An evaluation of PhlexNeuron, an internal, proprietary artificial intelligence (AI) tool for systematic literature review (SLR) screening

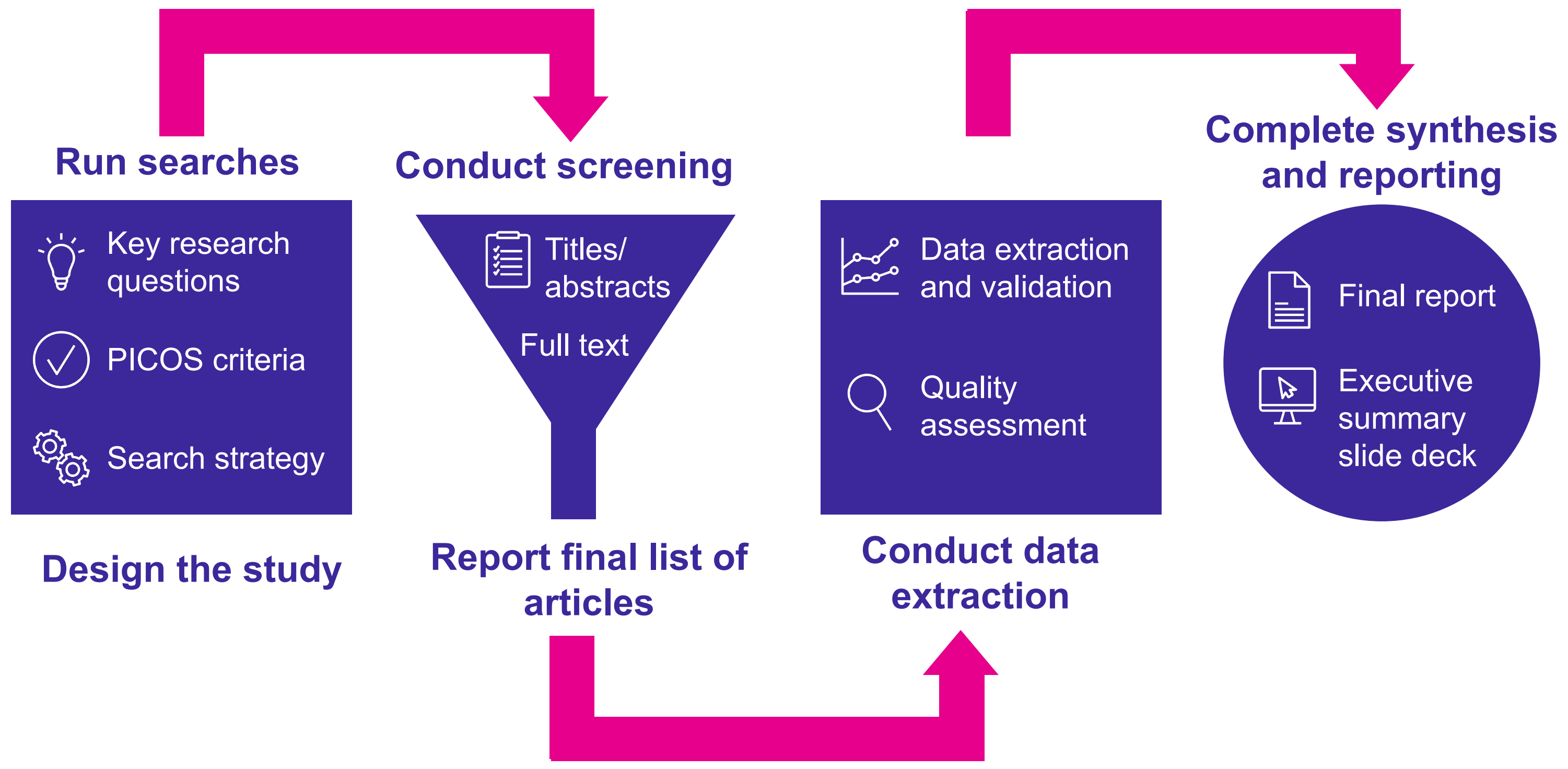
Szydlowski N[1,a]; Galloway B[1]; Gill M[1]; Swiger D[1]; Koppers D[1]; Shankar S[1]; Ruiz K[1]; Fusco N[1]

[1]Cencora, Conshohocken, PA, USA [a] Employed with Cencora at the time of this research.

