Piloting a GenAI Agent for Structured Quality Appraisal of Network Meta-Analyses Using the ISPOR-AMCP-NPC Checklist

MSR167

Sahil Sharma[1], George Xydopoulos[2], Larisa Gofman[3], Nils Fisher[4], Molebedi Segwagwe[5]

[1]ZS Associates, New Delhi, India, [2]ZS Associates, Cambridge, UK, [3]ZS Associates, Princeton, NJ, USA, [4]ZS Associates, Boston, MA, USA, [5]ZS Associates, London, UK

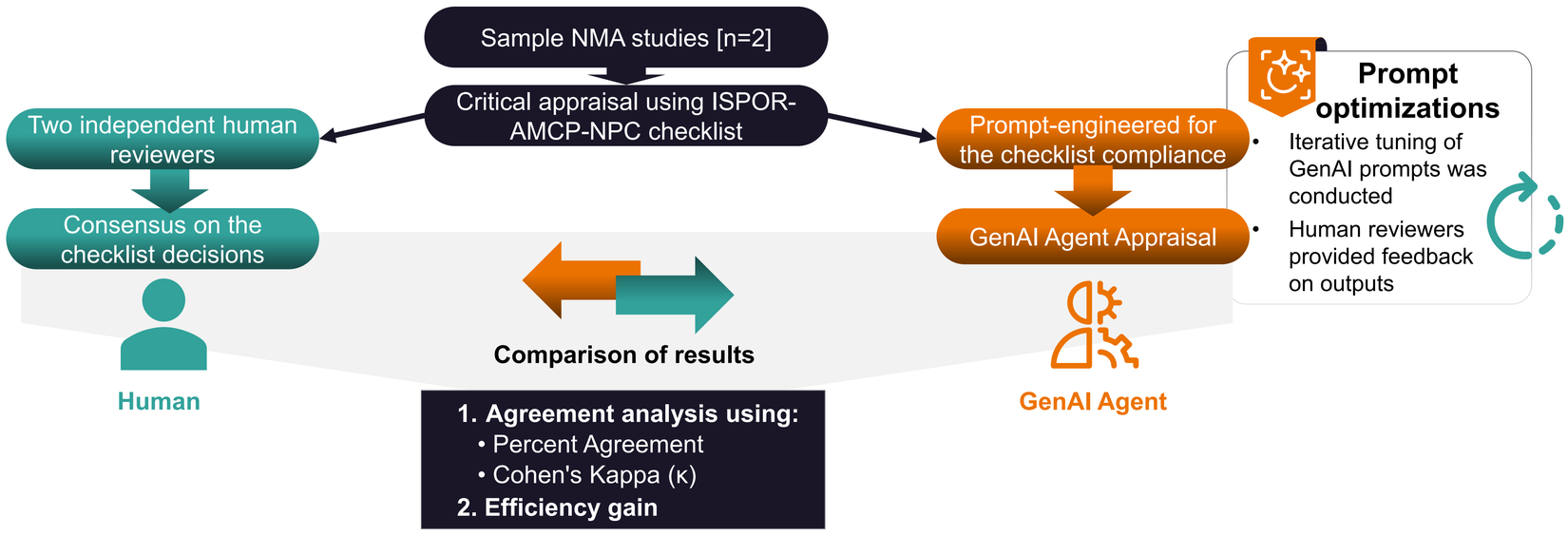
Background & Objectives

- **HTAs** and **EU JCAs** emphasize on **transparent and credible evidence**. Given the importance of **Indirect Treatment Comparisons (ITCs)** and **Network Meta-Analyses (NMAs)** in comparative effectiveness research, **structured, high-quality appraisal** of these studies is essential for reliable decision-making
- The vast and complex nature of NMA literature presents a challenge to ensuring consistent evaluation
- Our primary objective was to overcome the hurdles of scaling and standardizing NMA appraisal. We **developed and piloted a GenAI agent** trained on the **ISPOR-AMCP-NPC Checklist**[1]
- The goal was to use this agent to **support consistent and scalable appraisal** by testing its ability to reliably match human expert judgment, thus streamlining the evidence synthesis process

Methods

- The GenAI agent was developed using prompt engineering, embedded with guardrails, and informed by detailed training on the checklist's structure and interpretative guidance
- The agent was piloted on two published NMA studies in hormone receptor-positive (HR+)/HER2-negative advanced breast cancer (ABC)[2,3]
- Independent assessments by an experienced human reviewer served as the gold standard
- The agent's responses were compared with human evaluations across six checklist domains using percent agreement and Cohen's kappa[4]

Fig 1: Methodology for pilot testing of GenAI agent



Results

- **High overall agreement:** The GenAI agent demonstrated strong concordance with human reviewers, achieving an overall raw agreement of **84%** across all checklist items
- **Moderate inter-rater reliability:** The **Cohen's Kappa coefficient was 0.581 (95% CI: 0.336–0.827)**, indicating moderate agreement after accounting for chance
- **Consistent agreement across domains:** Domain-level agreement ranged from **80% to 90%**, with the highest alignment observed for **conflict of interest (90%)** and **interpretation (85%)** domains
- Most overlap was observed in the "Yes–Yes" category (**72 out of 104 observations**), indicating strong consensus between GenAI and human reviewers, with minimal disagreement across "No" and "Other" responses

