Artificial intelligence (AI)-based screening: Exploration of differences in two health technology assessment (HTA)-compliant systematic literature reviews (SLRs)

MSR163

Cichewicz, Allie, MSc[1]; Slim, Mahmoud, PharmD, PhD[2]; Deshpande, Sohan, MSc[3]

[1]Evidence Synthesis, Modelling & Communication, Evidera. Waltham, MA, USA; [2]Evidence Synthesis, Modelling & Communication, Evidera. St. Laurent, Quebec, Canada; [2]Evidence Synthesis, Modelling & Communication, Evidera. London, UK.

Background

- Advances in artificial intelligence (AI) present an opportunity to significantly enhance the efficiency of the study-selection process in systematic literature reviews (SLRs). This transformation is underway with the introduction of web-based tools designed to harness the power of machine learning (ML) algorithms in attempts to facilitate the selection of relevant records for inclusion in SLRs.
- We previously established a relationship between training set volume, performance, and time savings; however, variations in performance across topics and models may have subsequent implications for the minimum standards to be considered by health technology assessment (HTA) bodies.

Objectives

- We aimed to assess the performance of two different AI-assisted screening platforms versus humans in predicting screening decisions for titles and abstracts of two HTA-compliant SLRs.

Methods

SLRs

- Previously completed, HTA-compliant SLRs on psoriasis (PsO) and endometrial cancer (EC) were used as a training set to explore the performance of AI as a second screener of newly identified records from a search refresh. The eligibility criteria and volume of records of the original SLRs are detailed in **Table 1.**

Table 1. PICOS Criteria

	PsO SLR (n=4,000)	EC SLR (n=3,319)
Population	Adults with moderate-to-severe plaque PsO who are candidates for systemic therapies	Adults with primary advanced (stage III or IV) or first recurrent EC eligible for first-line systemic treatments
Interventions	Systemic biologic and non-biologic therapies approved for use in the US or Europe	Chemotherapies, hormonal therapies, targeted therapies, or immunotherapies recommended, marketed, or currently under investigation for the treatment of EC in the A/R setting, alone or in combination
Comparators	Any of the above, placebo	Any of the above, placebo
Outcomes	Efficacy, safety, HRQoL	Efficacy, safety, HRQoL
Study Design	RCTs	RCTs

Abbreviations: A/R = advanced or recurrent; EC = endometrial cancer; HRQoL = health-related quality of life; PsO = psoriasis; RCT = randomized controlled trial; SLR = systematic literature review; US = United States

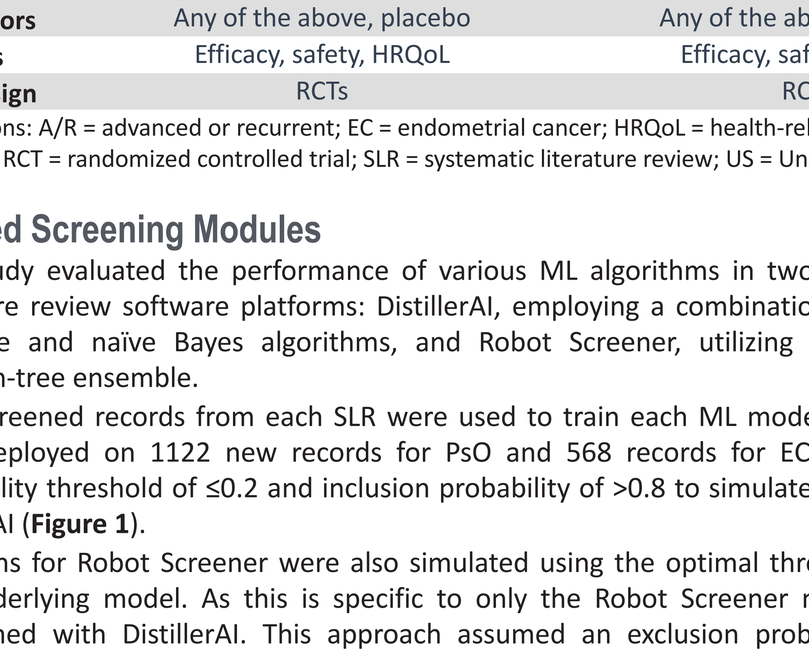
AI-based Screening Modules

- This study evaluated the performance of various ML algorithms in two distinct AI-assisted literature review software platforms: DistillerAI, employing a combination of support vector machine and naïve Bayes algorithms, and Robot Screener, utilizing a gradient-boosted, decision-tree ensemble.
- Fully screened records from each SLR were used to train each ML model. The models were then deployed on 1122 new records for PsO and 568 records for EC using an exclusion probability threshold of ≤0.2 and inclusion probability of >0.8 to simulate screening decisions by the AI (**Figure 1**).
- Decisions for Robot Screener were also simulated using the optimal threshold suggested by the underlying model. As this is specific to only the Robot Screener model, this was not performed with DistillerAI. This approach assumed an exclusion probability threshold of ≤ the optimal threshold and inclusion probability of >0.8.

