Automating Systematic Literature Review (SLR) Updates: A Comparative Validation Study of Artificial Intelligence (AI) versus Human Screeners

Allie Cichewicz, Apurva Pande, Katarzyna Borkowska, Lalith Mittal, Priscilla Wittkopf, Mahmoud Slim

Evidera Inc.

Poster No. MSR22

Background

- Systematic reviewers are becoming increasingly inundated by the growing body of literature.
- The time and resource-intensive nature of conducting systematic literature reviews (SLRs) often results in searches being outdated by the time the reviews are completed.
- Market access strategies rely heavily on the most up-to-date evidence, which is crucial for making informed decisions regarding product development, regulatory approvals, and healthcare policies.
- This growing need for current evidence has sparked an interest in the concept of living SLRs and artificial intelligence (AI) integration to expedite the review process and more effectively inform healthcare decisions.

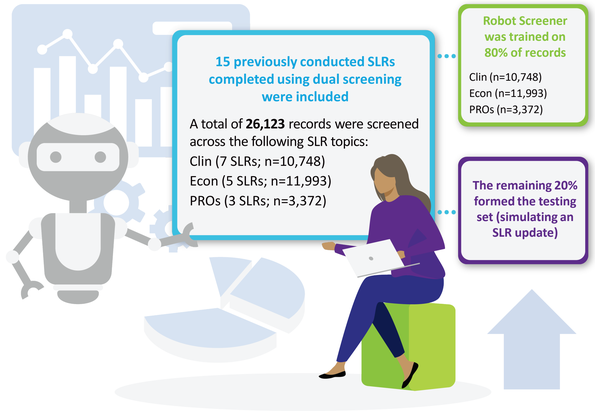
Objectives

- We aimed to validate the performance of an AI algorithm against human reviewers in predicting screening decisions for new records identified by refreshing the searches for previously completed SLRs on health economics and outcomes research (HEOR) topics used to inform reimbursement decisions by health technology assessment (HTA) bodies.

Methods

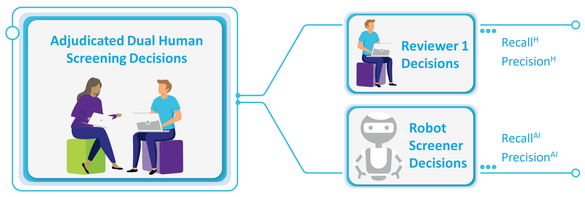
- Previously completed reviews on the following topics were used for this study: seven clinical efficacy and safety (Clin) SLRs; five economic burden and evaluation (Econ) SLRs; and three humanistic burden and utilities (PRO) SLRs.
- These SLRs encompassed a multitude of therapeutic areas including infectious diseases (COVID), neurodegenerative disorders (Alzheimer's disease and Friedreich's ataxia), oncology (ovarian cancer), autoimmune conditions (idiopathic thrombocytopenia, atopic dermatitis, relapsing-remitting multiple sclerosis, and urticaria), and metabolic disorders (diabetes).
- Robot Screener was trained with a subset of 80% of records from each SLR (training set). The remaining 20% of records, simulating an SLR update, formed the testing set (**Figure 1**).

Figure 1. Summary of Included SLR Topics and Training/Testing Approach



- Title/abstract screening decisions (inclusion or exclusion) made by Robot Screener on the testing set were compared against the final adjudicated screening decision following dual human review from the previously completed SLR to compute the recall, precision, false positives, and false negatives between humans and AI.
- To establish a direct benchmark for decisions made by AI, the same measures were then computed for human reviewer 1 from the previously completed SLR (prior to adjudication) vs. the final adjudicated screening decision (**Figure 2**). Differences in the mean recall and precision, as well as underlying false positives and false negatives, between Robot Screener and human reviewer 1 screeners were assessed using Mann-Whitney U-test.

