HEALTH TECH

As artificial intelligence shows off diagnostic chops, scientists reckon with the way forward

A prominent study in Science prompts physicians to call for rigorous clinical trials

By Katie Palmer – STAT News
April 30, 2026
Health Tech Correspondent

Getting a paper published in Science is a highlight of many researchers' careers. But for internist and clinical artificial intelligence researcher Adam Rodman, it's also been a source of some agita.

On Thursday, Rodman and his colleagues published a compilation of experiments, including one using real-world data from a Boston emergency department, that show a large language model from OpenAI can outperform physicians in case-based diagnostic and clinical reasoning evaluations. To Rodman, the paper's co-senior author, it's a response to a gauntlet thrown down in Science in 1959. That paper "described how you would know that a clinical decision support system was capable of doing diagnosis better than humans," he said. "And they can do it."

But as generative AI tools like chatbots are heavily marketed — both to patients and clinicians — it makes him worried that the science experiments, all based on simulated and historical cases, will be misconstrued as proof of AI's safety and efficacy when used to treat real patients.

"I worry that my research agenda — which is not to replace doctors, not to get people to stop seeing their doctors and talk to their chatbots; my research agenda is to use a new type of technology to improve medical care — is going to get used by companies that are heavily financed and are looking to skip some of these essential safe pieces of medicine," said Rodman, an assistant professor of medicine at Beth Israel Deaconess

Medical Center and visiting researcher at Google. “You can understand why I’m reticent.” That concern is mirrored by other clinical AI researchers who view the experiments’ results with more skepticism.

Since consumer-facing LLMs burst onto the scene in 2022, researchers have been chucking a variety of diagnostic tests their way. Multiple-choice medical licensing exams. Tricky case studies published in the New England Journal of Medicine. Models appeared to do well on most of them, but often without comparing their performance to large groups of physicians, said co-senior author Arjun Manrai. Their experiments, detailed in the Science paper, aimed to fill that gap by testing OpenAI’s 2024 o1-preview, one of a new generation of so-called “reasoning” models.

“We were really trying to throw everything that we could at the model,” said Manrai, an assistant professor of biomedical informatics at Harvard Medical School. They replicated a number of tests, some previously conducted on GPT-4, that measure an aspect of clinical reasoning against a baseline of performance from dozens of physicians.

Most of those experiments spoon-fed structured, curated case studies to the LLM. But in a novel approach, the group — which includes some of the best-known names testing the bleeding edge of clinical AI — also asked how the model would perform when prompted with messier data from 76 randomly selected real cases in the emergency department at Beth Israel.

Emergency room physicians and patients never interacted with the AI. But after the patients were seen, “we copied and pasted text straight from the electronic health record with no curation of data, including the random noise,” said co-first author Peter Brodeur, an internal medicine resident at Beth Israel, and showed them to the model and two internists. Then they were asked to offer second opinions about the patient’s diagnosis at three points along the patient’s ER journey.

At the point of triage, when information in the record dump was limited, the LLM performed better than two physicians at identifying a correct or very close diagnosis. It was still better when given all the data collected by the end of the emergency department encounter. By the point the patient would have been admitted to the hospital, when the most

information is available to inform a diagnosis, the AI and human scores finally converged.

In one case, Rodman said, a patient came to the ER with a blood clot that had traveled to their lungs, and after treatment with anticoagulants, they got worse. While the two physicians reviewing the records after the fact initially thought the drugs might have failed, the model honed in on the patient's history of lupus. Perhaps that was a unifying cause for the blood clot and the other symptoms? It turned out to be right: The patient had lupus pleuritis, inflammation of the heart and lungs caused by the autoimmune disease.

The study's other experiments — which tested o1-preview on documentation of clinical reasoning, management reasoning, and diagnostic reasoning — similarly showed the model outperformed average scores from dozens or hundreds of clinicians. "Long story short, the model outperformed our very large physician baseline," said Manrai.