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

Confusion Matrix (Human vs GenAI Agent)	Yes	No	Other*	Total
Yes	**72**	4	2	78
No	7	**10**	**1**	18
Other*	1	2	**5**	8
Total	80	16	8	**104**

*Other responses included 'not applicable' , 'not enough information '& 'not reported'

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

Cohen's Kappa (κ)	**0.581**	**Moderate agreement**
Standard error	0.125	
95% CI	0.336 to 0.827	

High prevalence of "Yes" responses by both GenAI and human reviewers' limits variability, lowering Kappa despite strong agreement[5]

Fig 2: Raw agreement between GenAI agent and human reviewers on critical appraisal of the NMA studies using ISPOR-AMCP-NPC checklist

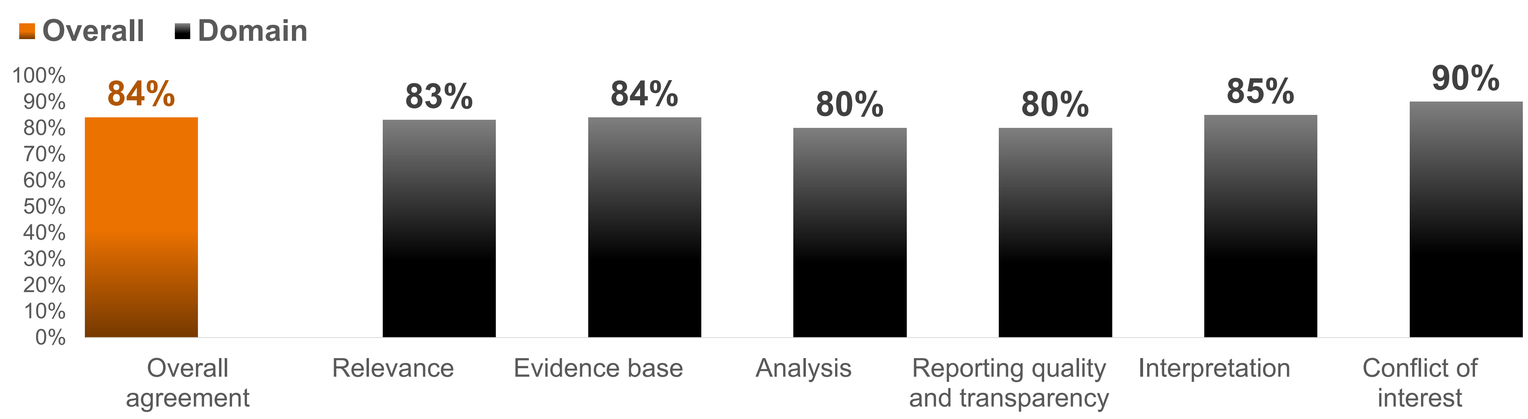


Fig 3. Time Investment Comparison – Human Reviewers vs. GenAI Agent for Critical Appraisal (2 Studies)

Efficiency Gain

- **Human review time:** A skilled NMA expert required **~16 hours in total** to complete the ISPOR–NPC–ITC checklist with rationale documentation for each item
- **GenAI agent efficiency:** The GenAI agent required **~1 hour for one-time prompt optimization** and **~4 minutes total** for automated appraisal and rationale generation

98%

Efficiency gain

Human	vs	GenAI Agent
~16 hours		~1 hour

The optimized agent could be applied to multiple articles without additional prompt development

Conclusion

- This pilot demonstrates the feasibility of deploying a GenAI agent to support quality assessment of NMAs using established checklists
- While early results are promising in core domains, further optimization is underway to improve performance in complex or assumption-prone areas
- The approach shows strong potential to enhance efficiency and consistency in HTA/JCA evidence review processes, ultimately supporting timely access to innovative therapies

References

1. Berger ML, Martin BC, Husereau D, et al. A questionnaire to assess the relevance and credibility of observational studies to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health. 2014;17(2):143-156. doi:10.1016/j.jval.2013.12.011
2. Kappel C, Elliott MJ, Kumar V, Nadler MB, Desnoyers A, Amir E. Comparative overall survival of CDK4/6 inhibitors in combination with endocrine therapy in advanced breast cancer. Sci Rep. 2024;14(1):3129. Published 2024 Feb 7. doi:10.1038/s41598-024-53151-8
3. Jhaveri K, O'Shaughnessy J, Fasching PA, et al. Matching-adjusted indirect comparison of PFS and OS comparing ribociclib plus letrozole versus palbociclib plus letrozole as first-line treatment of HR+/HER2- advanced breast cancer. Ther Adv Med Oncol. 2023;15:17588359231216095. Published 2023 Dec 14. doi:10.1177/17588359231216095
4. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282.
5. Zec S, Soriani N, Comoretto R, Baldi I. High Agreement and High Prevalence: The Paradox of Cohen's Kappa. Open Nurs J. 2017;11:211-218. Published 2017 Oct 31. doi:10.2174/1874434601711010211