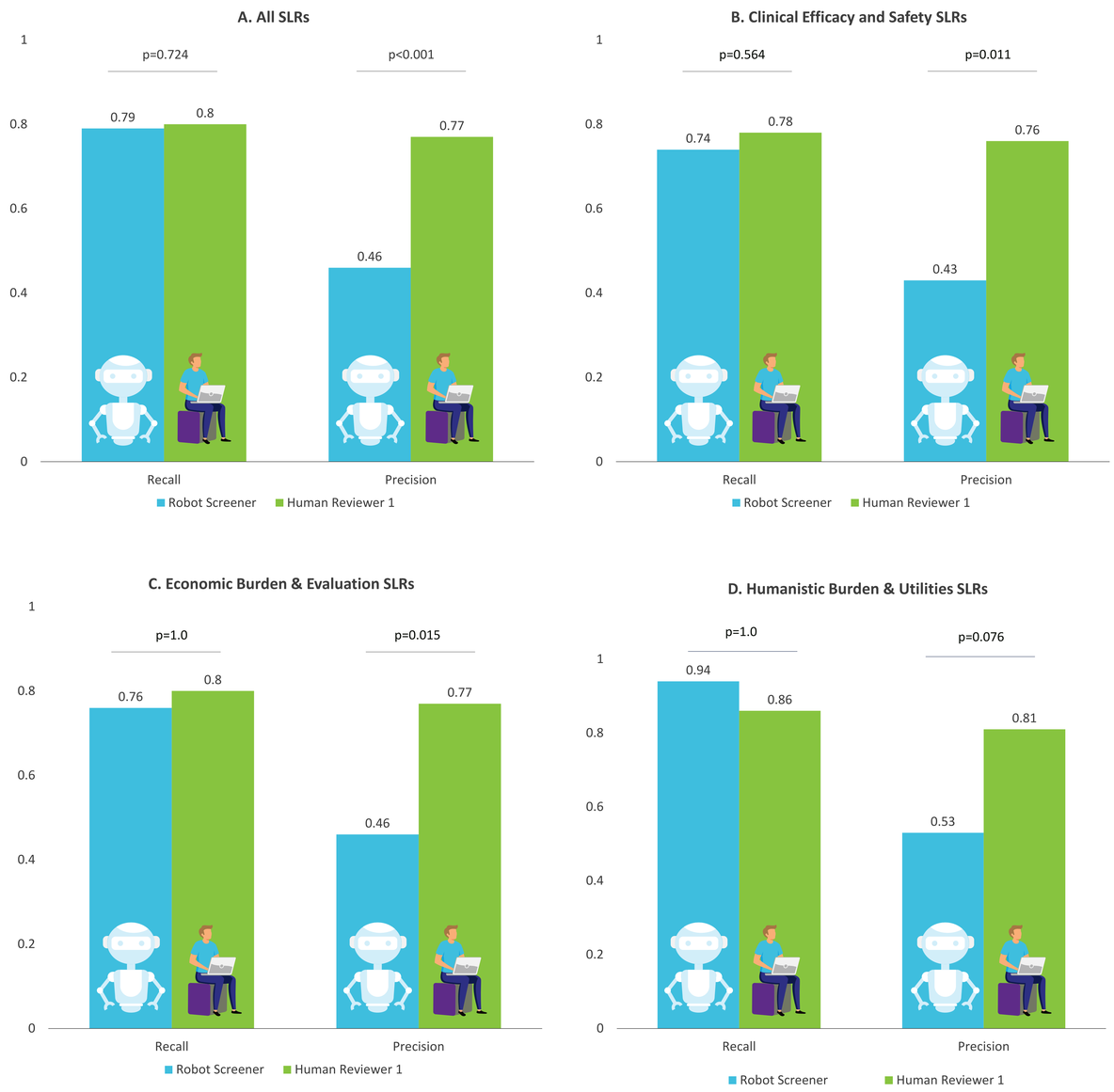
Figure 2. AI Validation Approach



Results

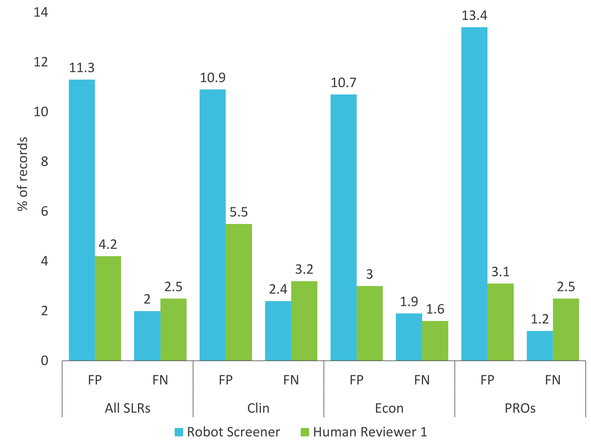
- The mean ± SD of recall rates for Robot Screener across all SLRs was 0.79 ± 0.18 compared to 0.80 ± 0.20 with Human Reviewer 1 (p=0.724; **Figure 3a**).
  - Across the different SLR topics, recall rates ranged from 0.74 to 0.94 for Robot Screener and from 0.78 to 0.86 for Human Reviewer 1, with the highest recall rates achieved for the PRO SLRs (0.94 ± 0.08 vs. 0.86 ± 0.24, respectively).
  - Differences between reviewers were not statistically significant in any of the SLR topics (**Figures 3b-3d**).
- The mean ± SD of precision rates for Robot Screener across all SLRs was 0.46 ± 0.13 compared with 0.77 ± 0.19 with Human Reviewer 1 (p<0.001; **Figure 3a**).
  - Across the different SLR topics, precision rates ranged from 0.43 to 0.53 for Robot Screener and from 0.76 to 0.81 for Human Reviewer 1, with statistically significant differences observed between reviewers for each SLR topic (**Figures 3b-3d**).
- Overall, the false negatives between Robot Screener and Human Reviewer 1 were comparable (2% vs. 2.5%; **Figure 4**). However, Robot Screener registered higher false positives.

Figure 3 Comparative Performance Metrics Between Robot Screener and Human Reviewer 1



Performance metrics are reported as a mean and SD of recall and precision across the SLRs. Recall and precision were computed separately for Robot Screener vs. the adjudicated final decision following dual human review and Human Reviewer 1 versus the final adjudicated decision. Values range from 0 to 1, with higher values indicating better recall or precision. Statistical hypothesis testing was carried out to assess the performance of Robot Screener vs. Human Reviewer 1, with a p<0.05 indicating statistically significant differences.

Figure 4. False Positive and False Negative Rates for Robot Screener and Human Reviewer 1



False negative (FN) represents potentially relevant records that were excluded whereas false positive (FP) represents irrelevant records that were included.

Discussion

- Robot Screener exhibited comparable recall rates to Human Reviewer 1, suggesting that the AI algorithm effectively captured potentially relevant records at similar rate to humans (prior to adjudication). Alongside the low probability (2%) of erroneously excluding relevant records at the title/abstract stage, these findings demonstrate the validity of Robot Screener as an alternative to a second human reviewer.
