MSR33

# Utilizing generative artificial intelligence in network meta-analysis: Assessing the effectiveness of GenAI as a tool in feasibility assessments

Pierce P[1], Kraan CW[1], Petersohn S[1], Bennison C[2], Kroep S[1], Nickel K[3]

[1]OPEN Health HEOR & Market Access, Rotterdam, Netherlands, [2]OPEN Health HEOR & Market Access, London, UK , [3]OPEN Health HEOR & Market Access, Berlin, Germany

OPEN HEALTH

## INTRODUCTION

- Network meta-analyses (NMAs) are routinely used in health technology assessments (HTA) to assess the comparative efficacy, and safety, of treatments lacking head-to-head trials.
- The feasibility assessment (FA) stage, which involves synthesizing the numerous qualitative and quantitative data sources, can be time-consuming.
- The application of generative artificial intelligence (GenAI), through large language models (LLMs), in optimizing manual evaluation is rapidly expanding:
  - GPT-4, a LLM, extracted population, intervention, comparator, outcomes, and study design (PICOS) data from 682,667 abstracts with 98% accuracy in a recent study.[1]
  - Reason et al. (2024) utilized python script and application program interface (API) to prompt GPT-4 to automate the extraction of binary data and time-to-event outcomes in four case studies, resulting in an accurate extraction of over 99% of the data across 20 runs for each case study.[2]
- There is potential to streamline the FA stage of the NMA process using GenAI.

**Objective:** This study investigates the reliability of Gemini 1.5 Pro (Gemini), an LLM with a 1 million token context length, in expediting and enhancing the FA stage of an NMA compared to traditional, manual evaluation.

## METHODS

Three previously conducted, manually produced FAs for NMAs in urothelial carcinoma (UC), renal cell carcinoma (RCC) and chronic lymphocytic leukemia (CLL) were used as case studies.

### Prompt generation in Gemini

Each stage of the FA process was carried out in a separate 'chat', that refers to the user's input (prompt) and the conversation that unfolds (chat), as depicted in **Fig.1**. The following three stages from the FA process were carried out for each case study:

**1.** Treatment effect modifier (TEM) and prognostic variable (PV) identification: Generate a relevant list of TEMs and PVs for the respective indication.

**2.** Data extraction & synthesis: Generate separate tables for baseline characteristics (BLC) and PICOS by extracting and synthesizing data from the included trial publications for each case study.
  - The tabulated data was assessed by the analysts for accuracy using a stoplight rating system[a]

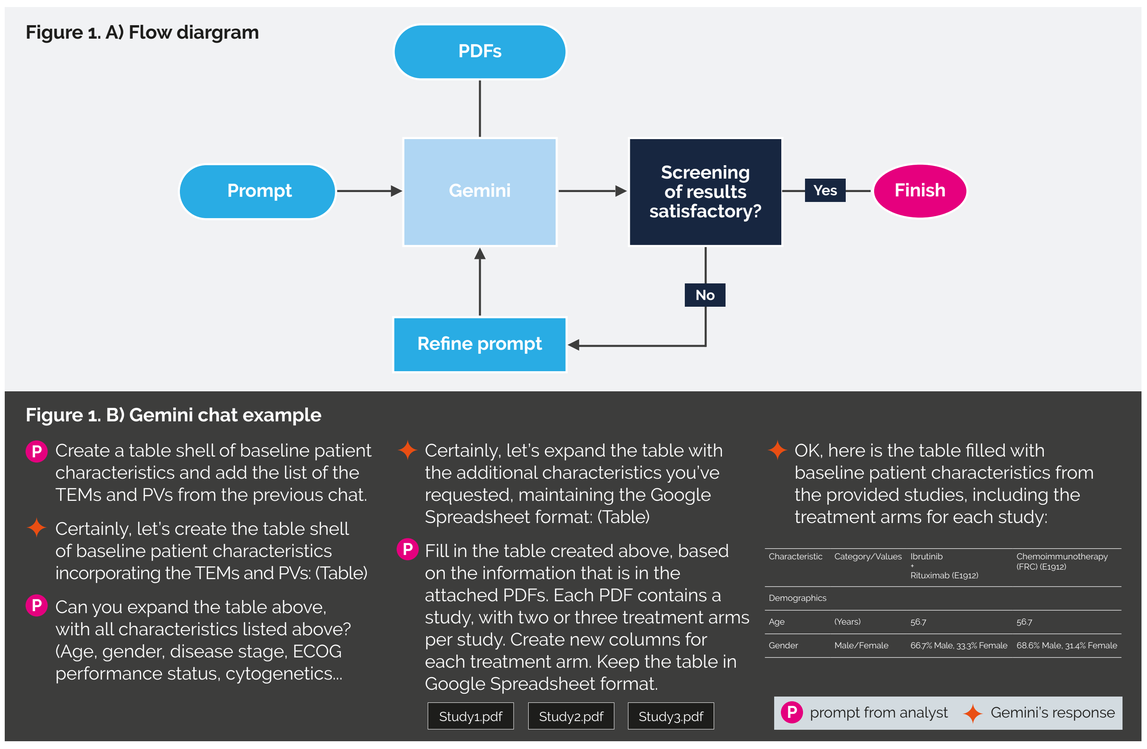
**3.** Similarity assessment: Assess the heterogeneity of the trials used in the case study.
  - The original case study's BLC and PICOS were uploaded to Gemini for assessment.
  - A ranking system with corresponding explanations per variable and overall summary was generated by Gemini.

