Are Classifiers the Future of Artificial Intelligence in Systematic Literature Reviews?

Poster No. 118497

Cichewicz, A, MSc[1]; Huelin, R, BA[1]; Kadambi, A, PhD[2]

[1]Evidence Synthesis, Modeling & Communication, Evidera. Waltham, MA, USA; [2]Evidence Synthesis, Modeling & Communication, Evidera. San Francisco, CA, USA

Background

- Artificial intelligence (AI) algorithms are being increasingly utilized when conducting literature reviews, primarily to facilitate expedited title and abstract screening. AI can be a valuable asset in reducing the screening burden for human reviewers
  - However, as the approach is not currently widely accepted by review bodies such as Health Technology Assessment (HTA) agencies, it is most often used to predict study eligibility and rank references according to relevance.
  - It may also be used as a quality control measure (e.g., employed as a secondary reviewer).
- While AI is currently far from fully replacing the need for humans in the conduct of systematic literature reviews (SLRs), it could be further optimized to better support the review process and keep up to date on the latest advances in the published literature.

Is there a need for more advanced applications of AI in systematic reviews?'

- In SLRs, references are reviewed for eligibility based on pre-specified inclusion and exclusion criteria.
- Human investigators review the title and abstract of each reference and assess whether the reference meets the criteria to include in the SLR or cannot be confidently excluded based on the available information.
- Selection criteria for SLRs can be very complex and multi-faceted, based on the Population, Interventions, Comparators, Outcomes, Study Design (PICOS) framework (**Table 1**). Reviewers must decide whether the reference meets each of the PICOS components to be eligible for the SLR – failure to meet at least one component renders the reference ineligible to proceed to the next stage of the review.

Table 1. Example of complex PICOS criteria

PICOS	Class	Inclusion Criteria
Population	Age group	Adults
	Disease type	Head and neck cancer
	Disease site	Larynx, oral cavity, oropharynx, pharynx tumors
	Disease stage	Locoregionally advanced
	Line of therapy	Previously treated
Interventions	Systemic therapy	Chemotherapy, immunotherapy, targeted therapy
	Radiotherapy	Intensity-modulated radiotherapy, volumetric modulated radiotherapy, image-guided radiotherapy, stereotactic radiosurgery
	Surgery	Transoral robotic assisted surgery, neck dissection, laryngectomy
Comparator	Best supportive care	High-dose cisplatin-based radiotherapy
Outcomes	Efficacy	Overall survival, progression-free survival, complete response, time to progression
	Safety	Adverse events, serious adverse events, discontinuation due to adverse events
	Costs	Costs associated with diagnostic tests, hospitalizations, outpatient visits, surgical procedures
	Resource use	Use of outpatient and hospital healthcare services
	Health-related quality of life	Scores on the following measures: EORTC QLQ-30, H&N-35, UW-QoL, EQ-5D
Study Design	Clinical trial	Randomized controlled trial, non-randomized controlled trial, single-arm trial
	Observational study	Prospective cohort, retrospective cohort, case-control, cross-sectional

