AI PROGNOSIS

Are we adopting AI scribes based on… vibes?

You're reading the web edition of STAT's AI Prognosis newsletter

Adobe

By Brittany Trang Feb. 26, 2025
Health Tech Reporter

You're reading the web edition of STAT's AI Prognosis newsletter, our subscriber-exclusive guide to artificial intelligence in health care and medicine. Sign up to get it delivered in your inbox every Wednesday.

AI scribe studies: Does a 'happy doctor' = good care?

Ambient AI medical scribes are the hottest tool in health care AI. The pitch is alluring: Instead of typing away at a computer during a patient visit, a clinician can turn their full attention to the patient while an AI tool listens, records, and summarizes the conversation, reducing the amount of time that the clinician spends documenting the appointment.

It's the perfect use case: It plays to AI's strengths in transcribing audio and summarizing text, decreases clinician burnout, makes the patient experience better, and offers opportunities to increase revenue by increasing patient appointments and upcoding with better documentation. And it doesn't affect clinical decision-making or outcomes, thus avoiding FDA regulation and enabling quick adoption. A win-win-win.

But…are all of those things actually true?

Last week, a group of researchers at the University of Pennsylvania published a study in JAMA Network Open examining whether AI ambient scribes actually decrease clinical note burden. The study followed 46 clinicians at UPenn's health system who used Microsoft-owned Nuance's DAX Copilot AI ambient scribe for 5 weeks in spring 2024.

You can read the whole paper for the detailed breakdown, but the study combined electronic medical record tracking data with a clinician survey to determine both quantitatively and qualitatively whether the AI tool saved clinicians time. (It did; about 2 minutes per note and 15 minutes per day in after-hours "pajama time" — which is a far cry from the Nuance-endorsed statistic of a two-hour reduction in pajama time, which presenters repeated in a Microsoft-sponsored Medical Group Management Association webinar from January.)

There's one qualitative finding from the UPenn study I'd especially like to point out: The researchers found that "the need for substantial editing and proofreading of the AI-generated notes, which sometimes offset the time saved" was a recurring theme in the clinicians' comments. The product received a net promoter score of 0 on a scale of -100 to 100, with an equal number of people recommending (13) and not recommending (13) the product, and the rest of the survey respondents (11) responding passively.

Some clinicians in the study commented that they were entirely satisfied with the notes: "I quickly became comfortable that it would capture all critical elements of the conversation."

But some were dissatisfied with the tool's level of accuracy: "It tries to paraphrase the conversation, and often does it in a way that utilizes layman's terms rather than medical terms; and often incorrectly documents what was discussed. This means that I must edit the content substantially because it cannot be used as-is in my closed note."

The only ways to reconcile these two different experiences are: 1) The clinicians have different definitions of what's acceptably accurate, 2) The AI tool is performing differently for different people, or 3) Some clinicians are checking the output carefully and some are not. To me, either of those latter two (very likely) explanations are alarming and show the need for studies looking at the accuracy of the output of AI scribes, not just whether physicians and patients are happier using the tool.

But the burnout and efficiency studies are the ones people are calling for — just a few weeks ago, Tina Shah argued in a Health Affairs Forefront that we need more studies about whether AI tools actually decrease provider burnout. Why? As disclosed at the very end of the article, Shah is the chief clinical officer of Abridge, one of the biggest AI ambient scribe companies, which only stands to gain from claims of efficiency and burnout reduction.

While we do need those studies, have you heard similar calls for studies of AI scribe accuracy and safety? MedStar Health Research Institute recently conducted an anemic version of this where they transcribed 11 real patient encounters, de-identified them, and then had staff re-enact them for two different ambient scribes (the exact products were not identified in the study) and concluded that there were "frequent errors," often of omission.

These studies aren't just pandering to overly worried patients — doctors should be calling for these quantitative studies too because they are betting their livelihoods on AI being accurate every time they click "OK" without thoroughly checking. In an invited commentary that was co-published with the UPenn study, Harvard bioethicist and health law expert I. Glenn Cohen and colleagues lay out legal and ethical issues with medical AI scribe tools. They echo an FDA advisory committee's conclusion that these AI tools aren't without risk to patients: "The possibility of that product hallucinating can present the difference between summarizing a health care professional's discussion with a patient and providing a new diagnosis that was not raised during the interaction."

The electronic health record is of paramount importance in any medical malpractice case, write Cohen and colleagues. The law traditionally holds clinicians responsible for the accuracy of patient records, and it's all too easy, as the clinician in the UPenn study said, to "quickly [become] comfortable" that the AI-generated note is accurate and not carefully check the record every time. The difference between a "point 5 milligram" and "5 milligram" prescription is small to an ambient listening tool but big to a physician who doesn't catch the mistake.

To date, I've not seen any large-scale, independent studies of the accuracy of these AI scribes, despite the fact that large health systems with the ability to do these studies, like Mass General Brigham and Cleveland Clinic, have piloted multiple scribes. Let me know if I'm wrong about the lack of studies or if you want to talk about your health system's experience (reply to this email or hit up aiprognosis@statnews.com).

