

[Skip to Main Content](#)

[Save 25% on up to 10 users with a STAT+ Group](#)

STAT [Reporting from the frontiers of health and medicine](#)
STAT



[Video](#)

Don't miss out

Subscribe to STAT+ today, for the best life sciences journalism in the industry

[Learn more](#)

Epic's overhaul of a flawed algorithm shows why AI oversight is a life-or-death issue



By [Casey Ross](#)

Oct. 24, 2022



Molly Ferguson for STAT

Epic, the nation's dominant seller of electronic health records, was bracing for a catastrophe.

It was June 2021, and a [study](#) about to be published in the Journal of the American Medical Association had found that Epic's artificial intelligence tool to predict sepsis, a deadly complication of infection, was prone to missing cases and flooding clinicians with false alarms. Reporters were clamoring for an explanation.

Epic executives quickly drafted a statement knocking down the findings and rushed to reassure worried customers. It followed up with a [blog post](#), its first in six months, pushing back against the implication that it wasn't being forthcoming: "Tens of thousands of clinicians have access to the sepsis model and transparency into how it works," Epic wrote.

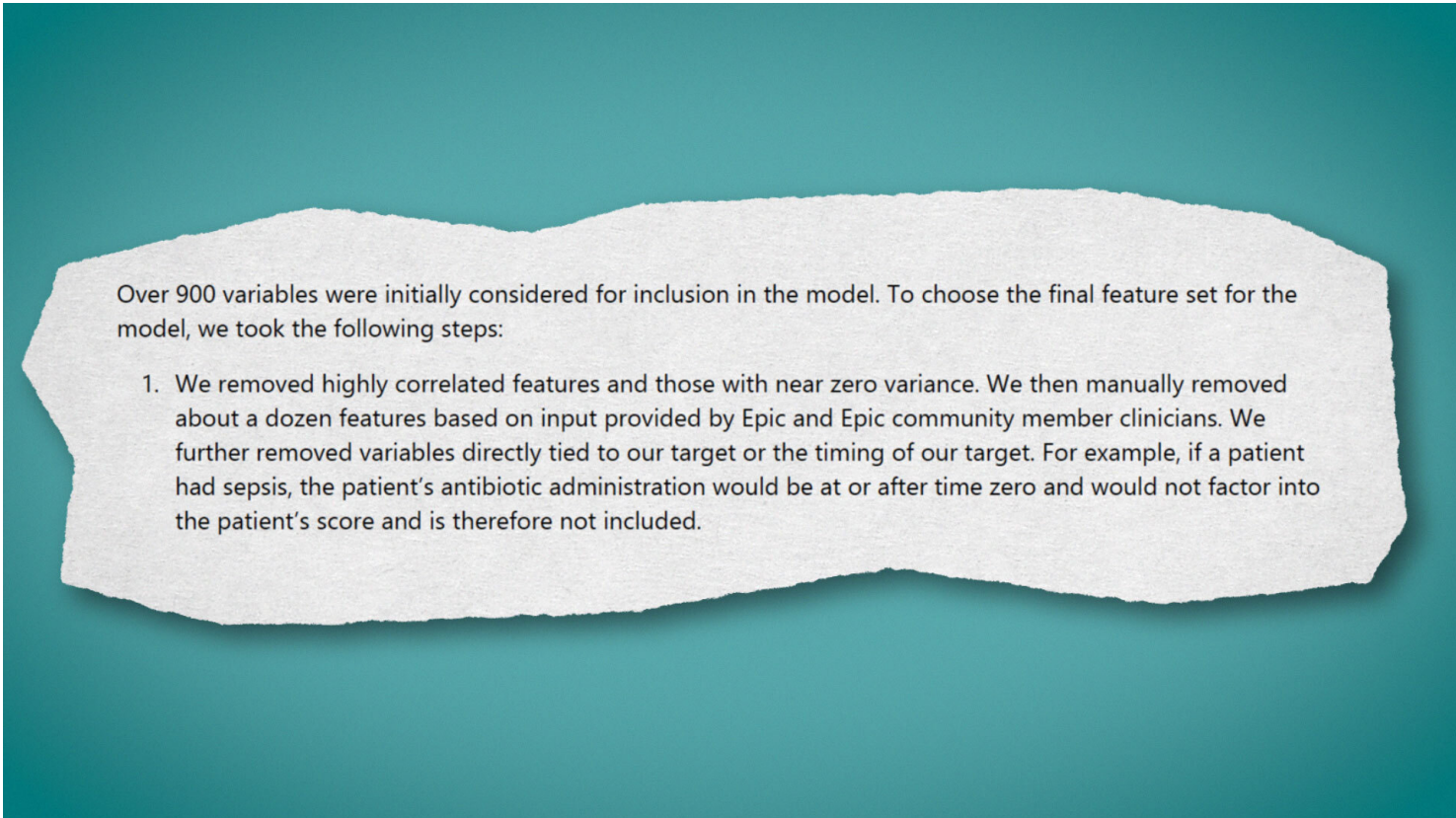
A full-scale crisis was averted. But barely.

