Study finds ChatGPT fails at heart risk assessment

May 1 2024, by Sara Zaske

Credit: MART PRODUCTION from Pexels

Despite ChatGPT's reported ability to pass medical exams, new research indicates it would be unwise to rely on it for some health assessments, such as whether a patient with chest pain needs to be hospitalized.

In a study involving thousands of simulated cases of patients with chest pain, ChatGPT provided inconsistent conclusions, returning different heart risk assessment levels for the exact same patient data. The generative AI system also failed to match the traditional methods physicians use to judge a patient's cardiac risk. The findings were published in the journal *PLOS ONE*.

"ChatGPT was not acting in a consistent manner," said lead author Dr. Thomas Heston, a researcher with Washington State University's Elson S. Floyd College of Medicine. "Given the exact same data, ChatGPT would give a score of low risk, then next time an intermediate risk, and occasionally, it would go as far as giving a high risk."

The authors believe the problem is likely due to the level of randomness built into the current version of the software, ChatGPT4, which helps it vary its responses to simulate natural language. This same randomness, however, does not work well for health care uses that require a single, consistent answer, Heston said.

"We found there was a lot of variation, and that variation in approach can be dangerous," he said. "It can be a useful tool, but I think the technology is going a lot faster than our understanding of it, so it's critically important that we do a lot of research, especially in these high-stakes clinical situations."

Chest pains are common complaints in emergency rooms, requiring doctors to rapidly assess the urgency of a patient's condition. Some very serious cases are easy to identify by their symptoms, but lower risk ones can be trickier, Heston said, especially when determining whether someone should be hospitalized for observation or sent home and receive outpatient care.

Currently medical professionals often use one of two measures that go

by the acronyms TIMI and HEART to assess heart risk. Heston likened these scales to calculators with each using a handful of variables including symptoms, health history and age. In contrast, an AI neural network like ChatGPT can assess billions of variables quickly, meaning it could potentially analyze a complex situation faster and more thoroughly.

For this study, Heston and colleague Dr. Lawrence Lewis of Washington University in St. Louis first generated three datasets of 10,000 randomized, simulated cases each. One dataset had the seven variables of the TIMI scale, the second set included the five HEART scale variables and a third had 44 randomized health variables.

On the first two datasets, ChatGPT gave a different risk assessment 45% to 48% of the time on individual cases than a fixed TIMI or HEART score. For the last data set, the researchers ran the cases four times and found ChatGPT often did not agree with itself, returning different assessment levels for the same cases 44% of the time.

Despite the negative findings of this study, Heston sees great potential for generative AI in health care—with further development.

For instance, assuming privacy standards could be met, entire medical records could be loaded into the program, and an in an emergency setting, a doctor could ask ChatGPT to give the most pertinent facts about a patient quickly. Also, for difficult, complex cases, doctors could ask the program to generate several possible diagnoses.

"ChatGPT could be excellent at creating a differential diagnosis and that's probably one of its greatest strengths," said Heston.

"If you don't quite know what's going on with a patient, you could ask it to give the top five diagnoses and the reasoning behind each one. So it

could be good at helping you think through a problem, but it's not good at giving the answer."

More information: Thomas F. Heston et al, ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain, *PLOS ONE* (2024). DOI: 10.1371/journal.pone.0301854

Provided by Washington State University

Citation: Study finds ChatGPT fails at heart risk assessment (2024, May 1) retrieved 1 June 2024 from https://medicalxpress.com/news/2024-05-chatgpt-heart.html