

# How Much Time Does Artificial Intelligence Really Save in Evidence Synthesis? A Systematic Literature Review

Bobrowska A<sup>1</sup>, Lunn L<sup>2</sup>, Chan K<sup>3</sup>, Murton M<sup>1</sup>

<sup>1</sup>Costello Medical, Cambridge, UK; <sup>2</sup>Costello Medical, Manchester, UK; <sup>3</sup>Costello Medical, London, UK



## Objective

To understand the extent to which Artificial Intelligence (AI) can save time and workload in the conduct of literature reviews (LRs).

## Background

- Traditional LRs are time- and resource-intensive, process-driven projects.
- The emergence of generative AI models has sparked considerable interest in whether they could expedite the conduct of LRs.
- At the same time, concerns about hallucinations and quality (critical in LRs) mean evidence synthesists approach AI with caution.
- A human-in-the-loop is required to verify the outputs of AI models, which raises questions about whether meaningful time savings can actually be achieved.
- To explore this, we conducted a systematic LR to assess whether AI can truly save time in the review process while maintaining sufficient levels of quality.

## Methods

- MEDLINE and Embase were searched in June 2025. Records were reviewed at title and abstract by two experienced reviewers and at full text by a single reviewer. We included primary research studies that reported time or workload saved from applying AI to a specific aspect of a LR, compared with humans. LRs were hand-searched and excluded.
- Data was extracted and synthesised qualitatively due to heterogeneity in outcome reporting. Where possible, hours saved per study were calculated; if ranges were reported, midpoints were used. Authors' conclusions were subjectively categorised as "positive", "cautiously positive" or "neutral/negative" regarding the AI-generated efficiencies in LRs.

## Results

- Searches produced 2,091 unique hits; 2,011 records were removed after title/abstract review. Ultimately, 56 studies were included (Figure 1). Studies used proprietary tools (19 tools in 29 studies), widely available general AI tools like ChatGPT (12 tools in 16 studies) or a trained, bespoke algorithm (n=11 studies) (Figure 2).

## Conclusion

Most benefits of AI are currently observed at the screening stage of a LR, with far fewer demonstrated at data extraction or quality assessment stages. However, comparisons across studies are hampered by the lack of a unified outcome measure to assess AI performance, both in terms of precision and efficiencies gained. There is also a risk that AI benefits could be inflated by assuming an unrealistically long time taken for tasks by humans.

- Most time savings were reported for study selection at title/abstract stage (n=45 studies), with fewer studies reporting time saved on quality assessments (n=6), extractions (n=2) or deduplication, feasibility assessment, or search strategy generation (n=1 each). Authors were more often positive (n=27) or cautiously positive (n=17) than negative (n=12) about the potential of AI to help conduct LRs (Figure 3).
- The median workload saved was 65% (n=27 data points) (Figure 4).
- The median time saved was 60% (n=8 data points, data not shown) and median time saved per study was 1.02 minutes (n=30 data points) (Figure 5).
- Assumptions on time taken by a human reviewer appear unrealistic in some studies, creating outliers. E.g. one study reported saving 15.5 hours per a risk of bias assessment,<sup>1</sup> when the average for the RoB 2.0 tool is no more than 1–2 hours by a reviewer with 1–2 years' experience (internal experience).

