// ARTICLE · PUBLISHED 27 OCTOBER 2026

Reading Healthcare-AI Evidence: One Arm Is Not Evidence

Three careful healthcare-AI results from October 2026 share one gap, and five questions show what each can and can't support

Jason Newell · MAX Research Collective · Article ART-HLC-019 · Published 27 October 2026
Licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/
Original: https://www.maxresearchcollective.com/articles/reading-healthcare-ai-evidence.html
Cite as: Newell, J. (2026). Reading Healthcare-AI Evidence: One Arm Is Not Evidence. MAX Research Collective, ART-HLC-019. https://www.maxresearchcollective.com/articles/reading-healthcare-ai-evidence.html

Three results, one missing group

Three healthcare-AI results made news in October 2026. Each is honest about what it did. A mental-health chatbot cut depression and anxiety scores. A startup's virtual drug trials called five of six real outcomes. Google's diagnostic chatbot, AMIE, interviewed patients before their primary-care visits, and The Lancet published the study. None of the three shows that the AI improves care. Not one had a comparison group.

The reasons repeat every time: no comparison group, few people or events, averages that hide the people left out of the benefit, and authors who built the product. Those reasons turn into five questions a reader can put to any claim.

The questions leave a study standing. They say what it can be used for, and that's the skill a buyer, a clinician or a patient needs when the evidence arrives as a preprint or a press release, months before any controlled trial. These three were chosen because they're good work, and if the best early evidence needs the questions, so does the rest.

The mental-health pilot

A preprint first posted to arXiv in November 2025, and revised in September 2026, reports a pilot of Ash, a model for mental health built by Slingshot AI, which followed 299 US adults with at least moderate depression or anxiety for up to 12 months. The early numbers are big: by 10 weeks, symptoms had fallen with effect sizes of 0.93 and 0.79. Effect size here is Cohen's d, the change measured in standard deviations, and anything above about 0.8 is conventionally called large.

The design is single-arm. Everyone got the AI.

No group got something else, or nothing, so there's no way to tell which part of the change the AI made. People who join a study at a bad moment tend to improve anyway. A before-and-after change in one arm mixes several causes, and the AI is only one.

