GenAI for critical appraisal of evidence for systematic literature reviews (SLRs): A face-off between GenAI and Human reviewers

Sahil Sharma[1], Sayyeda Anam[1], Rashi Tomer[1], Ashish Pandey[1], Sheetal Sharma[1], Larisa Gofman[2]
[1]ZS Associates, Gurugram, HR, India; [2]ZS Associates, Princeton, NJ, USA

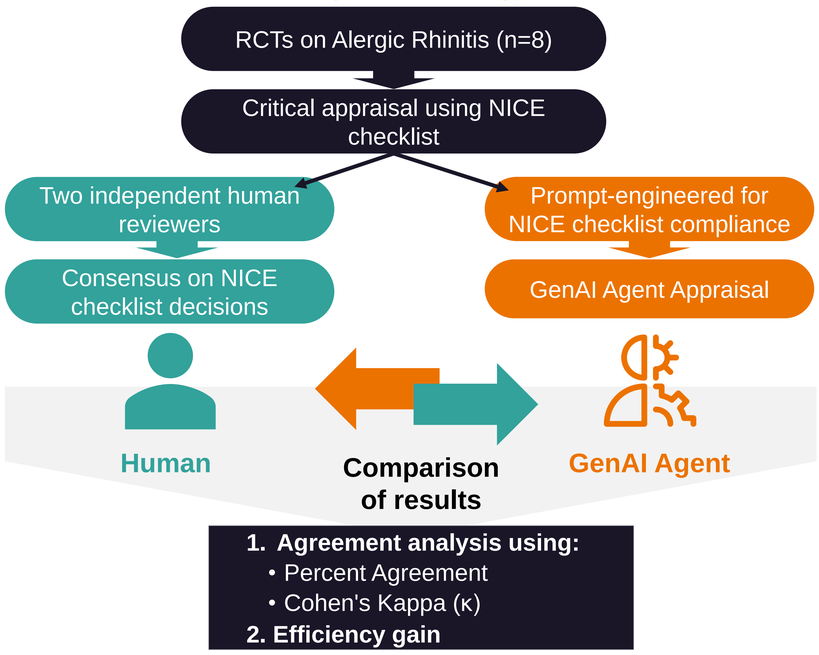
Background & Objectives

- Systematic literature reviews are crucial for guiding decision-making and ensuring the accuracy and reliability of research findings[1].
- In Health Economics and Outcomes Research (HEOR), the use of recommended checklists for critical appraisal is crucial to ensure that SLRs are based on high-quality, reliable evidence[2,3,4].
- Critical appraisal of studies allows researchers to identify potential biases, evaluate the robustness of study methodologies, and confirm the validity and generalizability of findings, thereby supporting evidence-based decision-making[2,3,4].
- The growing volume of scientific data highlights the need to incorporate Gen AI to analyze large datasets effectively in SLRs[5].
- GenAI, an advanced computer system capable of performing tasks that typically require human intelligence, can handle time-consuming activities with cognitive abilities, adaptability, and decision-making. Automating SLR stages with GenAI enables researchers to expedite reviews, reduce bias, and enhance transparency[5].
- This study compares the performance of human reviewers with a trained GenAI agent in appraising RCTs using the NICE checklist.

Methods

- A trained GenAI agent and two independent human reviewers appraised a set of eight RCTs focused on allergic rhinitis using the NICE checklist.
- The GenAI agent was developed and trained through prompt engineering to ensure adherence to the checklist. Key evaluation domain's as per NICE checklist included: performance bias, detection bias, selection bias, and attrition bias.
- The inter-rater agreement was assessed using Cohen's Kappa (κ) and as percent agreement between GenAI agent and the human reviewer[6]. Time investment was also calculated to assess efficiency gain.

Fig 1: Methodology



Results

- The agreement between the GenAI agent and human reviewers on the NICE checklist responses (Yes/No/Unclear) was 87.5% indicating promising accuracy.
- Domain-wise agreement was as follows: performance bias: 100%, detection bias: 85%, selection bias: 79.17%, and attrition bias: 75% (Figure 2). Cohen's Kappa (κ) was 0.401 (SE: 0.15, 95% CI: 0.108 to 0.695), indicating fair to moderate agreement (Table 2).
- The low Kappa value was due to both the GenAI agent and human reviewers' response as "Yes" for many NICE questions resulting in less variation in responses.
- This limited Kappa's ability to accurately capture the true level of agreement despite strong alignment.

