Leveraging Generative Artificial Intelligence for Assessing the Quality of Network Meta-Analysis: Methodological Considerations and Early Findings

Nevière Agathe[1], Friedrich Gabriela[2], Papadimitropoulou Katerina[3], Le Nouveau Pauline[1], Gauthier Aline[2]
[1]Amaris Consulting, Saint-Herblain, France, [2]Amaris Consulting, Barcelona, Spain, [3]Amaris Consulting, Lyon, France

#146529
MSR184

INTRODUCTION

- In recent years, we have seen the rise of **large language models** (LLMs), such as the Generative Pre-trained Transformer 4 omni (GPT-4o), which complement human efforts across a wide range of scientific tasks.
- Recent studies have evaluated the use of generative artificial intelligence (AI) in conducting and assessing **literature reviews**.(1,2)
- This approach could also be expanded to support the critical review of statistical manuscripts, including **network meta-analyses** (NMAs).

OBJECTIVES

- This study aimed to **evaluate the ability of GPT-4o to critically assess the quality of NMAs and compare the assessment to experts' conclusions**.

METHODS

- **A checklist** was developed following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) - NMA guidelines (3) and Pacou *et al.,*(4) to **assess the robustness of published NMAs**. The checklist is presented in Table 1 below.
- This checklist was provided to GPT-4o with a **generic zero-shot prompt** requesting a response to each item, along with a global assessment.
- The **performance of the LLM was evaluated based on its level of agreement** with experts' assessment as follows: "*agreement*" was defined when expert and GPT-4o provided similar answers, "*some agreement*" when overall assessment were aligned but details in responses were different, and "*disagreement*" when experts and GPT-4o were not aligned.

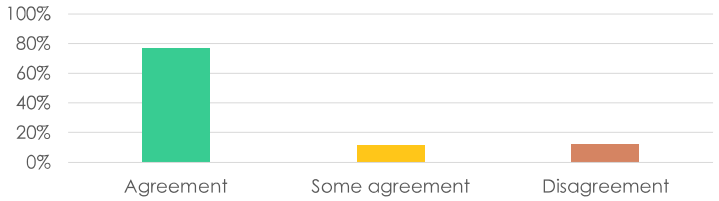
Table 1. *Checklist assessing the quality of published NMAs*

Checklist	Importance
Methods	
Q1. Is the protocol registered in PROSPERO?	**Low**
Q2. Was a systematic literature review (SLR) conducted to inform the indirect treatment comparison (ITC)?	**High**
Q3. Are the PICOS (population, intervention, comparator, outcome, study design) detailed?	**High**
Q4. Was any risk of bias assessment conducted?	**Low**
Q5. Were the population characteristics at baseline compared across trials?	**High**
Q6. Were heterogeneity and inconsistency assessed from both a statistical and clinical standpoint?	**High**
Q7. Were treatment effect modifiers identified and how? Was their distribution assessed across trials?	**High**
Q8. Was the statistical method for the ITC a non-naïve method?	**High**
Q9. Was the rule to select a fixed- vs. random-effect model reported?	**Medium**
Q10. Was a sensitivity analysis conducted?	**Medium**
Q11. Is the software reported?	**Low**
Q12. Are the codes available?	**Low**
Results	
Q13. Are the number of studies screened, assessed for eligibility, and included in the review reported? Is the PRISMA diagram reported?	**Low**
Q14. Is there at least one reference for each included trial?	**Low**
Q15. Are the data inputs presented?	**Medium**
Q16. Is there the same number of trials in data inputs and in the network?	**Medium**
Q17. Is there any heterogeneity detected? If yes, are results from sensitivity analyses or adjustment methods reported?	**High**
Q18. Is there any loop in the network? If yes, is there any inconsistency detected? If yes, are results from sensitivity analyses or adjustment methods reported?	**High**
Q19. Is the risk of bias reported? If yes, is there any bias detected?	**Medium**
Q20. If a Bayesian network is used, was the convergence assessed?	**High**
Q21. Were results of the meta-analysis presented as point estimate with 95% CrI/CI?	**Low**
Discussion	
Q22. Are the results compared to published NMAs?	**High**
Q23. Are there any comments on the validity of the assumptions (transitivity, consistency)?	**High**
Q24. Are there any comments or any concerns regarding network geometry or convergence?	**High**

