The car is here

Clinical AI has crossed the threshold.
The limiting factor is no longer the technology.

AUTHOR
Ali Khan

RESEARCHER
Dorothea Dwomoh

FEATURING
Bill Golden
Rob Duffy
Aman Bhasin
Grant Tarbox, D.O.
Peter Long

READING TIME
7 min read

Clinical AI has crossed the threshold. The limiting factor is no longer the technology.

A 36-year-old patient complains of abdominal pain. He has a spiking fever and falling oxygen: a bloodstream infection with no obvious source.

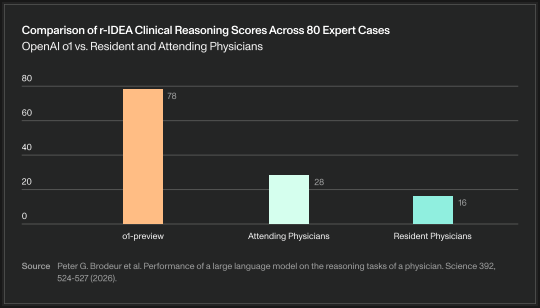
Researchers at Harvard Medical School ask two diagnosticians to solve the case, and in October 2025, the New England Journal of Medicine publishes their reasoning side by side.

The first is Dr. Gurpreet Dhaliwal of the University of California, San Francisco, whose colleagues call him a "real-life Dr. House."[1] Working from scratch, he reasons his way to an improbable answer: the patient, a heavy drinker, has swallowed a wooden toothpick without noticing. It has pierced his small intestine and let gut bacteria pour into his blood.

The second is Dr. CaBot, a clinical AI reasoning system. It reaches nearly the same conclusion—minus the toothpick.[2]

This was not a one-off. In April 2026, the same Harvard researchers pitted OpenAI's o1-preview model against hundreds of residents and attending physicians to test their clinical reasoning over a set of 80 cases.

The o1 model got a perfect Revised-IDEA score (a method to assess clinical reasoning quality) on 78 out of the 80 cases: 50 more than the best-performing attendings' mean results.[3]



After years of false starts and dead ends, machines that can reason like a human have finally arrived.

You are right to be skeptical

If you have led a health plan through the last decade of technology promises, you have good reason to be skeptical. The machine learning wave seemed promising, but ultimately fell short of the nuanced clinical reasoning our industry needs.

Starting in 2012, IBM Watson developed an oncology decision support tool from a small number of synthetic (fictional clinicals created from real-life scenarios) cancer cases. In 2017, STAT discovered internal documents that showed the tool recommended "unsafe and incorrect" cancer treatments.[4] In a separate event, MD Anderson and IBM Watson partnered to roll out another oncology-related tool. After spending more than four years and $62 million, MD Anderson

1 *The Short Coat*, episode 27, "A Real-Life Dr. House: Gurpreet Dhaliwal," University of Iowa, podcast, December 2, 2013.

2 Gurpreet Dhaliwal et al., "Case 28-2025: A 36-Year-Old Man with Abdominal Pain, Fever, and Hypoxemia," *New England Journal of Medicine* 393, no. 14 (2025): 1421–34.

3 Peter G. Brodeur et al., "Performance of a Large Language Model on the Reasoning Tasks of a Physician," *Science* 392, no. 6797 (2026): 524–27 (o1-preview achieved perfect Revised-IDEA scores on 78 of 80 responses, compared with 28 of 80 for attending physicians and 16 of 80 for residents).

4 Casey Ross and Ike Swetlitz, "IBM's Watson Supercomputer Recommended 'Unsafe and Incorrect' Cancer Treatments, Internal Documents Show," *STAT*, July 25, 2018.

ended the partnership. The system was never used on a patient.[5]

The payer side saw the same limitations play out: one national plan deployed a length-of-stay forecaster that was not robust enough for individual coverage decisions; another ran claims through a system that produced conclusions without rationales. Both episodes ended in class-action litigation.[6] The plans that deployed these tools were working with the best technology that existed, and in good faith, but the technology was simply not good enough.

Previous machine learning tools built one model per task, trained on curated historical examples of that task. This created two limitations. First, these models recognized patterns rather than reasoned: they matched the surface or descriptive features of what they had seen in training, and predicted an outcome. Second, they could not explain themselves: the prediction did not come with a rationale (only a confidence score), so there was nothing for a clinician to inspect or question.

Things are different now. Bill Golden, the CEO of UnitedHealthcare's Employer & Individual business during the machine learning wave, warns current executives that they ignore AI at their peril.

"The transformation is not only happening—it's happened...The question is, do you embrace it?" Golden asks. "The companies that figure out how to use it right—to make experiences better for members and providers, to organize data better—are the ones that will win. If you don't get that right, you probably won't catch up, because the moment you're playing catch-up in technology, someone else is already working on the next thing, and you'll never close the gap. That's a bad place to be when you're providing coverage to ten, twenty, thirty, forty million Americans."