A year later, after a [series of investigations](#) by STAT, the company [released a re-engineered version](#) of the model it had steadfastly defended. Epic changed the data variables it uses, its definition of sepsis onset, and its guidance for tuning the algorithm to local patients. Even the user guide for the hundreds of hospitals it serves nationwide was entirely different — and twice as long.

While it is common for software companies to upgrade products after their initial release, this was a wholesale remaking of an algorithm used to guide decisions about millions of seriously ill patients in U.S. hospitals at any given time. A missed case of sepsis doesn't cause annoyance or irritation. It very often leads to death.

Epic is not the only company moving aggressively to sell AI tools to health systems. But the precarious rollout of its popular sepsis algorithm has become a case study in the challenges of ensuring such algorithms are used safely and effectively at the bedside. It also underscores shortcomings in procedures for evaluating and regulating AI products, which risk giving faulty advice to doctors and nurses trying to make time-sensitive decisions about very sick people.

Federal regulators and hospital leaders are scrambling to anticipate problems likely to arise with AI systems without knowing exactly how the lines of responsibility should be drawn. The Food and Drug Administration published a [recent guidance](#) that put the regulation of sepsis alerts and other AI predictors squarely within its purview. But it has not made clear whether it will require a review process before those products go on the market, as is required of many algorithms that interpret medical images.



Over 900 variables were initially considered for inclusion in the model. To choose the final feature set for the model, we took the following steps:

1. We removed highly correlated features and those with near zero variance. We then manually removed about a dozen features based on input provided by Epic and Epic community member clinicians. We further removed variables directly tied to our target or the timing of our target. For example, if a patient had sepsis, the patient's antibiotic administration would be at or after time zero and would not factor into the patient's score and is therefore not included.

A model brief distributed to users of version 2 of Epic's sepsis model describes the rationale for the removal of antibiotic orders as a variable to predict the onset of sepsis. The use of antibiotics in the original version of the model often resulted in late alarms.

The White House, meanwhile, cited the initial concerns raised about Epic's sepsis model in its "[AI Bill of Rights](#)," a document that broadly calls for greater transparency and quality controls around the use of automated systems. And many experts in critical care medicine are saying that evaluation and oversight of AI in medicine must be more rigorous.

"The lack of standard empiric evidence supporting these algorithms is really bothersome for me," said Derek Angus, a physician at the University of Pittsburgh Medical Center and expert in treating sepsis. "We don't really know with strong enough cause when these algorithms are helping, and similarly we're not really sure when they're hurting."

That's particularly problematic, he said, when the algorithm developer asserts — as Epic does in marketing materials — that its product is saving lives. "That's like the use of a drug," Angus said. "It's fundamentally changing the way care is delivered in a hospital."

