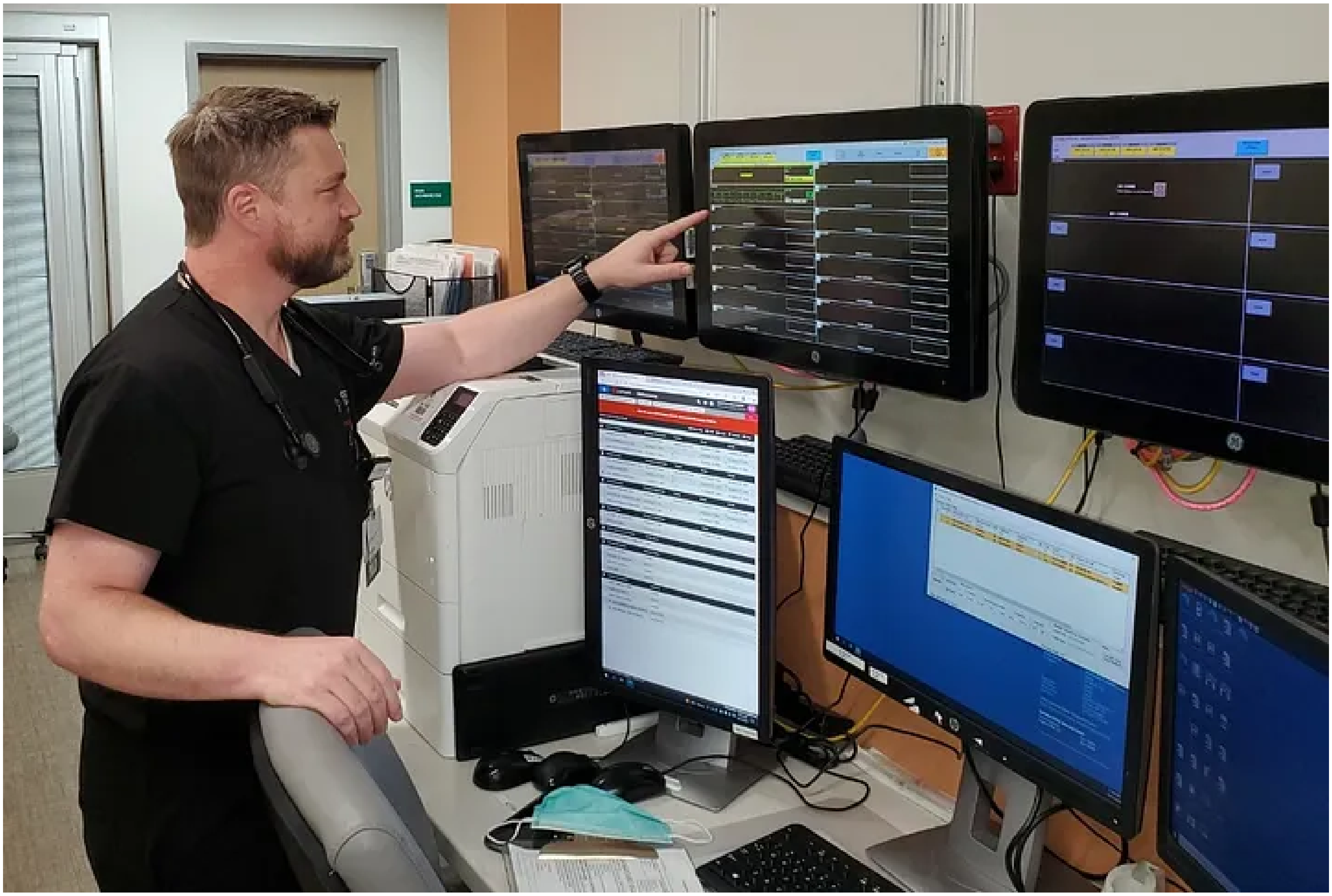
Inflect Health Follow

Apr 5 · 6 min read

I'm an ER doctor: Here's what I found when I asked ChatGPT to diagnose my patients

ChatGPT recently passed the U.S. Medical Licensing Exam, but using it for a real-world medical diagnosis would quickly turn deadly.



Dr. Tamayo-Sarver on duty

by Josh Tamayo-Sarver, MD, PhD

With news that ChatGPT successfully "passed" the U.S. Medical Licensing Exam, I was curious how it would perform in a real-world medical situation. As an advocate of leveraging artificial intelligence to improve the quality and efficiency of healthcare, I wanted to see how the current version of ChatGPT might serve as a tool in my own practice.

So after my regular clinical shifts in the emergency department the other week, I anonymized my History of Present Illness notes for 35 to 40 patients — basically, my detailed medical narrative of each person's medical history, and the symptoms that brought them to the emergency department — and fed them into ChatGPT.

The specific prompt I used was, "What are the differential diagnoses for this patient presenting to the emergency department [insert patient HPI notes here]?"

The results were fascinating, but also fairly disturbing.

OpenAI's chatbot did a decent job of bringing up common diagnoses I wouldn't want to miss — as long as everything I told it was precise, and highly detailed. Correctly diagnosing a patient as having nursemaid's elbow, for instance, required about 200 words; identifying another patient's orbital wall blowout fracture took the entire 600 words of my HPI on them.

