Understanding expectations for evidence synthesis when using AI: Survey results

July 2025

Ella Flemyng, Head of Editorial Policy & Research Integrity, Co-Convenor AI Methods Group, Responsible AI in Evidence Synthesis (RAISE) management group

Trusted evidence.
Informed decisions.
Better health.

Cochrane

Introduction

- How 'correct' evidence synthesis should aim to be with AI is an area of uncertainty
- Need to understand community expectations to inform benchmarks for accuracy standards

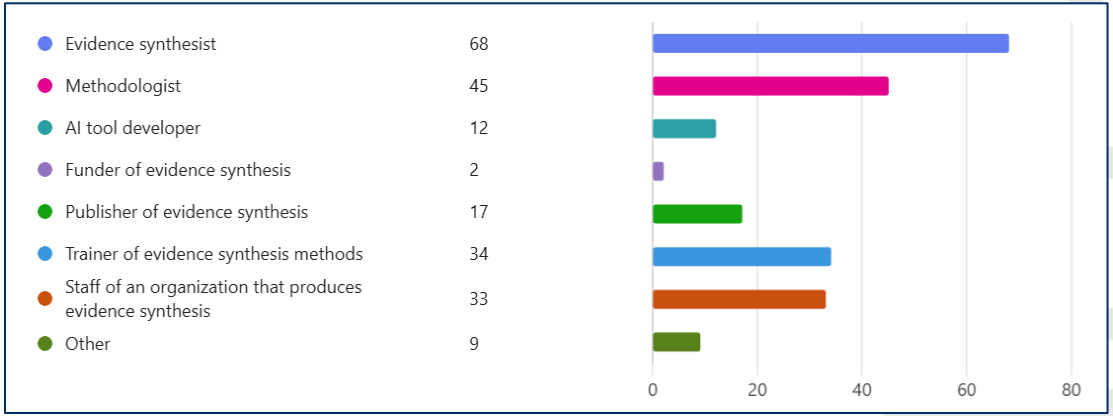
Cochrane

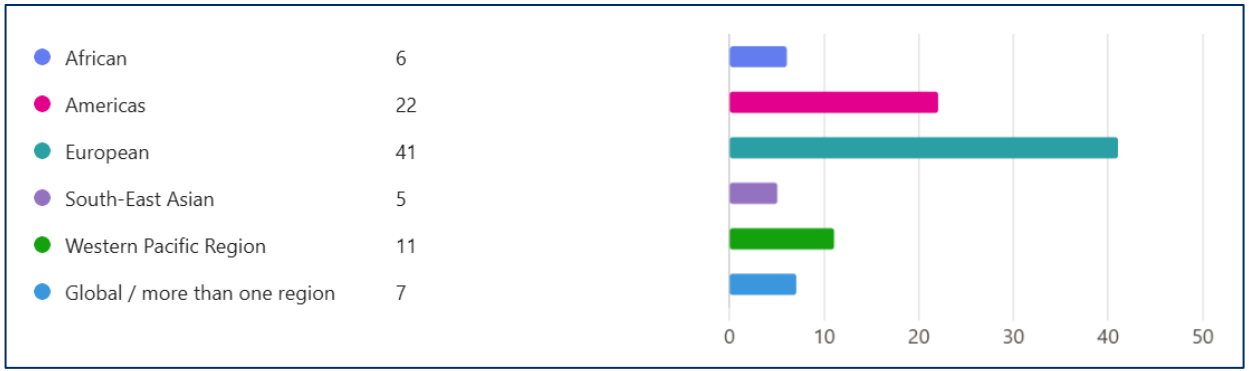
Methodology

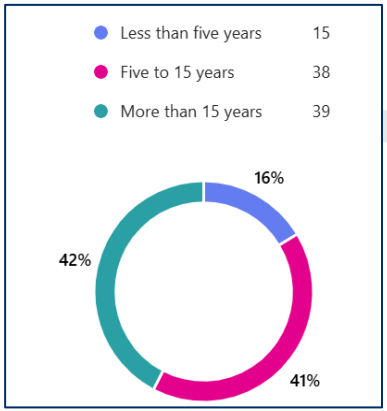
- Anyone with an interest or experience in evidence synthesis and AI
- Open 3 June to 2 July 2025, using Microsoft Forms
- Focused on screening and searching, tagging classifications, data extraction, and risk of bias
- Explored whether responses changed depending on:
  - The confidence in the result
  - The type of evidence synthesis

