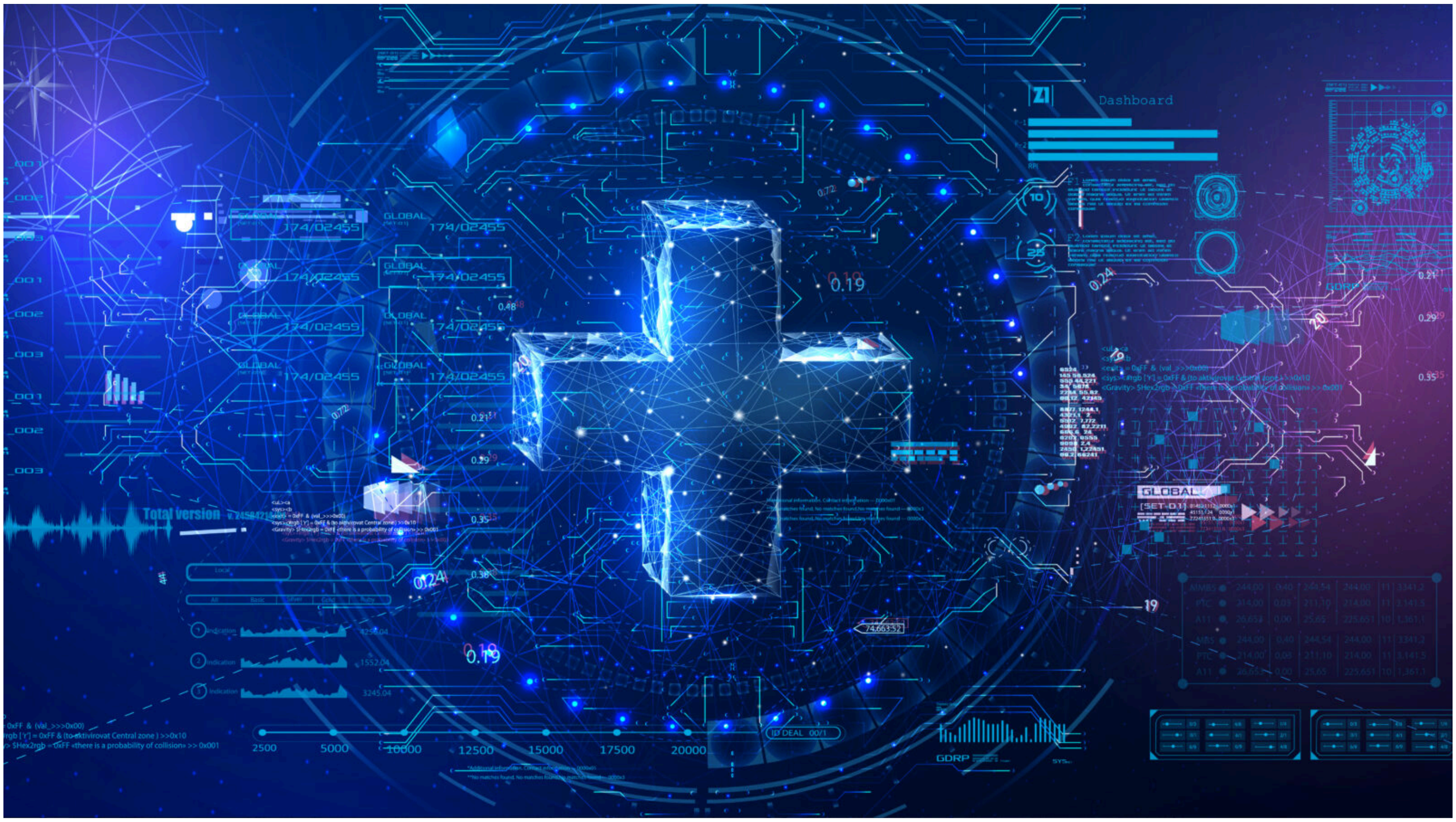
OPINIONS+

After FDA’s pivot on clinical AI, we need AI safety research more than ever

Research must keep up with the deployment of AI



Adobe

By Katherine E. Goodman and Daniel Morgan Jan. 15, 2026

Goodman is an assistant professor of epidemiology and public health at the University of Maryland School of Medicine. Morgan is a physician and AI researcher, and professor of epidemiology and medicine at the University of Maryland School of Medicine.

On Jan. 6, the Food and Drug Administration released updated guidance for clinical decision support (CDS) tools, relaxing key medical device requirements. With this change, many generative artificial intelligence tools that provide diagnostic suggestions or perform supportive tasks like medical history-taking — tools that probably would have required FDA sign-off under the prior policy — could reach clinics without FDA vetting.

Soon after the FDA's guidance release, two other major AI announcements also hit the news: Utah began a first-in-the-nation pilot with Doctronic for autonomous AI prescription refills, and OpenAI debuted ChatGPT Health, which will tailor responses to users' uploaded medical records and wearables data.