Proving that these tools provide a return on investment and decrease doctor burnout is important. But accuracy and safety often gets lost in the kind of math that clinicians and consultants presented in that Microsoft-sponsored call about DAX: "Happy doctor equals happy patient, happy patient equals good patient experience and good care, right?"

From the STAT archives:

- Last year, STAT's Katie Palmer looked at head-to-head pilots (the "'Pepsi challenge' of health care") and outlined all the ways ambient AI medical scribes are trying to differentiate their products.
- Now that most AI scribes are fully automated, the costs have decreased. But in 2023, I found that hospitals were paying thousands of dollars per doctor per month to simply buy doctors' happiness with the DAX AI scribe. Read more here.

Google's 'co-scientist' AI tool: A collaborator, or Reviewer #2?

Last week, Google announced an AI "co-scientist" with training inspired by the scientific method that's supposed to be able to help researchers surface new hypotheses. The system is more than just a regurgitative LLM, with several AI "agent" systems embedded within it to allow it to "reason."

However, when I think back to the multi-dimensional streams of information I had to sort through in my PhD to come up with hypotheses for what was going on with my samples, I remain skeptical that an AI would be able to pull that together in a meaningful way. As I understand it from the co-scientist paper, Google's tool is limited to evaluating text and not raw data, more like getting a peer review back than sitting down with a colleague.

Are you part of Google's Trusted Tester Program? Have you used the tool? Let me know about your experience: aiprognosis@statnews.com

A STAT exclusive: Elicit raises $22M for an AI research tool for scientists

Relatedly, AI research startup Elicit has raised a $22 million series A at a $100 million valuation for its AI research tool specifically designed for researchers, in addition to its $9 million in seed funding. Company executives told me that Elicit already has grown to 400,000 monthly users through word of mouth. Those include research teams at the NIH, NHS, and major pharma companies, which they couldn't name publicly.

The Elicit tool summarizes scientific literature and provides quotations from the sources it cites. The company has a free tier as well as individual and team subscriptions. Its ambitions are a little more down-to-earth than the Google's or OpenAI's tools for researchers.

How is Elicit different from competing deep research AI tools? Tools like OpenAI's and Perplexity's are trained using reinforcement learning, said Elicit co-founder and chief operating officer Jungwon Byun, where a human rates the model's output and that rating is used as feedback for further training.

The problem with that approach "is that they are basically saying, 'Hey, powerful AI system, here is a reward metric, a feedback signal. Do whatever it takes to make this number go up or down and use whatever techniques you want to get there,'" said Byun. "We think this is going to get riskier and riskier for really high-stakes research and reasoning, because already what happens is you start to see that the models do things that don't make sense," such as switching between languages and code, which makes it hard for humans to trace the AI's logic, she said.

How was Elicit designed? Byun said it's designed with research in mind, asking how humans put together meta-analyses and systematic literature reviews, and automating those processes with AI.

How does Elicit deal with the fact that lots of scientific research is paywalled? Users can add papers from their library to Elicit's tool, and the company is looking into ways to leverage researchers' subscriptions to journals. But right now, the tool operates off of open access papers and abstracts.

"You still have to figure out which paywalled papers are you going to pay for and read, and the ability to go through hundreds or thousands of them is still going to save you a lot of time," said Byun. "I think there's a lot more to be done before these tools can really lead to incredible breakthroughs. But I think right now they're already saving people a lot of time and are probably superhuman in being able to summarize all the relevant literature, to figure out what's most worth digging into more."

A 'ChatGPT for medicine' unicorn

Last week, AI medical information startup OpenEvidence announced a $75 million investment from Sequoia Capital that valued the company at $1 billion. If you're not familiar, OpenEvidence is like an AI-powered UpToDate (alternatively, a ChatGPT for medicine) that summarizes clinical evidence for health care providers. It recently announced a partnership with the New England Journal of Medicine's publisher that allows it to pull from all of the NEJM family content from 1990 forward.

Is the $1 billion valuation realistic, especially for a tool that's free for health care professionals? Of course, UpToDate leader Peter Bonis thinks "that there is a bit of a hype cycle going on, at least for the valuations just across the [AI] sector," though he wouldn't comment on OpenEvidence specifically.

UpToDate last week announced an integration with health care AI assistant Corti, which is similar to a partnership the company announced with Abridge last fall. Though UpToDate's content is still written by humans, getting that information pulled in to AI-generated clinical note platforms helps clinicians make better decisions, Bonis told me. Eventually, the company hopes that this information can even surface in the middle of a visit where an AI scribe is being used.

Song of the Week: "How Can I" by Dead Gowns

When I hit play on "How Can I," I got a shiver like I was listening to Julien Baker or Phoebe Bridgers for the first time: *I'm going to love this*. The song is the first track off of *It's Summer, I Love You, and I'm Surrounded by Snow*, the debut LP from Maine singer-songwriter Genevieve Beaudoin, aka Dead Gowns. It's my first "album of the year" contender for 2025.

Genevieve's voice reminds me of Julia Jacklin's and somehow takes me back to a simpler time in my life: the first Trump administration. Please enjoy, and let me know if you think this will be on your year-end albums list too.

Do you have questions about what's in this week's newsletter, or just questions about AI in health in general? Suggestions? Story tips? Ideas for song of the week? Simply reply to this email or contact me at AIPrognosis@statnews.com.