MSR72

An evaluation of the Rayyan artificial intelligence tool for systematic literature review screening

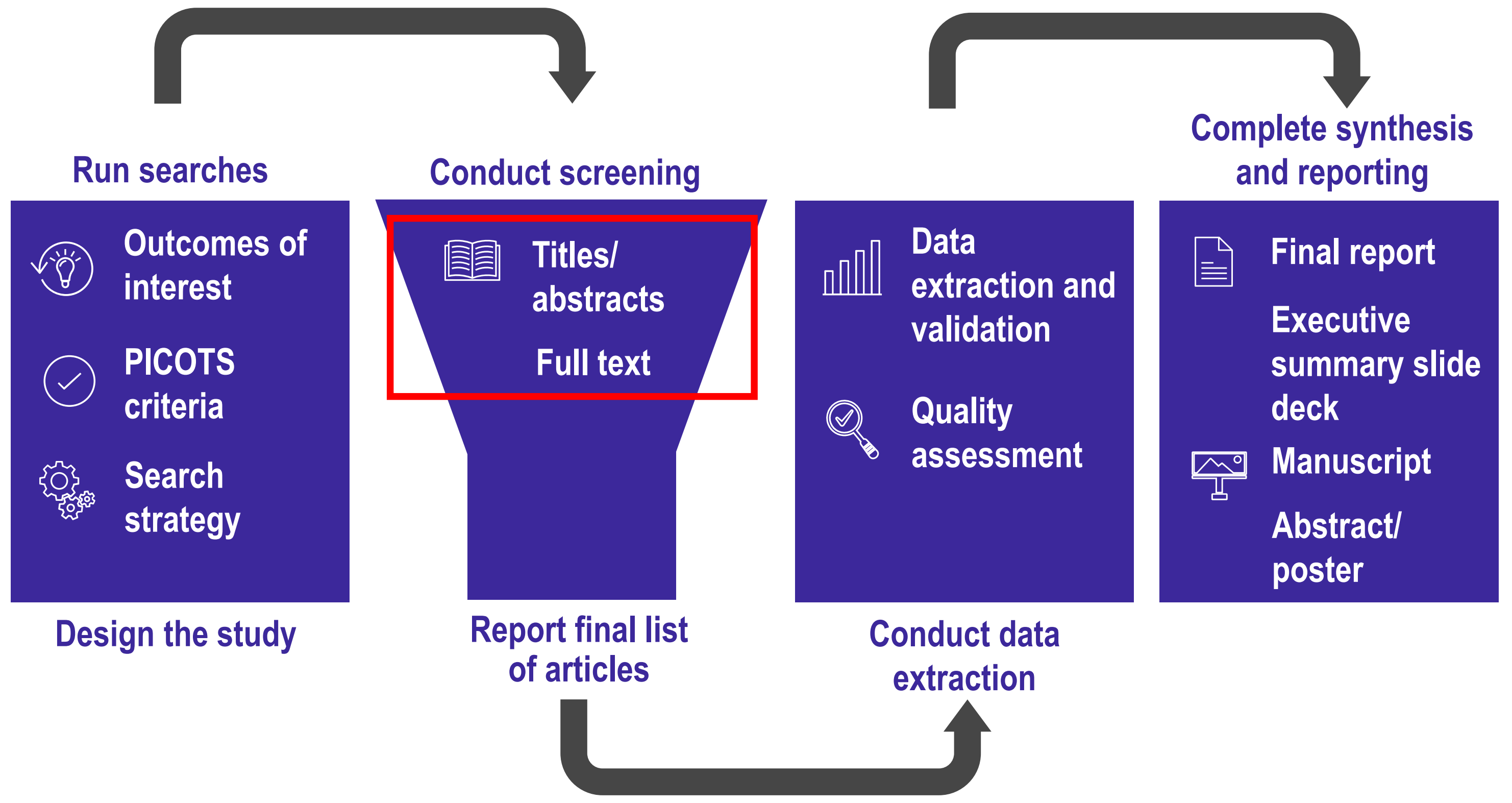
Ng J[1]; Szydlowski N[1]; Gill M[1]; Fusco N[1]; Ruiz K[1]

[1]Cencora, Conshohocken, PA, USA

Background

- Systematic literature reviews (SLRs) are known as the "gold standard" for evidence-based medicine and have long been placed at the top of the evidence-based medicine hierarchy.[1,2]
- SLRs are conducted per-protocol and are designed to be transparent and reproducible. Due to the stringent methods used to conduct SLRs, these reviews are labor- and time-intensive.[3,4]
- Limitations of SLRs include the possible bias inadvertently introduced by human reviewers, as well as the risk of the SLR being out of date upon completion due to the rapidly increasing number of articles published.[5]
- An area of interest to improve efficiencies in conducting an SLR is title/abstract (TIAB) screening.[6] In TIAB screening, researchers review the titles and abstracts of references to determine their eligibility for inclusion in the SLR. Eligible references are then reviewed in full text. TIAB and full text screening in the SLR process are shown in **Figure 1**.
- Research on the use of artificial intelligence (AI) with SLRs has frequently focused on the TIAB screening phase of the process.[7-9]