Demographics

- 92 respondents

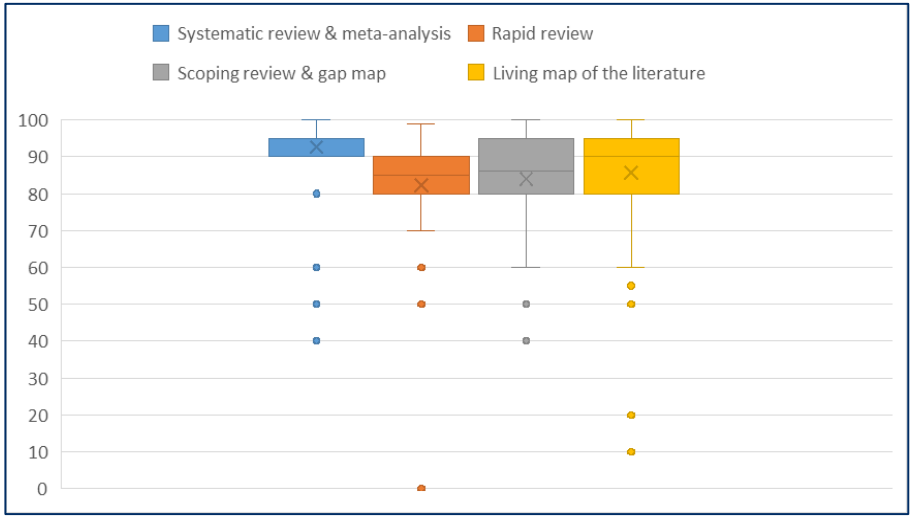






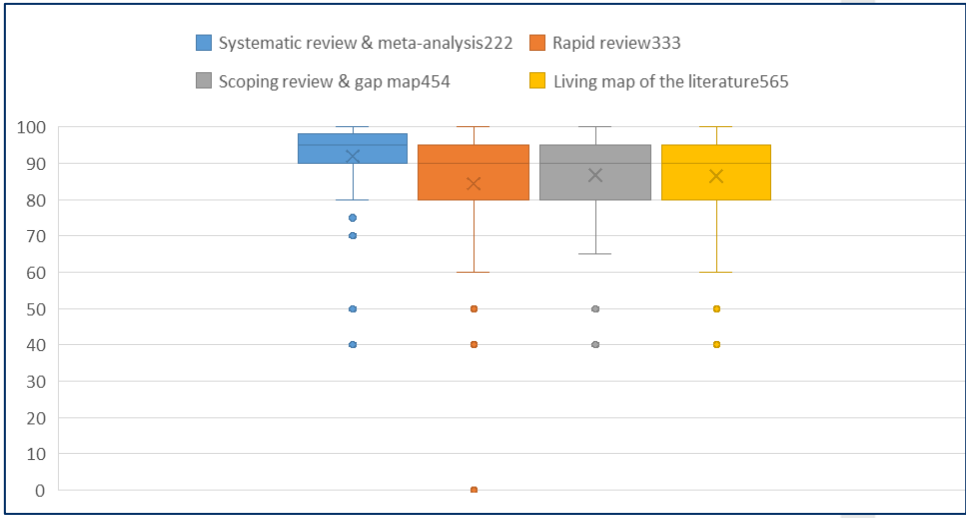
Searching - recall

Human only



95 85 86 90

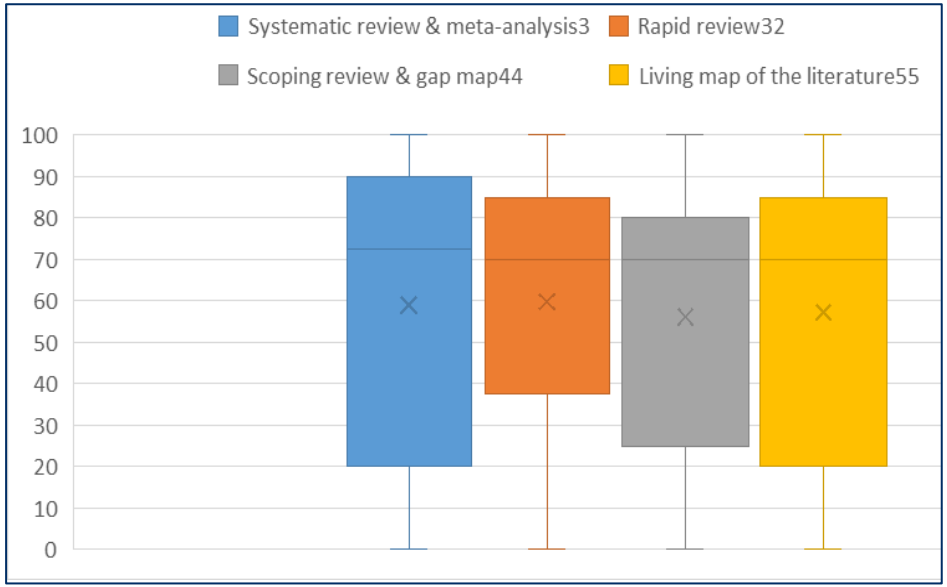
With AI



95 90 90 90

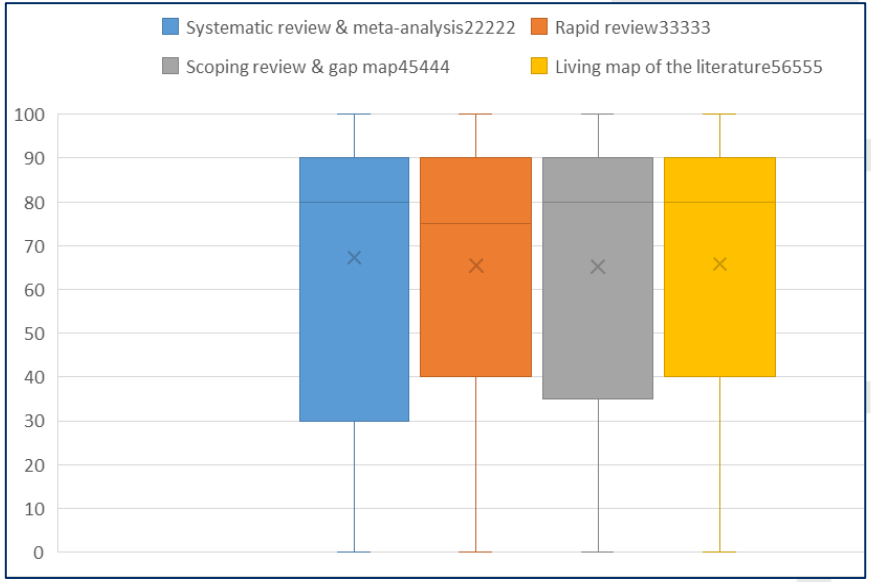
Searching - precision

Human only



72.5 70 70 70

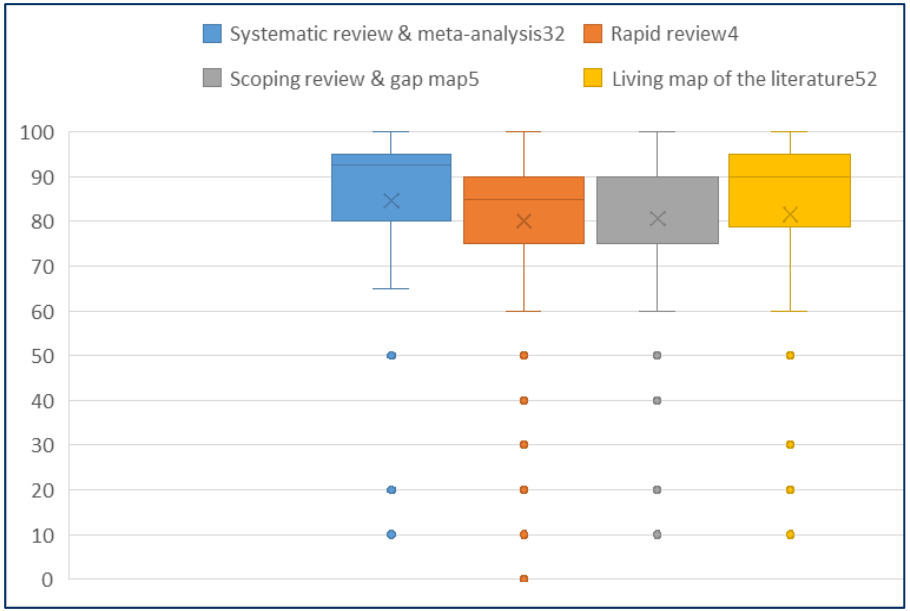
With AI



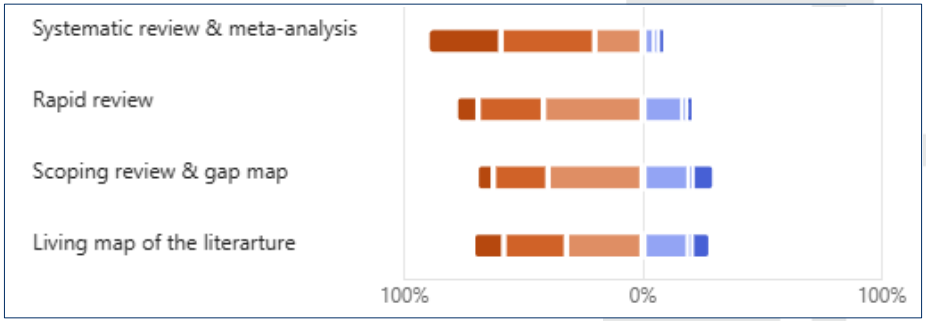
