MSR219

Can Artificial Intelligence (AI) Accurately Screen Abstracts in Systematic Literature Reviews?

Dmitry Gultyaev[1], **Noemi Hummel**[1], Linnea Koller[1], Lee Stern[2], Ananth Kadambi[2], Carl de Moor[3]

[1]Certara GmbH, Lörrach, Germany, [2]Certara Inc., Radnor, PA, USA, [3]GlaxoSmithKline, Philadelphia, PA, USA

Introduction

- One of the most time-consuming tasks in systematic literature reviews (SLRs) is the screening of abstracts according to PICOS (Population, Intervention, Comparator, Outcome, Study design) criteria.
- Publicly available generative pre-trained transformers such as ChatGPT4 can be prompted to answer questions on the content of abstracts,[1] thereby potentially facilitating abstract screening.

Objective

- In this study, we aimed to assess the accuracy and potential efficiency gains of abstract screening in SLRs supported by ChatGPT4.

Methods

- We evaluated ChatGPT4's performance in screening abstracts for previously performed SLRs in two indications: achondroplasia and advanced metastatic renal cell carcinoma (RCC).
- The PICOS schemes for inclusion and exclusion of the abstracts are shown in **Table 1**.
- Possible screening decisions were inclusion ('Yes'), exclusion ('No'), or 'Unclear', where ChatGPT4 was unable to determine inclusion or exclusion due to insufficient information provided in the abstract.
- ChatGPT4 and human decisions ('Yes'/'No') were compared, and precision, recall, F1 score (harmonic mean of precision and recall), and specificity were calculated
- Assuming abstracts included or excluded by ChatGPT4 will not require human verification, the proportion of abstracts determined as 'Yes' or 'No' by ChatGPT4 compared to the total number of abstracts screened was calculated to estimate maximum time savings of AI-supported abstract screening.

Table 1 PICOS schemes for the two SLRs

	Achondroplasia	RCC
Population	✓ Patients with achondroplasia	✓ Patients with renal cell carcinoma
Intervention	✓ Vosoritide	✓ Cabozantinib + nivolumab, cabozantinib, sunitinib, pazopanib, tivozanib, ipilimumab + nivolumab, axitinib + avelumab, axitinib + pembrolizumab, lenvatinib + pembrolizumab (all as first-line treatment)
Comparator	-	✓ Any of the intervention
Outcome	-	✓ Efficacy: OS, PFS, ORR, CR, PR, stable disease, time to discontinuation, time to deterioration ✓ PROs: EQ-5D-3L, EQ-5D-5L, NCCN FKSI-19, EORTC-QLQ-c30, FACT-G
Study design	✓ RCT, observational study ✓ Review ✗ Case report	✓ RCT ✓ SLR/meta-analysis of RCTs

CR: complete response, EORTC-QLQ-c30: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30, EQ-5D-3L/5L: EuroQoL-5 dimensions-3/5 levels questionnaire, FACT-G: Functional Assessment of Cancer Therapy - General NCCN FKSI-19: National Comprehensive Cancer Network Functional Assessment of Cancer Therapy Kidney Cancer Symptom Index - 19 Item Version, ORR: overall response rate, OS: overall survival, PFS: progression-free survival, PR: partial response, RCT: randomized controlled trial

Funding This study was funded by Certara.
Conflicts of interest None.

Results

- Among the 179 abstracts screened for achondroplasia, 38 were categorized as 'Unclear'. For the remaining 141 abstracts, precision was 0.81, recall was 0.95, F1 score was 0.88, accuracy was 0.91, and specificity was 0.90 (**Table 2**).
- Among the 551 abstracts screened for RCC, 83 were categorized as 'Unclear'. For the remaining 468 abstracts, precision was 0.72, recall was 0.73, F1 score was 0.72, accuracy was 0.87, and specificity was 0.91 (**Table 2**).
- Maximum time savings amounted to 79% for achondroplasia, and to 85% for RCC (**Figure 1**).