Figure 1. SLR process



Run searches

Outcomes of interest

PICOTS criteria

Search strategy

Design the study

Conduct screening

Titles/ abstracts

Full text

Report final list of articles

Data extraction and validation

Quality assessment

Conduct data extraction

Complete synthesis and reporting

Final report

Executive summary slide deck

Manuscript

Abstract/ poster

Key: PICOTS – population, interventions, comparators, outcomes, timing, study design; SLR – systematic literature review.

Objective

- The objective of this research is to evaluate the quantitative efficiencies and performance of the Rayyan AI tool (ie, Rayyan) in SLR TIAB screening.

Methods

- Two SLRs previously screened by 2 human reviewers covering the therapeutic areas of ophthalmology (ie, SLR 1) and oncology (ie, SLR 2) were identified.
- The SLRs assessed clinical, economic, and humanistic outcomes.
- Rayyan was trained on 20% of the total references for both SLR 1 and SLR 2, which were used to predict the relevance of the remaining references using a 5-level relevancy rating system from "most likely to exclude" to "most likely to include."
- Rayyan's relevancy ratings were compared to the original SLR inclusion/exclusion decisions by human reviewers to calculate the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and time-savings (**Table 1**).

Methods (cont.)

Table 1. Equations for calculated measurements to characterize the Rayyan AI tool

Measurement		Equation[a]
Sensitivity[b] (%)		$\frac{\text{\# of references included by both Rayyan and human reviewers}}{\text{Total \# of references included by human reviewers (excluding training set)}} \text{ x } 100$
Specificity[b] (%)		$\frac{\text{\# of references excluded by both Rayyan and human reviewers}}{\text{Total \# of references excluded by human reviewers (excluding training set)}} \text{ x } 100$
PPV[b] (%)		$\frac{\text{\# of references included by both Rayyan and human reviewers}}{\text{Total \# of references included by Rayyan}} \text{ x } 100$
NPV[b] (%)		$\frac{\text{\# of references excluded by both Rayyan and human reviewers}}{\text{Total \# of references excluded by Rayyan}} \text{ x } 100$
Accuracy[b] (%)		$\frac{\text{\# of references included by both Rayyan and human reviewer} + \text{\# of references excluded by both Rayyan and human reviewers}}{\text{Total \# of references (excluding test set)}} \text{ x } 100$
Time-savings[b,c] (percentage difference, %)	1 AI-assisted reviewer (single screening)	$\frac{\frac{\text{Total \# of references}}{50}\text{ hr} - \left(\frac{(0.2 \text{ x total \# of references}) + \text{total \# of references included by Rayyan}}{50}\text{ hr}\right)}{\frac{\text{Total \# of references}}{50}\text{ hr}} \text{ x } 100$
	1 human reviewer + 1 AI- assisted reviewer (double screening)	$\frac{\frac{\text{Total \# of references x 2}}{50}\text{ hr} - \left(\frac{\text{total \# of references} + [(0.2 \text{ x total \#references}) + \text{total \#of references included by Rayyan}]}{50}\text{ hr}\right)}{\frac{\text{Total \# of references x 2}}{50}\text{ hr}} \text{ x } 100$

Key: AI – artificial intelligence; hr – hour; PPV – positive predictive value; NPV – negative predictive value; TIAB – title/abstract.

[a] "Human reviewers" refers to the original decisions made in the SLRs when TIAB screening was completed by humans.

[b] The Rayyan inclusion category included the following relevancy ratings: "no recommendation," "likely to include" and "most likely to include."

[c] Assumes that an experienced human reviewer screens an average of 50 TIAB references per hour (ie, 50/hour).

Results

- There were a total of N=7,238 references for SLR 1 and N=5,494 references for SLR 2. **Table 2** displays the measurements calculated by the outcome of interest in both SLR 1 and SLR 2 using the equations displayed in **Table 1**.

