Cochrane Evaluation of (Semi-) Automated Review Methods: CESAR

National Academy of Sciences

Gerald Gartlehner
Cochrane Austria, University Krems
RTI International

Funding and Conflicts of Interest

Funding: CESAR is supported by the Cochrane Collaboration and the Wellcome Trust.

Institutional conflict: RTI International maintains an institutional partnership with Nested Knowledge, including an equity investment made in 2024.

Personal disclosure: I have no personal financial or non-financial relationships with Nested Knowledge or Laser AI and do not use them in my evidence synthesis work.

Cochrane Austria

Call for Proposal and Objectives

Call for proposals: AI tools to transform evidence synthesis

Cochrane is looking for AI tool developers to propose tools for consideration for the pilot initiative

Wednesday, November 19, 2025

- Evaluate AI tools for potential recommendation to Cochrane review authors.
- Collaborate with AI tool developers to create solutions aligned with Responsible use of AI in evidence SynthEsis (RAISE) standards and the Cochrane workflow.
- Establish clear standards for evaluation methods, accuracy thresholds, and validation procedures when integrating AI into systematic review processes.

Cochrane Austria

Assessment of Submissions

- 48 submissions
- Cochrane scored the AI tools using predefined criteria, including:
  - Alignment with the RAISE initiative.
  - Consistency with Cochrane's mission, vision and values.
  - Validation and compliance with data protection and copyright standards.
  - Technical feasibility, such as potential interoperability with RevMan, scalability, and usability.
- The Cochrane AI Methods Group then ranked nine shortlisted AI tools.
- The top-ranked AI tools: **Laser AI** and **Nested Knowledge**

Tool name	Alignment with RAISE /10
Nested Knowledge	8
otto-SR	7
Loon Lens™	7
Elicit	7
ASReview	7
Sysrev	6
DistillerSR	6
Laser AI	6
SYMPRO AI	5

Courtesy of Jo-Ana Chase and Sean Gardner, Cochrane Collaboration

Cochrane Austria

Cochrane Methods
Artificial Intelligence

Objectives

To compare the performance of conventional human-only approaches with semi-automated and fully automated methods for completing individual review tasks within ongoing Cochrane review updates.

Challenges

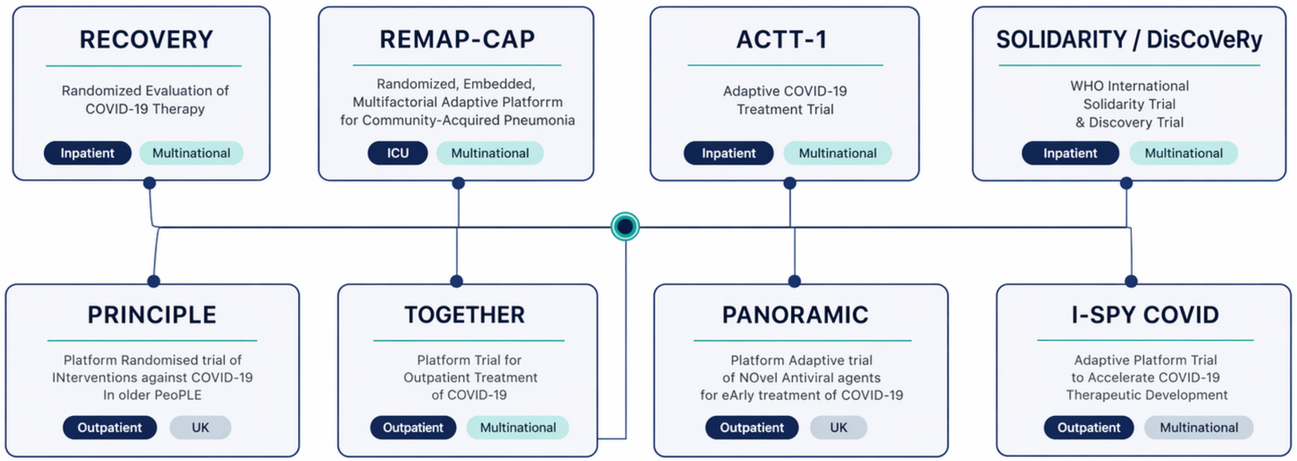
- To be applicable, the study must be embedded in Cochrane workflows.
- To engage Cochrane review groups, the study needs to be pragmatic.
- Traditional evaluation methods for AI tools are often static and do not adapt as new data and insights accumulate.
- Flexible methodologies are needed to incorporate new insights and replace AI tools that fall short of performance standards.