A zero-shot approach was used for stage 1 and 2, wherein the LLM was prompted to perform a task without task-specific training data. In stage 3, Gemini was provided with guidance on similarity assessments.[3,4]

### Assessment of results

- Results from each chat were assessed quantitatively and qualitatively by comparing to the publications and the respective manual FAs for accuracy and completeness.
- The process was completed by three analysts, who documented the evolution of their prompts and results for each of the stages.

**Figure 1.** Process flow of generating results for the stages of FA in individual Gemini chats




## METHODS

### TEMs and PVs

- The TEMs and PVs generated were largely in line with those obtained in the corresponding case study, with Gemini being able to correctly identify 60-80% of the TEMs and PVs (**Table 1**).
- Gemini generated detailed responses, including categorization of TEM or PV, and specifications of nuances such as ECOG being favored over Karnofsky for measuring performance status.

**Table 1.** Overview of the identified TEMs and PVs per case study

| | UC Case Study | RCC Case Study | CLL Case Study | Total |
|---|---|---|---|---|
| Manual, total reported (N) | 10 | 10 | 8 | 28 |
| Gemini, total identified (N) | 15 | 14 | 7 | 36 |
| Matching Gemini & Manual (N) | 6 | 8 | 6 | 20 |

## RESULTS

### Data extraction & synthesis

BLC tables for all three case studies had inconsistent accuracy, particularly for demographics reported diversly across publications (i.e. tumor location, previous therapies), see **Fig. 2**.

Continuous or binary variables (i.e. age, gender, gene expression) were reported more reliably in BLC, and PICOS tables were more successfully synthesized in both accuracy and completeness, see **Fig. 3**.

In total, across all case studies, baseline characteristics were extracted by Gemini 58% correct, 16% partially correct and 26% incorrect. For PICOS, 87% correct, 5% partially correct, and 8% incorrect.

**Figure. 2** Accuracy of synthesized baseline characteristics data, by case study

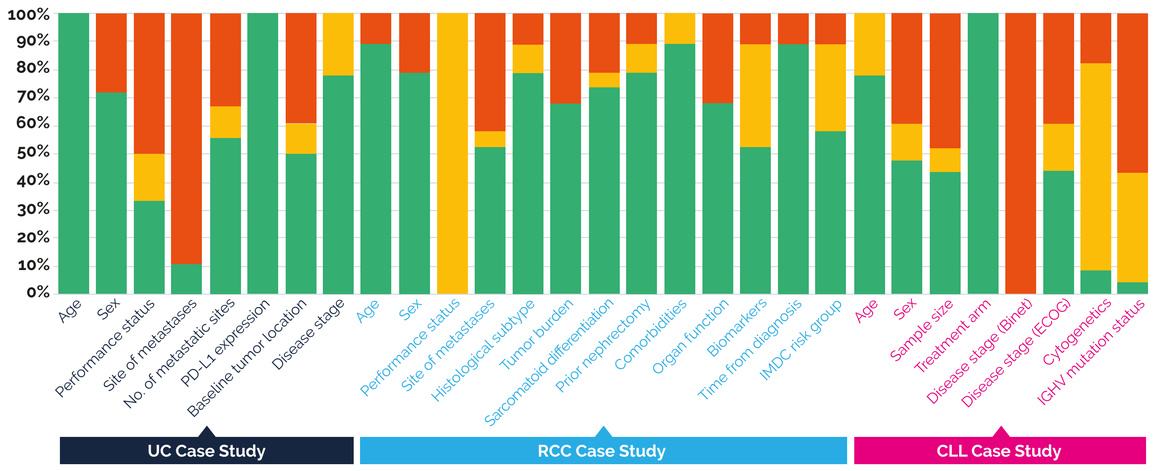
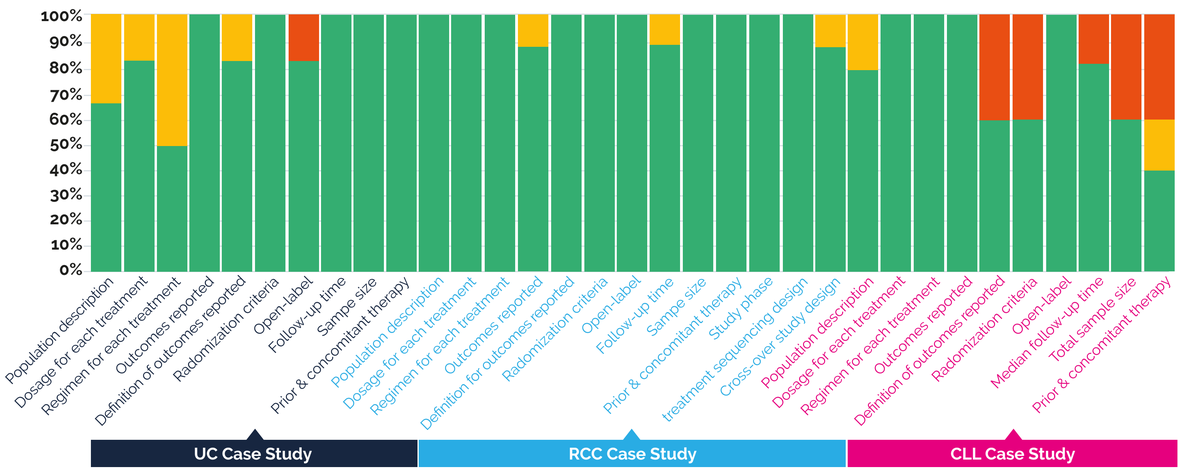



