# Utility of Artificial Intelligence in Abstract Screening: A Case Study of Multiple Systematic Literature Reviews

POSTER 120585

Allie Cichewicz[1], Ananth Kadambi[2], Louis Lavoie[3], Lalith Mittal[4], Vicki Pierre[3], Renuka Raorane[5]

[1]Evidera, Waltham, MA, USA; [2]Evidera, San Francisco, CA, USA; [3]Evidera, Montreal, QC, Canada; [4]Evidera, Bengaluru, India; [5]Evidera, London, England

## Introduction

- With the increasing amount of published literature, there is a need for more efficient methods to screen larger volumes of references when conducting systematic literature reviews (SLRs).
- Integrating artificial intelligence (AI) into screening of SLRs may reduce time and effort needed to conduct robust, fully comprehensive reviews. However, there remains some uncertainty around the universal utility of AI when applied to various SLRs related to health economics and outcomes research (HEOR).

## Objective

- To determine the ability of AI to accurately identify relevant literature on the following topics: epidemiology, SLRs and/or network meta-analyses (NMA), treatment patterns, treatment guidelines, and health utilities.

## Methods

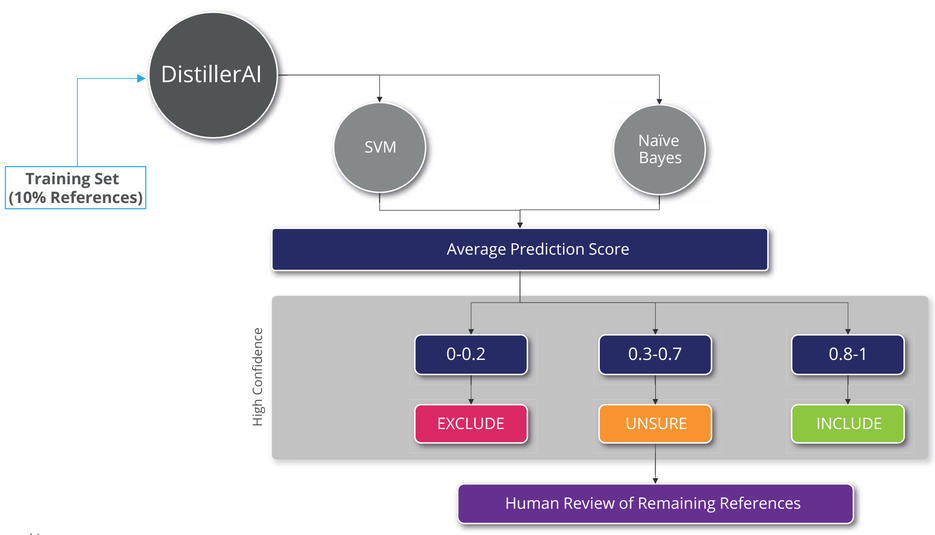
- Ten unique SLRs were conducted across the five HEOR topics of interest in various disease indications; two SLRs per topic (**Table 1**). All records were screened by two independent reviewers, with disagreements resolved by a third reviewer. DistillerAI was employed to replicate these SLRs with AI as the second reviewer (**Figure 1**).
- Reference sets used to train the AI reviewer consisted of 10% of total screening yield for each SLR (range: 50–203 references) and were hand selected to provide a variety of examples of included and excluded references based on each individual SLR's eligibility criteria based on the Population, Intervention, Comparison, Outcomes and Study Design (PICOS) framework.
- Prediction scores ranged from 0–1, with 0 indicating a definite exclude and 1 indicating a definite include. For the purposes of this study, the following scores were set as AI reviewer thresholds: ≤0.2 for excludes and ≥0.8 for includes.
- Screening decisions were compared between AI and human reviewers and used to calculate inter-rater reliability (IRR) based on Cohen's kappa statistics.
- Several test scenarios were carried out to assess how the AI reviewer performed when adjusting various parameters as outlined in **Figure 2**.

### Table 1. SLR topics and disease areas selected as test cases

| | SLRs for test scenarios #1-4 | SLRs for test scenario #5 |
|---|---|---|
| Epidemiology | Peripheral T-cell lymphoma | Diffuse large B-Cell lymphoma |
| SLRs/NMAs | Non-valvular atrial fibrillation | Osteoarthritis |
| Treatment Patterns | Major depressive disorder | Attention-deficit/hyperactivity disorder |
| Treatment Guidelines | Amyloidosis | Endometrial cancer |
| Health utilities | Cardiovascular disease | Prostate cancer |

NMA = network meta-analysis; SLR = systematic literature review.