Background

- Systematic literature reviews (SLRs) play an important role in performing evidence-based decision-making in drug development and approvals.
- Due to the vast amount of available scientific evidence and the rigorous methodology necessary, SLRs require significant time and effort in the process of title/abstract (TIAB) and full-text (FT) review.
- An SLR necessitates that researchers evaluate hundreds or even thousands of potentially relevant publications. This process begins with TIAB screening to assess the relevance of each publication for inclusion in the review. Then, eligible references are reviewed at the FT level (**Figure 1**).
- Artificial intelligence (AI) is a promising technology that could be used to reduce time and workload burden by increasing the efficiency of SLRs.
- One possible application of AI for SLRs is to identify relevant studies during TIAB screening at a speed considerably faster than humans.[1]
- Most AI tools that are currently available for literature screening require a training set for each individual SLR, which requires human reviewers to review a portion of the references prior to using the AI tool.[1,2]

Figure 1. SLR process



Key: PICOS – population, intervention, comparator, outcome, and study design; SLR – systematic literature review.

Objective

- The objective of this research was to assess the performance of PhlexNeuron, a proprietary, internal AI tool for literature screening that does not require a training set for each new SLR, which will likely result in time savings.

Methods

- Four SLRs that had previously undergone screening by human reviewers were identified. The clinical, costs/healthcare resource utilization (HCRU), economic evaluation, and humanistic sets consisted of 4,233, 1,908, 1,289, and 2,637 references, respectively.
- Eligibility questions were generated using the population, intervention, comparator, outcome, and study design (PICOS) criteria from the original SLRs.
- The title, abstract, and publication date for all references were uploaded and, without the use of a training set, PhlexNeuron was prompted to answer the PICOS questions for each reference with 3 possible responses: "Yes," "No," and "Uncertain." PhlexNeuron also provided an explanation for its response to each PICOS question.
- Then, a screening recommendation was generated based on the PICOS question answers. If the PICOS question answers were all "Yes" or "Uncertain," then the screening recommendation was "Include." If PhlexNeuron provided a "No" response to any of the PICOS questions, then the screening recommendation was "Exclude."
- PhlexNeuron's ratings, which were based only on TIAB review, were compared to the original TIAB inclusion/exclusion decisions of human reviewers to calculate sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) (**Table 1**). In addition, PhlexNeuron's ratings were compared to the studies included in the completed SLR.

Table 1. Equations for calculated measurements to characterize the PhlexNeuron tool

Measurement[a]	Equation[b]
Sensitivity, %	$\frac{\text{\#of references included by both PhlexNeuron and human reviewers}}{\text{Total \# of references included by human reviewers}} x\ 100$
Specificity, %	$\frac{\text{\# of references excluded by both PhlexNeuron and human reviewers}}{\text{Total \# of references excluded by human reviewers}} x\ 100$
PPV, %	$\frac{\text{\#of references included by both PhlexNeuron and human reviewers}}{\text{Total \# of references included by PhlexNeuron}} x\ 100$
NPV, %	$\frac{\text{\# of references excluded by both PhlexNeuron and human reviewers}}{\text{Total \# of references excluded by PhlexNeuron}} x\ 100$
Accuracy, %	$\frac{\text{\# of references included by both PhlexNeuron and human reviewers} + \text{\# of references excluded by both PhlexNeuron and human reviewers}}{\text{Total \# of references}} x\ 100$

Key: AI – artificial intelligence; NPV – negative predictive value; PPV – positive predictive value.
[a] The PhlexNeuron inclusion categories include "include," "exclude," and "uncertain."
[b] "Human reviewers" refers to the original decisions made in the SLRs when screening was completed by humans.

