Evaluation of Machine Learning–Assisted Title and Abstract Screening in 5 Clinical Systematic Literature Reviews

MSR46

Jose Marcano-Belisario,[1] Michaela Lunan-Taylor,[1] Emma Hawe,[1] Eugene M. Farrelly[2]

[1] RTI Health Solutions, Manchester, United Kingdom; [2] RTI Health Solutions, Research Triangle Park, NC, United States

BACKGROUND

- Artificial intelligence (AI), including machine learning (ML) techniques, have the potential to automate some manual tasks in systematic literature reviews (SLRs).
- By leveraging AI, researchers could make SLRs less resource intensive and more comprehensive and timely.
- Screening of titles and abstracts is often the focus of many SLR programmes that offer ML-assisted screening. Robot Screener,[1] a built-in ML feature in Nested Knowledge, calculates the probability of each citation being advanced at the abstract and title stage.
- Advancement probabilities, generated by Robot Screener after a training phase involving human decisions, are then used by Nested Knowledge to either sort citations according to likelihood of inclusion or make recommendations about advancement or exclusion.
- Advancement probabilities could also reduce the manual screening workload if a suitable threshold for making bulk exclusions could be identified.

OBJECTIVE

- The objective of this project was to assess if an appropriate threshold of advancement probabilities could be determined for 5 therapeutic areas by modifying the size of the training datasets used for Robot Screener.

METHODS

- Screening decisions from 5 clinical SLRs were compared with the advancement probabilities generated by Robot Screener across 4 training scenarios: 20% (T1), 30% (T2), 40% (T3), and 50% (T4) of randomly selected citations screened manually prior to Robot Screener.
- These SLRs covered systemic lupus erythematosus (SLE), non-small cell lung cancer (NSCLC), breast cancer (BC), amyotrophic lateral sclerosis (ALS), and allergic rhino-conjunctivitis (ARC).
- For each scenario, cross-validation metrics obtained after training Robot Screener were recorded: recall, area under the curve (AUC), precision, F1 score, and accuracy.
- For each scenario, citations assigned advancement probabilities between 0.00 and 0.01 were assessed across all studies identified by the model to determine how many had advanced from the title-abstract stage as well as how many had been identified as correctly included. The upper limit of the advancement probabilities was increased by 0.05 if all the citations within the previous range were excluded from the original SLRs.

Robot Screener Terms

Recall: Proportion of relevant studies that are correctly identified by the model out of all relevant studies available. It is calculated as:

$$Recall = \frac{True\ Includes}{True\ Includes + False\ Excludes}$$

Precision: Proportion of relevant studies among the studies identified by the model. It is calculated as:

$$Precision = \frac{True\ Includes}{True\ Includes + False\ Includes}$$

Accuracy: Proportion of correctly identified studies out of all studies evaluated by the model. It is calculated as:

$$Accuracy = \frac{True\ Includes + True\ Excludes}{Total\ Number\ of\ Records}$$

F1 score: Harmonic mean of precision and recall. It is calculated as:

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The power of **knowledge**
The value of **understanding**

RESULTS

- Across all scenarios, recall probabilities ranged from 0.67 (ARC T1) to 0.92 (SLE T3 and SLE T4). Table 1 summarises recall probabilities and the other cross-validation metrics obtained for each SLR and training scenario.
- Overall, recall probabilities tended to increase with larger training datasets; however, AUC, precision, F1 score, and accuracy were not strongly correlated with training dataset size (see Table 2). When considering disease area, the correlation between training dataset size and precision, F1-score, and accuracy showed the greatest variability.

Table 1. Summary of Cross-Validation Metrics by Disease Area and Training Scenario