But Epic's tool was treated nothing like a drug. It wasn't subjected to an FDA review process or third-party testing before its initial release. No independent systems were put in place to monitor its use across hospitals or report problems. And its benefits and side effects were not measured in a randomized trial before it was plugged in at facilities around the country.

A spokesperson for Epic declined to answer questions about the changes to its sepsis model or the push for stepped-up oversight of AI products.

The new version [corrects several problems](#) raised by STAT's reporting. Epic is now recommending that its model be trained on a hospital's own data prior to clinical use, a major shift aimed at ensuring its

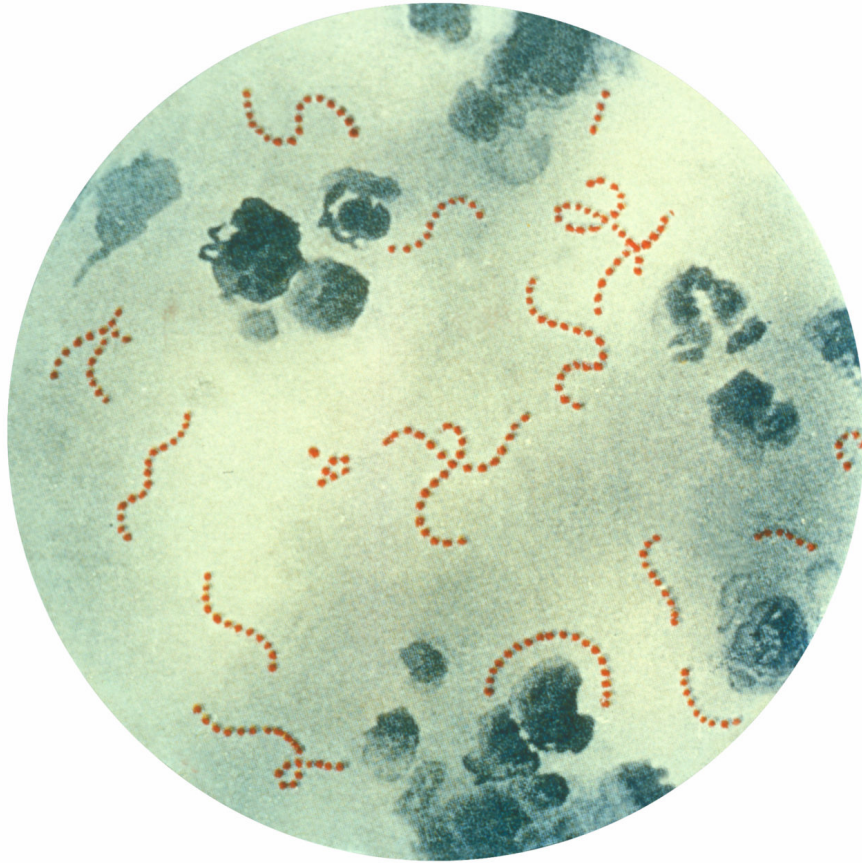
recommendations are relevant to the patients a hospital sees. Several hospitals found the model's accuracy to be much lower than Epic advertised in its initial user guide, suggesting it struggled to deal with differences in patient populations and the variability in sepsis symptoms.

The company also switched its definition of sepsis onset to a more commonly accepted standard and removed clinician orders for antibiotics as an input variable. STAT found the model's use of antibiotics to be [particularly problematic](#), creating a kind of circular logic that often resulted in [late alarms](#). It was essentially using a doctor's response to sepsis to predict that the condition was about to take hold. While that may have boosted accuracy in testing, it meant in real life, the algorithm was often telling doctors something they already knew.

“It didn't have enough lead time; in fact, sometimes it didn't have any lead time,” said Vincent Major, a professor of population health at New York Langone Health. He said the health system tested the Epic model, but opted to develop its own alert systems for adult and pediatric patients that could be tailored to its own data and practices.