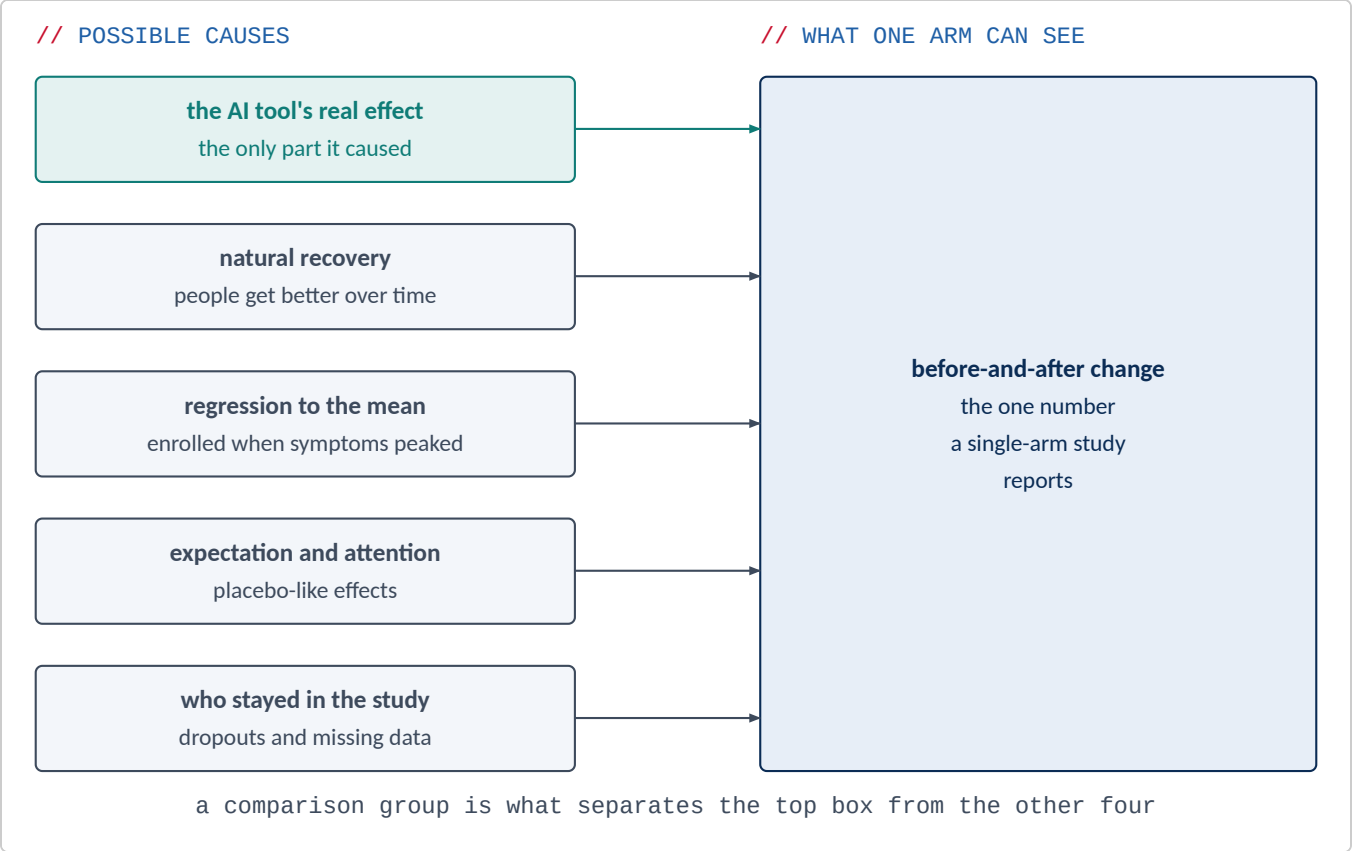


Figure 1. Five causes feed the one change a single-arm study can see, and only the top one belongs to the AI.

The preprint also reports what most summaries left out. Sorted by how their symptoms moved, 57.2% of participants were "non-responding," 37.1% improving and 5.7% rapidly improving. Most people stayed where they were. A large average effect sits beside a non-responding majority, and the minority carry the average.

Clinicians checked the automated safeguards, the rules that hand a conversation to a person, and found that they had escalated the conversations they were supposed to escalate. That's a real safety finding, and a separate matter from effectiveness.

The disclosures sit in the full paper: three of the five authors are current or former Slingshot employees, a fourth holds advisor shares in the company, and the fifth has a research grant from it.

The virtual drug trials

On 7 October 2026 Reuters reported that BioinvestGPT, a Copenhagen startup, had lodged detailed predictions for several high-profile drug trials with the news agency in July. Five of six had matched the outcome so far. The miss was Novartis's pelacarsen. The company puts it down to not accounting for variation in lipoprotein size. One correct call was a weak benefit for Novartis's del-desiran, which then missed its goal.

Lodging predictions with a newspaper before the results is good practice: nobody can choose afterwards which predictions to remember. The open question is the baseline. Six is a small number, and many late-stage trials are fairly predictable to a well-informed observer. A fair coin lands heads at least five times in six about one time in nine. Without knowing how an informed analyst with no AI would have scored on the same six, 5 of 6 has nothing to beat.

The same story quotes Takeda's Andy Plump: "I don't think we are near being able to use tools like AI to make definitive predictions." BioinvestGPT claims nothing definitive, so the quote reads as a reminder rather than a rebuttal: a short public record is a reason to keep watching.

AMIE in the clinic

Google and Beth Israel Deaconess Medical Center published a prospective study of AMIE (Articulate Medical Intelligence Explorer) in The Lancet on 8 October 2026. Patients enrolled between April and November 2025 typed with AMIE before urgent primary-care visits, while physicians watched each conversation live against stop rules written in advance. Of 114 enrolled, 98 finished both the AI interview and the appointment. That's careful work.

No conversation had to be stopped.

Supervisors noted one hallucination (a confident statement with no basis in the conversation), and they added clinical information in five. The accuracy numbers are good. Google's summary says AMIE's list of possible diagnoses contained the doctors' final diagnosis 90% of the time. The earlier preprint of the same study, at 100 patients, reports 75% top-3 accuracy, meaning the right answer was among AMIE's first three suggestions. It also reports that primary-care physicians beat AMIE on how practical and cost-effective their management plans were.

Physicians answered the follow-up survey in 60 of the 98 cases, had reviewed the transcript beforehand in 44 of those, and in 33 of the 44 found it helpful for preparing the visit (a share of the cases where a physician looked, rather than of all 98).

One physician judged an interaction somewhat harmful: AMIE had listed lymphoma as a possible diagnosis, and the patient may have left anxious.

The authors call this a feasibility study. That's fair. It shows a supervised AI interview can run safely in one clinic. It had no comparison group of patients who skipped AMIE. It wasn't built to show that patients do better. And its authors work for the company that built the system — normal for a first study, and the reason independent replication matters.

Five questions that sort a claim

The same five questions sort all three results.

Table 1. Five questions, applied to the three October results.