Disease area	Rate of advancement*	Rate of inclusion*	Training scenario	Recall	AUC	Precision	F1	Accuracy
SLE	13.4%	3.3%	T1	0.87	0.94	0.5	0.62	0.83
			T2	0.89	0.93	0.49	0.63	0.85
			T3	0.92	0.96	0.47	0.61	0.87
			T4	0.92	0.96	0.53	0.67	0.88
NSCLC	18.7%	3.7%	T1	0.8	0.78	0.3	0.43	0.62
			T2	0.84	0.82	0.34	0.48	0.66
			T3	0.84	0.8	0.36	0.5	0.68
			T4	0.85	0.83	0.33	0.48	0.67
BC	18.9%	4.3%	T1	0.81	0.85	0.38	0.5	0.7
			T2	0.81	0.86	0.41	0.54	0.74
			T3	0.83	0.85	0.42	0.55	0.72
			T4	0.88	0.88	0.41	0.55	0.73
ALS	21.9%	5.6%	T1	0.83	0.8	0.4	0.54	0.68
			T2	0.83	0.83	0.48	0.61	0.76
			T3	0.85	0.87	0.44	0.57	0.72
			T4	0.88	0.89	0.49	0.63	0.77
ARC	22.2%	7.9%	T1	0.67	0.6	0.29	0.41	0.58
			T2	0.73	0.76	0.34	0.46	0.64
			T3	0.8	0.79	0.38	0.51	0.7
			T4	0.88	0.85	0.42	0.56	0.71

* Rates refer to the rates in the original SLRs (based on the decisions made by researchers).

Table 2. Correlation Coefficients of Cross-Validation Metrics and Training Dataset Size, Overall, and by Disease Area

Category	Recall	AUC	Precision	F1	Accuracy
Overall	0.618	0.434	0.229	0.320	0.288
Disease area					
SLE	0.949	0.774	0.360	0.638	0.990
NSCLC	0.864	0.812	0.473	0.663	0.772
BC	0.899	0.731	0.745	0.868	0.531
ALS	0.929	0.993	0.722	0.737	0.722
ARC	0.998	0.943	0.998	1.000	0.965

- Across all scenarios, the percent of unscreened citations that had been included in the original SLRs that had an advancement probability less than 0.85 ranged from 25% (SLE T2) to 76.47% (BC T1).
- The percent of these citations with an advancement probability less than 0.8 ranged from 25% (SLE T2) to 73.53% (BC T1).

Figure 1. Percentage in the Original SLRs of Unscreened Relevant Citations by Advancement Probability

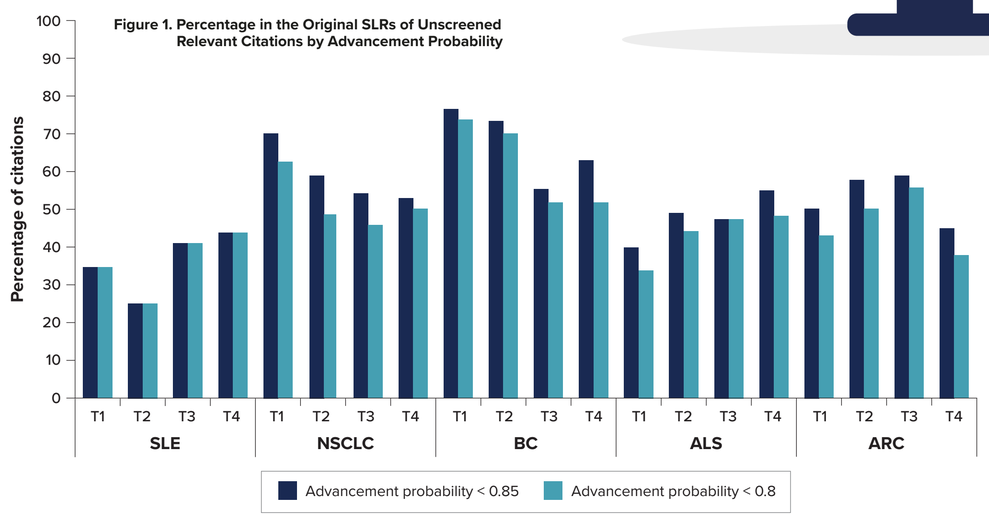

Table 3. Number of Unscreened Relevant Articles by Advancement Probability Threshold