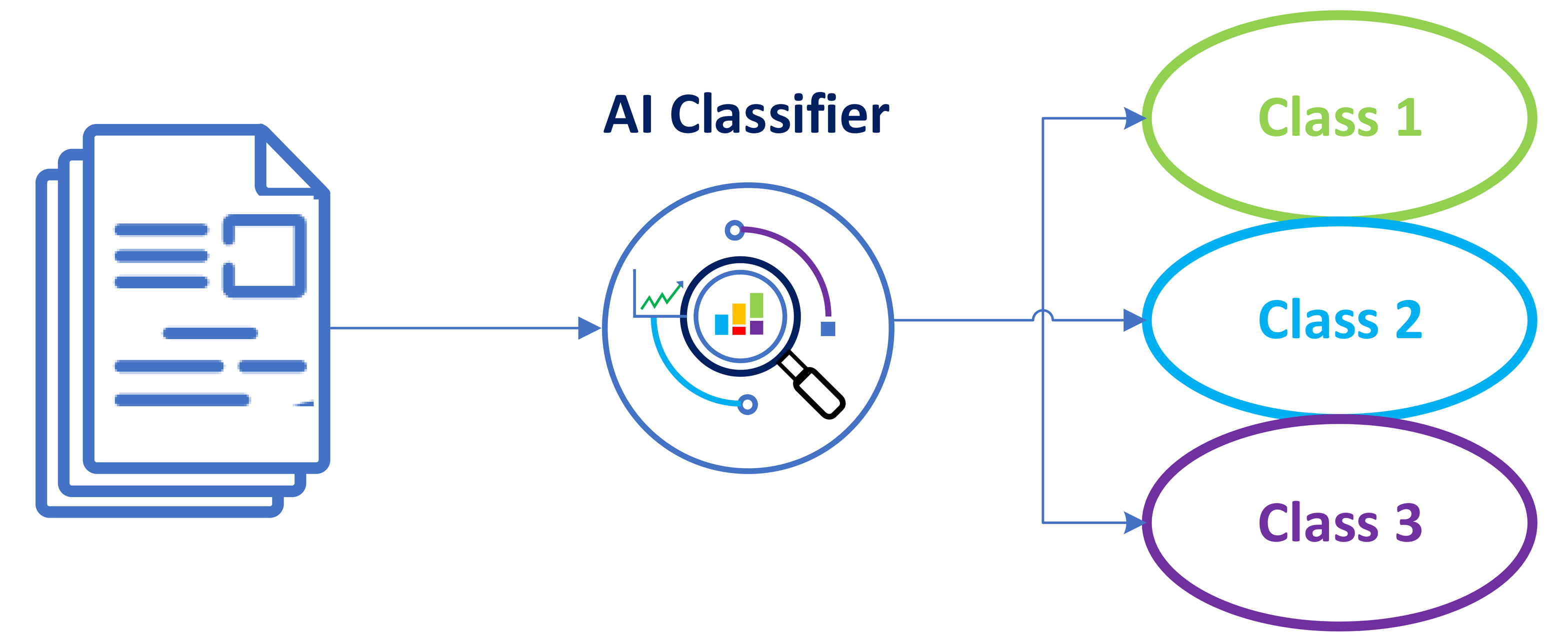
PICOS = Population, Interventions, Comparators, Outcomes, Study Design

- The screening process is often resource intensive and time consuming, particularly when the complex PICOS criteria are compounded by tight timelines for completion.
- With the speed at which literature is published, the increasing volume of references increases the challenge for researchers to conduct robust and comprehensive reviews in a timely manner.
- The use of AI reduces the screening burden for human reviewers and can help expedite the screening process. However, AI is primarily employed in SLRs to only make binary decisions (yes or no) regarding study eligibility and cannot make more granular decisions that indicate the reason for inclusion or exclusion. When selection criteria are multi-faceted, as with the above example, the lack of AI's ability to provide a rationale for its decision making can be a substantial obstacle to its more widespread adoption, particularly when conducting systematic reviews for submission to HTA agencies.
- In this sense, it is imperative to improve transparency with regards to the accuracy of the algorithm's decision-making process to more closely align with the decisions and assumptions that are typically made by human reviewers during the screening process.

What are classifiers?

- Natural language processing text classification algorithms, also referred to as classifiers, are capable of labeling or organizing data into different categories or "classes" (**Figure 1**).
- A classifier can be built to sort, label, or categorize text into pre-defined discrete classes. In the context of systematic literature reviews (SLRs), the algorithm reads text from the title and abstract of each reference for keywords and text patterns and assigns the reference to a corresponding class.
- Classifiers can be customized by the user to categorize references broadly or to identify very narrow, more specific categories.

