AI gets high marks from doctors in answering medical questions

October 3 2023, by Dennis Thompson

The ChatGPT artificial intelligence (AI) program could grow into a

source of accurate and comprehensive medical information, but it's not quite ready for prime time yet, a new study reports.

ChatGPT's responses to more than 280 medical questions across diverse specialties averaged between mostly to almost completely correct, according to a report published online Oct. 2 in *JAMA Network Open*.

"Overall, it performed fairly well as far as both accuracy and completions," said senior researcher Dr. Douglas Johnson, director of the Melanoma Clinical Research Program at Vanderbilt-Ingram Cancer Center in Nashville, Tenn.

"Certainly, it was not perfect. It was not completely reliable," Johnson continued. "But at the time we were entering the questions, it was actually pretty accurate and provided, relatively speaking, reliable information."

Accuracy improved even more if a second AI program was brought in to review the answer provided by the first, the results showed.

Johnson and his colleagues set out to test ChatGPT by peppering the AI with health questions between January and May 2023, shortly after it came online.

People and doctors already lean on search engines like Google and Bing for answers to health questions, Johnson said. It makes sense that AI programs like ChatGPT will be the next frontier for researching medical issues.

Such AI programs "provide almost an answer engine for many types of questions across different fields, certainly including medicine, and so we realized that patients as well as potentially physicians would be using those," Johnson said. "We wanted to try to understand across medical

disciplines how accurate, how complete the information that they provided was going to be."

Researchers recruited 33 physicians across 17 specialties to come up with 284 easy, medium and hard questions for ChatGPT.

The accuracy of ChatGPT's responses to those questions averaged 4.8 on a 6-point scale, the researchers said. A score of 4 is "more correct than incorrect" and 5 is "nearly all correct."

Average accuracy was 5 for easy questions, 4.7 for medium questions and 4.6 for difficult questions, the study authors said.

ChatGPT also provided fairly complete answers, scoring 2.5 on a 3-point scale, according to the report.

"Even at the relative infancy of the programs, it was short of completely reliable but still provided relatively accurate and comprehensive information," Johnson said.

The program performed better regarding some specialties. For example, it averaged 5.7 accuracy on questions regarding common conditions, and 5.2 on questions regarding melanoma and immunotherapy, the investigators found.

The program also did better responding to "yes/no" questions than open-ended questions, with an average accuracy score of 6 versus 5, respectively.

Some questions ChatGPT knocked out of the park.

For example, the AI provided a perfectly accurate and complete response to the question, "Should patients with a history of acute

myocardial infarction [AMI] receive a statin?"

"Yes, patients with a history of AMI should generally be treated with a statin," the response begins, before rolling on to provide a flurry of context.

Other questions the program struggled with, or even got wrong.

When asked "what oral antibiotics may be used for the treatment of MRSA infections," the answer included some options not available orally, the researchers noted. The answer also omitted one of the most important oral antibiotics.

However, misses like that might be as much the fault of the doctor, for not phrasing the question in a way the program could easily grasp, said Dr. Steven Waldren, chief medical informatics officer for the American Academy of Family Physicians.

Specifically, the program might have stumbled over the phrase "may be used" in the question, Waldren said.

"If this question would have been 'what oral antibiotics are used,' not may be used, it may have picked up that (omitted) drug," he said. "There wasn't much conversation in the paper about the way that the questions need to be crafted, because right now, where these large language models are, that is really important to be done in a way that will get the most optimal answer."

Further, researchers found that ChatGPT's initially poor answers became more accurate if the initial question was resubmitted a week or two later.

This shows that the AI is quickly growing smarter over time, Johnson said.

"I think it's most likely improved even further since we did our study," Johnson said. "I think at this point physicians could think about using it, but only in conjunction with other known resources. I certainly wouldn't take any recommendations as gospel, by any stretch of the imagination."

Accuracy also improved if another version of the AI was brought in to review the first response.

"One instance generated the response to the prompt, and a second instance became kind of the AI reviewer that reviewed the content and asked, 'is this actually accurate?'" Waldren said. "It was interesting for them to use that to see if it helped solve some of these inaccurate answers."

Johnson expects accuracy will further improve if AI chatbots are developed specifically for medical use.

"You can certainly imagine a future where these chatbots are trained on very reliable medical information, and are able to achieve that kind of reliability," Johnson said. "But I think we're short of that at this point."

Both Johnson and Waldren said it's very unlikely that AI will replace physicians altogether.

Johnson thinks AI instead will serve as another helpful tool for doctors and patients.

Doctors might ask the AI for more information regarding a tricky diagnosis, while patients could use the program as a "health coach," Johnson said.

"You can certainly imagine a future where somebody's got a cold or something and the chatbot is able to input vital signs and input symptoms

and so forth and give some advice about, OK, is this something you do need to go see a doctor for? Or is this something that is probably just a virus? And you can watch out for these five things that if those do happen, then go see a doctor. But if not, then you're probably going to be fine," Johnson said.

There is some concern that cost-cutting health systems might try using AI as a front-line resource, asking patients to refer to the program for advice before scheduling an appointment with a doctor, Waldren said.

"It's not that the physicians are going to be replaced. It's the tasks that physicians do are going to change. It's going to change what it means to be a physician," Waldren said of AI. "I think that the challenge for patients is going to be that there's going to be financial pressures to try to push those tasks away from the highest-cost implementations, and a physician can be pretty costly."

So, he predicted, it's likely more patients will be pushed to a nurse line with AI chat.

"That could be a good thing, with increased access to care," Waldren added. "It also could be a bad thing if we don't continue to support continuity of care and coordination of care."

More information: Rachel S. Goodman et al, Accuracy and Reliability of Chatbot Responses to Physician Questions, *JAMA Network Open* (2023). DOI: 10.1001/jamanetworkopen.2023.36483