News and Perspectives

Generalist Versus Specialist: What Is the Best AI Model for Health Care?

Simon Spichak, JMIR Correspondent

Abstract

Recent research pitting general large language models against specialized models widely used to support clinical decision-making in health care has caused heated debate. In this *News and Perspectives* article, JMIR Correspondent Simon Spichak unpacks the controversy and more critical question of the impact of these models on real patient outcomes.

Key Takeaways:

- A contested study fueled controversy when it found that frontier models outperform OpenEvidence on benchmarks.
- Although clinicians regularly use both frontier and specialized AI models, there's a lack of independent evaluations and real-world data assessing whether they ultimately improve patient outcomes.

More than 80% of physicians are using AI tools in their medical practice. According to the company, an estimated 40% are using OpenEvidence—a large language model (LLM) trained on medical literature that can provide suggestions for care and medication options—and many hospitals are rolling out chatbot co-pilots for physicians built atop general-purpose LLMs, which are advertised as Health Insurance Portability and Accountability Act–compliant.

Among AI adopters, there is palpable tension over the potential advantages and trade-offs of using a generalized model instead of a specialized medical model. The tension boiled over as Eric K Oermann, MD, and colleagues, researchers from NYU Langone, published a study benchmarking and comparing the specialized medical LLMs to generalized LLMs. The frontier models outperformed the specialized models developed by UpToDate and OpenEvidence across three metrics. The findings came as no surprise to many experts.

"Once general models get large enough and complex enough that they can incorporate text, images, or data from specialty domains, they often outperform smaller bespoke models," says MIT researcher Marzyeh Ghassemi, PhD, adding that despite their limitations, the benchmarks are commonly used to test models when researchers "want to verify that a model does well in a controlled environment."

Still, the research was swiftly criticized by UpToDate and OpenEvidence. In public statements, OpenEvidence cited "a massive undisclosed conflict of interest and irredeemable methodological flaws" and, over email, shared a letter sent to *Nature Medicine*'s editors calling for retraction and a public apology.

The research raises questions about how LLMs are assessed before they're integrated into the medical clinic and whether real-world evidence supports their use.

Judging AI

Oermann and his colleagues evaluated OpenEvidence and UpToDate Expert AI against GPT-5.2, Gemini 3.1 Pro, and Claude Opus 4.6. They used two common benchmarking tools, HealthBench and MedQA, to assess how the models answered 500 questions using a panel of 3 AI models to assess the answers.

Peer reviewers pointed to a major limitation. Generalized models are trained on the answers to the questions in the databases, biasing them to judge frontier models more favorably than other models.

In response to reviewer feedback, the researchers then developed a set of 100 real clinical questions (RCQs) based on real physician queries, with each response evaluated by 3 of 12 randomized, blinded, peer-reviewer physicians. Physicians ranked RCQ responses on clinical correctness, clarity, completeness, and safety/harm avoidance. The holistic ranking, says Ghassemi, would put frontier models at an advantage since they're more fluent and conversational. "It's not surprising to me that, you know, clinicians had a better sort of experience interacting with these agents," she says.

Still, OpenEvidence criticized the use of HealthBench and MedQA. "Many academics have been forced to resign their academic posts for far lesser cases of academic malpractice and naked intellectual dishonesty—especially when outside reviewers highlighted the obvious data contamination issue," Daniel Nadler, PhD, cofounder of OpenEvidence, noted over email.

Meanwhile, UpToDate wrote in a statement that "The study's limited, non-reproducible sample and lack of a rigorous clinical reference standard make it difficult to draw meaningful conclusions about real-world care." In the supplemental data, the researchers had provided the rubric used for RCQs and sample questions, but not the entire dataset of questions.

Adam Rodman, MD, MPH, Director of AI Programs at the Shapiro Center for Research and Education, and Ghassemi, neither of whom were involved in the study, both found the comparisons on RCQ data were the most compelling parts of the study. While he agrees there are limitations, "it is a study that is in line with a lot of the evaluation literature," says Rodman.

On X, OpenEvidence also claimed that there was an undisclosed conflict of interest by the authors, who had asked the company for an API to power their in-house medical AI to build a competing product, but when OpenEvidence declined, "this paper coincidentally appeared." The study, which was published on June 12, 2026, was first submitted for review on December 1, 2025.

"We continue to stand by our study and its results and disclosed relevant author interests in accordance with institutional and journal policies and procedures," Oermann said over email. He declined to comment further.

Since then, a preprint conducting a similar analysis found that OpenEvidence outperformed generalized models. The study's disclosure statement notes that, while none of the authors are affiliated with the company, OpenEvidence was involved in developing and implementing the data collection and paid physicians to complete blinded surveys assessing the model answers.

Real-World Use and Evidence

Despite studies showing how well these products work on benchmarks, there is a lack of data linking them to real-world clinical outcomes.

Gilles Frydman, BSc, who started one of the earliest patient communities in 1985 and is the founder of Synambix LLC and PatientsUseAI, says that study "doesn't correlate or even relate to real care," since it uses benchmarks rather than real-world outcomes. "And anything in medicine that doesn't directly connect with the patient and patient care is, in my opinion, either a distraction or, worse, toxic."

Other experts also pointed to the lack of real-world data collected on these tools. "My view as a researcher in this area is that we have insufficient evidence to claim that any of these systems will lead to improvements in patient outcomes," says Ghassemi.

Rodman sees OpenEvidence and UpToDate as reference materials that provide clinical decision support. "I don't think anyone's expectation is that a reference material necessarily affects patient outcomes," he says.

Joel Selanikio, MD, a pediatrician at Georgetown University Hospital, says that, compared to frontier models, "I get better answers from OpenEvidence" in part because its layout is designed with clinicians in mind, and it provides references for its assertions. But he does run into issues from time to time. "I think it would be crazy if I actually followed every recommendation of OpenEvidence for every patient, and a huge waste of resources," he says.

He anticipates that whatever company spends the most on training their medical AI will eventually come out on top. He still wants to see more independent evaluations. "What we don't have is a framework for continued evaluation that includes both the medical ones against the general ones and also the medical ones against each other."

> *"And anything in medicine that doesn't directly connect with the patient and patient care is, in my opinion, either a distraction or, worse, toxic.*
>
> Gilles Frydman, BSc

Rodman is also a regular user. "For a simple evidence lookup, like a drug dose, I'll use OpenEvidence," he says. "If I have a tricky case where I need a second opinion, I will reframe it to a foundation model."

In a recent perspective, Rodman and his colleagues wrote that they foresee that medical reasoning AI could act as a collaborative aid and help with decision-making.

Meanwhile, Ghassemi urges for more research as generative AI continues to be rapidly adopted in health care. "AI has entered our lives in an unevaluated, under-regulated way," she says. "We need much, much more research to be done on whether these administrative, or informative, or informal uses of AI in health actually lead to better patient health outcomes, because I don't think we're there yet."

Keywords: artificial intelligence; large language models; clinical decision support systems; decision-making; computer-assisted; evidence-based medicine; patient outcome assessment; benchmarking; reproducibility of results; AI

Conflicts of Interest

None declared.

Please cite as:
Spichak S
Generalist Versus Specialist: What Is the Best AI Model for Health Care?
J Med Internet Res 2026;28:e108066
URL: https://www.jmir.org/2026/1/e108066
doi: 10.2196/108066