Bill Golden spent 3 years as the CEO of UnitedHealthcare's Employer & Individual business. Photograph: Cody Cutter

Large language models (LLMs) represent the major capability leap our industry needed. They reason generally, including through the messy, ambiguous records that broke ML models. And unlike the black-box nature of ML models, their reasoning is auditable—they output language and logic—enabling trust in AI systems not possible before.

> "The companies that figure out how to use it right—to make experiences better for members and providers, to organize data better—are the ones that will win."
>
> **Bill Golden**
> Former CEO, Employer & Individual, UnitedHealthcare

We have the right of way

The care moment is when a doctor delivers a hard diagnosis, or the small interactions over time that build enough trust for a patient to say yes to a treatment. It is a fundamentally human

5 Charlie Schmidt, "M. D. Anderson Breaks with IBM Watson, Raising Questions about Artificial Intelligence in Oncology," *Journal of the National Cancer Institute* 109, no. 5 (2017): djx113.

6 Jakob Emerson, "UnitedHealth, Cigna Face Lawsuits over Alleged Automated Claims Denials," *Becker's Payer Issues*, November 27, 2023.

exchange between people: a job that no artificial intelligence can, or should, automate. Yet not all jobs in healthcare are like this.

The back office of a health plan does a lot of work to support the care moment, built around documents and diagrams like medical policies, contracts, operating procedures, and decision frameworks. The work usually has a right, and verifiable, answer that makes it extremely well-suited to LLMs.

Every day, "nurses are going through thousands of pages to try and figure out which parts are relevant to the request in front of them," recalls Grant Tarbox, DO, a senior executive medical director at Health Care Service Corporation. For decades, this was the reason the work needed a human; no software could read or reason through scribbled, handwritten notes and fax scan errors. The latest models are different.

"A lot of claims still come in on paper—sometimes handwritten, sometimes scribbles. For a long time we used older models to try to read them, and it was hit and miss. Two years ago we pointed a frontier model at it, and it was kind of accurate. Now the latest versions are incredibly accurate, even with really bad writing—and they understand the structure within them," says Rob Duffy, chief technology officer at HealthEdge.

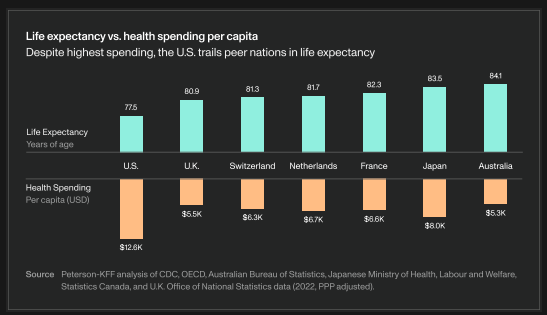
These models can apply written criteria and use them to reason over patient clinicals. Because the criteria (or the contract, or SOP) are codified, a clinician reviewing an AI determination can inspect a visual chain of reasoning against standards they are familiar with.

"Now the latest versions are incredibly accurate, even with really bad writing—and they understand the structure within them."

Rob Duffy CTO at HealthEdge

Not only is the work well-suited to this new technology, but it is also unusually worth transforming, and the stakes go beyond efficient operations. The United States spends more per person on healthcare than any comparable country and lives shorter lives for it.[7] Few industries can claim that automating their paperwork improves both their margins and their members' experience of care, with the proceeds reinvested thoughtfully.

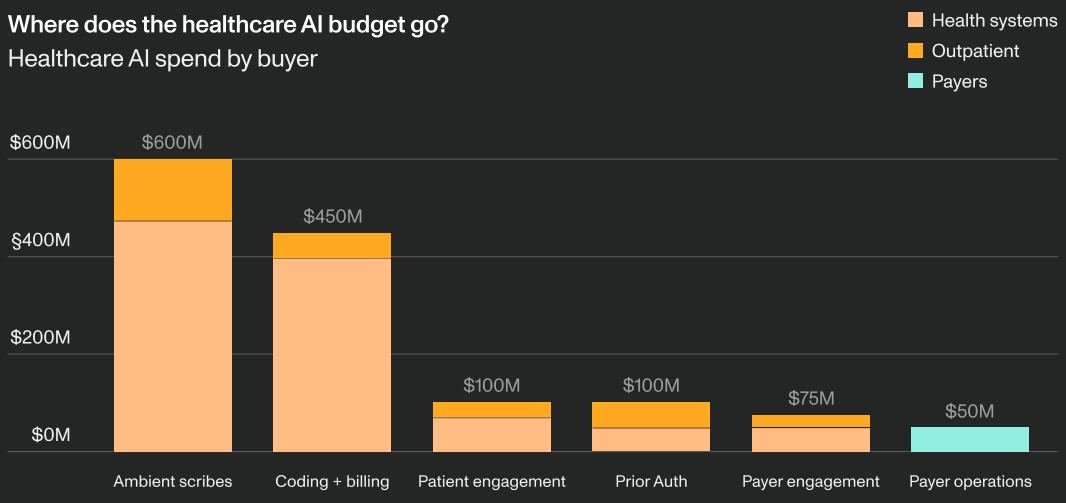
"Health plans have all the levers: the data, the payment, the network. They're the only actor set up to manage population health," said Peter Long, former Chief Strategy Officer of Blue Shield California. If done right, health plans have the most to gain.