80 75 80 80

Screening - recall

With AI



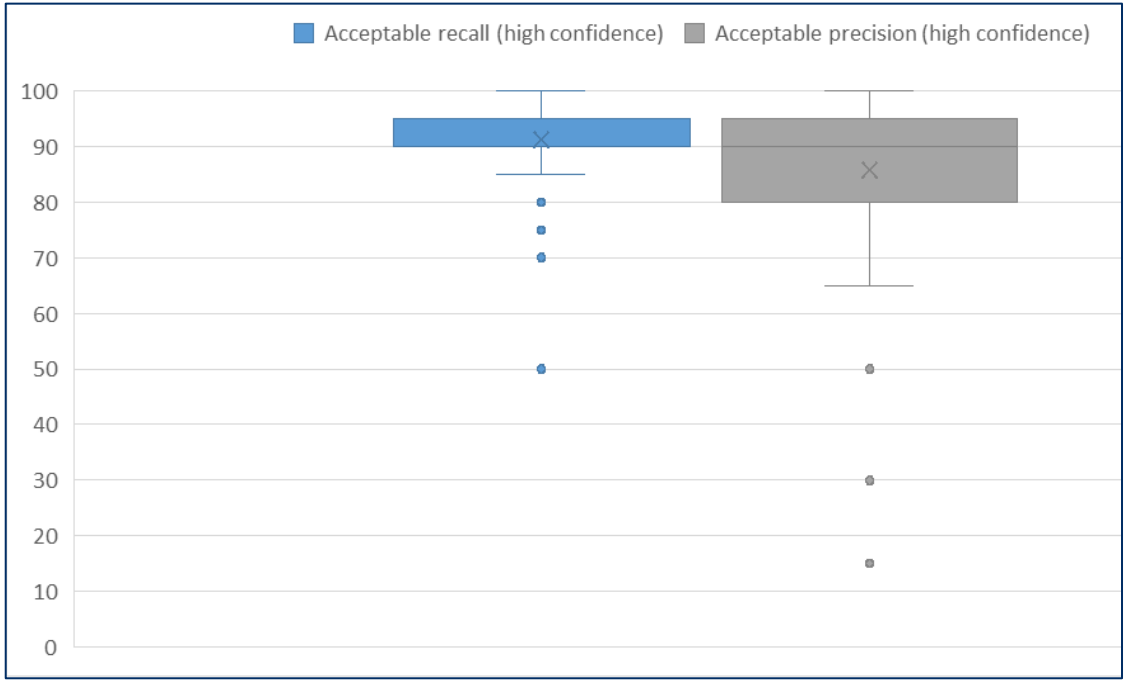
92.5 85 90 90



- Errors are not acceptable
- Some errors made, but no difference in review findings
- Some errors made, and minor differences in results (such as slight changes in risk of bias judgements) but overall, no changes in the findings or conclusions
- Some errors made, and differences in results that impact the findings (such as minor differences in the assessments of the certainty or quality of the evidence included), but no change in conclusions
- Some errors made, and differences in results that impact the findings (such as major differences in the assessments of the certainty or quality of the evidence included), which changes the conclusions
- Don't know / don't feel qualified to answer

