Generative artificial intelligence (GenAI)-based screening for systematic literature reviews – exploratory analysis of a novel repeatability score

Valerie Moss[a], George Miles[a], Lucia Giles[a], Gillian Sibbring[b], Brian Norman[a] and Bernard Kerr[a]
[a]Prime, London, UK, [b]Prime, Knutsford, UK
Presented by Tom Gegeny, Prime, New York, New York, USA

No. 24

Explore research in more detail by scanning the QR code. A video describing this research can be accessed on the mobile-friendly website version of this poster.

Conclusions

- Our analysis suggests that a repeatability metric may be useful to guide refinement of prompts, as well as identify records at increased risk of inaccurate screening due to unreliable responses
- GenAI responses with poor repeatability occurred frequently in our analysis
  - There are likely multiple underlying causes; for example, the variable language and complexity of literature may result in poor parsing of text, and there is potential for incomplete, inconsistent, and incorrect interpretation of complex prompts
- Screening performance for 3/4 PICOS domains was decreased in records with low repeatability decisions, suggesting that factors associated with reproducible responses also contribute to improved accuracy of outputs
- The GPT4o model showed higher repeatability for "yes" responses than "no" responses on most prompts; "unclear" was underutilized and had low repeatability
  - This suggests that the presence of specific relevant text in source material leads to a single response consistently having a high probability, whereas the absence of relevant text results in some or all possible responses having similar probabilities
- Supplementing categorical screening with additional approaches, such as text summarization, and implementing newer models with reasoning capabilities may help to overcome challenges associated with poor repeatability of responses
- Studies are needed to assess the incidence and impact of poor repeatability in other GenAI use cases

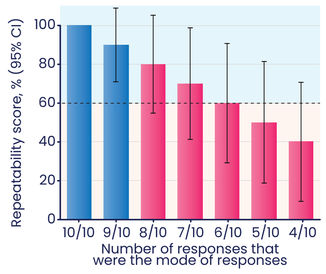
Introduction

- GenAI may allow large literature searches to be screened more quickly than by human reviewers alone
- Previously, we evaluated the performance of categorical screening of titles/abstracts by GenAI against population, intervention/comparator, outcome and study design (PICOS) inclusion criteria in terms of agreement with screening by a human reviewer[1]
  - Using the most common result from 10 screening runs resulted in a small improvement in performance relative to a single run
- This analysis assessed the prevalence of responses with poor repeatability, and investigated whether repeatability was associated with screening outcomes and agreement of GenAI responses with human review decisions

Methods

- A proprietary GPT4o-based tool was used to screen 279 titles/abstracts with 16 question prompts designed to assess aspects of PICOS domains
- Responses for each prompt were restricted to "yes", "no", or "unclear", and agreement with human review was assessed as previously reported[1]
  - Screening criteria for each domain were split into individual prompts to reduce complexity, and the GenAI screening decision for each domain was determined using Boolean logic
  - The population and intervention/comparator domains required a "yes" response from all component prompts for inclusion, and the study design domain required a "yes" response from 2 prompts (Q8 AND either Q6 OR Q7); the outcome domain required a "yes" response for any prompt (i.e. OR)
- Screening was repeated for a total of 10 runs, and the modal response for each prompt (the most common response from all runs) was used as the screening decision
- A repeatability score (RS) was calculated per prompt for each record, defined as the percentage of total responses that were the most common response
  - An RS of 100% indicated that all 10 responses had the same outcome (e.g. 10 yes)
  - An RS of 50% indicated that 5/10 responses were the most common response (e.g. 5 no, 4 yes, 1 unclear)
  - The lowest possible score was 40% (e.g. 4 unclear, 3 yes, 3 no)
- Responses were categorized as high-RS (≥90%) and low-RS (≤80%), with scores ≤80% having a lower bound of the 95% CI that encompassed <6 out of 10 responses (Figure 1)
- Agreement versus human review was assessed in high- and low-RS subgroups for each PICOS domain to investigate relationships between repeatability and accuracy, precision, and recall
  - Subgroups for PICOS domain decisions were based on the lowest RS score of the component prompts for that domain

Figure 1. Cutoff selection for high (blue) vs low (red) repeatability

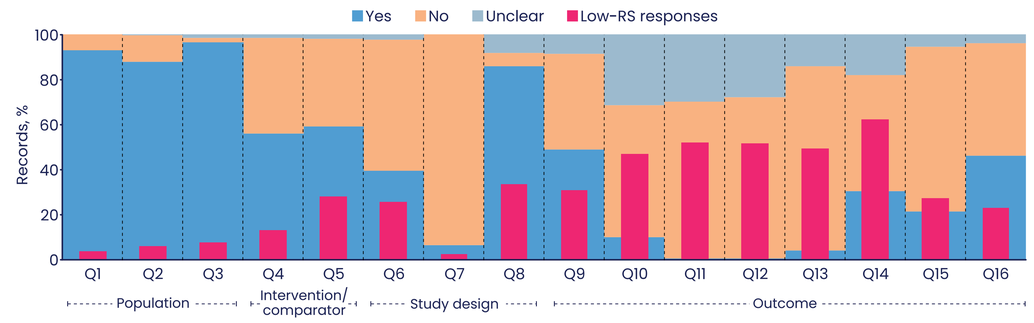


The dotted line indicates a majority of responses (6/10). The lower bound of the 95% CI for scores ≤80% crossed this threshold and were classified as low.

