Accuracy and Reliability of Artificial Intelligence Language Models in Providing Patient Education About Anti-hypertensive Medications: A Comparative Analysis

Abdelrahman M. Tawfik[1], Mohamed El-mezayen[1], Mohamed Habashy Abozaid[1], Abdalla Ali Elmansoury[1], Mohamed Hany Afify[1], Marwa Farouk[1], and Azza Mohamed Baraka[2]

[1]Faculty of Medicine, Alexandria University, Alexandria, Egypt
[2]Pharmacology Department, Faculty of Medicine, Alexandria University, Alexandria, Egypt

Background

Artificial intelligence (AI) language models are increasingly used by patients as sources of medical information. However, concerns persist regarding their reliability. Hypertension requires precise and accessible patient education to ensure medication adherence and safety.

Aim and objectives

To evaluate the performance of leading AI chatbots in responding to patient queries about anti-hypertensive medications.

Methods

A set of 50 commonly asked questions on anti-hypertensive medications, including lifestyle recommendations, side effects, and drug interactions, was prepared from patient forums. As of March 2025, each question was entered separately into six AI models: ChatGPT-3.5, ChatGPT-4, Gemini-2.5, DeepSeek, Copilot-balanced mode, and Grok-2. Responses were evaluated by an experienced clinical pharmacologist using a standardized scoring system assessing accuracy (alignment with new evidence), clarity (ease of understanding and simplicity), completeness (inclusion of essential details), neutrality (absence of bias), and appropriateness (relevance for a layperson).

Results

DeepSeek demonstrated the highest overall performance scoring (94.1%), ranking the highest in all categories: accuracy (100%), clarity (89.3%), completeness (95.3%), neutrality (94%) and appropriateness (92%). Gemini-2.5 ranked second with an overall score of (86.9%), excelling in accuracy (96.7%) and lacking in completeness (79%). ChatGPT-4 achieved a moderate performance scoring (79.3%) with a similar performance to ChatGPT-3.5. The newly launched Grok-2 ranked fifth, scoring (70%) overall but achieved the lowest in completeness. Lastly, Copilot scored the lowest overall score of (62%).

Conclusion

There is noticeable variability in AI models' ability to provide accurate and patient-friendly education on anti-hypertensive medications. DeepSeek performed best, followed by Gemini. ChatGPT-4 had moderate performance, while Grok-2 and Copilot lagged. This variability necessitates continuous refinement and clinician oversight. Clinicians should guide patients toward dependable AI resources while remaining cautious about inaccuracies. Further research should test evolving model updates.

Keywords

Artificial Intelligence, Chatbots, blood pressure, hypertension, Patient education