Tagging studies with classifications

With AI

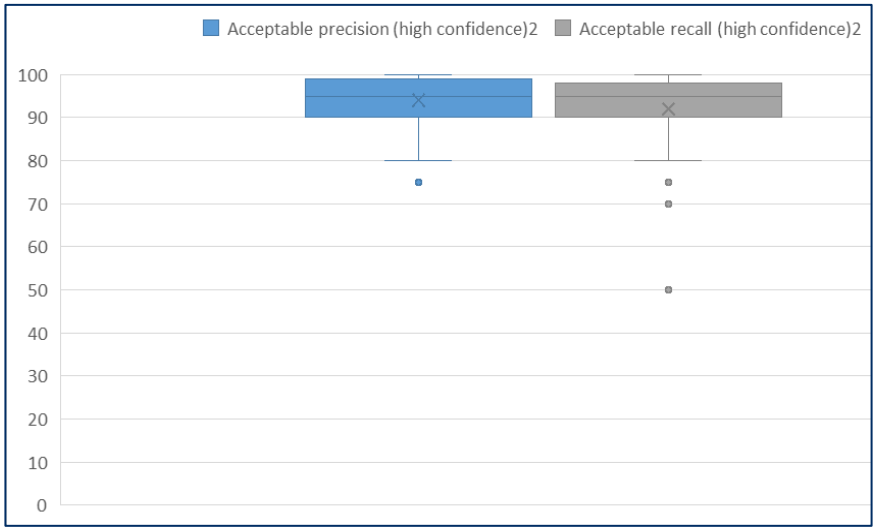


95 90

Cochrane

Data extraction

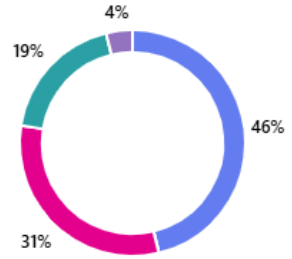
With AI

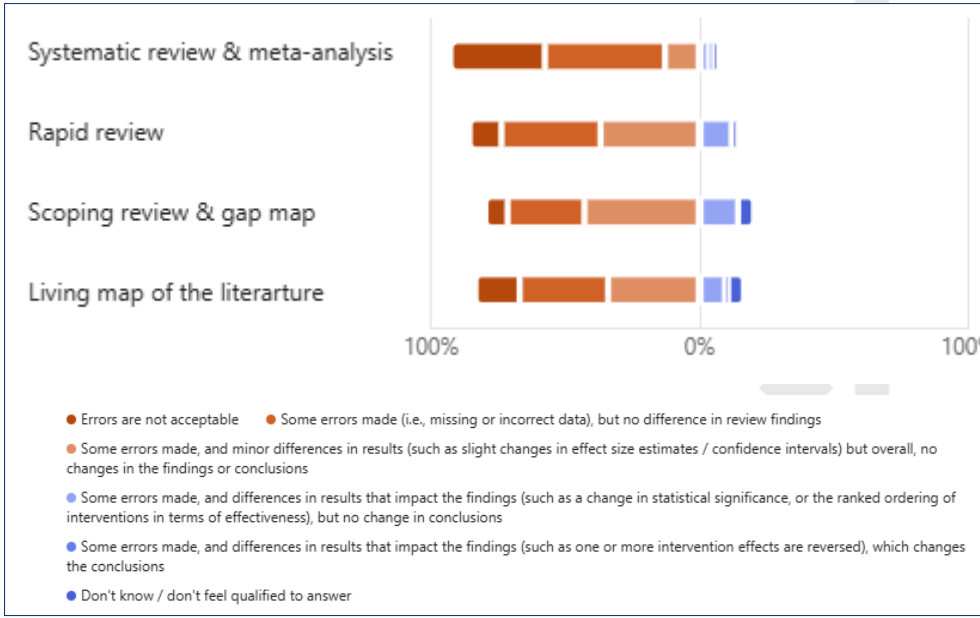


95 95

Would your responses change if it was for another type of evidence synthesis, e.g., rapid or scoping review?