Together, these three developments have dramatically shifted the health care AI landscape. Given the abysmal state of patient access and notorious challenges navigating U.S. health care, there is a lot to be cautiously enthusiastic about. With no or minimal expense, patients could soon receive better health advice, possibly more involved medical support, and easier access to medication refills.

But for clinical safety researchers like us, this trifecta has also rapidly made theoretical safety concerns real, slingshotting two particular safety gaps to the forefront.

First, medicine is unprepared for "agentic" AI tools and chatbots that operate autonomously. Between the Utah announcement, FDA Commissioner Marty Makary's statements that the FDA is "here to promote AI" and "get out of the way," and ARPA-H's just-announced agentic AI program, we expect to see more companies releasing these products. For example, beyond Doctronic's prescription renewals chatbot, the company also offers an "AI doctor" that chats with users about medical symptoms, suggests diagnoses, and sometimes refers them to telehealth visits with human physicians.

From a safety perspective, agentic AI is challenging because it relies on a different approach than traditional "human-in-the-loop" CDS tools (where, at least in theory, a human checks the AI's outputs and takes any final action). Instead, agentic AI relies on "human-*on*-the-loop" oversight — the AI operates independently, but with human monitoring and experts taking over for complex or "edge" cases.

While human-on-the-loop may be a comforting metaphor, it's less akin to an all-seeing factory manager looking out over the factory floor and more like the school principal who doesn't know there's a problem until a student appears at his or her office. In other words, the system relies on AI identifying edge cases, recognizing its own limits, and escalating up the chain to a doctor.

This is worrisome, because for all of AI's success at acing medical exams and solving complex diagnostic challenges, it's much worse at knowing when to ask for help or admit uncertainty. For clinical agentic AI, we'll probably need AI that is less confident and more cautious. Developers can train for these qualities, but without regulatory sticks companies may lack incentives to build them.

Second, whether it is agentic AI escalating cases or non-agentic AI providing recommendations, at some point the AI must perform a hand-off to doctors. While hand-offs are not a new software challenge — CDS tools have always "handed off" suggestions, in the narrow sense — in this new relaxed regulatory climate coupled with AI's rapid technical advances, they'll be much more clinically critical.

For example, expect a not-too-distant future in which chatbots routinely go through pre-visit checklists with patients or take histories, then hand off summaries to clinicians. (Even patient-facing tools like ChatGPT Health will presumably offer summaries or scripts for users to present to doctors.) In another example, tools embedded directly in clinical workflows such as AI scribes may soon provide comprehensive diagnostic and management suggestions based on clinical conversations that clinicians themselves do not closely remember.

Hand-offs and information exchanges are some of the highest-risk moments in health care and "where safety often fails first." It is why shift changes are generally the most dangerous hour of the hospital day and explains why roughly 80% of adverse events result from inadequate handovers. We need studies to understand better and worse ways to execute AI hand-offs, recognizing that what is optimal between human providers may differ for AI-to-human hand-offs.

Critically, this need goes beyond simply making AI outputs explainable and traceable, one of the FDA's key criteria for bypassing CDS tools from medical device review. Rather, it's about determining how much information to provide (and when) and how to best present it. For example, which is better: bullet points or short summaries with bolded information? Should AI suggest lengthy differential diagnoses, or only top diagnostic possibilities until common diagnoses are ruled out? Too little information risks patients falling through the cracks. But too much could drown out salient details or trigger harms for patients like over-testing.

So, where do we go from here? Researchers must rapidly ramp up the types of studies described above, focusing on optimizing safety for agentic AI and AI-to-human hand-offs. Some of this research might involve smaller, controlled studies of clinician behavior using synthetic data, but others would require larger pragmatic clinical trials to track real physician decisions and patient outcomes. And more generally, real-world evaluations will become increasingly important as products are deployed, to catch problems quickly.

Fundamentally, however, the field needs a cultural shift to prioritize clinical AI safety research and pressure companies to collaborate. Otherwise, much of this research will never get off the ground without funding to support it and, more critically, data to study. Given that much of these data will exist outside of traditional clinical environments, researcher access will depend on voluntary data sharing by companies.

Absent such access, health care researchers will be sidelined, which is bad for patients, but also companies. Without independent research, careful review may not occur until patients suffer harm, and AI health care companies could find themselves in a similar situation to Waymo — probably safer overall, but struggling to gain public trust.

Last week’s AI developments are promising, but rapid progress without clear guardrails should keep doctors and patients awake at night. Quickly, medicine’s focus must pivot to critical clinical AI safety gaps, specifically around agentic AI and AI hand-offs. Academic researchers are ideally suited to perform this work, but not without data access and adequate investment. Meeting this challenge requires independent safety research, an essential safety net in this new world where the FDA is “out of the way.”

Katherine E. Goodman, Ph.D., J.D., is an assistant professor of epidemiology and public health at the University of Maryland School of Medicine whose research focuses on novel AI evaluation and regulation. Daniel J. Morgan, M.D., M.S., is a physician and AI researcher, and professor of epidemiology and medicine at the University of Maryland School of Medicine. He directs the Center for Innovation in Diagnosis.

Letter to the editor

Have an opinion on this essay? Submit a letter to the editor.