Enhancing Systematic Literature Reviews with Generative Artificial Intelligence: Development, Applications, and Evaluation

MSR234

Ying Li, PhD[1], Surabhi Datta, PhD[2], Kyeryoung Lee, PhD[2], Hunki Paek, PhD[2], Eefje Bergraph, MS[1], Julie Glasgow, PhD[2], Chris Liston, PharmD[2], Long He, MS[2], Majid Rastegar-Mojarad, PhD[2], Xiaoyan Wang, PhD[2] , Yingxin Xu, BPharm, PhD[1]

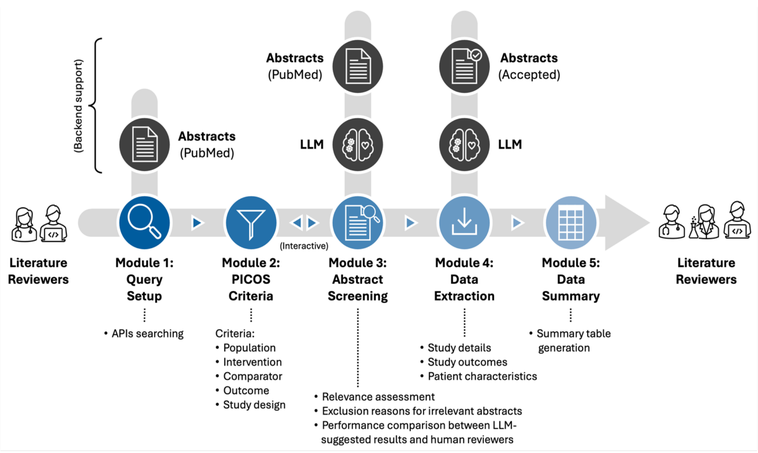
[1]Regeneron Pharmaceuticals, Inc., Tarrytown, NY, USA; [2]Intelligent Medical Object, Inc. Rosemont, IL, USA

BACKGROUND

- Health Technology Assessment (HTA) agencies evaluate the properties, effects and impacts of health technologies by requiring manufacture to submit Systematic Literature Reviews (SLRs) of ***clinical***, cost-effectiveness, and humanistic data to inform their decisions.
- The current approach to SLR is generally time-consuming, labor-intensive, and costly.
- The rapidly growing literature, diverse requirements from different HTAs across countries, and the need to conduct searches 3 months to 1 year before submission have made SLRs increasingly challenging, consequently placing a tremendous burden on manufactures striving to make healthcare products available in these markets.
- To address this need, we explored a Large Language Model (LLM) based AI-assisted SLR (AI-SLR) system to facilitate the clinical SLR process. We compared the performance of the system to humans, evaluating their accuracy and ability to reproduce results generated by human experts.

METHODS

Figure 1. Overview of the AI-Assisted SLR System. The PICOs criteria (module 2) are an input for the LLM prompt (module 3). The data field descriptions (module 4) form part of the prompt for data extraction.



- Users can specify PICO criteria and data elements of interest.
- Users can provide background knowledge related to disease areas to guide the LLM in screening and extraction.
- Users can iterate between modules 2 and 3 until screening performance meets their expectations.
- The PICOs framework for relapsed/refractory multiple myeloma (RRMM) is presented (**Table 1**). A similar framework was developed for the advanced melanoma SLR review.
- During abstract screening, the AI system evaluates each abstract and recommends inclusion or exclusion.

Table 1. Descriptions of PICOs Criteria Used for Relapsed/ Refractory Multiple Myeloma

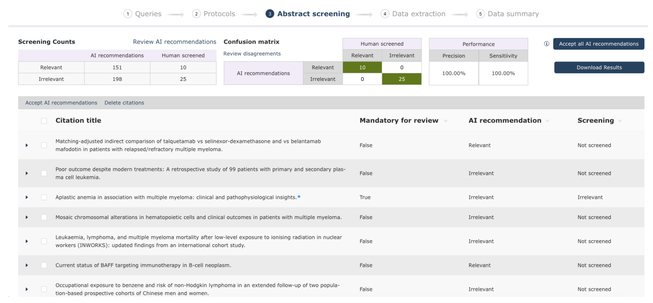
PICOs	I/E	Summary of Eligibility Criteria
P	I	▪ Studies including relapsed/refractory multiple myeloma. ▪ The target population should comprise individuals who have exhibited progression after undergoing at least two prior lines of therapies, with a preference for those exposed to triple-class treatments (PI, IMiDs, anti-CD38).
	E	▪ Studies exclusively involving patients under the age of 18. ▪ Studies exclusively centered on newly diagnosed or treatment-naive multiple myeloma patients and did not include relapsed/refractory multiple myeloma (RRMM) patients. ▪ Studies not targeting multiple myeloma (MM) patients.
I/C	I	▪ All interventions currently available for multiple myeloma are eligible for consideration.
	E	▪ Studies that do not mention treatment for multiple myeloma. ▪ Studies primarily involving stem cell transplantation (SCT) and total body irradiation before SCT as interventions when not for 2nd line of therapy.
O	I	▪ The study results must include at least one of the specified outcomes including safety, adverse events (AE), hospitalization information regardless of the cause, efficacy, or patient-reported outcomes).
	E	▪ Studies that lack reporting on any outcomes mentioned in the inclusion criteria.
S	I	▪ Original research study - Clinical trial study ▪ Original research study - Real world evidence study