- Although low precision suggests that Robot Screener is more conservative, resulting in advancing irrelevant records to full-text review, this will not compromise the overall quality of the SLR. Prioritizing recall over precision during model fine-tuning is intended to mitigate the risk of missing potentially relevant records, despite additional human effort needed for adjudication.
  - Lower precision with Robot Screener may stem from the nuances of complex eligibility criteria, such as stringent treatment settings or restrictive disease staging requirements.[1] Human reviewers skilled in systematic review also have the ability to make screening assumptions beyond the title and abstract text.
- Despite the need for additional screening efforts to adjudicate the conflicts caused by the false positives, the potential time-saving benefits and rapid access to up-to-date evidence offered by employing Robot Screener as a second screener outweigh these potential challenges. The consistent performance of Robot Screener compared with Human Reviewer 1 across the different HEOR topics also highlights the robustness and generalizability of using Robot Screener across various SLRs.

Conclusions

- The promising findings of our study pave the way for the semi-automation of SLR updates, leveraging a "human-in-the-loop" approach. By demonstrating the validity of the Robot Screener in capturing potentially relevant records against human reviewers, our study suggests that AI-assisted title and abstract screening, when integrated with human expertise, can play a pivotal role in expediting the process of updating SLRs.
- While our study focused on validating the performance of Robot Screener with a robust training set simulating an SLR update, future research studies should explore its validity vs. human reviewers in performing de novo SLRs to obtain an optimal training set threshold.

References

1. Cichewicz A., Slim M., Deshpande S. MSR163 Artificial Intelligence (AI)-Based Screening: Exploration of Differences in Two Health Technology Assessment (HTA)-Compliant Systematic Literature Reviews (SLRS). Value Health 2023;26(12 Supplement):S425.

Disclosures/Acknowledgments

Editorial support and graphic design were provided by Michael Grossi and Richard Leason of Evidera, a business unit of PPD, part of Thermo Fisher Scientific.

Funding provided by Evidera Inc., a business unit of PPD, part of Thermo Fisher Scientific