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

Ying Li, PhD¹, Surabhi Datta, PhD², Kyeryoung Lee, PhD², Hunki Paek, PhD², Eefje Bergraph, MS¹, Julie Glasgow, PhD², Chris Liston, PharmD², Long He, MS², Majid Rastegar-Mojarad, PhD², Xiaoyan Wang, PhD², Yingxin Xu, BPharm, PhD¹ ¹Regeneron Pharmaceuticals, Inc., Tarrytown, NY, USA; ²Intelligent 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*, costeffectiveness, and humanistic data to inform their decisions.
- The current approach to SLR is generally time-consuming, laborintensive, 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

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

		AI recommendations	Human screened	Review disagreements		Human screene Relevant Irrel		Performance Precision Sensitiivity	Accept all AI recommendati
R	Relevant	151	10		Relevant		0		Download Results
In	relevant	198	25	AI recommendations	Irrelevant	0	25	100.00% 100.00%	
Accept	AI recommenda	ations Delete citations							
C	Citation	title				Mandatory	for review	AI recommen	dation < Screening <
•			on of talquetamab vs seline refractory multiple myelom	xor-dexamethasone and vs bela a.	antamab	False		Relevant	Not screened
•	Poor outcor ma cell leul		ments: A retrospective stud	y of 99 patients with primary ar	nd secondary plas-	False		Irrelevant	Not screened
•	Aplastic an	emia in association with i	multiple myeloma: clinical a	nd pathophysiological insights.*	•	True		Irrelevant	Irrelevant
•	Mosaic chro	omosomal alterations in h	nematopoietic cells and clinio	cal outcomes in patients with m	ultiple myeloma.	False		Irrelevant	Not screened
•			e myeloma mortality after lo Igs from an international col	w-level exposure to ionising rac hort study.	diation in nuclear	False		Irrelevant	Not screened
•	Current sta	atus of BAFF targeting im	munotherapy in B-cell neopl	asm.		False		Relevant	Not screened
•		al exposure to benzene a prospective cohorts of Cł		phoma in an extended follow-up	p of two popula-	False		Irrelevant	Not screened
nethasone (Vd) in HODS:Post hoc , bortezomib-naïv JLTS:At a media was longer in all s). The lenalidomi / good partial resp CLUSIONS:With :tory, PI-naïve, b	in 402 patients with analysis of progress ive, and one prior lin an follow-up of over subgroups (lenalidou ide-refractory subgr sponse rates were hig h over 2 years of fo	relapsed/refractory multiple my sion-free survival (PFS), overa te of therapy (1LOT) patient sul r 28 months, clinically meaning mide-refractory: 10.2 vs. 7.1 m roup had longer OS with SVd gher with SVd. The manageable Jlow-up, these clinically mean	eloma (RRMM) in the phase 3 BC II survival (OS), and safety for le ogroups. ful improvements in PFS were no onths, PI-naïve: 29.5 vs. 9.7; borte (26.7 vs. 18.6 months; HR 0.53; safety profile of SVd was similar	nalidomide-refractory, proteasome inhi ted across all groups with SVd. The mr romib-naïve: 29.5 vs. 9.7; 1LOT; 21.0 v p = .015). In all subgroups, overall res to the overall patient population. the use of SVd in patients who are lena	tomib and bitor (PI)- edian SVd ws. 10.7; p ponse and alidomide-	meeting the popula are treatments for	patients with re ation criteria. The MM, meeting the	e interventions mentioned are intervention criteria. The out	eloma (RRMM) who have undergone prior t selinexor, bortezomib, and dexamethasons iccomes reported include progression-free s tieria. The study is in English and is a clini
ortezomib.		Is this umends this us relevant	citation relevant?	anovi na obsi na jjangj oviven	o selinexor	study, meeting the			
) Light chai	citation a	s relevant 🤹 Relevan	t Irrelevant		3				☑ Show AI recommendation Edit Protoc
) Light chai Raffaella Ca Buda	citation a in deposition dis assano Cassano,4	sease: pathogenesis, clinie Angelo Giovanni Bonadio,M	t Irrelevant	iment strategies.	3	study, meeting the	other criteria.		
) Light chai Raffaella Ca Buda 2024-Aug-2 PMID: 39196: Light chain d organs: Togett explore treat summary of It multiple myel multiple myel mu	citation a in deposition dis assano Cassano, A 28 pii:10.1007/sC 28 pii:10.1007/sC 3376 leposition disease (LC ther with renal impai mark are recommed bistological and clini loma are recommed organ responses. Al corgan responses. Al ses, no statistically sig bortezonib-based the responses as a recove	Relevant Relevant ss relevant Relevant ssease: pathogenesis, clinic Angelo Giovanni Bonadio,M 00277-024-05911-9 doi:10 CDD) is a rare hematologic disor riment is being the primary mort here are no approved or univer- ical aspects of LCDD and treatr fed when LCDD patients also pri ic stem cell transplantation (4) bordy and treagents that the patien gaificant superiority can be demort rapy appears to be favorable str y of renal function. Encouraging	t Irrelevant cal characteristics and treat aria Livia Del Giudice,Domeni D.1007/s00277-024-05911-9 der characterized by the deposition bidity associated with this disease. Isally accepted standard of care trea nent options of available literature i esented multiple mytlema. Anywa h followed by ASCT appears to be an is undergoing ASCT seem to achie' strated over non-transplant or standa rategy as long as no dose modific idata were also demonstrated by tre	iment strategies.	riele	Study, meeting the Study, meeting the Study type Other/unspecified revi Exclusion details Population Studies not targeting i Interventions/comparator Studies that do not m Outcomes Studies that do not m Outcomes Studies that lack repo Other L/E Criteria AI Explanation	ew - Review MM patients. s ention treatment rting on any outco	for MM. osition disease (LCDD) and not	Show AI recommendation Edit Protoc