Table 2 Performance of AI-supported abstract screening

	Achondroplasia	RCC
# abstracts	179	551
% unclear	21.2	15.1
# in- or exclude	141	468
% false negative	1.1	5.6
Precision	**0.81**	0.72
Recall	**0.96**	0.73
F1 Measure	**0.88**	0.72
Accuracy	**0.92**	0.87
Specificity	0.90	**0.91**

Bolded numbers highlight the **best** performance among the tested case studies.

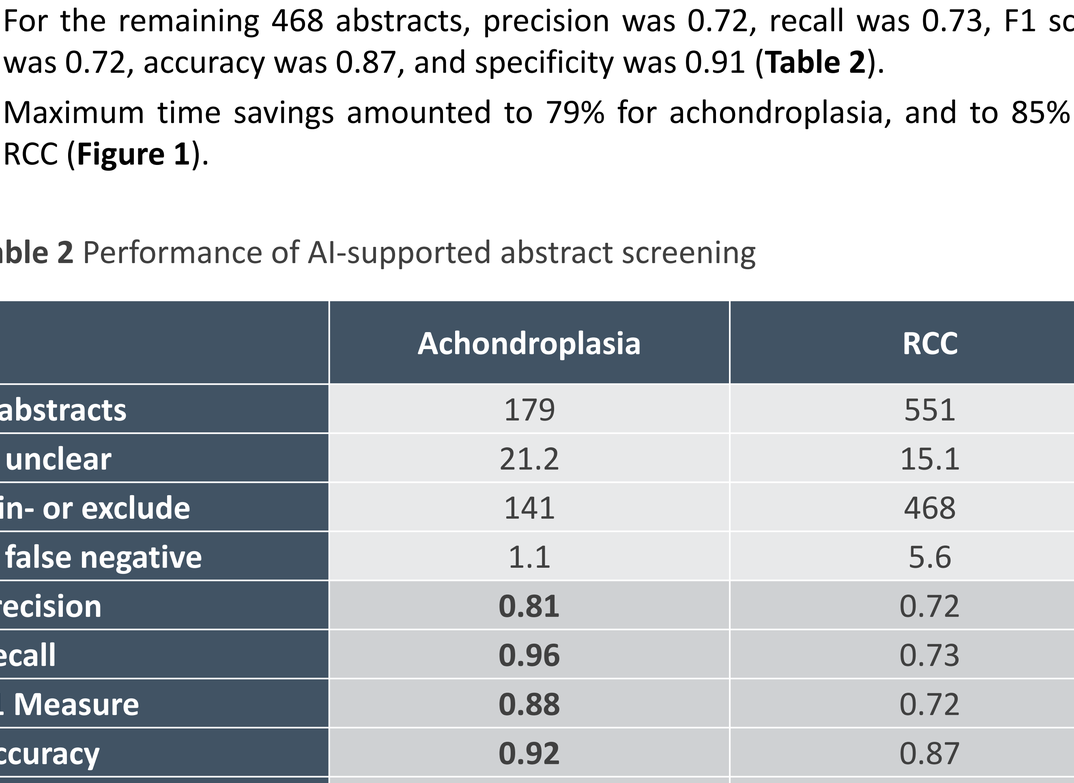

Figure 1 Time savings of AI-supported abstract screening

Time savings were defined as the proportion of abstracts determined to be included or excluded by ChatGPT4 of the total number of screened abstracts.

Conclusions

- ChatGPT4 is accurate (F1 > 0.7) and highly specific (specificity > 0.9) in abstract screening.
- Additionally, ChatGPT4 offers considerable potential time and subsequent cost savings and can be used to efficiently assess available evidence for multiple HEOR applications outside of health technology assessment (HTA) submissions, e.g., epidemiology, manuscript preparation, competitive intelligence, and maintenance of living SLRs.

References 1. Alshami A, Elsayed M, Ali E, Eltoukhy AEE, Zayed T. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems*. 2023; 11(7):351. https://doi.org/10.3390/systems11070351

Want to learn more?
<< Scan Here
Scan me