Still, these results are not proof, emphasized both the authors and outside experts, to support the use of AI in clinical care. There is still a gaping chasm between diagnostic reasoning based on text-based prompts and the actual practice of medicine, and researchers raised concerns about the reliability of the findings.

The risk of a prominent paper like this, said NYU AI researcher Eric Oermann, is that "everyone says, 'Oh, look, there's a paper in Science on language models in the emergency department, they're ready to use for patients,' when nothing remotely like that has been shown here."

For one, diagnosis is just one tiny slice of the work of a doctor — and it isn't always their top priority. In the emergency department, it's about triage and immediate symptom management. In primary care, it's about tracking and treating chronic disease. But researchers often focus on LLMs' diagnostic accuracy because it's relatively easy to test.

In the real world, said Manrai, "doctors talk to patients. They counsel them, they listen to their values. They interpret images, they read EKGs and ECGs. And they integrate all of that together to guide a patient through challenging decisions."

Text-based LLMs can't process a lot of the information a physician uses to inform all those aspects of care, the authors acknowledge. "A clinician in real life would be able to look at a patient and get some gestalt about how sick they are," said Emily Alsentzer, assistant professor of biomedical data science at Stanford. The doctor would incorporate other visual information into their assessment — all things that could narrow any performance gaps between LLMs and humans.

For these reasons and more, said Alsentzer, "I don't think we should read this as language models outperforming humans."

Clinical AI researchers pointed out issues in the study's methodology that could temper its findings. Model performance, for example, was based on subjective rating scores from experienced clinicians, including what counts as a "very close" diagnosis to the ground truth. "To some degree the results here are obscured by the clinician's judgment of what is a good differential versus not," said Alsentzer. The ratings' objectivity can be called into question when they are delivered by senior authors, said Oermann, a criticism Rodman called "fair."

They also emphasized that the study, while it included physician baselines, used relatively small comparison groups. In the emergency department experiment, diagnostic scores came from only two physicians — "and really, one physician's low scores drives the difference," noted Oermann. Rodman defended the comparison and the skill of the physician participants, emphasizing that diagnostic accuracy isn't a primary goal of emergency medicine.

In the end, clinical AI researchers agree that large language models' diagnostic performance points toward one thing: It's time to test AI in prospective clinical trials, especially to see how they perform when used alongside and by doctors.

"Despite this really strong performance we're seeing in these simulated settings, how do we get the best out of both of us within the real clinical setting?" said Brodeur. "That's really what we are after over the next couple of years." That means exposing LLMs to incomplete information about the patient, data inputs beyond text, and all the kinds of patients clinicians would see in typical practice, said Alsentzer.

Those kinds of trials will have to be carefully designed to protect patient safety when LLMs make inevitable mistakes. And they will have to confront difficult questions about when and how AI is most likely to influence patient care and outcomes. The physicians who made up the baseline scores in each of the Science experiments were largely internists and generalists who work at large academic medical centers in Massachusetts and California. “I have every reason to believe there would be different performance,” said Rodman, if models were compared to specialists or in the context of specialist care.

It’s easier said than done. “Often in the tech industry, people don’t have an appetite for running a 50,000 person, one or two year study,” said Oermann. “In one or two years, the technology’s totally changed.” But the field is starting to get there.

OpenAI and Penda Health, a network of primary care clinics in Nairobi, last year published a preprint of a prospective study showing fewer treatment errors with AI support. A randomized trial in Pakistan showed diagnostic reasoning improvements for physicians using AI.

And the researchers on the Science paper have long since moved on from the results being published this week. In March, Rodman, Brodeur, and others published the results of a prospective study of Google’s conversational clinical agent AMIE in primary care as a preprint, work that Oermann called “phenomenal:” “In my mind, this is exactly what we need,” he said.

“I’m not being a hypocrite here,” said Rodman. “I did this experiment, and then I frickin’ ran a clinical trial.”