

Artificial Intelligence in Diagnosing and Managing Vascular Surgery Patients: An Experimental Study Using the GPT-4 Model

Vangelis G. Alexiou, MD, Ioannina University Hospital, Ioannina, Greece

Bauer E. Sumpio, MD, Yale University School of Medicine, New Haven, Connecticut, United States of America

Areti Vassiliou, MD, Ioannina University Hospital, Ioannina, Greece

Stavros K. Kakkos, MD, University Hospital of Patras, Patras, Greece

George Geroulakos, MD, National and Kapodistrian University of Athens, Athens, Greece

Objective:

The introduction of artificial intelligence (AI) has led to groundbreaking advancements across many scientific fields. Machine learning algorithms have enabled AI models to learn, adapt, and solve complex problems in previously unimaginable ways. Natural language processing (NLP) allows these models to comprehend and respond to inquiries in a natural and humanly understandable way. We sought to investigate the application and performance of an AI chatbot in the diagnosis and management of vascular surgery patients.

Methods:

An experimental study was undertaken to evaluate the performance of GPT-4 AI model across 57 clinical scenarios derived from a textbook in vascular surgery. Specific prompts were devised to address the AI model and task it to identify symptoms, diagnose conditions, and select appropriate therapeutic approaches. Answers were scored, descriptive statistics were produced and means were compared across topics. The reasoning and evidence used in the cases in which AI performed poorly were critically reviewed.

Results:

The AI model correctly answered over 65% of the 385 questions. Performance variation between and within 13 vascular surgery topics did not show any statistically significant differences. Analysis of the questions where the model failed by more than 50% suggests a gap in the ability to interpret and process multifaceted medical information. 27% of these errors were attributed to potential lack of understanding of complex clinical scenarios. The AI model also quoted incorrect or outdated information in 14% of cases and showed an inability to comprehend context, nuances, and medical classification systems in 11% of the cases.

Conclusion:

GPT-4 demonstrated potential to provide clinically relevant answers for most of the tested scenarios. However, its reasoning must still be carefully analyzed for exactitude and clinical validity. While language models show promise as valuable tools for clinicians, it is essential to recognize their role as supportive mechanisms rather than standalone solutions.