epidemiological study. J Clin Epidemiol 2018;102:1−11. https://doi.org/10.1016/j.jclinepi.2018.05.022.

[32] Carlini N, Ippolito D, Jagielski M, Lee K. Quantifying memorization across neural Language models. International conference on learning representations. arXiv 2023. https://doi.org/10.48550/arXiv.2202.07646.