A.I. Chatbots Defeated Doctors at Diagnosing Illness

A small study found ChatGPT outdid human physicians when assessing medical case histories, even when those doctors were using a chatbot.

▶ Listen to this article · 9:25 min Learn more

Share full article

In an experiment, doctors who were given ChatGPT to diagnose illness did only slightly better than doctors who did not. But the chatbot alone outperformed all the doctors. Michelle Gustafson for The New York Times

By Gina Kolata

Nov. 17, 2024

Dr. Adam Rodman, an expert in internal medicine at Beth Israel Deaconess Medical Center in Boston, confidently expected that chatbots built to use artificial intelligence would help doctors diagnose illnesses.

He was wrong.

Instead, in a study Dr. Rodman helped design, doctors who were given ChatGPT-4 along with conventional resources did only slightly better than doctors who did not have access to the bot. And, to the researchers' surprise, ChatGPT alone outperformed the doctors.

“I was shocked,” Dr. Rodman said.

The chatbot, from the company OpenAI, scored an average of 90 percent when diagnosing a medical condition from a case report and explaining its reasoning.

Doctors randomly assigned to use the chatbot got an average score of 76 percent. Those randomly assigned not to use it had an average score of 74 percent.

The study showed more than just the chatbot’s superior performance.
It unveiled doctors’ sometimes unwavering belief in a diagnosis they made, even when a chatbot potentially suggests a better one.

And the study illustrated that while doctors are being exposed to the tools of artificial intelligence for their work, few know how to exploit the abilities of chatbots. As a result, they failed to take advantage of A.I. systems’ ability to solve complex diagnostic problems and offer explanations for their diagnoses.

A.I. systems should be “doctor extenders,” Dr. Rodman said, offering valuable second opinions on diagnoses.

But it looks as if there is a way to go before that potential is realized.

Case History, Case Future

The experiment involved 50 doctors, a mix of residents and attending physicians recruited through a few large American hospital systems, and was published last month in the journal JAMA Network Open.

The test subjects were given six case histories and were graded on their ability to suggest diagnoses and explain why they favored or ruled them out. Their grades also included getting the final diagnosis right.

The graders were medical experts who saw only the participants’ answers, without knowing whether they were from a doctor with ChatGPT, a doctor without it or from ChatGPT by itself.

The case histories used in the study were based on real patients and are part of a set of 105 cases that has been used by researchers since the 1990s. The cases intentionally have never been published so that medical students and others could be tested on them without any foreknowledge. That also meant that ChatGPT could not have been trained on them.

But, to illustrate what the study involved, the investigators published one of the six cases the doctors were tested on, along with answers to the test questions on that case from a doctor who scored high and from one whose score was low.

That test case involved a 76-year-old patient with severe pain in his low back, buttocks and calves when he walked. The pain started a few days after he had been treated with balloon angioplasty to widen a coronary artery. He had been treated with the blood thinner heparin for 48 hours after the procedure.

The man complained that he felt feverish and tired. His cardiologist had done lab studies that indicated a new onset of anemia and a buildup of nitrogen and other kidney waste products in his blood. The man had had bypass surgery for heart disease a decade earlier.

The case vignette continued to include details of the man's physical exam, and then provided his lab test results.

The correct diagnosis was cholesterol embolism — a condition in which shards of cholesterol break off from plaque in arteries and block blood vessels.

Participants were asked for three possible diagnoses, with supporting evidence for each. They also were asked to provide, for each possible diagnosis, findings that do not support it or that were expected but not present.

The participants also were asked to provide a final diagnosis. Then they were to name up to three additional steps they would take in their diagnostic process.

Like the diagnosis for the published case, the diagnoses for the other five cases in the study were not easy to figure out. But neither were they so rare as to be almost unheard-of. Yet the doctors on average did worse than the chatbot.

What, the researchers asked, was going on?

The answer seems to hinge on questions of how doctors settle on a diagnosis, and how they use a tool like artificial intelligence.

The Physician in the Machine

How, then, do doctors diagnose patients?

The problem, said Dr. Andrew Lea, a historian of medicine at Brigham and Women's Hospital who was not involved with the study, is that "we really don't know how doctors think."

In describing how they came up with a diagnosis, doctors would say, "intuition," or, "based on my experience," Dr. Lea said.

That sort of vagueness has challenged researchers for decades as they tried to make computer programs that can think like a doctor.

The quest began almost 70 years ago.

"Ever since there were computers, there were people trying to use them to make diagnoses," Dr. Lea said.