For roughly half of my patients, ChatGPT suggested six possible diagnoses, and the "right" diagnosis — or at least the diagnosis that I believed to be right after complete evaluation and testing — was among the six that ChatGPT suggested.

Not bad. Then again, a 50% success rate in the context of an emergency room is also not good.

ChatGPT's worst performance happened with a 21-year-old female patient who came into the ER with right lower quadrant abdominal pain. I fed her HPI into ChatGPT, which instantly came back with a differential diagnosis of appendicitis or an ovarian cyst, among other possibilities.

But ChatGPT missed a somewhat important diagnosis with this woman.

She had an ectopic pregnancy, in which a malformed fetus develops in a woman's fallopian tube, and not her uterus. Diagnosed too late, it can be fatal — resulting in death caused by internal bleeding. Fortunately for my patient, we were able to rush her into the operating room for immediate treatment.

Notably, when she saw me in the emergency room, this patient did not even know she was pregnant. This is not an atypical scenario, and often only emerges after some gentle inquiring:

“Any chance you’re pregnant?”

Sometimes a patient will reply with something like “I *can’t* be.”

“But how do you know?”

If the response to that follow-up does not refer to an IUD or a specific medical condition, it’s more likely the patient is actually saying they don’t *want* to be pregnant for any number of reasons. (Infidelity, trouble with the family, or other external factors.) Again, this is not an uncommon scenario; about 8% of pregnancies discovered in the ER are of women who report that they’re not sexually active.

But looking through ChatGPT’s diagnosis, I noticed not a single thing in its response suggested my patient was pregnant. It didn’t even know to ask.

My fear is that countless people are already using ChatGPT to medically diagnose themselves rather than see a physician. If my patient in this case had done that, ChatGPT’s response could have killed her.

ChatGPT also misdiagnosed several other patients who had life-threatening conditions. It correctly suggested one of them had a brain tumor — but missed two others who also had tumors. It diagnosed another patient with torso pain as having a kidney stone — but missed that the patient actually had an aortic rupture. (And subsequently died on our operating table.)

In short, ChatGPT worked pretty well as a diagnostic tool when I fed it perfect information and the patient had a classic presentation.

This is likely why ChatGPT “passed” the case vignettes in the Medical Licensing Exam. Not because it’s “smart,” but because the classic cases in the exam have a deterministic answer that already exists in its database. ChatGPT rapidly presents answers in a natural language format (that’s the genuinely impressive part), but underneath that is a knowledge retrieval process similar to Google Search. And most actual patient cases are not classic.

My experiment illustrated how the vast majority of any medical encounter is figuring out the correct patient *narrative*. If someone comes into my ER saying their wrist hurts, but not due to any recent accident, it could be a psychosomatic reaction after the patient’s grandson fell down, or it could be due to a sexually transmitted disease, or something else entirely. The art of medicine is extracting all the necessary information required to create the right narrative.

Might ChatGPT still work as a doctor's assistant, automatically reading my patient notes during treatment and suggesting differentials? Possibly. But my fear is this could introduce even worse outcomes.

If my patient notes don't include a question I haven't yet asked, ChatGPT's output will encourage me to keep missing that question. Like with my young female patient who didn't know she was pregnant. If a possible ectopic pregnancy had not immediately occurred to me, ChatGPT would have kept enforcing that omission, only reflecting back to me the things I thought were obvious — enthusiastically validating my bias like the world's most dangerous yes-man.

None of this suggests AI has no potentially useful place in medicine, because it does.

As a human physician, I'm limited by how many patients I can personally treat. I expect to see roughly 10,000 patients in my lifetime, each of them with a unique body mass, blood pressure, family history, and so on — a huge variety of features I track in my mental model. Each human has countless variables relevant to their health, but as a human doctor working with a limited session window, I focus on the several factors that tend to be the most important historically.

So for instance, if I review a patient's blood test and see high levels of hemoglobin A1C, then I diagnose them as likely to have the early stages of diabetes. But what if I could keep track of the countless variables about the person's health and compare them with other people who were similar across all the millions of variables, not just based on their hemoglobin A1C? Perhaps then I could recognize that the other 100,000 patients who looked just like this patient in front of me across that wide range of factors had a great outcome when they started to eat more broccoli.

This is the space where AI can thrive, tirelessly processing these countless features of every patient I've ever treated, and every other patient treated by every other physician, giving us deep, vast insights. AI can help do this eventually, but it will first need to ingest millions of patient data sets that include those many features, the things the patients did (like take a specific medication), and the outcome.

In the meantime, we urgently need a much more realistic view from Silicon Valley and the public at large of what AI can do now — and its many, often dangerous, limitations. We must be very careful to avoid inflated expectations with programs like ChatGPT, because in the context of human health, they can literally be life-threatening.

Originally published in FastCompany

Dr. Josh Tamayo-Sarver works clinically in the emergency department of his local community and is a vice president of innovation at Inflect Health, an innovation incubator for health tech.

Chatgpt Generative Ai Artificial Intelligence Healthtech Healthcare