Utility of Artificial Intelligence in Systematic Literature Reviews for Health Technology Assessment Submissions

Allie Cichewicz[1], Heather Burnett[2], Rachel Huelin[1], Ananth Kadambi[3]

[1]Evidera, Waltham, MA United States; [2]Evidera, Montreal, QC, Canada; [3]Evidera, San Francisco, CA

#116574

Background

- With the continually growing body of published literature, there is an increasing burden on reviewers to screen larger volumes of references within submission timelines, and to meet Health Technology Assessment (HTA) requirements for robustness and recency.
- The use of artificial intelligence (AI) software in health economics and outcomes research has been explored, but its utility in aiding screening of systematic literature reviews (SLRs) on the humanistic and economic burden of disease has neither been well established nor adopted by HTA bodies.

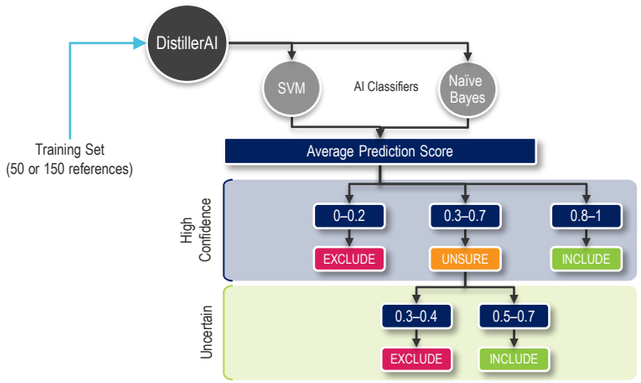
Objective

- Given the uncertainty of accuracy and precision with AI and inherent challenges with its implementation and integration into rigorously established frameworks for SLRs, this study aimed to assess the ability of AI to accurately identify evidence required for HTA submissions including costs, healthcare resource use (HCRU), economic evaluations (EE), and patient-reported outcomes (PRO).

Methods

- Two SLRs were conducted in attention-deficit/hyperactivity disorder covering a 10-year time frame (2008–2018). One search was conducted on economic studies (cost/HCRU outcomes and economic evaluations) and a second on PROs. Two independent reviewers screened all records, with disagreements resolved by a third reviewer.
- DistillerAI was employed to replicate these SLRs with AI as the second reviewer **(Figure 1)**.

Figure 1. Flow diagram of the AI reviewer training and decision process



SVM = support vector machine

Methods (Cont'd)

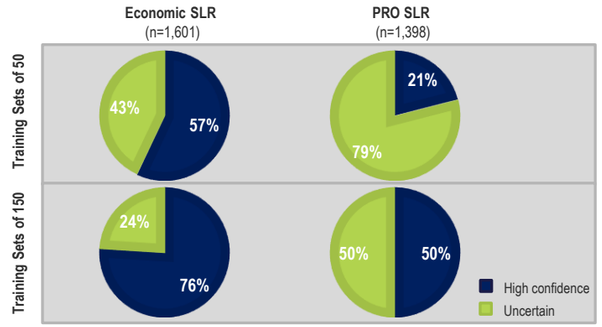
- Sets of 50 references (2–3% of the total search yields) and 150 references (9–10% of the total search yields) from the title/abstract screening decisions of each SLR were used to train the AI reviewer. The training sets of 50 included the same references that were completed as a calibration exercise by the human reviewers prior to beginning the screening phase. The training sets of 150 included these same references along with a random sample of an additional 100 references.
- Prediction scores ranged from 0–1, with 0 being the highest confidence for excludes and 1 being the highest confidence for includes. For the purposes of this study, the following scores were set as AI reviewer thresholds: ≤ 0.2 for excludes and ≥ 0.8 for includes.
- For references falling within the prediction score range of 0.3–0.7, the AI reviewer was unsure whether to include or exclude (i.e., the two AI classifiers did not agree) and deferred to a human reviewer to adjudicate. For these references, a prediction score of < 0.5 was considered an exclude and a score ≥ 0.5 was considered an include.
- Screening decisions for 3,201 references were compared between AI and human reviewers to determine AI accuracy and error rates (i.e., overly inclusive or overly exclusive). Both the final human decision and the AI reviewer decision for each reference were used to calculate inter-rater reliability (IRR) based on Cohen's kappa statistics
- The studies ultimately included in the SLRs were also used to investigate rates of erroneous excludes at the title/abstract level by the AI reviewer.

Results

How Training Sets Influence the Results

- In comparison to the training set of 50 references, the larger training set of 150 references resulted in more references being screened by the AI reviewer within the high confidence range of prediction scores, increasing by 19% for the economic SLR and 29% for the PRO SLR **(Figure 2)** .