One of the most ambitious attempts began in the 1970s at the University of Pittsburgh. Computer scientists there recruited Dr. Jack Myers, chairman of the medical school's department of internal medicine who was known as a master diagnostician. He had a photographic memory and spent 20 hours a week in the medical library, trying to learn everything that was known in medicine.

Dr. Myers was given medical details of cases and explained his reasoning as he pondered diagnoses. Computer scientists converted his logic chains into code. The resulting program, called INTERNIST-1, included over 500 diseases and about 3,500 symptoms of disease.

To test it, researchers gave it cases from the New England Journal of Medicine. "The computer did really well," Dr. Rodman said. Its performance "was probably better than a human could do," he added.

But INTERNIST-1 never took off. It was difficult to use, requiring more than an hour to give it the information needed to make a diagnosis. And, its creators noted, "the present form of the program is not sufficiently reliable for clinical applications."

Research continued. By the mid-1990s there were about a half dozen computer programs that tried to make medical diagnoses. None came into widespread use.

“It’s not just that it has to be user friendly, but doctors had to trust it,” Dr. Rodman said.

And with the uncertainty about how doctors think, experts began to ask whether they should care. How important is it to try to design computer programs to make diagnoses the same way humans do?

“There were arguments over how much a computer program should mimic human reasoning,” Dr. Lea said. “Why don’t we play to the strength of the computer?”

The computer may not be able to give a clear explanation of its decision pathway, but does that matter if it gets the diagnosis right?

The conversation changed with the advent of large language models like ChatGPT. They make no explicit attempt to replicate a doctor’s thinking; their diagnostic abilities come from their ability to predict language.

“The chat interface is the killer app,” said Dr. Jonathan H. Chen, a physician and computer scientist at Stanford who was an author of the new study.

“We can pop a whole case into the computer,” he said. “Before a couple of years ago, computers did not understand language.”

But many doctors may not be exploiting its potential.

Operator Error

After his initial shock at the results of the new study, Dr. Rodman decided to probe a little deeper into the data and look at the actual logs of messages between the doctors and ChatGPT. The doctors must have seen the chatbot’s diagnoses and reasoning, so why didn’t those using the chatbot do better?

It turns out that the doctors often were not persuaded by the chatbot when it pointed out something that was at odds with their diagnoses. Instead, they tended to be wedded to their own idea of the correct diagnosis.

“They didn’t listen to A.I. when A.I. told them things they didn’t agree with,” Dr. Rodman said.

That makes sense, said Laura Zwaan, who studies clinical reasoning and diagnostic error at Erasmus Medical Center in Rotterdam and was not involved in the study.

“People generally are overconfident when they think they are right,” she said. But there was another issue: Many of the doctors did not know how to use a chatbot to its fullest extent.

Dr. Chen said he noticed that when he peered into the doctors’ chat logs, “they were treating it like a search engine for directed questions: ‘Is cirrhosis a risk factor for cancer? What are possible diagnoses for eye pain?’”

“It was only a fraction of the doctors who realized they could literally copy-paste in the entire case history into the chatbot and just ask it to give a comprehensive answer to the entire question,” Dr. Chen added.

“Only a fraction of doctors actually saw the surprisingly smart and comprehensive answers the chatbot was capable of producing.”

Gina Kolata *reports on diseases and treatments, how treatments are discovered and tested, and how they affect people.*

https://www.nytimes.com/2024/11/17/health/chatgpt-ai-doctors-diagnosis.html

I have been waiting for this, and more to come. Just a Nurse Practitioner to deliver the AI diagnosis and treatment? Or just an aide to wheel in the AI robot? To strap on the electrodes, take blood pressure, and prick a finger for blood tests? A voice analyzer to judge responses. Maybe, for a gentler, friendly touch, a dog trained to sniff out common diseases, deficiencies, or illnesses?

How will this affect medical liability? “You mean you over-rode the AI diagnosis and my child died because of it?” Medical egos are going to have to give way, but without God grandiosity how could they practice or justify their fees? What will Med School training look like?

Medicine, being a doctor, seems to me a strange profession. Dealing with death “objectively” every day, making mistakes like as in any other human enterprise. In 15 minutes, or after years of the same old noncompliance with directives by patients set in unhealthy ways, they encounter unreal expectations and do a dance with patients that gets them out the door “safely treated.” Doctors must all feel they could be prosecuted for malpractice. No wonder they settle into hide-bound ways to get through a day and a career.

As with so many professions, thank goodness we all have met a doctor we like and trust. Sort of. To which medicos reply “The guy who doctor’s himself has a fool for a doctor.” TJB