Figure 1. Process of classifying data

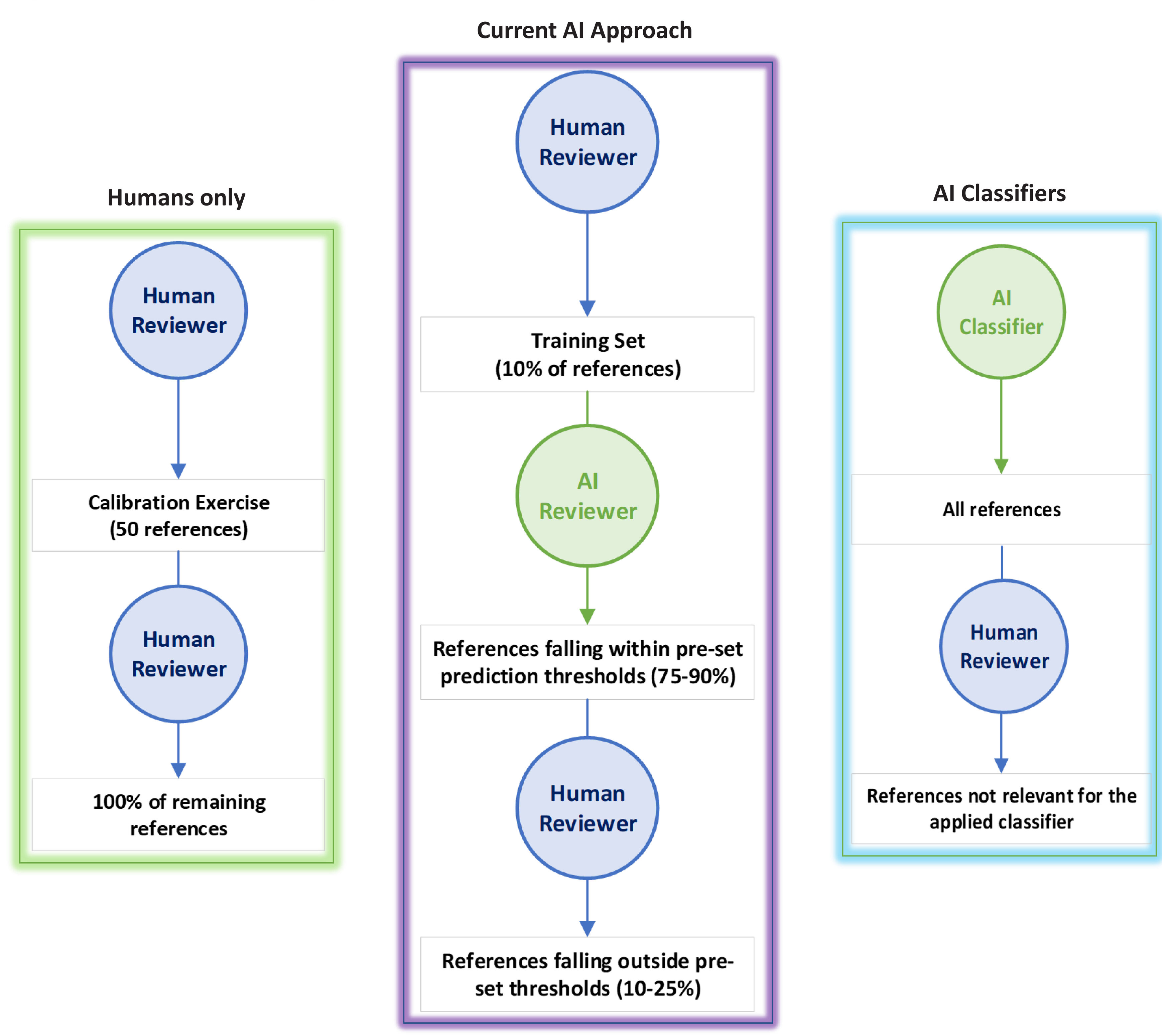


AI = artificial intelligence

How can classifiers enhance the review process?

- Classifiers can be customized to be as broad or narrow as needed based on the PICOS criteria of the SLR. For example, a classifier to determine the study design as either a clinical trial or observational study is widely applicable across many SLRs and may have a high return on investment following the initial development process.
- On the contrary, classifiers can also be developed for specific SLRs to identify references related to a particular disease area or specific interventions.
  - By customizing classifiers for each component of the PICOS framework, this type of categorization could enhance the existing binary include/exclude approach to provide the human investigators with reasons for inclusion or exclusion (**Figure 2**).
  - However, there are also important limitations to consider when applying classifiers on criteria that may not be well defined in the title or abstract such as patient age or geographic region.
- In addition to using classifiers to make screening decisions on references without further requiring a human reviewer, they can also indirectly inform the screening process by labeling references with their assigned class. When labels are visible to reviewers, they can help to validate the human reviewer's decision which can also reduce potential screening errors.

Figure 2. SLR screening process with and without use of AI



AI = artificial intelligence

Conclusions

- Greater visibility into AI decisions may lead to increased uptake by researchers, particularly within highly complex reviews.
- With classifiers, AI may edge closer to being able to replicate human reviewers, improve efficiency in the screening process and reduce time and resources required to systematically review large bodies of literature.
- Future guidance on validation standards for classifier algorithms may also make a step towards acceptance of AI technology within the stringent review frameworks required by HTA bodies.
- Further confirmation of situations in which classifiers provide the highest return on investment are being explored to generate quantitative insights.

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