FIGURE 1

Prisma flow diagram

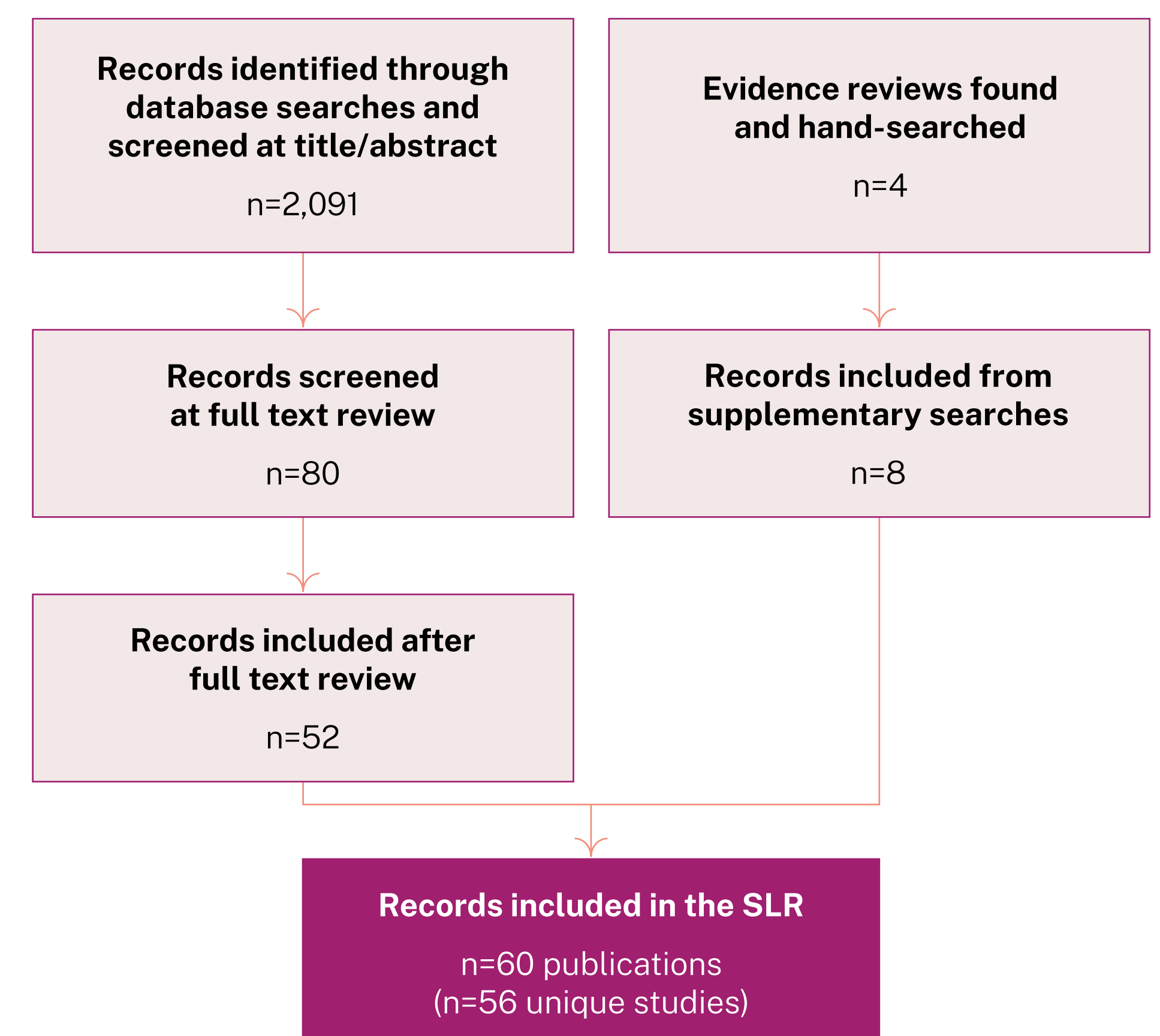