Table 2. Performance of Rayyan AI by outcome of interest in SLR 1 and SLR 2[a]

SLR[b]	Outcome of interest	Number of references[c]	Sensitivity	Specificity	PPV	NPV	Accuracy
1	Clinical	2,919	96%	44%	44%	96%	60%
1	Costs	1,389	94%	20%	24%	93%	36%
1	Economic evaluation	964	79%	35%	18%	90%	41%
1	Humanistic	1,966	96%	41%	30%	98%	52%
2	Clinical	4,206	88%	62%	15%	99%	64%
2	Humanistic	951	100%	8%	3%	100%	10%
2	Economic evaluation	337	Could not be assessed due to the low number of included studies[d]				

Key: AI – artificial intelligence; PPV – positive predictive value; NPV – negative predictive value; SLR – systematic literature review.

[a] The following Rayyan relevancy ratings were included in the calculations: "no recommendation," "likely to include," and "most likely to include."

[b] SLR 1 covered ophthalmology; SLR 2 covered oncology.

[c] Includes a 20% training set.

[d] Only 7 studies were included during the original TIAB screening of economic evaluations for SLR 2, and only 1 included study was part of the 20% training set. Economic evaluations for SLR 2 could not be assessed because a Rayyan training set requires 50 references, 5 of which must be references marked as included.

- Sensitivity was highest for the humanistic outcome of interest (96%–100%).
- Sensitivity was also generally high for the clinical (88%–96%) and the cost outcomes of interest (94%).
- Sensitivity for the economic evaluation outcome of interest was the lowest across all outcomes of interest (79%).
- The greatest time-savings for the AI-assisted reviewer were seen with SLR 2, where the amount of time spent screening clinical references was reduced by 47% (**Table 3**).
- Although using Rayyan resulted in time-savings for the AI-assisted reviewer across all outcomes of interest (6%–47%), time-savings for dual screening with 1 human reviewer and 1 AI-assisted reviewer were modest (3%–23%) (**Table 3**).

Results

Table 3. Time-savings with AI-assisted screening by SLR and outcome of interest[a]

SLR[b]	Outcomes of interest	Time-savings with AI-assisted reviewer[a]			Time-savings for dual screening (1 human review + 1 AI-assisted reviewer[a])		
		Screening hours for 1 human reviewer	Screening hours for 1 AI-assisted reviewer	% time saved	Screening hours for 2 human reviewers	Screening hours for 1 human reviewer + 1 AI-assisted reviewer	% time saved
1	Clinical	58	44	25%	116	102	13%
1	Costs	28	24	14%	56	52	7%
1	Economic evaluation	19	14	26%	38	34	13%
1	Humanistic	39	29	27%	78	68	13%
2	Clinical	84	45	47%	168	128	23%
2	Humanistic	19	18	6%	38	37	3%

Key: AI – artificial intelligence; SLR – systematic literature review.

[a] The AI-assisted reviewer used in this study was the Rayyan AI tool.

[b] SLR 1 covered ophthalmology; SLR 2 covered oncology.

Limitations

- The results from this analysis were generated by testing Rayyan with 2 SLRs. Therefore, the results cannot necessarily be applied to other AI tools or other types of literature reviews.
- Rayyan's ability to differentiate between the inclusion and exclusion criteria of references is dependent on the training set. If few references that meet the SLR eligibility criteria are included in the training set, Rayyan may not be able to effectively identify relevant literature.

Conclusions

- AI screening with Rayyan resulted in high sensitivity (88%–100%) and potential time-savings.
- Sensitivity (accurately including relevant references) is extremely important for SLRs, where the purpose is to include all references meeting eligibility criteria.
- Despite these promising results, low-to-moderate specificity values may limit time-saving benefits, due to the additional time required to sort through irrelevant references.
- AI tools have the potential to increase the efficiency of SLRs, but it is important to validate new AI tools to maintain methodological rigor and accuracy. Additional research is needed to investigate the most effective ways to incorporate AI in SLR processes.

References: 1. Wallace SS, et al. *Hosp Pediatr*. 2022;12(8):745-750. **2.** Howick J, et al. Explanation of the 2011 Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (Background Document). Oxford Centre for Evidence-Based Medicine. Accessed March 7, 2024. https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence **3.** Michelson M, et al. *Contemp Clin Trials Commun*. 2019;16:100443. **4.** Borah R, et al. *BMJ Open*. 2017;7(2):e012545. **5.** Atkinson CF. *Soc Sci Comput Rev*. 2024;42(2):376-393. **6.** Gates A, et al. *Syst Rev*. 2020;9(1):272. **7.** Odintsova VV, et al. *Psychiatr Genet*. 2019;29(5):170-190. **8.** Gartlehner G, et al. *Syst Rev*. 2019;8(1):277. **9.** Tsou AY, et al. *Syst Rev*. 2020;9(1):73.

Presented at ISPOR 2024
May 5-8, 2024 | Atlanta, GA
Funded by Cencora

Please direct questions to:
Joanna Ng | Joanna.Ng@cencora.com