RESULTS

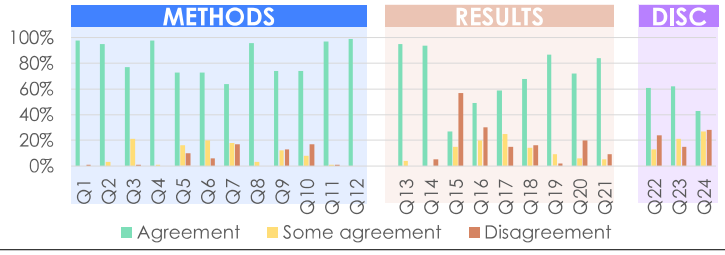
- **100 published NMAs** from 2016 to 2021 identified in a previous work (5) were independently assessed according to the checklist by GPT-4o and by four independent ITC experts.
- Overall, we obtained **77% of agreement** between GPT-4o and experts, 11% of partial agreement, and 12% of disagreement (see Figure 1). Those results showed a benefit in using LLM for review.

Figure 1. *Overall agreement results between GPT-4o and ITC experts*



- To understand GPT-4o performance, agreements are summarized by item (see Figure 2). **GPT-4o was found to perform well on closed and straight-forward questions** such as Q1, Q2, Q4, Q12, Q13 and Q14. However, **difficulties were observed** for Q15, Q16, Q17 and discussion points which are open questions or referring to figures.
- Moreover, it was observed that GPT-4 **struggled to identify low-quality NMA**, which was characterized by the absence of heterogeneity assessment, inconsistency checks, and sensitivity analyses. In 14 cases where experts marked heterogeneity assessment as "not reported," GPT-4 consistently reported an assessment.
- Experts also warn that GPT-4o can display **excessive confidence** in its responses, even when providing incorrect information. This can lead to overconfidence in reviews generated by LLMs and may diminish users' perspective and critical insight.

Figure 2. *Agreement results between GPT-4o and ITC experts per question*



CONCLUSIONS

- The current work aimed to evaluate the quality for LLM in the quality assessment of published NMA following a checklist.
- Overall, **results were aligned** between ITC experts and GPT-4o. However, **limitations** of assessment were exhibited when the points were associated with an open question in the provided checklist.
- The study currently uses zero-shot prompting, but exploring other **prompt engineering strategies** could improve the performance of the experiment by better customizing the LLM to cover the limitations found.
- For example, **few-shot prompting** could provide the model with examples to increase accuracy, and **chain-of-thought prompting** could enhance reasoning capabilities by guiding the AI step-by-step.

REFERENCES

(1) Shuai W, et al. 2023. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23). Association for Computing Machinery, New York, NY, USA. **(2)** Haman M, et al. (2023). Using ChatGPT to conduct a literature review. Accountability in Research, 31. **(3)** Hutton B, et al. (2015) The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-analyses of Health Care Interventions: Checklist and Explanations. Ann Intern Med. **(4)** Pacou M, et al. (2016). Proposed Checklist For Non-Statisticians To Assess The Quality Of A Network Meta-Analysis In The Context Of A Nice Submission. Value in Health. 19. A98. **(5)** Spineli L.M., et al. (2024) Low awareness of the transitivity assumption in complex networks of interventions: a systematic survey from 721 network meta-analyses. BMC Med 22, 112.

DISCLOSURES

AN, GF, KP, PLN and AG are employees of Amaris Consulting. Authors have no conflict of interest to declare.

Presented at: ISPOR Europe 2024; 17-20 November 2024; Barcelona, Spain