**Figure. 3** Accuracy of synthesized PICOS data, by case study




[a]Green indicates completely correct extraction, yellow indicates partially correct (i.e. presenting a portion of the reported data correctly), and orange indicates entirely incorrect (i.e. reported that data was missing when it was present, or reporting incorrect values)

### Similarity assessment

- Gemini's heterogeneity grading system included a rank from 1-3 by variable (no description of the ranking system was provided), corresponding assessment of the heterogeneity, suggestions how to interpret (i.e. consider with caution due to it being a known TEM), and an overall summary.
- The content of each case study's assessment was frequently inaccurate, even after prompt refinement.
  - For example, when prompted to update the results with descriptions of any outliers, in the UC case study an incorrect publication was flagged for median age.

## DISCUSSION

### Limitations & future directions for Gemini use case

- A major limitation is the inconsistency in results, both within single prompts and re-running of chats:
  - While Gemini can process and understand language, it struggles to interpret nuances and subtleties, resulting in different or incorrect extractions for the same variable across different publications.
  - Replication of results is challenging; the same prompt can generate different results at different moments.
- Gemini generated incorrect or misleading results for all case studies. This "hallucination" occurred primarily in the BLC extraction and similarity assessment stages.
  - For example, the BLC extraction results for the RCC case study appeared acceptable, presenting plausible values for each variable, however upon inspection, some values were not only incorrect, but nonexistent in the publication.
- Based on the limitations of this study and the findings from previous, similar, studies, future directions for research should focus on:
  - Training the LLM by providing examples or training data, specific to the task.
  - Different methods for data extraction and synthesis, i.e. uploading documents in different formats like markdown, generating results for individual publications rather than several at once.

## CONCLUSIONS

- Gemini is a promising GenAI tool for optimizing FA stages compared to manual processes, especially for the identification of TEMs and PVs, and data extraction of binary and continuous variables.
- Results generated should be carefully reviewed as hallucinations are persistent, thus in its current stage it is not a reliable tool, however different methodological strategies should be explored in future studies.

## REFERENCES

**1.** Reason, T., J. Langham, and A. Gimblett, *Automated Mass Extraction of Over 680,000 PICOs from Clinical Study Abstracts Using Generative AI: A Proof-of-Concept Study*. Pharmaceut Med, 2024. 38(5): p. 365-372; **2.** Reason, T., et al., *Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models*. PharmacoEconomics - Open, 2024. 8(2): p. 205-220; **3.** Patel, D., et al. *A practical approach to concepts and methods used to assess heterogeneity and inconsistency in network meta-analyses*. in *ISPOR*. 2015. Philadelphia, PA, USA; **4.** Cope, S., et al., *A process for assessing the feasibility of a network meta-analysis: a case study of everolimus in combination with hormonal therapy versus chemotherapy for advanced breast cancer*. BMC medicine, 2014. 12(1): p. 93.

**ACKNOWLEDGMENTS**
We would like to thank Claire Simons, Claire Ainsworth, and Elisabeth Fenwick for their valuable feedback and contributions to the development of this poster.

**Presented at: ISPOR Europe 2024**