Evaluating the Influence of AI-Language Models on Scientific Research and Clinical Decision-Making: A ChatGPT Study on Squamous Cell Carcinoma

Xin Mu
post-graduate, Peninsula Health, The University of Melbourne, Melbourne, Victoria, 3199, Australia

Ishith Seth
Peninsula Health, The University of Melbourne, Melbourne, Victoria, 3199, Australia
Central Clinic School, Monash University, Melbourne, Victoria, 3004, Australia
Department of Surgery, Bendigo Health, Victoria, 3500, Australia

Yi Xie
Peninsula Health, The University of Melbourne, Melbourne, Victoria, 3199, Australia

Angus Lee
Department of Surgery, Bendigo Health, Victoria, 3500, Australia

Abstract:

Background: Large language models (LLMs) have the potential to transform scientific research and medical management. This study explored the role of three leading LLMs, ChatGPT, BARD and BingAI in identifying squamous cell carcinoma (SCC) in providing safe medical advice. SCC was used as a mere example of LLM efficacy in clinical decision-making for patient education.

Methods: A series of simulated clinical questions on SCC diagnosis and management were posed to ChatGPT, BARD, and Bing AI. The responses were analyzed qualitatively by a panel of doctors utilizing a Likert scale for comparative evaluation with existing literature and guidelines. Quantitative evaluation of readability was performed using three renowned algorithms: the Flesch Reading Ease Score, the Flesch-Kincaid Grade Level, and the Coleman-Liau Index. Information quality and relevance were assessed through the modified DISCERN score. A two-tailed t-test was performed for determination of statistical significance.

Results: Comparative analysis revealed similar reliability and readability in the medical advice offered by the three LLMs, with ChatGPT demonstrated superior performance with a DISCERN score of 72.0 (±4.00). BARD's responses were found to be the most readable, achieving a Flesch Reading Ease Score of 72.7 (±5.96), a Flesch-Kincaid Grade Level of 6.37 (±1.11), and a Coleman-Liau Index of 7.7 (±0.58). Notably, the only comparison deemed statistically significant with $p<0.05$ was between the readability of ChatGPT and BARD.

Conclusion: All three LLMs demonstrated the ability to provide reliable and comprehensible medical advice on SCC. ChatGPT marginally surpassed BARD and BingAI in delivering the most precise information, whereas BARD emerged as the most user-friendly LLM. This study highlights the potential of artificial intelligence-driven frameworks within the healthcare sector, particularly in the realm of dermatology.