Yet, the investment from health plans has been slow. Of the $1.4 billion spent on healthcare AI in 2025, per Menlo Ventures, payers accounted for roughly

7 Emma Wager, Lynne Cotter, Nirmita Panchal, and Cynthia Cox, "What Drives Differences in Life Expectancy between the U.S. and Comparable Countries?," Peterson-KFF Health System Tracker, February 14, 2025.

Where does the healthcare AI budget go?
Healthcare AI spend by buyer



Source Recreated from data released by Menlo Ventures in 2025

$50 million. Not even 5 percent.[8] Health systems have implemented AI at nearly twice the rate of health plans; spending on prior authorization AI grew more than tenfold—overwhelmingly on the provider side. It is ironic that the area of the industry best suited to this technology, with the most to gain from it, and with the greatest potential for impact on the member, is currently the slowest to adopt it.

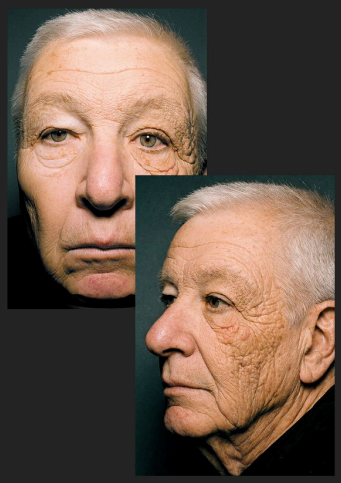
So far on the road

Despite the underinvestment, clinical AI is live across the different axes of a health plan, and showing promise.

Across workflows. In utilization management, Anterior's clinical AI, Florence, helps Geisinger Health Plan now complete cancer care approvals in roughly 155 seconds—often while the patient is still in the consultation room—with three-quarters auto-approved, clinical accuracy above 99 percent under independent verification. The system is built such that it can never auto-deny claims. In payment integrity, AI-assisted audit has lifted recovered dollars by 24 percent for a regional plan. In case management, the case research that used to take MedWatch nurses 45 minutes now takes around two minutes.

Across specialties. The judgment holds in complex specialties where it is truly clinical. For example, in complex rehabilitation, asked whether a therapy plan contained at least one goal tied to real function, AI parsed goals about walking and returning to work, concluded the documentation could not support a determination either way, and escalated to a clinician with its reasoning attached. It knew what it did not know. In outpatient surgery, where imaging is important, the reasoning can be just as fine-grained: AI can check whether full-face and lateral photographs are present, at neutral position, and sufficient to support medical necessity before any determination is made.[9]

8 Menlo Ventures, "2025: The State of AI in Healthcare," 2025.
9 Anterior, internal research, 2026



Are full face and lateral photographs at neutral gaze submitted?

Yes, multiple photographs are present in the record showing the patient's face from various angles including frontal and lateral/side views. The images show full face photographs with clear visualization of eyelid position, brow position, and periorbital anatomy.

Lateral/profile views are also present. The photographs appear to be taken at neutral gaze with the patient looking straight ahead. The images are well-lit, clear, in focus, and of sufficient quality to evaluate eyelid position, margin placement, and dermatochalasis.

The latest frontier models are capable of analyzing and understanding across a range of media.

And across modalities. Google's LLM-based system AMIE conducted the patient interview through live text chat and outperformed primary care physicians on diagnostic accuracy.[10] In ambient transcription and voice, Abridge is drafting clinical notes in exam rooms at national scale.[11] In text, scanned images, and unstructured information, Anterior reasons through the messy faxes, PDFs, and charts our industry receives, at accuracy above 99 percent.[12]

So, the technology is live, but if you look closely, you see that it's all pointed at completing the same processes faster and cheaper. Plans process the same submissions at lower cost; providers generate the same submissions with less effort; members marginally benefit.

As an industry, we risk getting stuck, as we have before, in point solution purgatory: a tool for every problem, each with its own data and its own new workflow, multiplying the administrative sprawl they were bought to reduce.

There's an old line that if you'd asked people what they wanted before the automobile, they'd have said faster horses. Healthcare is at that same moment. The instinct is to optimize the horse—but the car is already here. The question is whether we'll make the best of it.

The problems facing our industry are numerous. If we only use AI as a vehicle for faster work, we miss a generational opportunity to transform our operating models and improve the lives of members across the US. We can't afford to do that.

10 Tao Tu et al., "Towards Conversational Diagnostic Artificial Intelligence," *Nature* 642, no. 8067 (2025): 442–50.

11 Abridge, "Johns Hopkins Medicine Deploys Abridge AI Platform for Clinical Documentation Across Enterprise," December 18, 2024.

12 HealthEdge, "HealthEdge GuidingCare Launches Decision Intelligence Ecosystem with Anterior, Latitude Health and Case Health AI," February 25, 2026.