RESULTS

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

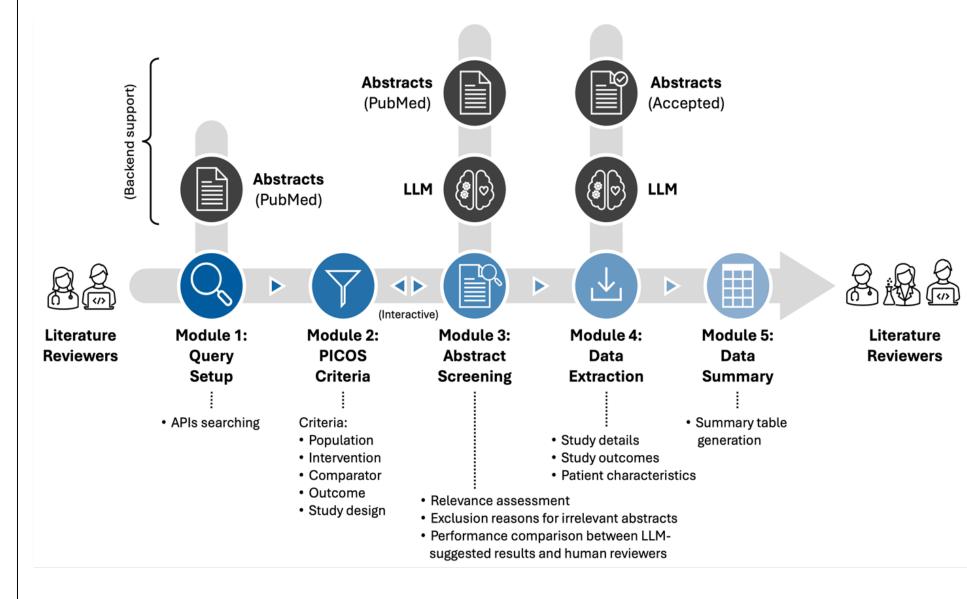
MSR234

	Total	Human- expert							
Evaluation Set	Abstracts, 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	o) Performa	nce	90	76	84	82	89	0.71	0.78

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.