Results

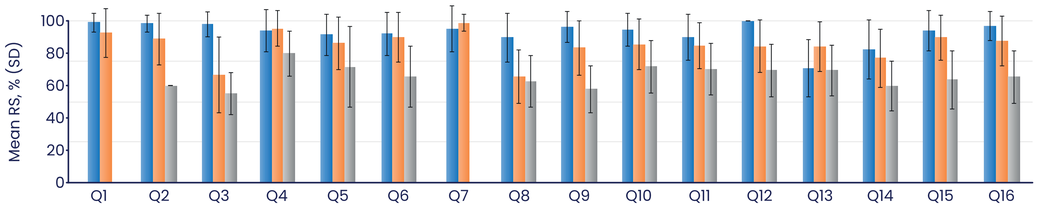
- Low-RS results occurred in 269 of the 279 records (96%); records had a median of 5 prompts with low RS results, and low RS results occurred in two or more PICOS domains in 164 (59%) records
- The percentage of low RS results varied considerably across prompts, with a median of 27% (range 3%–62%)
- The percentage of records with low RS was modestly associated with the percentage of records with "yes" responses for each prompt (Spearman's r = -0.6)

Figure 2. Percentage of low-RS responses (bars) and distribution of screening outcomes (area) across the 16 prompts



- "Yes" responses had higher mean RS than "no" for 14/17 prompts; responses of "unclear" had markedly lower RS than "yes" or "no" responses
- Utilization of the "unclear" response was generally low, comprising 2% of total responses across the population, intervention/comparator, and study design domains, and 17% of total responses in the outcomes domain

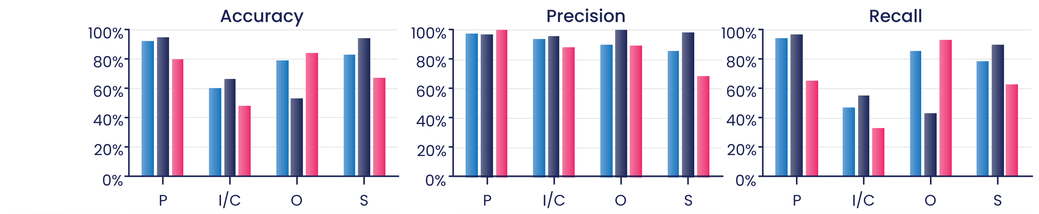
Figure 3. Mean RS by response outcome



	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16
Yes, n	259	245	269	156	165	110	19	239	137	28	2	2	12	85	60	129
No, n	20	33	6	118	108	162	260	17	118	163	194	199	228	144	203	139
Unclear, n	0	1	4	5	6	7	0	23	24	88	83	78	39	50	16	11

- Subgroups with a low-RS response on ≥1 prompt for each domain had reduced accuracy and recall versus subgroups with high minimum RS scores on the population, intervention/comparator, and study design domains
- Agreement results for the outcomes domain were not associated with low-RS responses due to combining results with a Boolean OR (versus AND for other domains)

Figure 4. Accuracy, precision and recall for domain-level decisions in subgroups by high vs low minimum RS



		Accuracy				Precision				Recall			
		P	I/C	O	S	P	I/C	O	S	P	I/C	O	S
Overall	%	93%	60%	80%	83%	98%	94%	90%	86%	94%	47%	86%	78%
	n/N	259/279	168/279	222/279	231/279	237/243	94/100	209/231	109/128	237/251	94/199	209/244	109/139
High RS	%	95%	67%	53%	94%	97%	96%	100%	99%	97%	55%	43%	90%
	n/N	227/239	122/183	24/45	152/161	222/228	71/74	16/16	72/73	222/228	71/129	16/37	72/80
Low RS	%	80%	48%	85%	67%	100%	88%	90%	69%	65%	33%	93%	63%
	n/N	32/40	46/96	198/234	79/118	15/15	23/26	193/215	37/54	15/23	23/70	193/207	37/59

FN, false negative; FP, false positive; TN, true negative; TP, true positive. Accuracy = (TP + TN)/(TP + TN + FP + FN); Precision = TP/(TP + FP); Recall = TP/(TP + FN).

Limitations

- Human decisions were made at the domain level whereas GenAI decisions were based on multiple prompts, precluding evaluation of performance at an individual prompt level
- The relatively small number of screening runs limits sensitivity to discriminate high- versus low-RS GenAI outputs, due to the small numerical difference between high repeatability and a chance result
- The impact of poor repeatability on overall screening outcome across all PICOS domains could not be assessed because low-RS decisions occurred in one or more domain for almost all records
- The small number of records in some screening decision and RS subgroups limits interpretability of results

References: 1. Miles G, Giles L, Kerr BC, Norman B, Sibbring GC. MSR201 Optimising Performance of Generative Artificial Intelligence (GenAI) in Systematic Literature Review (SLR) Screening Using PICOS Criteria. *Value Health*. 2024;27(12 Suppl): S478.

Acknowledgements: The authors thank the Synapse Studio and Editor teams at Prime for their support with developing this poster.

Disclosures: The research and interpretation provided here represent the views of the authors and are not necessarily those of their employer. Valerie Moss, George Miles, Lucia Giles, Brian Norman, and Bernard Kerr are employees of Prime, London, UK. Gillian Sibbring is an employee of Prime, Knutsford, UK.