Figure 2. References screened by AI reviewer based on prediction score and size of the training set

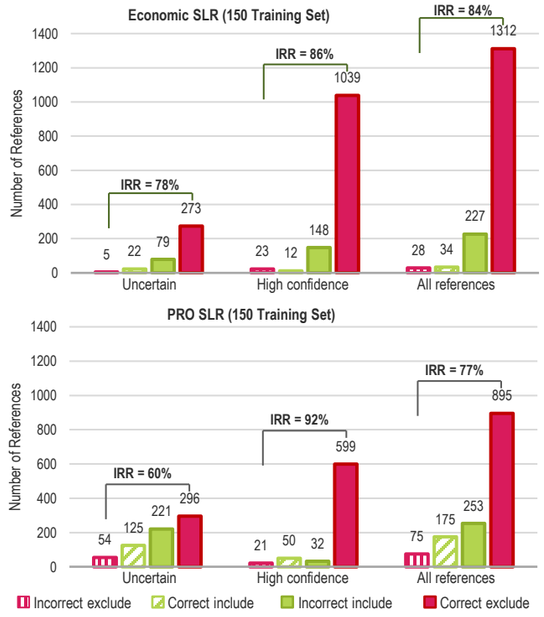


SLR = Systematic literature review

Results (Cont'd)

- Overall, agreement rates between AI and human reviewers based on IRR ranged from 77–84% for all references across SLRs and training sets. When restricting to references screened according to the high confidence prediction score threshold, this range improved to 82–92%.
  - Increasing the training sets from 50 to 150 references improved the IRR across all references from 77 to 84% for the economic SLR (costs/HCRU and EE), but decreased IRR from 79 to 77% for the PRO SLR.
  - Among the larger training sets, IRR was higher overall for the economic SLR demonstrating the AI reviewer was better at identifying relevant economic studies. Agreement levels were best (92%) when the AI reviewer made decisions within the high confidence prediction threshold. However, the large variation in IRR between high confidence and uncertain thresholds (92% vs 60%, respectively) for the PRO SLR suggests that the training set may have been inadequate. **(Figure 3)**

Figure 3. Inter-rater reliability and accuracy of AI reviewer decisions in comparison with human screeners across all prediction score thresholds



References screened with high confidence fell within the prediction score range of 0–0.2 (excludes) and 0.8–1 (includes) and those categorized as uncertain fell within 0.3–0.7. IRR = interrater reliability

The Impact of AI Decisions on Study Attrition

- AI reviewer decisions erred on the side of being overly inclusive (87–94% of incorrect decisions were includes) for the economic SLR, whereas decisions were more overly exclusive for the PRO SLR (20–40% of incorrect decisions were excludes) despite higher IRR % with the larger training set, suggesting that the AI reviewer was less accurate when reviewing PRO studies (**Table 1**).
- While title/abstract screening that is overly inclusive is less detrimental to the SLR process – relevant studies are less likely to be missed – this still adds to the human screening burden at the full-text level of review.

Comparison of Error Rates Between Human and AI Reviewers

- Several exclusions made by the AI reviewer were references that were ultimately included in the SLRs by human reviewers. These exclusion errors were identified within the references screened by AI that fell within the uncertain threshold of prediction scores and occurred most often with EE (85.7%) followed by costs/HCRU (58.3%), and PRO (46.8%).
- By contrast, none of the human screening errors impacted the overall set of included references once the SLR was completed.

Table 1. Disagreements between human and AI reviewers: Distribution of incorrect AI decisions

	Economic SLR (n=255)		PRO SLR (n=328)	
	Incorrect include	Incorrect exclude	Incorrect include	Incorrect exclude
High confidence	148/171 (86.5%)	23/171 (13.5%)	32/53 (60.4%)	21/53 (39.6%)
Uncertain	79/84 (94%.0)	5/84 (6.0%)	221/275 (80.4%)	54/275 (19.6%)
All references	227/255 (89.0%)	28/255 (11.0%)	253/328 (77.1%)	75/328 (22.9%)

Data are presented for results when DistillerAI was trained with 150 references. Includes or excludes were considered incorrect when the AI reviewer decision differed from the human decision. Percentages are calculated out of the number of total disagreements between reviewers (n). High confidence: prediction score 0–0.2 or 0.8–1. Uncertain: prediction score 0.3–0.7.

Conclusions

- Larger training sets impacted the ability of AI reviewer to accurately identify EE and cost/HCRU studies, and to a lesser extent PRO studies.
  - Given the wide range of available tools and instruments to assess PROs, the impact of posing a more focused research question should be considered.
- Increasing confidence levels may improve screening accuracy but result in fewer references screened by the AI reviewer, thus limiting its value in reducing human screening burden.
- There are significant technical challenges associated with the use of AI that need to be overcome to meet the strict evidentiary requirements of HTA bodies.
- Exploration of alternative AI tools to aid with SLR reference screening is ongoing.

Acknowledgments

Graphics and editorial assistance was provided by Kawthar Nakayima of Evidera, Inc.

Funding provided by Evidera Inc.

Presented at ISPOR US 2022 • May 15–18, 2022 • Washington, DC, United States