“What makes sepsis a challenging use case for AI is that it's almost the scariest clinical thing to be predicting,” Major said. “It's time-sensitive. It's a big deal with terrible patient outcomes, and there's already a whole bundle of interventions.”

To make automated alerts truly useful, experts said, algorithms must recognize trends in the data much earlier, and cut through the normal hospital bureaucracy to prompt urgent action.



Streptococcus pyogenes is one of many bacterial infections that can lead to sepsis. *CDC*

“If you wait for humans to recognize sepsis, that’s too late,” said Shamim Nemati, a professor of biomedical informatics at UC San Diego School of Medicine, who is preparing to launch a clinical trial to test a sepsis [alert system](#) he developed. The tool seeks to intervene at the diagnosis stage, automating the ordering of screening tests to confirm cases of sepsis and prompt faster treatment.

Nemati has also submitted the tool for approval from the FDA. He said evaluation of AI products, particularly for sepsis, should be carried out in stages, from an initial retrospective comparison of its predictions to patient outcomes, to validation studies prior to clinical use, to randomized testing, and finally ongoing surveillance after systems are deployed in live settings.

“We need rigorous protocols across the board to ensure that we’re evaluating these systems properly,” he said. “These algorithms are being used at the bedside. Patient safety is at stake.”

As Epic was preparing to launch its sepsis tool in 2017, several factors were pushing against that kind of slow and methodical evaluation. A new generation of products using machine learning, a subtype of AI, was generating heaps of publicity and hype.

Hospitals were also under increasing pressure to improve their treatment of sepsis, which kills nearly 270,000 Americans a year, often because it is not discovered in time. The federal government measures hospitals’ quality based in part on their sepsis outcomes and whether they are following standard

protocols for treating it. Furthermore, the Centers for Disease Control and Prevention [launched a campaign](#) that year emphasizing the need to more proactively treat patients for sepsis and develop better processes for doing so.

All of which made AI an alluring option. Epic's tool, though just one of many available, was the easiest to install because it was already embedded in the company's widely-used health records software.

[UPCOMING EVENT](#)

Tackle the biggest questions in health and medicine

From virtual conversations on AI and CAR-T to multi-day summits featuring the most innovative speakers across the life sciences, explore upcoming STAT Events.

At the time, however, there were few standards for evaluating machine-learning algorithms and testing them in individual hospitals. At a minimum, most hospitals ran it in the background of their data systems, so they could see how it would respond to individual patients before the alerts were turned on for clinical use. But the extent of the evaluations — whether based on prior data or live patients, and how the effects were analyzed — varied widely.

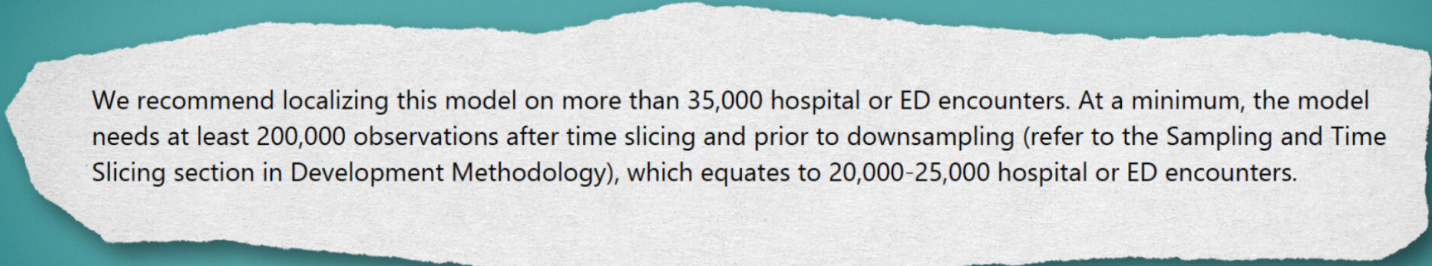
This approach is out of step with the high stakes for patients, experts said. “When you start to see concern raised that some of these algorithms could be having unwanted side effects, then I don't know why they are held to a lower standard,” Angus said.