QUESTION	WHY IT MATTERS	IN THESE THREE RESULTS
Is there a comparison group?	Without one, the size of the effect is unknown	None had one
Were outcomes or predictions fixed in advance?	It stops anyone picking the results to report afterwards	BioinvestGPT and AMIE fixed theirs
Are the people who didn't benefit reported?	An average can hide a majority who didn't improve	The mental-health preprint reported 57.2% non-responding
How many people or events?	Small numbers give wide uncertainty	299 people; 6 trials; 98 patients
Who did the work, and who paid?	Builders have reasons to see success	Company staff produced all three

A reader can walk the questions in order. A "no" changes what the claim can support; it doesn't end the reading, since a study without a comparison group can still prove something can be done safely.

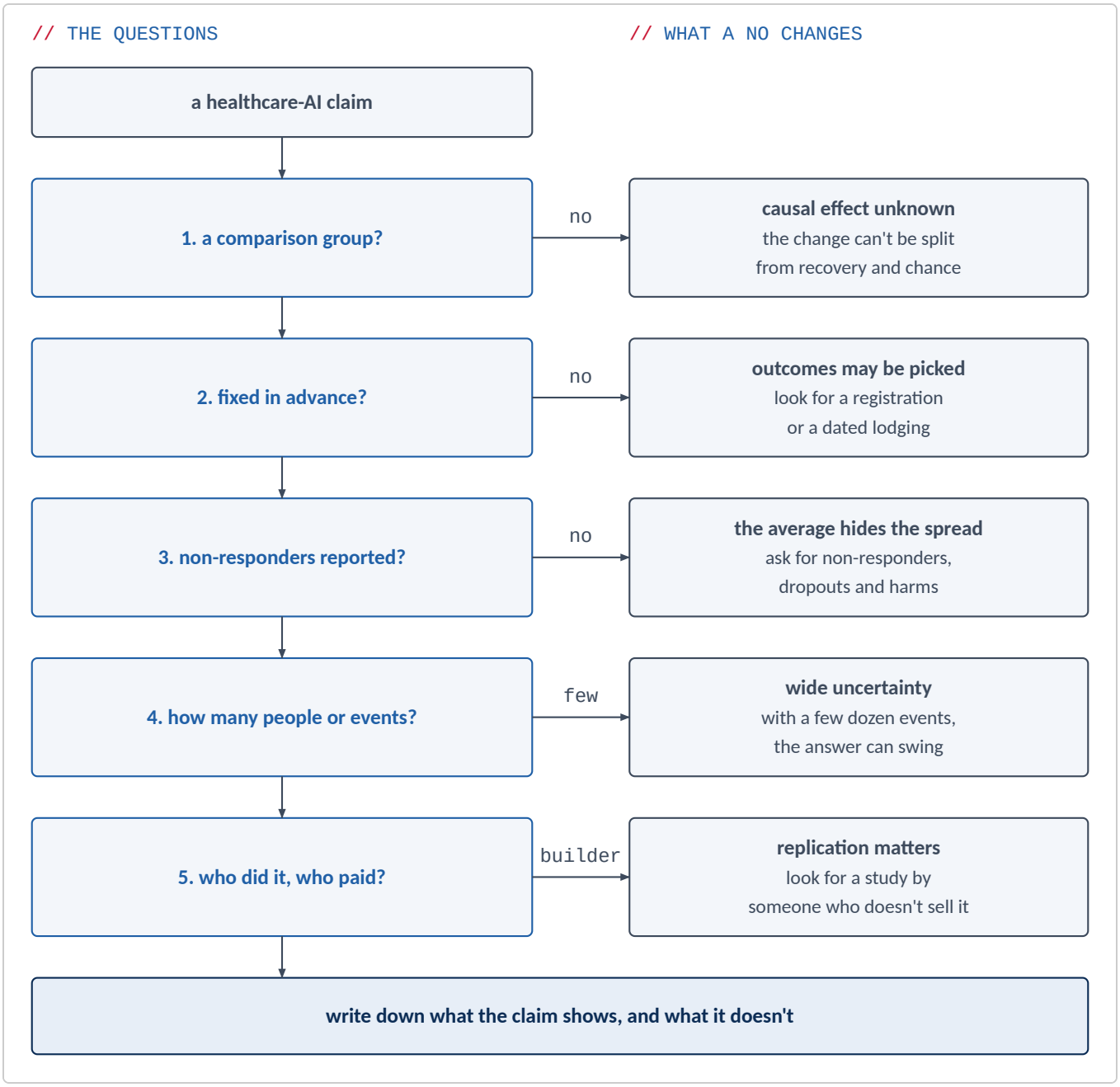


Figure 2. Each "no" on the left sends the reader right to what the claim can no longer support, and every path still ends at the bottom.

Smaller checks travel with them. Ask ”of what?” for every percentage. Note the setting: one clinic in one language travels less well than it reads. If a person was watching, the result describes the AI with a supervisor. Check whether the outcome is something patients feel, such as symptoms, harm or survival, or a proxy such as accuracy. And rank the format, since a press release says less than a preprint and a preprint says less than a peer-reviewed paper.