Cochrane
Austria

Adaptive Platform Trials

…. were a fast and flexible way to test treatments during the pandemic

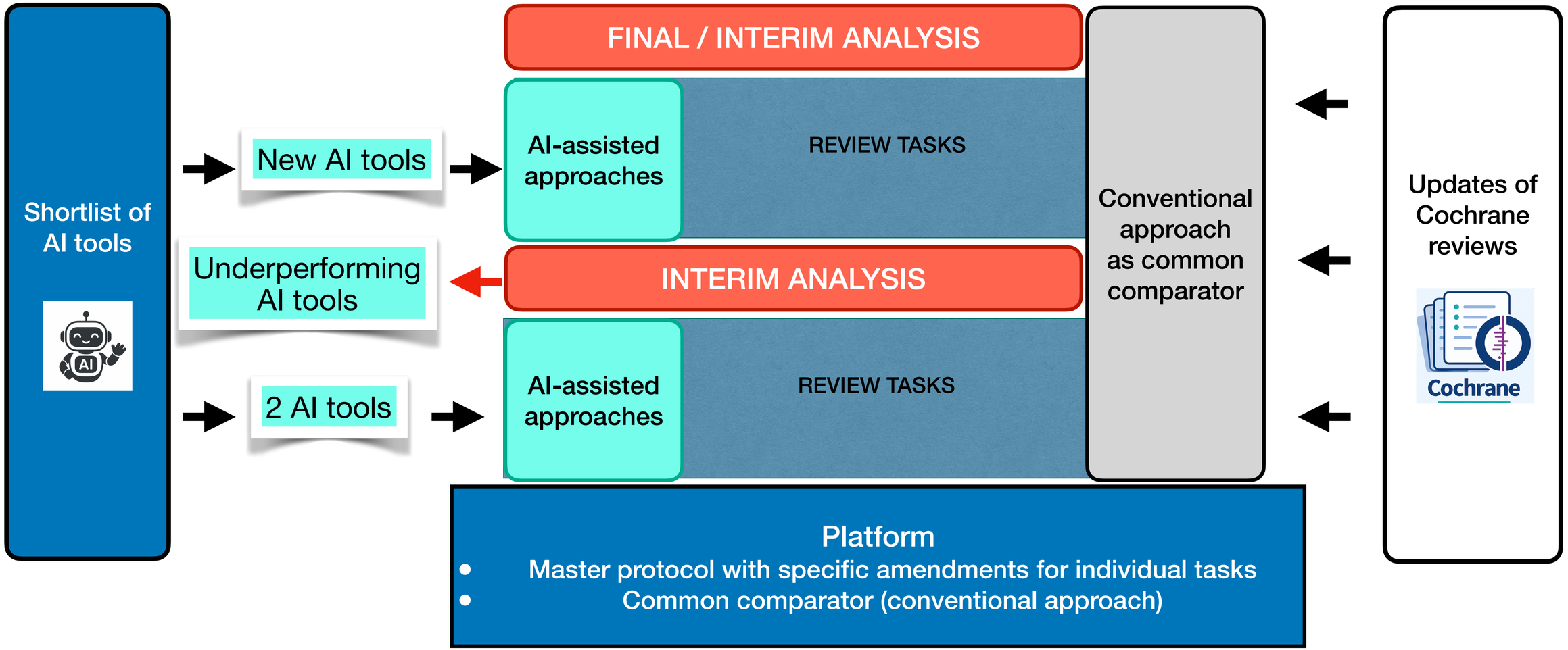
- Use a common master protocol instead of launching separate studies; specific amendments for different interventions
- Use a shared control group to improve efficiency
- Adapt over time as data accumulates and allow for continuous learning
- Drop ineffective treatments and add new promising treatments
- Support rapid evidence-based decisions



Created by DALL-E (2026)

Cochrane Austria

Adaptive Platform Study Within Reviews (adaptive platform SWAR)



