Completeness and Accuracy of Artificial Intelligence Chatbot Responses on Cardiovascular and Oncological Disease Information

Teodor-Marian BACIU*, Alexandra-Maria NEGUȚESCU, Iris NĂDĂȘAN, Dragoş AVRAM, and Valentin NĂDĂȘAN

George Emil Palade University of Medicine, Pharmacy, Science and Technology of Târgu Mureş, Gheorghe Marinescu Str., no. 38, 540142 Târgu Mureş, Romania.
E-mails: teodor.baciu08@yahoo.com; alexandranegutescu@yahoo.com; irisnadasan@gmail.com; valentin.nadasan@umfst.ro; dragosavram2013@gmail.com

* Author to whom correspondence should be addressed

Abstract

Background and Aim: AI-powered chatbots are expected to revolutionize patients' access to health information but their ability to provide comprehensive and scientifically accurate answers is insufficiently known. The study aimed to assess the completeness and accuracy of information regarding two cardiovascular and two oncological diseases of high interest to health information seekers. *Materials and Methods:* The completeness and accuracy of the information about acute myocardial infarction, peripheral artery disease, colorectal and gastric cancer provided by three AI chatbots (ChatGPT–Open AI, Gemini–Google, Llama–Meta) were evaluated against an evidence-based information quality benchmark on a scale ranging from 0 to 10. Chatbot prompting followed two basic scenarios, plausible for most users: (A) a single broad-scoped question; (B) a series of focused questions covering basic information about disease definition, causes, risk factors, symptoms, treatment and prevention. Responses were rated against evidence-based, disease-specific quality benchmarks following a predefined procedure. Data were collected between October 2023 and May 2024. Overall and chatbot-specific mean completeness, and accuracy scores were calculated. *Results:* Scenario A yielded an overall completeness score of 5.4, while bot-specific scores were 6.3 for ChatGPT, 5.1 for Gemini, and 5.0 for Llama. The overall accuracy score was 6.6, and bot-specific scores were 7.0 for ChatGPT, 6.1 for Gemini, and 6.5 for Llama. Scenario B showed an overall accuracy score of 8.1 with the following bot-specific accuracy scores: ChatGPT 8.6, Gemini 8.2, and Llama 7.6. The completeness score was not applicable within the second scenario. *Conclusions:* The overall completeness of the information provided by the three studied AI-powered chatbots about the four investigated diseases was moderate. The overall accuracy scores were high in case of the first scenario and very high in the second one. ChatGPT performed slightly better than the other two bots on both quality measures.

Keywords: Consumer Health Informatics; Internet Use; Data Accuracy; Artificial Intelligence; Disinformation