Disease area	Training Scenario	Advancement probability range	Unscreened studiesprogressed for full-text screening in the original SLRS	Unscreened studies included in the original SLRs	Description of studies included in SLR
SLE	T1	0.00-0.01	4	1	Trial record (n = 1)
	T2	0.00-0.01	8	1	Title with no abstract (n = 1)
	T3	0.00-0.01	4	1	Title with no abstract (n = 1)
	T4	0.00-0.01	2	1	Title with no abstract (n = 1)
NSCLC	T1	0.00-0.01	29	4	Trial records (n = 3); journal article (n = 1)
	T2	0.00-0.01	34	1	Trial record (n = 1)
	T3	0.00-0.01	27	2	Trial record (n = 1); journal article (n = 1)
	T4	0.00-0.01	20	3	Trial records (n = 2); journal article (n = 1)
BC	T1	0.00-0.01	19	3	Journal articles (n = 2); abstract only (n = 1)
	T2	0.00-0.01	13	2	Journal articles (n = 2)
	T3	0.00-0.01	13	1	Journal article (n = 1)
	T4	0.00-0.01	13	2	Journal articles (n = 2)
ALS	T1	0.00-0.01	14	1	Trial record (n = 1)
	T2	0.00-0.01	5	0	Not applicable
		0.00-0.05	11	0	Not applicable
		0.00-0.1	17	0	Not applicable
		0.00-0.15	24	2	Trial records (n = 2)
	T3	0.00-0.01	17	4	Trial records (n = 4)
	T4	0.00-0.01	18	3	Trial records (n = 3)
ARC	T1	0.00-0.01	8	2	Trial record (n = 1); journal article (n = 1)
	T2	0.00-0.01	7	1	Journal article (n = 1)
	T3	0.00-0.01	11	1	Journal article (n = 1)
	T4	0.00-0.01	5	0	Not applicable
		0.00-0.05	8	1	Journal article (n = 1)

- A non-zero number of articles with very low advancement probabilities based on Robot Screener were previously determined through manual screening to meet the SLR inclusion criteria.
- For scenarios T1 and T3 and across all 5 SLRs, at least 1 citation with a low advancement probability had been included in the original SLR (see Table 3). These citations corresponded to records from trial registries (e.g., EUCTR, ClinicalTrials.gov) and journal articles.
- For ALS T2, the advancement probability range was increased to 0.00 to 0.15 before citations that had been included in the original SLR could be identified (these corresponded to trial records).
- For ARC T4, the range of advancement probabilities had to be increased to 0.00 to 0.05 before a relevant citation could be identified.

DISCUSSION

- No advancement probability threshold that can be safely applied to SLRs was identified in this project.
- This project identified a key aspect of the training phase that could have led Robot Screener to assign low advancement probabilities to relevant citations: representativeness of the training dataset. Only 1 of the relevant studies that was assigned a low advancement probability was a title with no abstract. This is an obvious example of a ML algorithm not having access to sufficient data to make a decision (after copying an excerpt from the article into the abstract field in Nested Knowledge, the advancement probability of this citation increased to 0.3). Other examples of relevant studies with low advancement probability suggest that the training datasets may have not been representative of all possible inclusions:
  - For example, trial records differ from regular titles and abstracts in layout and content. By not having enough examples in the training dataset, Robot Screener may have assumed that these records were irrelevant as they did not conform to the standard title and abstract format, particularly, if the trial records in the training dataset were excluded for other reasons.
  - Relevant journal articles with low advancement probabilities captured certain aspects of complex PICOS (population, intervention, comparison, outcome, and study) criteria that were not represented in the training dataset. For example, a novel treatment in the BC SLR was not represented in the training dataset. This also has implications for the type of SLRs that researchers may encounter (e.g., inclusion criteria in economic SLRs vary depending on the type of study being considered).
  - Future research should test this hypothesis by comparing how the selection of examples for training datasets (e.g., random selection vs. deliberate selection) affects the performance of ML algorithms across different types of SLRs (e.g., economic, clinical, humanistic).
- This project provided practical examples of the impact that cross-validation measures have on the reliability of SLR findings. As shown in Figure 1, a considerable percentage of relevant studies were assigned advancement probabilities below 0.8 by Robot Screener. The implication for SLRs is that, at the current time, ML-assisted screening may not be suitable to reduce the volume of citations for manual screening. However, this approach to screening may be helpful for targeted literature reviews, which tend to be exploratory.
- With continuing development of AI capabilities, it will be important to continue to reassess performance and suitability for use across different scenarios.

Reference

1. Nested Knowledge. Robot Screener. 2024. https://about.nested-knowledge.com/docs/robot-screener/. Accessed 9 April 2025.

Contact: Jose Marcano-Belisario (jmarcano@rti.org)

Presented at: ISPOR 2025; 13-16 May 2025; Montreal, Quebec, Canada