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




SVM = support vector machine

### Figure 2. Test scenarios across the five SLR topics

1. One SLR per HEOR topic
2. #1 with a narrower AI reviewer threshold of 0-0.1 for excludes and 0.9-1 for includes
3. #1 with a broader AI reviewer threshold of 0-0.3 for excludes and 0.7-1 for includes
4. #1 with a random selection of references for the training set
5. A second SLR per HEOR topic with different PICOS criteria from #1

AI = artificial intelligence; HEOR = health economics and outcomes research; PICOS = Population, Intervention, Comparison, Outcomes and Study Design; SLR = systematic literature review

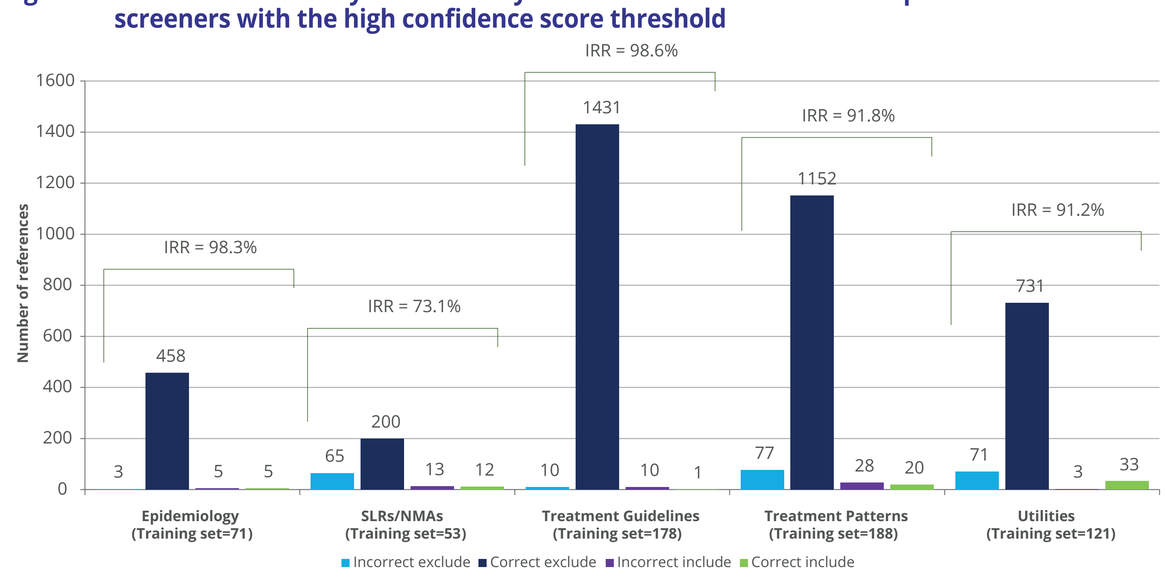
## Results

**How selection criteria by HEOR topic influence the results (Scenario #1)**

- In the first test scenario, agreement rates between AI and human reviewers based on IRR were >90% (91.2%-98.6%) for all SLR topics except SLRs/NMAs which had an IRR of 73%, indicating only slight agreement between reviewers (**Figure 3**).
- The AI reviewer consistently screened >70% of references for each SLR with a median of 75% and range of 73% for SLRs/NMAs to 91% for Treatment Guidelines.
- When there were disagreements between reviewers, the AI was more often overly exclusive (median [range]: 6% [0.6–19%]) than overly inclusive (1.1% [0.4–3.8%]) across SLRs, with the highest uncertainty of decisions for the SLR of SLRs/NMAs.

## Results (Cont'd)

### Figure 3. Inter-rater reliability and accuracy of AI reviewer decisions in comparison with human screeners with the high confidence score threshold




References screened with high confidence fell within the prediction score range of 0–0.2 (excludes) and 0.8–1 (includes). IRR = inter-rater reliability; NMA = network meta-analysis; SLR = systematic literature review .