Table 1. Confusion Matrix Comparing GenAI Agent and Human Reviewers Judgments

Confusion matrix			
Human vs GenAI Agent	Yes	No	Unclear
Yes	**92**	0	8
No	0	**2**	0
Unclear	6	0	**4**

- This confusion matrix (Table 1) compares the GenAI agent's responses (columns) to human reviewer judgments (rows) across the NICE checklist.
- Majority of responses (n=92) were consistent for "Yes" judgments. Minor discrepancies were observed in the "Unclear" category.

Table 2. Agreement between GenAI and human reviewers on NICE checklist response

Cohen's Kappa (κ)	**0.401**	**Fair to moderate agreement**
Standard error	0.15	
95% CI	0.1083 to 0.6948	

High prevalence of "Yes" responses by both GenAI and human reviewers' limits variability, lowering Kappa despite strong agreement

- The total time investment to complete critical appraisal of eight studies was notably lower for the GenAI agent compared to a human reviewers (Figure 3).
- The GenAI agent demonstrated a 57.3% efficiency gain compared to the human reviewer, when accounted for one-time GenAI agent setup, prompt engineering and optimization.
- The efficiency gain for completing the critical appraisal checklist was 96.7% (without the one-time GenAI setup) and the tool was suitable for other indications/projects without the need for re-training.

Fig 2: Raw agreement between GenAI agent and human reviewers on critical appraisal of RCTs using NICE checklist

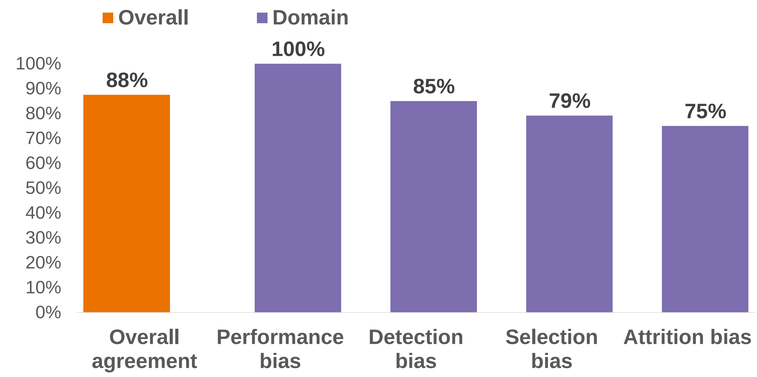
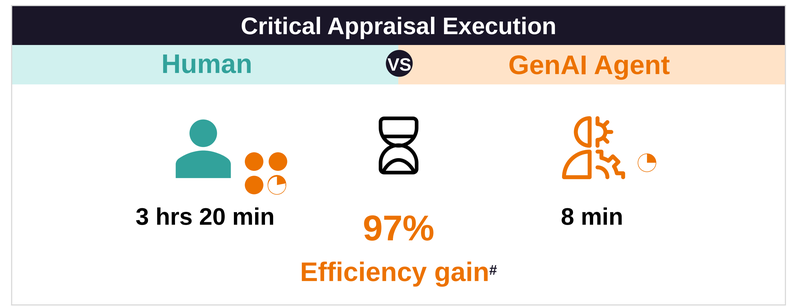


Fig 3. Time Investment Comparison – Human Reviewers vs. GenAI Agent for Critical Appraisal (8 Studies)



#Not inclusive of time spent on prompt engineering and GenAI setup

Conclusions

- The GenAI agent, trained on the NICE checklist, can serve as a reviewer for critical appraisal of RCTs, streamlining the SLR process.
- Human quality assurance remains essential to validate AI outputs and handle study-specific complexities.
- The agent's alignment with a standardized framework like NICE enables potential use across multiple disease areas.
- With further refinement, this approach can drive broader automation in HEOR, including future integration into data extraction workflows.

References

1. Boren SA, Moxley D. Systematically reviewing the literature: building the evidence for health care quality. Mo Med. 2015;112(1):58-62.
2. University College London. Describing and appraising studies - Systematic reviews - Library guides and databases. UCL Library Services. https://library-guides.ucl.ac.uk/systematic-reviews/describing-appraising. Published 2024. Accessed April 28, 2025.
3. Clapton J, Sami J. Dissecting the literature: the importance of critical appraisal. Royal College of Surgeons of England.https://www.rcseng.ac.uk/library-and-publications/library/blog/dissecting-the-literature. Published March 31, 2023. Accessed April 28, 2025.
4. Tod, D., Booth, A., & Smith, B. (2021). Critical appraisal. International Review of Sport and Exercise Psychology, 15(1), 52–72. https://doi.org/10.1080/1750984X.2021.1952471
5. Atkinson, C. F. (2023). Cheap, Quick, and Rigorous: Artificial Intelligence and the Systematic Literature Review. Social Science Computer Review, 42(2), 376-393. https://doi.org/10.1177/08944393231196281
6. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282.