Yes	37
No	25
Maybe	15
Don't know / don't feel qualified to answer	3

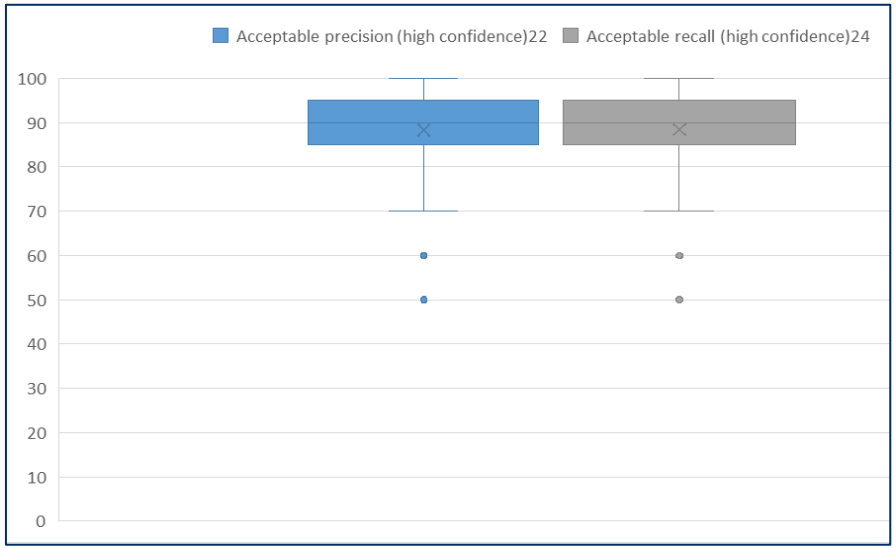




Cochrane

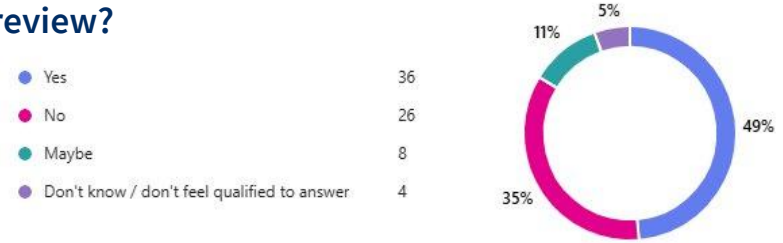
Risk of bias

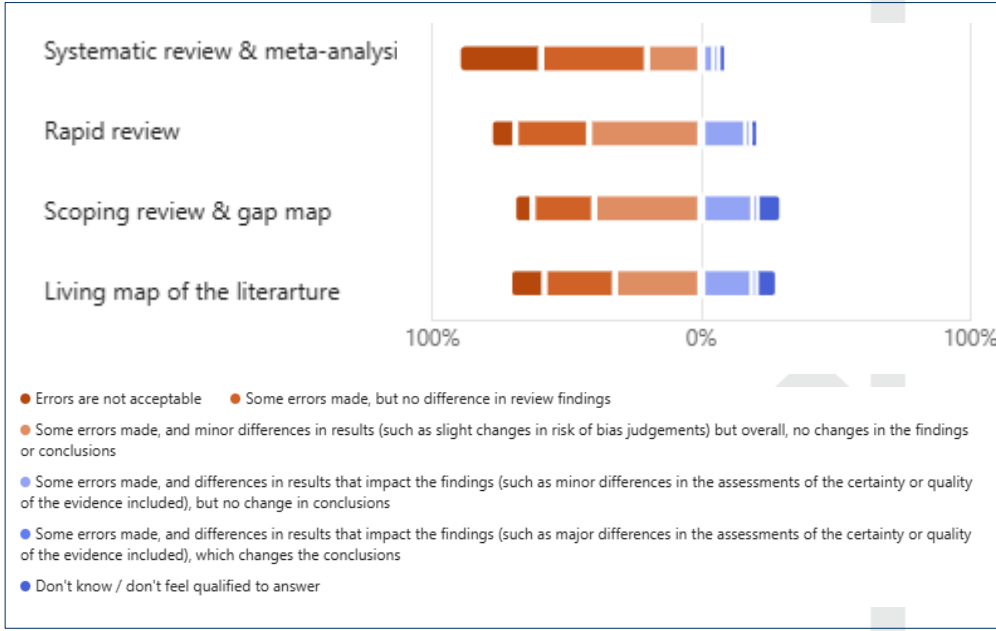
With AI



90 90

Would your responses change if it was for another type of evidence synthesis, e.g., rapid or scoping review?