Table 2. Key Data Extraction Information

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

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

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

Evaluation Category	Criterion	ТР	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 2 nd 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

 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

Ρ

I/C

Ο

S

- 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).
- 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.
- All interventions currently available for multiple myeloma are eligible for consideration.
- 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.
 - 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).

< Data extraction			PMID:39226081	Approve cit
Matching-adjusted indirect comparison of talquetamab vs selinexor-dexamethasone and vs belantamab mafodotin in patients with relapsed/refractory multiple myeloma.	 Study Details 			
Donna Reece,Joris Diels,Suzy Van Sanden,Lixia Pei,Eric Ammann,Christoph Heuck,Colleen Kane,Anil Londhe,Steve Peterson,Ajai Chari	Details		Value	
2024-Sep-03 doi:10.1080/03007995.2024.2391553	Supplementary information			
PMID: 39226081	Registry number		NCT03399799, NCT04636552, NCT023 NCT0325678	36815,
OBJECTIVE:Falquetanab is the first GPRCSD-targeting bispecific antibdoy approved for the treatment of triple-class exposed (TCE) relapsed/refractory multiple myelome (RRMM). This matching-adjusted indirect comparison (MAIC) study was conducted to compare the effectiveness of talquetamab vs selinexor-dexamethasone (sel-dex) and vs belantamab mafodotin (belamaf) in patients with TCE RRMM. METHODS:An unanehored MAIC was performed using individual patient-level data from patients treated with subcutaneous talquetamab 0.4 mg/kg weekly (QW) and 0.8 mg/kg every other week (QW) from MonumenTAL-1 (WCT0339799/NCT0436552) and published summary data for sel-dex from STORM (NCT0233618) and belamaf from DREAMM-2 (NCT0325678). Patients from MonumenTAL-1 who net key eligibility criteria for STORM and DREAMM-2 were included. Outcomes of interest were overall resonse rate (ORR), complete resonse or better (cRC), duration of response (DOR).	Study cohort		Patients with triple-class exposed (TCE relapsed/refractory multiple myeloma (·
ogression-free survival (PFS), and overall survival (OS). ESULTS: After adjustment for cross-trial differences, patients treated with both dosing schedules of talquetamab showed significantly better ORR, \geq CR, d DOR vs. sel-dx and significantly higher ORR and 2 CR vs belamaf; DOR was relatively similar to belamaf. PFS was significantly improved with	Trial phase			
talquetamab Q2W and numerically in favor of talquetamab QW vs sel-dex and significantly improved with both dosing schedules of talquetamab vs behamaf. QS was significantly improved with both dosing schedules of talquetamab vs sel-dex and was numerically in favor of both dosing schedules of talquetamab vs belamaf. CONCLUSION/These analyses show superior effectiveness of both talquetamab dosing schedules vs sel-dex and vs belamaf for most outcomes and highlight talquetamab as an effective treatment option for patients with TCE RRMM.	Interventions		Talquetamab, Selinexor-Dexamethason Mafodotin	ne, Belantamat
	Publication type		Original Research Article	
	✓ Outcome			
	Study Outcome	Study Group	Number of Patients	Percen
	overall response rate (ORR)	talquetamab 0.4 mg/kg weekly		
		talquetamah 0.8 mg/kg		

EVALUATION

		AI-SLR	Prediction	
		Positive	Negative	
Human expert-derived	Positive	True Positive (TP)	False Negative (FN)	Recall/Sensitivity (Rec; %) TP / (TP + FN)
Human exp	Negative	False Positive (FP)	True Negative (TN)	Specificity (Spec; %) TN / (TN + FP)
-		Precision (Pre; %) TP / (TP + FP)	Negative Predictive Value (NPV; %) TN / (TN+FN)	

- 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)

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, patientcharacteristics, 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
	Study Details	21	168	99	94	97
Advanced Melanoma	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.

- Studies that lack reporting on any outcomes mentioned in the inclusion criteria.
- 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.
- 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:
- 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**)
- Larger abstract screening (provided by two clinical SLR vendors typically used in HTA submissions)

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