Where MARA sits

MARA, the Medical Agentic Resident Assistant this project is designing, takes one rule from these three. No claim that she improves care rests on a single arm. Her results report the people she didn't help. Every result names who ran it and who paid. MAX4, the desktop resident where her methods are proved first, holds its own learning to the same rule: a skill is credited only when an outcome confirms it, and the misses are reported beside the hits.

What wasn't checked

The Lancet paper itself wasn't opened. Its figures here come from Google's two pages, the preprint and Becker's Hospital Review, and a "top-7" accuracy figure quoted in some coverage was left out because no source here confirms it. Reuters' own page refused access, so its story was read in an attributed syndication on KSL. BioinvestGPT's six predictions weren't re-scored independently.

Why it matters

Healthcare AI is moving from demonstrations into clinics, insurers and drug pipelines, and for the next few years the first evidence that a hospital's buyers, its clinicians and their patients see about almost any new tool will come from a single arm. A 0.93 effect size, 5 of 6 and 90% each read as proof on a slide. Each one, asked the five questions, turns into something smaller and more useful: a tool that some people responded to, a forecaster worth watching, an interview that ran safely under supervision. The purchase decisions being made this year will be made on numbers like these, and the comparison group that would settle them usually arrives after the contract is signed.

Sources

- Hull TD, Zhang L, Stamatis CA, Areán PA, Malgaroli M. Generative AI Purpose-built for Social and Mental Health: A Real-World Pilot. arXiv:2511.11689, v1 November

2025, v4 29 September 2026; affiliations and competing-interest statements read in the full text on 21 October 2026. https://arxiv.org/abs/2511.11689

- Beasley D. AI companies run virtual drug trials, aim to improve success of human studies. Reuters, 7 October 2026, read in the KSL syndication. https://www.reuters.com/legal/litigation/ai-companies-run-virtual-drug-trials-aim-improve-success-human-studies-2026-10-07/ and https://www.ksl.com/article/51633346/
- Google. AMIE clinical study published in The Lancet. 8 October 2026. https://blog.google/innovation-and-ai/technology/health/amie-clinical-study-lancet/
- Google Research. A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic, The Lancet, 2026, NCT06911398. https://research.google/pubs/a-prospective-clinical-feasibility-study-of-a-conversational-diagnostic-ai-in-an-ambulatory-primary-care-clinic/
- Brodeur P, Koshy JM, Palepu A, Saab K, et al. The same study as a preprint. arXiv:2603.08448, March 2026. https://arxiv.org/abs/2603.08448
- Becker's Hospital Review. How an AI chatbot performed in pre-visit primary care interviews: 6 notes. October 2026, opened 21 October 2026. https://www.beckershospitalreview.com/quality/patient-safety-outcomes/how-an-ai-chatbot-performed-in-pre-visit-primary-care-interviews-6-notes/

Glossary

TERM	DEFINITION
AMIE (Articulate Medical Intelligence Explorer)	Google's conversational diagnostic AI. The Lancet study tested it before real primary-care visits.
Cohen's d	an effect size measured in standard deviations, where about 0.8 or more is conventionally large. The mental-health pilot reported 0.93 and 0.79.
Comparison group	a group that gets usual care, another treatment or nothing, so a difference can be attributed. None of the three results had one.
Feasibility study	a study of whether something can be done safely and practically, as opposed to whether it works better. AMIE's authors call theirs one.
Hallucination	a confident statement from an AI with no basis in its input. AMIE's supervisors noted one in 98 conversations.
MARA (Medical Agentic Resident Assistant)	this project's design for a clinical assistant that recommends while a person decides. Her results will report non-responders beside averages.
MAX4	the desktop generation of the MAX AI resident, from October 2026 on, where MARA's methods are proved first.
Non-responder	a participant whose outcome didn't meaningfully improve. They were the majority in the mental-health pilot.
Preprint	a paper posted publicly before peer review. Two of the sources here are preprints on arXiv.
Pre-registration	recording the outcomes or the predictions before the results are known. BioinvestGPT's lodging with Reuters did that job.

TERM	DEFINITION
Regression to the mean	the tendency of people enrolled at an extreme to drift back toward average on their own. It's one reason a single arm improves without any help.
Single-arm study	a study in which everyone gets the same intervention and nobody is compared. All three results are single-arm in effect.
Top-3 accuracy	the share of cases in which the right answer is among a system's first three suggestions. AMIE's preprint reported 75%.

HEALTHCARE ART-HLC-019

// ABOUT THE AUTHOR

Jason Newell leads organizations and people in technology, and is a researcher, deep technologist and builder — from a CRAY to agentic AI. He founded MAX Research Collective to build AI for regulated industries that runs where its owner has control, with healthcare as its strongest focus. About Jason → Connect on LinkedIn →