Cochrane Austria

Study Characteristics

- **Convenience sample:** 10-15 updates of Cochrane systematic reviews
- **AI tool allocation:** Block randomization
- **Comparison:** Conventional, semi-automated and fully automated completion of review tasks
- **Unit of analysis:** Review tasks
    1. Abstract screening
    2. Full-text screening
    3. Data extraction
- **Study duration:** 6-8 months
- **Statistical analysis:** Exploratory and descriptive

Cochrane Austria

Outcomes

- **Accuracy metrics (vary by task):** Recall, precision, F1 score, proportion of included records
- **Stability metrics:** 10 independent iterations of fully automated tasks within a single day to assess recall
- **Efficiency metrics:** The total time required to complete a review task from initiation to verified output
- **Usability metrics:** System Usability Scale and open-ended questions to assess barriers and facilitators
- **Error analysis:** Proportion of errors/serious errors, downstream impact of errors on meta-analyses and certainty of evidence ratings

Potential Impact of Errors	
Major error	This error substantially compromises data correctness and, if uncorrected, **could lead to erroneous conclusions**
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
Inconsequential difference	This difference most likely would **not impact the interpretation** of data or conclusions

Cochrane Austria

Performance Thresholds

Futility boundary (25th percentile):
Sets a ***minimum bar*** (expectations of 75% of users are higher)***.*** A point estimate below the futility boundary indicates unambiguous underperformance.

Non-inferiority margin (75th percentile):
Sets a ***high bar*** (expectation of 75% of users are lower). If even the most optimistic estimate (upper CL) cannot reach the top quartile of user expectations, the tool lacks sufficient promise to justify continued use.

Performance Metrics	Futility Boundaries (Point Estimate)*	Non-inferiority Margins (95% CI Upper Limit)*	Decision Rule
Screening			
Sensitivity	<80%**	<95%**	Stop if either boundary crossed
Specificity (for full-text screening only)	<50%	<60%	Stop if either boundary crossed
Data Extraction			
Sensitivity	<90%**	<98%**	Stop if either boundary crossed
Major Error Proportion	>3%†	>2.5%†	Stop if either boundary crossed
Usability			
System Usability Scale (score)	<57‡	<75‡	Stop if threshold not met

**Based on a presentation by Flemyng et al., 2025. Understanding expectations for evidence synthesis when using AI: Survey results
** Based on Gartlehner et al. Artificial Intelligence-Assisted Data Extraction With a Large Language Model: A Study Within Reviews Ann Intern Med. 2025 Dec;178(12):1763-1771.
‡ Based on Sauro, Jeff, and James R. Lewis. Quantifying the user experience: Practical statistics for user research. Morgan Kaufmann, 2016.

Cochrane Austria

medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES

Cochrane Evaluation of (Semi-) Automated Review Methods (CESAR): Protocol for an adaptive platform study within reviews

Gerald Gartlehner, Susan Banda, Max Callaghan, Jo-Ana Chase, Andreea Dobrescu, Angelika Eisele-Metzger, Ella Flemyng, Sean Gardner, Ursula Griebler, Bartosz Helfer, Pawel Jemiolo, Biljana Macura, Jan C. Minx, Anna Noel-Storr, Noosheen Rajabzadeh Tahmasebi, Amin Sharifan, Joerg J. Meerpohl, James Thomas

doi: https://doi.org/10.64898/2026.04.13.26350802

Cochrane Austria

Acknowledgements

Cochrane Collaboration, UK: Ella Flemyng, Sean Gardner, Jo-Ana Chase

Cochrane Austria: Andreea Dobrescu, Bartosz Helfer, Ursula Griebler, Amin Sharifan, Susan Banda

Cochrane Germany: Jörg Meerpohl, Angelika Eisele-Metzger, Noosheen Rajabzadeh Tahmasebi

Cochrane Poland: Pawel Jemiolo

Stockholm Environment Institute: Biljana Macura

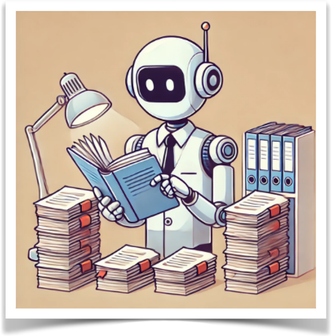
Potsdam Institute for Climate Impact Research: Jan Minx, Max Callaghan

University of Central London: James Thomas

Cochrane Austria

Thank you very much!

gartlehner@cochrane.at



Created by DALL-E (2025)

Cochrane Austria