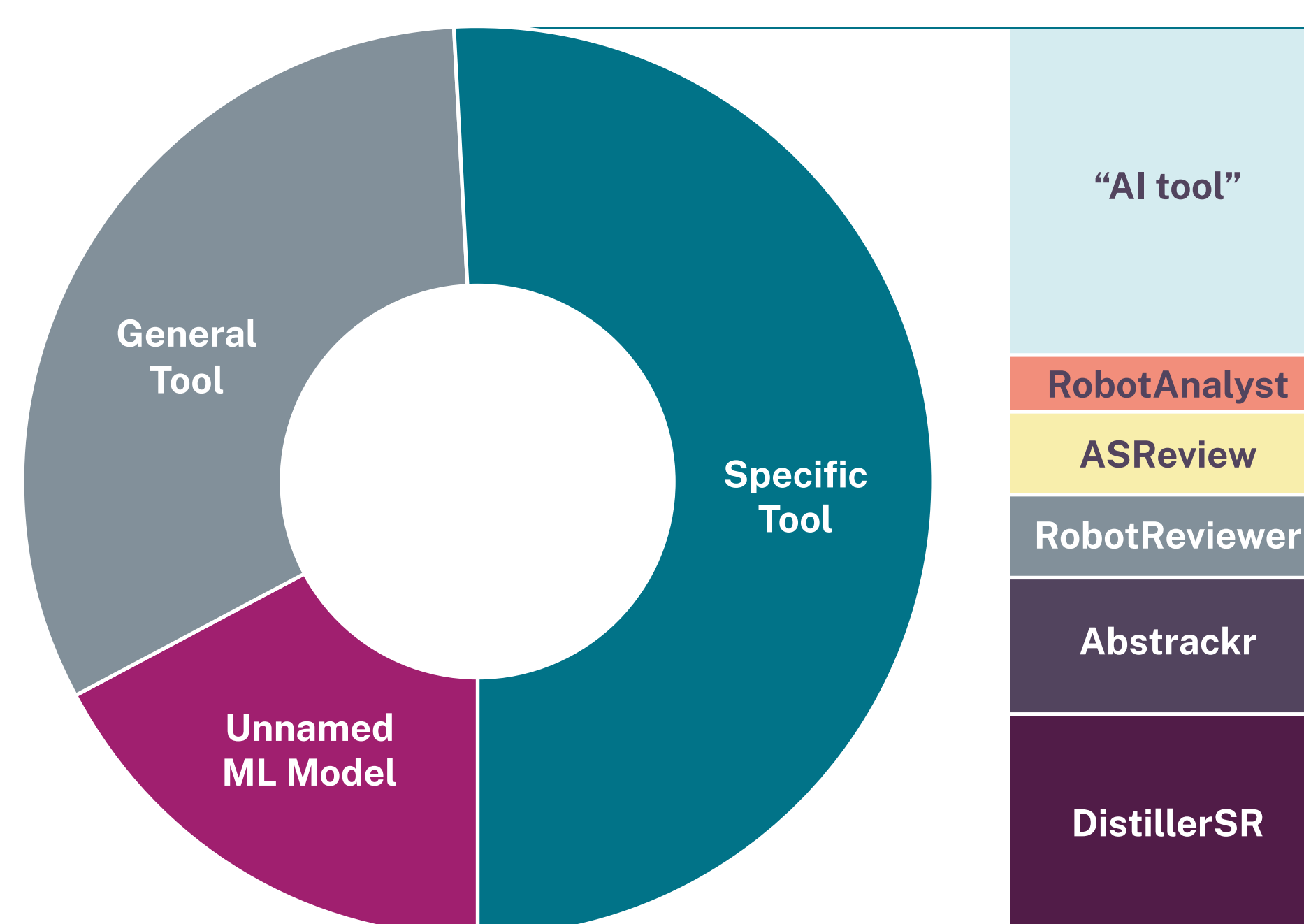


