Comparative Analysis of AI vs. Human Reasoning in Clinical Trial Interpretation: A Study Using Large Language Models

Arshjeet Singh[1], Gordon Mao[2], Barnabas Obeng-Gyasi[1], Anoop Chinthala[1], William Snyder[1] Ethan Brown[1], Kyle Ortiz[2]

Indiana University School of Medicine[1]; Department of Neurosurgery[2]

Introduction:

The emergence of advanced LLMs has created new possibilities for medical research interpretation and clinical decision support. While these models have demonstrated impressive capabilities in various medical tasks, their ability to independently reason through clinical trial data and arrive at sound medical recommendations remains understudied. Understanding the alignment between AI and human expert reasoning is crucial for the future integration of AI tools in medical research and practice.

Methods:

We found 20 landmark clinical trials that were published in the last 30 years in the New England Journal of Medicine. We took those articles and removed all the text from it leaving only the clinical trial data and graphs. We then plugged the data into four different AI platforms (ChatGPT, Gemini, Grok 3, and Claude) with the standardized prompt of : “With the tables and figures provided, interpret the following: evidence, statistics, clinical relevance, limitations, and practical applicability. Do not reference the original paper this was obtained from.” We then had two people compare the AI analysis of the data to the conclusions drawn in the article and rate each of the sections on a scale of 1-5.

Results:

ChatGPT and Gemini were successful in accurately analyzing the clinical data and drew similar conclusions as the article. Claude and Grok 3 were less accurate in its conclusions and oversimplified much of the data in its interpretation. ChatGPT was generally the most accurate in its conclusion with 97.9% accuracy followed by Gemini with an 88.3% accuracy. Grok3 was 78.2% accurate in its conclusions while Claude was only 64.5% accurate in its conclusions.

Conclusion:

ChatGPT and Gemini were accurate in analyzing the clinical data and drawing conclusions that were almost identical to the ones in the original article. This opens the door for AI to be used in healthcare to accurately summarize clinical data.