PICOs: Population, Intervention/Comparison, Outcome, Study Type; I: Inclusion; E: Exclusion

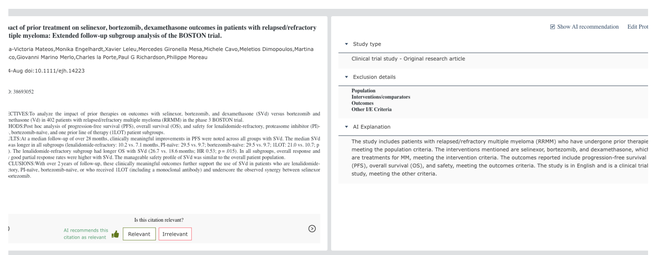
- The AI system mandates review of 10% of retrieved citations, or at least 30 abstracts for searches yielding fewer than 300 results. (**Figure 2A**).
- Abstracts are included (**Figure 2B**) or excluded (**Figure 2C**) based on PICOs criteria, with detailed explanations provided. Exclusion reasons may include wrong irrelevant population, intervention/comparison, outcomes, or study type.

Figure 2. AI-SLR abstract screening module in RRMM: (A) Summary page, (B) Example included abstract, (C) Example excluded abstract

(A)



(B)



(C)

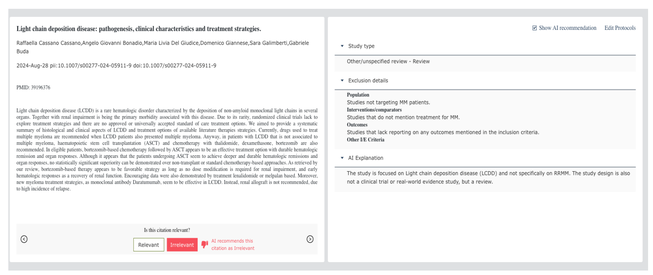
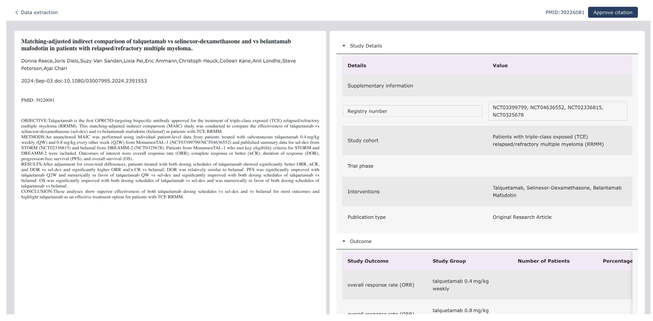


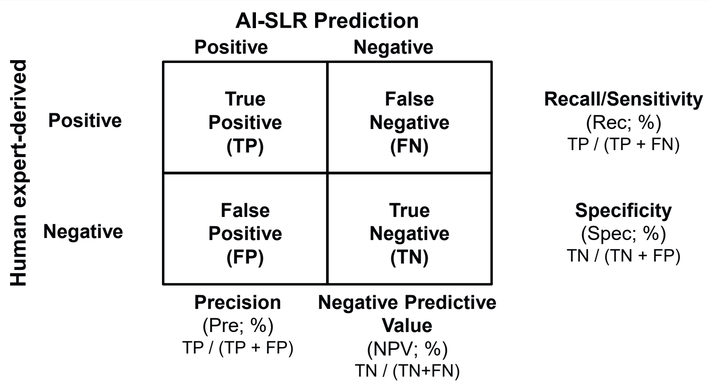
Table 2. Key Data Extraction Information

Information Type	Data Field
Study Details	Study cohort
	Interventions
	Publication type
	Study design
	Trial phase
	Supplementary information
Patient Characteristics	Age
	Gender
Study Outcomes	Study Outcome
	Group description
	Number of patients
	Percentage of patients
	Median
	Hazard ratio
	Other information