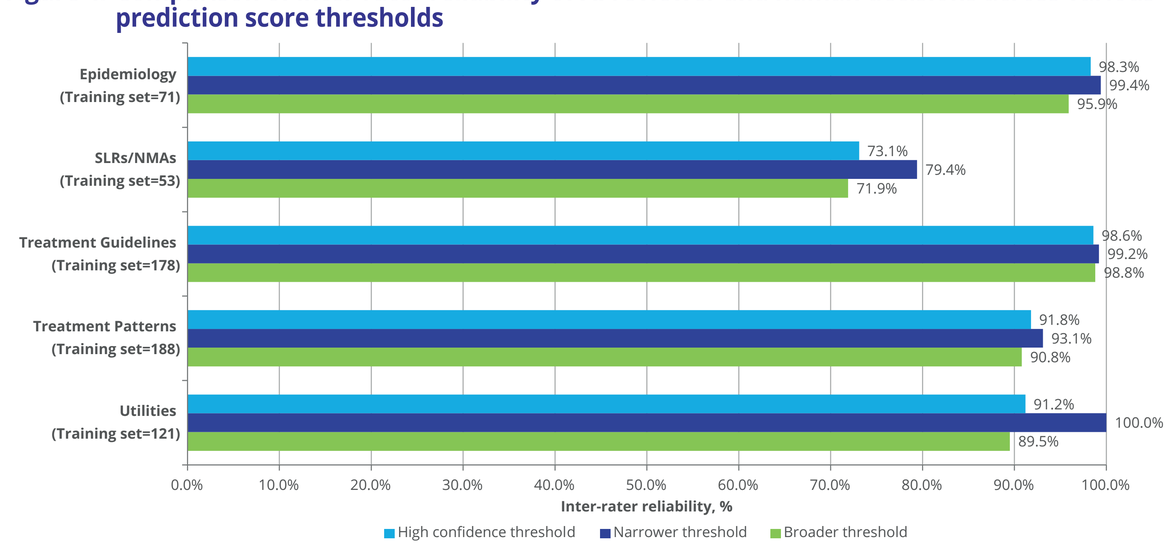
**Impact of varying the screening threshold for excludes and includes (Scenarios #2 and #3)**

- Scenario #2: Applying a narrower threshold (excludes 0–0.1, includes 0.9–1) led to a slight increase in agreement rates between reviewers for all SLR topics (**Figure 4**). However, this resulted in approximately 20% fewer references being screened by the AI reviewer with a median (range) of 53% (35-83%) of the total search yield across SLRs.
  - The narrower threshold followed the same trend as the first test scenario in that incorrect AI decisions primarily exclusive than inclusive, with a small reduction in incorrect decisions of approximately 1% across SLRs, except for the Utilities SLR for which all AI decisions were correct.
- Scenario #3: Applying a broader threshold (excludes 0–0.3, includes 0.7–1) led to a slight decrease in agreement rates between AI and human except for the Treatment Guidelines SLR (**Figure 4**).
  - This resulted in approximately 12% more references being screened by the AI reviewer with a median (range) of 87% (75–96%) of the total search yield across SLRs, but also more screening errors were made by the AI reviewer for most SLRs.

**Impact of changing training set from selected to random references (Scenario #4)**

- Using a random set of references for the training set rather than a hand-selected set had a minimal impact on the AI reviewer decisions for the Epidemiology, Treatment Guidelines, and Treatment Patterns SLRs but reduced the IRR by 18% and 15% for SLRs/NMAs and Utilities, respectively.
  - These two SLR topics also saw substantial shifts in the proportion of references screened by the AI reviewer and 20–40% increases in incorrect inclusions by the AI.

### Figure 4. Comparison of inter-rater reliability of AI reviewer and human decisions across various prediction score thresholds




References screened with high confidence fell within the prediction score range of 0–0.2 (excludes) and 0.8–1 (includes). A narrower screening threshold considered excludes of 0–0.1 and includes of 0.9–1, while a broader screening threshold considered excludes of 0–0.3 and includes 0.7–1. IRR = interrater reliability; NMA = network meta-analysis; SLR = systematic literature review

**Impact of changing the PICOS criteria for each SLR topic (Scenario #5)**

- When testing the AI with a second SLR per HEOR topic with a different set of PICOS criteria, results were consistent with the first test scenario in terms of IRR, proportion of references screened by the AI reviewer, and incorrect decisions for Treatment Guidelines, Treatment Patterns, and Utilities SLRs.
- In contrast, agreement between human and AI reviewers based on IRR differed considerably in the Epidemiology and SLR/NMAs SLRs, where IRR was 46% lower and 10% higher, respectively.
  - Such a disparity between Epidemiology SLRs could be attributed to the broader set of outcomes in the second SLR as well as a sample size requirement of >10 patients.

## Conclusions

- Decision-making between human and AI reviewers showed good agreement across most of the SLRs replicated by this case study in which AI was most accurately able to identify relevant treatment guidelines and epidemiologic outcomes, but struggled the most with objectively differentiating SLR/NMA study designs from topic-specific relevance.
- Further research is ongoing to provide additional insights into where AI implementation can most benefit the screening process, particularly in combination with the use of more complex tools (e.g., classifiers).

Presented at EU 2022 Annual Meeting • November 6–9, 2022 • Vienna, Austria