Analysis

- Inclusion and exclusion decisions suggested by each model were then compared with the final consensus made by two human reviewers.
- Performance metrics (inter-rater reliability [IRR], Cohen's kappa, recall, precision, and specificity) were calculated for each SLR and model.

Figure 1. Overview of ML Model Training and Decision Process



ML = machine learning

Results

Model Performance

An overview of the performance metrics of each model vs. the human reviewers is summarized in **Table 2** for the PsO SLR and **Table 3** for the EC SLR.

PsO SLR

- The ML models differed in their inclusion predictions between SLRs, with more records falling outside the preset threshold with DistillerAI; this resulted in 20% fewer records screened compared with Robot Screener.
- Both models performed well across all metrics using the ≤0.2 and >0.8 inclusion probability threshold. Agreement rates between human and AI reviewers were high (IRR >93%), and each model correctly identified most relevant records. Of the missed records that were eligible for full-text review, none was ultimately included in the SLR.
- Using the optimal threshold suggested by the underlying Robot Screener model (≤0.34 for exclusions), IRR, Cohen's k, and recall improved, exhibiting metrics in line with DistillerAI.

Table 2. Summary of Model Performance Metrics for PsO SLR Update (n=1,122)

	DistillerAI	Robot Screener	Robot Screener (optimal)*
Records with inclusion probability within AI threshold	708 (63%)	936 (83%)	936 (83%)
IRR	96.2%	93.1%	95.3%
Cohen's k	0.90	0.8	0.86
Recall	0.98	0.82	0.91
Precision	0.87	0.88	0.87
Specificity	0.97	0.96	0.96

*Optimal exclusion threshold of 0.34

Abbreviations: IRR = inter-rater reliability; PsO = psoriasis; SLR = systematic literature review

EC SLR

- Similar to the PsO SLR findings, more records fell outside the preset threshold with DistillerAI, resulting in 10% fewer records screened compared with Robot Screener. However, screening rates were similar when considering the optimal threshold approach for Robot Screener.
- Although DistillerAI demonstrated an improved recall rate when employing the inclusion probability thresholds of ≤0.2 and >0.8, the recall rates between the two models were comparable when utilizing the optimal threshold approach for Robot Screener.
- Agreement rates between human and AI reviewers were high (IRR >95%), and each model correctly identified most relevant records. Of the records missed by Robot Screener that were eligible for full-text review, only one was included in the SLR as opposed to none by DistillerAI.
- Using the optimal threshold suggested by the underlying Robot Screener model (≤0.08 for exclusions), IRR, Cohen's k, and recall improved, exhibiting metrics in line with DistillerAI.

Table 3. Summary of Model Performance Metrics for EC SLR Update (n=568)

	DistillerAI	Robot Screener	Robot Screener (optimal)*
Records with inclusion probability within AI threshold	458 (81%)	524 (92%)	459 (81%)
IRR	97.4%	95.9%	97.2%
Cohen's k	0.49	0.35	0.77
Recall	0.75	0.27	0.75
Precision	0.38	0.54	0.80
Specificity	0.97	0.99	0.99

*Optimal exclusion threshold of 0.08

Abbreviations: EC = endometrial cancer; IRR = inter-rater reliability; SLR = systematic literature review

Discussion

- When deploying ML models to support title/abstract screening for SLR updates, both models performed comparably for each SLR. However, key metrics such as recall and precision were higher with the PsO SLR. This may be attributed to the smaller sample size of the EC review and more restrictive treatment setting.
- In ML-assisted SLRs, model training/optimization should aim toward achieving high recall (i.e., the model's ability to identify potentially relevant records), even if compromising precision (i.e., the model's ability to exclude irrelevant records). This strategy is recommended given that there are alternative opportunities to exclude irrelevant articles at a later SLR stage.
- While our study was designed to test the performance of AI in simulating an SLR update, it is essential to identify the minimum threshold of records screened and included that optimizes the model's performance, and across various types of reviews. This is especially important if these models are to be employed in de novo SLRs.

Conclusions

- Both ML models performed comparably; however, this study demonstrates that the model performance may vary across SLRs with different PICOS criteria.
- While model performance was compared with final consensus reached by human reviewers, it is important to consider these metrics in the context that human screening errors are not uncommon.
- Future uptake by HTA bodies should consider the complexity of study selection criteria and set minimally acceptable model performance metrics.

Presented at the ISPOR Europe Conference ● 12–15 November 2023 ● Copenhagen, Denmark