Evaluating the efficacy of Artificial Intelligence models as suitable screening tools in systematic reviews: the example of Glucagon-like Peptide-1 Receptor Agonists in Heart Failure treatment

MSR93

Kyriacos Ioannou[1], Sotiris Christoforou[1], Maria Kounnafi[1], Theodoros Christophides[1,2], Gisella Figlioli[3], Daniele Piovani[3], Stefanos Bonovas[3], Georgios Nikolopoulos[1].
1.Medical School, University of Cyprus, Nicosia, Cyprus, 2.Department of Cardiology, Nicosia General Hospital, Nicosia, Cyprus, 3.Humanitas University, Pieve Emanuele, Italy

INTRODUCTION

Systematic reviews and meta-analyses are essential for:
- synthesizing evidence
- guiding healthcare decision-making

They are often limited by:
- time consuming processes
- vulnerable to human error processes

Recent advances in Artificial Intelligence (AI) have introduced powerful language models which may enhance:
- efficiency
- consistency

OBJECTIVE

This study aimed to evaluate the efficiency and accuracy of AI models for screening of studies in an ongoing systematic review and meta-analysis. For this study, ChatGPT and DeepSeek were utilized.

METHODS

- The AI models were used for the screening of articles regarding their inclusion in a systematic review and meta-analysis on the use of of Glucagon-like Peptide-1 Receptor Agoninsts (GLP-1 RAs) in Heart Failure treatment
- For a study to be considered eligible, it had to fulfill the following criteria:
  a) Be a Randomized Controlled Trial (RCT) investigating GLP-1 RAs,
  b) Involve patients diagnosed with heart failure,
  c) Report at least one relevant predetermined outcome.
- These criteria were converted to prompts
- The screening phase was divided into :
  a) the manual screening
  b) the AI-assisted screening based on the prompts
- The time taken for each response from the AI models was recorded
- The AI models had access to only publicly available information from the web.

CONCLUSIONS

This study highlights the potential of AI models in supporting systematic reviews for the screening phase, emphasizing:
- DeepSeek's slightly better accuracy
- ChatGPT's better speed

CONTACT INFORMATION

Georgios Nikolopoulos

nikolopoulos.georgios@ucy.ac.cy

RESULTS

Among 48 eligible studies:
- ChatGPT correctly identified 35 (including studies labelled as partially eligible)
- DeepSeek correctly identified 31 (including studies labelled as partially eligible)

Among 230 non-eligible studies:
- ChatGPT correctly classified 195
- DeepSeek correctly classified 201

In total:
- ChatGPT correctly identified 230 studies
- DeepSeek correctly identified 232 studies

Regarding time efficiency, the average time for the evaluation for:
- ChatGPT was 25.87 seconds
- DeepSeek was 32.80 seconds.