Results

- When PhlexNeuron ratings were compared to the results of human reviewers, sensitivity was consistently high, ranging from 89% to 93% across all topics for TIAB screening and 93% to 98% when compared to the completed SLRs.
- Specificity ranged from 45% to 71% for TIAB screening and 43% to 64% compared to the completed SLRs.
- PhlexNeuron provided rating justifications for all references in each SLR dataset.

Table 2. Performance of PhlexNeuron TIAB screening vs TIAB screening by human reviewers

Performance metric	Clinical	Costs/HCRU	Economic evaluation	Humanistic
Number of references evaluated by PhlexNeuron,[a] n	4,233	1,908	1,289	2,637
Sensitivity,[a] %	91	89	93	89
Specificity,[a] %	65	50	45	71
PPV,[a] %	45	29	20	40
NPV,[a] %	96	95	98	97
Accuracy,[a] %	71	57	51	74

Key: AI – artificial intelligence; HCRU – healthcare resource utilization; NPV – negative predictive value; PPV – positive predictive value; TIAB – title/abstract.
[a] PhlexNeuron inclusion relevancy ratings included in the calculations: "include," "exclude," and "uncertain."

Table 3. Performance of PhlexNeuron TIAB screening vs completed SLR by human reviewers[a]

Performance metric	Clinical	Costs/HCRU	Economic evaluation	Humanistic
Number of references evaluated by PhlexNeuron,[b] n	4,233	1,908	1,289	2,637
Sensitivity,[a] %	98	93	98	94
Specificity,[a] %	54	46	43	64
PPV,[a] %	10	15	10	18
NPV,[a] %	100	98	100	99
Accuracy,[a] %	56	50	46	67

Key: AI – artificial intelligence; HCRU – healthcare resource utilization; NPV – negative predictive value; PPV – positive predictive value; SLR – systematic literature review; TIAB – title/abstract.
[a] In the completed SLR by human reviewers, all references were reviewed at the TIAB level and the subset of references that were included by humans at the TIAB level were also reviewed by humans at the FT level.
[b] PhlexNeuron inclusion relevancy ratings included in the calculations: "include," "exclude," and "uncertain."

Conclusions

- AI-assisted TIAB screening using PhlexNeuron was highly sensitive for TIAB (89%-93%) across topics (clinical, humanistic, costs/HCRU, and economic evaluations).
- When AI-assisted TIAB screening by PhlexNeuron was compared to completed SLRs by human reviewers, sensitivity remained high (93%-98%). The higher sensitivity implies that some studies excluded by PhlexNeuron but included by human reviewers at TIAB were later excluded by human reviewers during FT review.
- High sensitivity (ie, the ability to include relevant references accurately) is extremely important to produce high-quality SLRs.
- AI-assisted TIAB screening is a promising method for increasing efficiency of SLRs. Using AI as a second reviewer for SLRs is of particular use, since this method can provide efficiency while integrating a quality check on AI responses.
- Future research should confirm the performance of AI-assisted screening across SLRs varying in size (number of references) and topics of interest to ensure that similar results can be reliably obtained without a training set.

Limitations

- The results from this analysis were generated by testing PhlexNeuron on 4 large SLRs. The generalizability to other types of reviews is uncertain.
- A human review of the AI-generated justifications is necessary to accurately interpret the results and ensure that the reasoning adequately justifies study inclusion and exclusion decisions. Therefore, AI-assisted screening with PhlexNeuron is best used as a second reviewer in the SLR process.

References

1. Bolaños F, Salatino A, Osborne F, et al. Artificial intelligence for literature reviews: opportunities and challenges. *Artif Intell Rev*. 2024;57(259). https://doi.org/10.1007/s10462-024-10902-3 **2.** Kumar Venkata S, Velicheti S, Jamdade V, et al. Application of artificial intelligence in literature reviews. Poster presented at: ISPOR 26th Annual European Congress; Copenhagen, Denmark; November 12-15, 2023.

Presented at: ISPOR 2025; May 13-16, 2025; Montreal, Quebec, Canada. Funded by Cencora.

The authors acknowledge Anke Huber of PharmaLex for project management and support.

cencora