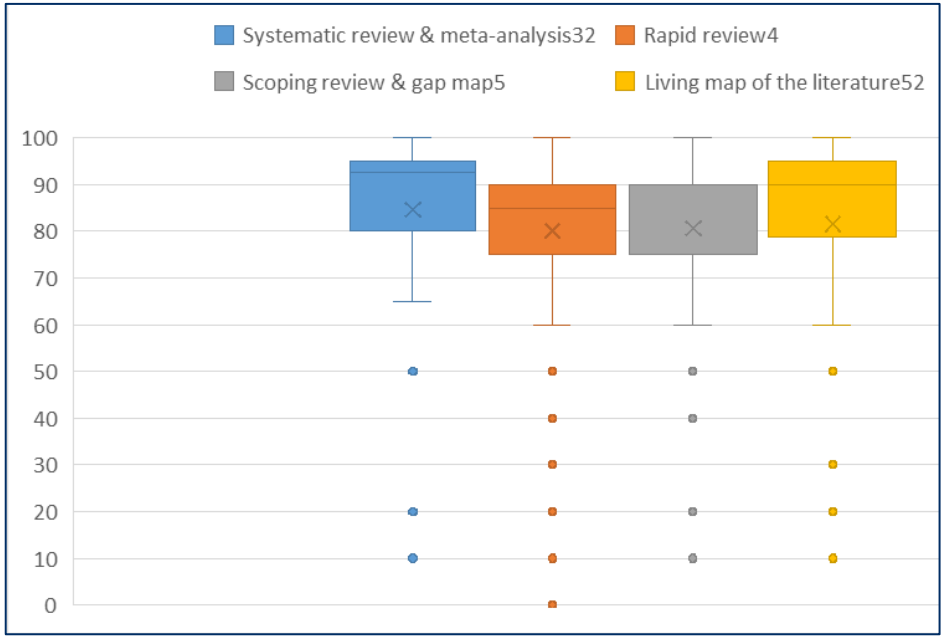
Cochrane

Reflections

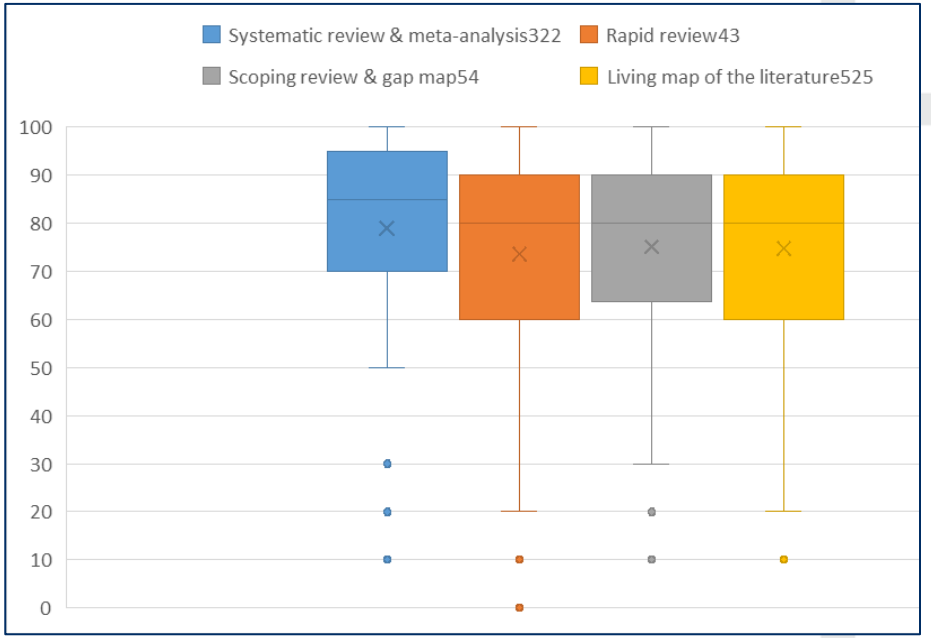
1. For screening and searching, accuracy is expected to be higher in systematic reviews compared to other review types (less clear for other stages of the review)
2. Support for the use of AI as long as there are no errors that impact the review's findings or conclusions
3. Some differences between expectations and what's feasible at this point, i.e., classification tagging
4. Some confusion about what the questions were asking, e.g., what precision and recall meant in context, the high/low confidence dichotomy

Example of the uncertainty issue

With high confidence in the result



With low confidence in the result



Cochrane

What next?

1. Deep dive into the feedback in the Potsdam hackathon, including:
   - Where does there seem to be consensus? Where are the gaps?
   - How can we define acceptable trade-offs ?
2. Define next steps for defining accuracy benchmarks:
   - Formal recommendations
   - Areas for further work