The side effects can take different forms. An alarm that is not sensitive enough may miss cases of sepsis, causing clinicians to delay necessary treatment or not deliver it at all. Conversely, an algorithm that is tuned to catch all the cases might also trigger false alarms, diverting clinicians from other patients in crisis. In some situations, a false alarm may also prompt providers to prescribe unnecessary antibiotics, which could cause the patient to have an adverse reaction or fuel antimicrobial resistance.

At UC Health in Colorado, which paired Epic's sepsis algorithm with its own prediction models, the ratio of false alarms to true positives was about 30 to 1, according to CT Lin, the health system's chief medical information officer.

“Our doctors' response to us was, ‘Do you not think we're running as fast as we can already treating patients who are trying to die in front of us right now?’” Lin said. They simply didn't have time to click through dozens of alerts to find the one that might require urgent attention.

To deal with the high rate of false alarms, UC Health started using a remote team of clinicians to monitor the model's output and examine patients through a live video feed. When a patient truly seemed to be deteriorating, a remote clinician would call the bedside nurses. But then the bedside nurses, annoyed by the perceived intrusion, began putting coats over the cameras.



We recommend localizing this model on more than 35,000 hospital or ED encounters. At a minimum, the model needs at least 200,000 observations after time slicing and prior to downsampling (refer to the Sampling and Time Slicing section in Development Methodology), which equates to 20,000-25,000 hospital or ED encounters.

A brief for version 2 of Epic's sepsis model tells users that it should be trained on a hospital's own data prior to clinical use. Some hospitals found the original model was less accurate than advertised when applied to their patients.

“The way we had to solve that was to have all the virtual health nurses rotate in person, shake hands and go, ‘Hi, my name’s Amy. I’m the person on the other side of the camera,’” Lin said. The formula for effectively embedding the sepsis alarms in the hospital ended up being “20% solving the math problem, and 80% relationship building.”

In the end, the process produced a positive outcome: Patients received antibiotics to treat sepsis in half the time compared to before the sepsis tool was installed, within an average 40 minutes rather than 80 minutes. The health system estimates that speedier sepsis care saves 211 lives annually.

But the process UC Health developed around the use of the tool was just as important as the tool itself. Since hospitals will inevitably make different decisions about how to embed sepsis alerts and other AI tools in their treatment practices, that will make it harder to create standards around AI products and objectively measure their value.

“People are quite wary of whether or not these tools will fundamentally improve our patient care without the risk of becoming clinically burdensome,” said Vincent Liu, a physician and research scientist at Kaiser Permanente who develops predictive analytics tools.

But Liu said throwing out the tools, or abandoning the quest altogether, is not the right move either. Just as hit-or-miss hurricane warning algorithms can still prove beneficial, imperfect sepsis alerts can still drive improvements in care.

The problem, he said, is that so many people have been seduced by the belief that there is such a thing as a perfect global sepsis model that will dramatically improve outcomes — but without special effort to

ensure it's implemented in the right way, at the right moment of treatment.

“Those who think that’s the goal are missing the point,” he said. “Sepsis is just very heterogeneous. Patients come with different symptoms. They have different organ failures. It’s not as universal as we would like it to be, and all of those challenges make it difficult to have a really high-performing prediction model that clearly improves outcomes.”

This story is part of a series examining the use of [artificial intelligence in health care](#) and practices for exchanging and analyzing patient data. It is supported with funding from the [Gordon and Betty Moore Foundation](#).

About the Author



[Casey Ross](#)

National Technology Correspondent

Casey Ross covers the use of artificial intelligence in medicine and its underlying questions of safety, fairness, and privacy.

casey.ross@statnews.com
[@caseymross](#)

STAT encourages you to share your voice. We welcome your commentary, criticism, and expertise on our subscriber-only platform, [STAT+ Connect](#)



To submit a correction request, please visit our [Contact Us page](#).