Figure 3. AI-SLR data extraction module: Example from an included abstract



EVALUATION



- **Accuracy (Acc, %):** (TP + TN) / (TP+FP+FN+TN)
- **F1 score (F1, %):** Harmonic mean of Precision and Recall
- **Cohen's κ:** Inter-rater agreement between two raters (accounts for agreement by chance)
- **Prevalence-adjusted bias-adjusted κ (PABAK):** Modified Cohen's κ adjusted for prevalence and bias
  - ≤ 0: No agreement
  - 0.01 – 0.20: None to Slight
  - 0.21 – 0.40: Fair
  - 0.41 – 0.60: Moderate
  - 0.61 – 0.80: Substantial
  - 0.81 – 1.00: Almost Perfect agreement
- Four evaluation sets compared expert-led reviews (ground truth) against AI-SLR in RRMM and advanced melanoma:
  1. Abstract screening (18 included / 49 randomly screened RRMM and 21/50 advanced melanoma abstracts)
  2. Exclusion reason validation
  3. Key data extraction from included abstracts (**Table 2**)
  4. Larger abstract screening (provided by two clinical SLR vendors typically used in HTA submissions)

RESULTS

Table 3. Performance of GPT-4 for screening titles and abstracts (Evaluation Sets 1 and 4)

Evaluation Set	Total Abstracts, N	Human-expert included, N	Rec (%)	Pre (%)	Spec (%)	F1 (%)	Acc (%)	Cohen's κ	PABAK
RRMM (Set 1)	49	18	89	80	87	84	88	0.74	0.76
Advanced Melanoma (Set 1)	50	21	90	90	93	90	92	0.84	0.84
RRMM (Set 4)	3665	2071	97	75	59	85	80	0.57	0.61
Advanced Melanoma (Set 4)	2753	145	82	60	97	69	96	0.67	0.92
Total (Macro) Performance			**90**	**76**	**84**	**82**	**89**	**0.71**	**0.78**

Table 4. Performance of GPT-4 in identifying specific exclusion criteria for RRMM abstracts (Evaluation Set 2)

Evaluation Category	Criterion	TP	FP	TN	FN	Rec (%)	Pre (%)	Spec (%)	NPV (%)	F1 (%)	Acc (%)
Population	Studies exclusively involving patients under the age of 18	47	0	1	1	97	100	100	50	98	97
Population	Studies exclusively centered on newly diagnosed or treatment-naïve MM patients and did not include RRMM patients	46	0	2	1	95	100	100	33	97	95
Population	Studies not targeting MM patients	45	0	3	1	97	100	100	75	98	97
Intervention/ Comparators	Studies that do not mention treatment for MM	44	1	4	0	100	97	80	100	98	97
Intervention/ Comparators	Studies primarily involving SCT and total body irradiation before SCT as interventions when not for 2nd line of therapy	45	0	3	1	97	100	100	75	98	97
Outcomes	Studies that lack reporting on any outcomes mentioned in the inclusion criteria	45	0	4	0	100	100	100	100	100	100
Study type	Studies not either clinical trials or real-world evidence study	22	3	24	0	100	88	88	100	93	93
Other	Studies not in English	49	0	0	0	100	100	N/A	N/A	100	100
					Macro Performance	**98**	**98**	**95**	**76**	**98**	**97**

N/A, Not Applicable: None of the abstracts were considered as either true negative or false negative in this criterion

Table 5. Performance of GPT-4 in extracting study details, patient characteristics, and study outcomes (Evaluation Set 3)

Evaluation Case		Abstracts, N	Data Fields, N	Rec (%)	Pre (%)	F1 (%)
RRMM	Study Details	18	144	100	100	100
	Patient Characteristics	18	36	100	100	100
	Study Outcomes*	18	95	83	88	86
Advanced Melanoma	Study Details	21	168	99	94	97
	Patient Characteristics	21	42	100	80	89
	Study Outcomes*	21	98	83	96	84
			Macro Performance	**94**	**91**	**93**

*Study Outcomes consist of capture of 7 data elements: Outcome, Group Description, Number of patients, % of patients, Hazard Ratio, Median, and Other relevant information. For AI-SLR system's extraction to be considered correct, it must include all 7 elements correctly

CONCLUSIONS

We developed a generalizable, end-to-end LLM based AI-SLR system. To our knowledge, this is the first time that PICO criteria, which are critical for any clinical SLR, have been introduced as a screening strategy to instruct an LLM. The system includes a human-in-the-loop module that displays LLM performance in real-time, allowing end users to adjust their prompts accordingly. The results showed high sensitivity, Cohen's κ, and PABAK for abstract screening, as well as a high F1 score for data extraction. Our system can potentially reduce the time, cost, and human errors associated with traditional SLRs, ultimately contributing to more timely and comprehensive evidence generation.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the following individuals for their contributions to the development of this project: Wenhui Wei and Andreas Kuznik for their valuable support; Stephen Curley and Saadia Alvi for kick-starting the internalization of the software; Luis Duarte, Narayan Ramakrishna, and the global IT team for enabling the internalization of the software; Deep Harthi for his work on poster development

REGENERON