

Health Care AI, Intended To Save Money, Turns Out To Require a Lot of Expensive Humans

Artificial intelligence algorithms are used to predict patients' risk of death, suggest diagnoses, approve insurance claims and more.

January 16, 2025 By Darius Tahir and KFF Health News

Preparing cancer patients for difficult decisions is an oncologist's job. They don't always remember to do it, however. At the University of Pennsylvania Health System, doctors are nudged to talk about a patient's treatment and end-of-life preferences by an artificially intelligent algorithm that predicts the chances of death.

But it's far from being a set-it-and-forget-it tool. A routine tech checkup revealed the algorithm decayed during the COVID-19 pandemic, getting 7 percentage points worse at predicting who would die, according to a 2022 study.

There were likely real-life impacts. Ravi Parikh, an Emory University oncologist who was the study's lead author, told KFF Health News the tool failed hundreds of times to prompt doctors to initiate that important discussion — possibly heading off unnecessary chemotherapy — with patients who needed it.

He believes several algorithms designed to enhance medical care weakened during the pandemic, not just the one at Penn Medicine. "Many institutions are not routinely monitoring the performance" of their products, Parikh said.

Algorithm glitches are one facet of a dilemma that computer scientists and doctors have long acknowledged but that is starting to puzzle hospital executives and researchers: Artificial intelligence systems require consistent monitoring and staffing to put in place and to keep them working well.

In essence: You need people, and more machines, to make sure the new tools don't mess up.

"Everybody thinks that AI will help us with our access and capacity and improve care and so on," said Nigam Shah, chief data scientist at Stanford Health Care. "All of that is nice and good, but if it increases the cost of care by 20%, is that viable?"

Government officials worry hospitals lack the resources to put these technologies through their paces. “I have looked far and wide,” FDA Commissioner Robert Califf said at a recent agency panel on AI. “I do not believe there’s a single health system, in the United States, that’s capable of validating an AI algorithm that’s put into place in a clinical care system.”

AI is already widespread in health care. Algorithms are used to predict patients’ risk of death or deterioration, to suggest diagnoses or triage patients, to record and summarize visits to save doctors work, and to approve insurance claims.

If tech evangelists are right, the technology will become ubiquitous — and profitable. The investment firm Bessemer Venture Partners has identified some 20 health-focused AI startups on track to make \$10 million in revenue each in a year. The FDA has approved nearly a thousand artificially intelligent products.

Evaluating whether these products work is challenging. Evaluating whether they continue to work — or have developed the software equivalent of a blown gasket or leaky engine — is even trickier.

Take a recent study at Yale Medicine evaluating six “early warning systems,” which alert clinicians when patients are likely to deteriorate rapidly. A supercomputer ran the data for several days, said Dana Edelson, a doctor at the University of Chicago and co-founder of a company that provided one algorithm for the study. The process was fruitful, showing huge differences in performance among the six products.

It’s not easy for hospitals and providers to select the best algorithms for their needs. The average doctor doesn’t have a supercomputer sitting around, and there is no Consumer Reports for AI.

“We have no standards,” said Jesse Ehrenfeld, immediate past president of the American Medical Association. “There is nothing I can point you to today that is a standard around how you evaluate, monitor, look at the performance of a model of an algorithm, AI-enabled or not, when it’s deployed.”

Perhaps the most common AI product in doctors’ offices is called ambient documentation, a tech-enabled assistant that listens to and summarizes patient visits. Last year, investors at Rock Health tracked \$353 million flowing into these documentation companies. But, Ehrenfeld said, “There is no standard right now for comparing the output of these tools.”

And that’s a problem, when even small errors can be devastating. A team at Stanford University tried using large language models — the technology underlying popular AI tools like ChatGPT — to summarize patients’ medical history. They compared the results with what a physician would write.

“Even in the best case, the models had a 35% error rate,” said Stanford’s Shah. In medicine, “when you’re writing a summary and you forget one word, like ‘fever’ — I mean, that’s a problem, right?”

Sometimes the reasons algorithms fail are fairly logical. For example, changes to underlying data can erode their effectiveness, like when hospitals switch lab providers.

Sometimes, however, the pitfalls yawn open for no apparent reason.

Sandy Aronson, a tech executive at Mass General Brigham’s personalized medicine program in Boston, said that when his team tested one application meant to help genetic counselors locate relevant literature about DNA variants, the product suffered “nondeterminism” — that is, when asked the same question multiple times in a short period, it gave different results.

Aronson is excited about the potential for large language models to summarize knowledge for overburdened genetic counselors, but “the technology needs to improve.”

If metrics and standards are sparse and errors can crop up for strange reasons, what are institutions to do? Invest lots of resources. At Stanford, Shah said, it took eight to 10 months and 115 man-hours just to audit two models for fairness and reliability.

Experts interviewed by KFF Health News floated the idea of artificial intelligence monitoring artificial intelligence, with some (human) data whiz monitoring both. All acknowledged that would require organizations to spend even more money — a tough ask given the realities of hospital budgets and the limited supply of AI tech specialists.

“It’s great to have a vision where we’re melting icebergs in order to have a model monitoring their model,” Shah said. “But is that really what I wanted? How many more people are we going to need?”

[This story](#) was published by KFF Health News on January 10, 2025. It is republished with permission.