News and Perspectives

From Hallucinations to Uncertainty: The Next Challenge for Clinical AI

Elena Ikonomovska, JMIR Correspondent

Abstract

Clinical AI continues to improve, but are existing models truly ready for the clinic? In this *News and Perspectives* article, JMIR Correspondent Elena Ikonomovska reports on the evolution of clinical AI evaluation, sharing her opinion on the next frontier that large language models must reach in order to support holistic, effective clinical care.

Key Takeaways:

- Accuracy isn't the same as reasoning, and while leading large language models are increasingly accurate, dealing with uncertainty continues to be a challenge.
- Clinical AI should be judged on whether it shows honest uncertainty, with every claim tied to evidence that a clinician can audit, and it should keep competing hypotheses visible.

Elena Ikonomovska, PhD, is a cofounder and the CEO of Diadia Health—a clinical reasoning health platform for clinicians managing complex chronic health cases. A serial founder with a PhD in machine learning, Elena holds 2 decades of data science experience.

For much of the last few years, the seemingly golden metric for evaluating clinical AI was its rate of hallucinations. A model that fabricated a drug interaction or invented a citation was unsafe, while clean, clear, accurate conclusions were the primary goal. From here, the field of clinical AI organized itself around driving hallucinations down. The effort has largely worked. As outlined in my recent preprint, today's large language models (LLMs) hallucinate far less than their predecessors, and on the surface, they look clinic-ready.

In April 2026, researchers at Mass General Brigham published a study in *JAMA Network Open* that complicates the picture. They ran 21 leading models through a series of clinical cases, and when handed complete patient information, the models reached the correct final diagnosis more than 90% of the time on average. On the flip side, when faced with the need for a differential diagnosis, the same models failed to produce an appropriate answer more than 80% of the time. Herein lies the crucial difference between clinicians and current LLMs; as the researchers noted, "clinicians preserve uncertainty and iteratively refine differential diagnoses, whereas LLMs collapse prematurely onto single answers...."

Flash forward to now, and clinical AI accuracy has improved while reasoning largely has not. Yet, reasoning is where so much of the energy poured into practicing medicine is spent. Rarely do clinicians work from a complete file and find only one right answer at the end of the road. Along the way, clinicians weigh competing explanations, sit with incomplete and evolving evidence, and hold several possibilities open until the picture resolves.

Frontier Models Handle Uncertainty Poorly

As a data scientist, I spent 2 decades building machine learning systems before turning my efforts toward medicine. As a patient, I was handed panels that all read "normal" while knowing that something was wrong with my body. On the design and build side, I understood why hallucination had established itself as a primary metric because, well, it is easy to measure. But as a patient, medicine was failing me. You can count false statements. Far harder to count is a more common failure: the right answer looking less plausible in light of missing data and being seen as negative or wrong. The trouble is that many models struggle to clearly communicate what is known, what is plausible and should be explored, and what remains uncertain in a way that supports nuanced clinical reasoning, leads to the right choices, and reflects how good medicine is practiced.

LLMs still handle uncertainty poorly. They overstate a conclusion; offer one without a chain of evidence that a clinician can follow; or, hedging against risk, fall back on the statistically average opinion or options that a clinician has already worked through. A recent *Nature Medicine* editorial warned that without a clear connection between claims and evidence, medical AI risks being adopted faster than its real value can be understood.

Simply stated, medicine doesn't advance when clinicians employ the exact same tactics for each complex patient case. Medicine advances when clinicians can take what is known and verified, what is supported by science, what is likely

but possibly untested, and what the pooled data show, and develop a plan for the patient in front of them.

Clinicians Need Honest Tools for the Deluge of Data

Traditional medicine is based on aggregate-level statistics—on what worked for the average patient in a population study—an approach that struggles with edge cases. A complex chronic illness case, for example, rarely resembles the average. Genetics, thyroid function, metabolism, and gut composition must be evaluated not just one panel at a time but also in terms of how they influence each other. Which signal is driving the others? How do these signals relate to other relevant clinical information? A single marker read in isolation or as part of an average cannot determine root causes nor reliably point toward effective treatment paths.

> *Anyone evaluating clinical AI must inspect its ability to respond with honest uncertainty where uncertainty exists, and be able to quantify the plausibility of its responses accordingly.*

Now, the need is to construct a new dataset for unique biological circumstances. Frequently, the data needed to parse these complex clinical cases exist but are scattered across systems, are separated in time, and are sometimes missing critical variables. Beyond this, we're seeing that patients are not waiting for the data to land in their laps. They are actively seeking these data out on their own, often in the form of consumer diagnostics and wearables, and are increasingly turning to LLMs for medical advice. Patients arrive at the clinic having already run their symptoms through a chatbot and their blood through a longevity panel. Clinicians need tools that can read the deluge of data and reason transparently over it, without overtaking the judgment that must stay with the clinician.

Solving these challenges is the next frontier for AI. Anyone evaluating clinical AI must inspect its ability to respond with honest uncertainty where uncertainty exists, and be able to quantify the plausibility of its responses accordingly.

Oversight Is Only Possible When a System Is Built to Be Overseen

In practice, honest uncertainty is concrete. Tie each claim to the evidence that supports or challenges it, so a clinician can see whether a recommendation rests on a randomized trial or a single mechanistic study. Separate a claim that the evidence contradicts from one that the data cannot yet answer. Keep competing hypotheses on the table instead of discarding them—the way a good differential diagnosis stays an option until testing excludes it. Mark the boundary where established knowledge ends and inference begins, rather than smoothing over it. When the evidence genuinely disagrees, let the disagreement show instead of resolving it toward whatever the model produced first.

These changes will only make our clinical AI models more useful and our clinicians more effective, allowing them to serve more patients with higher-quality care. A clinician who can see that a suggestion is a well-supported first option treats it differently from one who is told, in the same confident tone, that an unproven inference is worth trying. Both can have value, but a system that projects false confidence can unintentionally narrow the clinician's field of view and rule out the less obvious lines of inquiry that hard cases and medical progress depend on. The clinician stays the one who decides. What has to change is the quality of the information they have to decide with.

> *The clinician stays the one who decides. What has to change is the quality of the information they have to decide with.*

The Mass General Brigham authors reached a similar conclusion: off-the-shelf models are not yet ready for unsupervised clinical-grade deployment and still need close oversight. As someone who builds and continuously refines such models, I would take this a step further: oversight is only possible when a system is built to be overseen. You cannot audit what models don't show you. You can only audit a line of reasoning with its evidence attached. The model that can honestly evaluate what is known, what is plausible, and what remains unknown—all tied to evidence—is a model that rises to meet clinicians.

Keywords: artificial intelligence; large language models; machine learning; clinical decision support systems; health information; patient-centered care; chronic disease; medical complexity

Please cite as:
Ikonomovska E
From Hallucinations to Uncertainty: The Next Challenge for Clinical AI
J Med Internet Res 2026;28:e116270
URL: https://www.jmir.org/2026/1/e116270
doi: 10.2196/116270

© JMIR Publications. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 23.Nov.2026