FIGURE 2

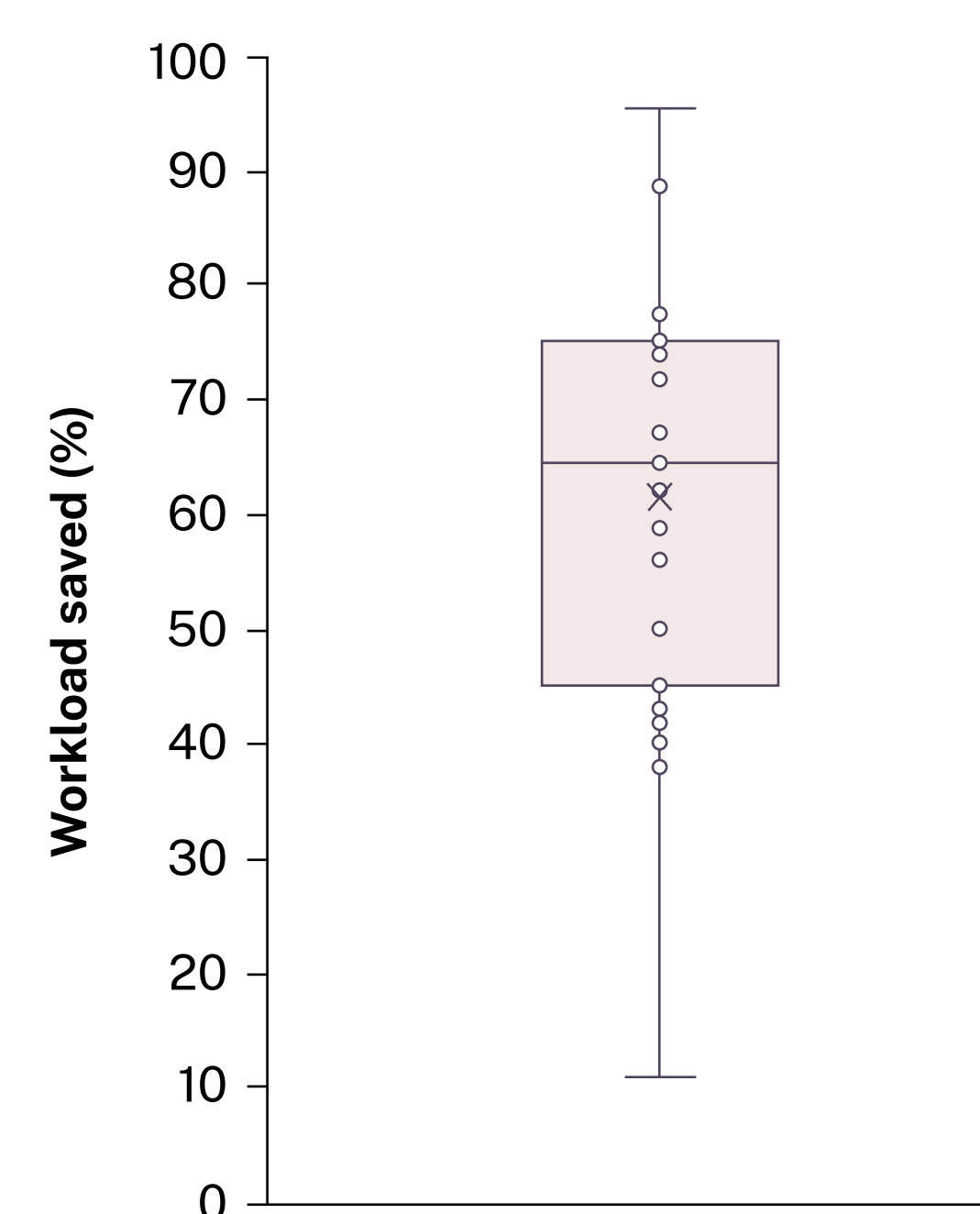
Tools used



AI tool category contains tools with only 1 study reporting each.

FIGURE 4

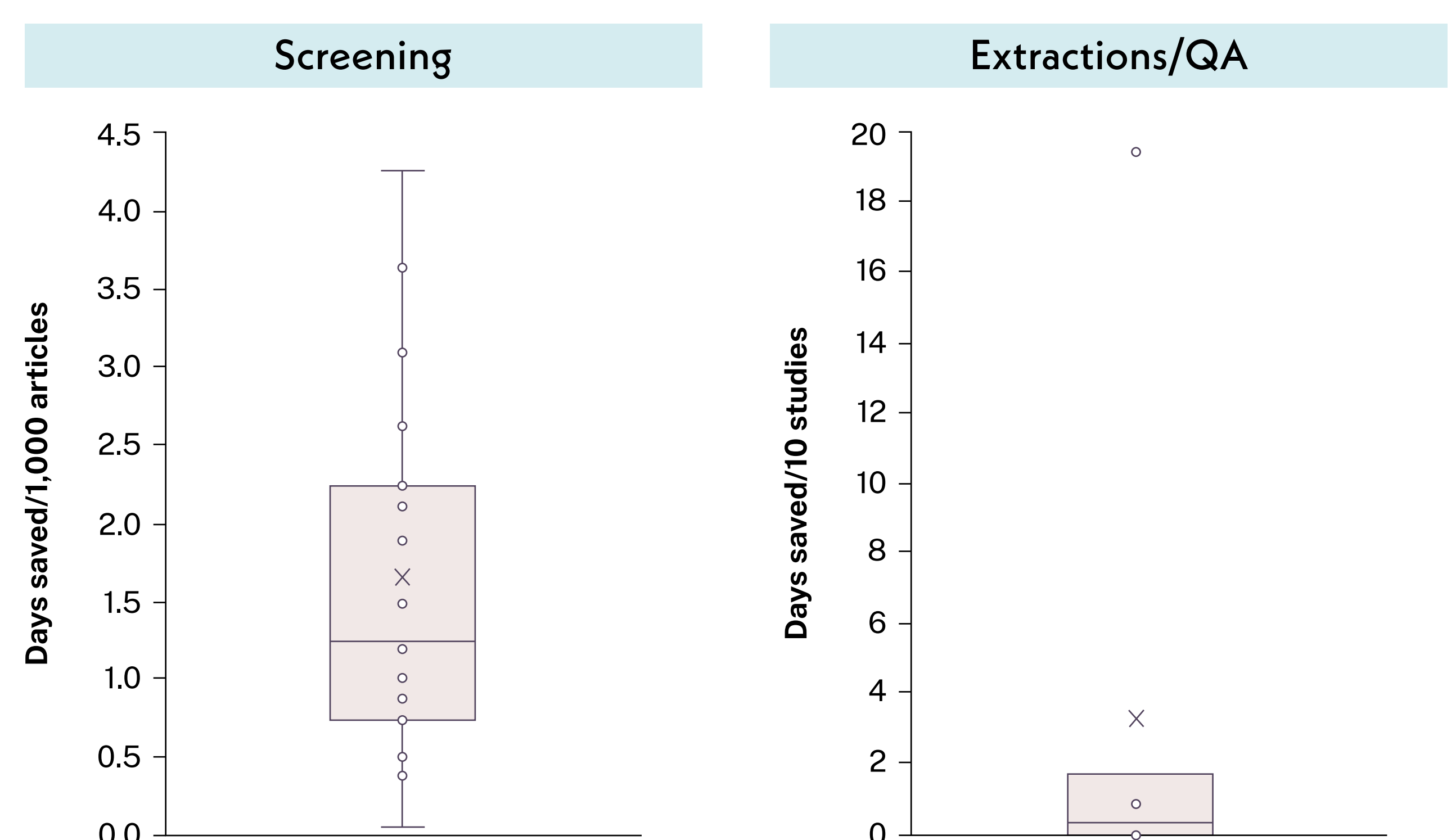
Percent workload saved



The mean is shown by the cross and the median by the horizontal line.

FIGURE 5

Time saved



Left panel: time saved with AI in screening stages of a review, presented as days per 1,000 articles screened. Right panel: time saved with AI in QA or extraction stages of a review, presented as days saved per 10 studies. The mean is shown by the cross and the median by the horizontal line.

Abbreviations: AI: artificial intelligence; FA: feasibility assessment; LR: literature review; PRISMA: Preferred Reporting in Systematic reviews and Meta-Analyses; QA: quality assessment; RoB 2.0: risk of bias version 2.0; SLR: systematic literature review.

References: Aggarwal S, Kumar S, Topaloglu O. MSR174 Leveraging Chat-GPT for Conducting Systematic Literature Reviews. Value in Health. 2024 Dec 1;27(12):S472. Acknowledgements: The authors thank Henny Zhan, Costello Medical, for graphic design assistance.