

AI HEALTH WEEKLY

What the week's evidence supports, and what it does not.

THE PRESENTING WEEK

The week AI in health got audited

Regulators, journals and hospitals asked the same question: not whether AI works, but where, for whom, and who checks.

ASSESSMENT

The strongest items this week were checks, not launches. One government agreed to watch AI devices for their whole working life. One journal published the first real-clinic test of a patient-facing chatbot. One hospital model showed uneven results across patient groups. Funding rounds were many, and mostly company-reported.

FINDINGS OF THE WEEK, GRADED BY EVIDENCE

POLICY

UK backs its AI commission in full

The government accepted every recommendation on regulating AI in healthcare. The MHRA opened AI Airlock phase 3 on post-market monitoring, with draft guidance on changing AI devices due by December.

Why it matters: regulators now expect evidence after launch, not only before it.

Source: GOV.UK, government response

FEASIBILITY

A patient-facing chatbot, tested in clinic

Patients used Google's AMIE before urgent primary-care visits, under physician supervision. One hallucination was noted. Single-centre, single-arm.

Why it matters: a safety signal, not proof of better care.

Source: The Lancet, AMIE feasibility study

BENCHMARK

More agent tools did not mean better diagnosis

Several GPT-5 setups were scored on real, clinician-validated cases. Image misreading and uneven web search held them back.

Why it matters: ask about input quality, not tool count.

PROSPECTIVE

Fuller notes, less factual

An outpatient-note assistant scored higher on integrity (3.50 v 3.21) but lower on factuality (3.34 v 4.01) than physicians.

Why it matters: writing time saved can return as checking time.

Source: Scientific Reports, outpatient-note study

RETROSPECTIVE

Epic's cleared model, uneven by group

Sensitivity at the medium-risk threshold was 26% for Black, 32% for Hispanic and 51% for white patients. Hospitals set their own alert thresholds.

Why it matters: validate by subgroup before relying on any alert.

Source: FDA 510(k) summary K260008

PRESS-REPORTED

India: rules for AI claims and devices

The NMC was reported to bar AI-made testimonials in medical ads. Device-safety guidance was reported to reach AI devices and primary health centres.

Why it matters: marketing and procurement now carry AI-specific compliance.

INSIDE THIS ISSUE

- 2 Editorial: Prove it in the clinic, then keep proving it
- 3 The deep read: the end of the one-off clearance
- 3 Mechanism of action
- 4 The radar: what we follow, and where it stands
- 5 Rapid rounds, by domain
- 5 Vital signs and follow-up

EVIDENCE KEY

PROSPECTIVE	real patients, planned
FEASIBILITY	small, one arm
BENCHMARK	tested on cases
RETROSPECTIVE	past data
POLICY / PRESS	rules and reports
COMPANY	self-reported

44 of 44

recommendations accepted: the baseline for this week's debate

QUOTE OF THE WEEK

“There's no finish line for what is good enough.”

REBECCA SOSKIN HICKS, OPENAI

NEXT ISSUE • SUNDAY, 18 OCTOBER 2026

Prove it in the clinic, then keep proving it

Dr. Hemant L. Patel

A clearance, a feasibility study, or a funding round is not the end of the evidence story for AI in health. It is usually the start.

This week's clearest signal came from the United Kingdom. The government accepted every recommendation of the National Commission into the Regulation of AI in Healthcare and opened Phase 3 of the MHRA's AI Airlock specifically to test how devices are watched after they leave the sandbox. Draft guidance on how AI devices may change after clearance is due by December; most detailed timelines arrive only with the spring 2027 roadmap. The direction is right: approval can no longer be treated as a one-time event when models drift, retrain, or are used on populations the original data never saw. Whether the regime actually catches problems will depend on the measures that are still being written—which outcomes, which subgroups, reported to whom, and who pays for the watching.

The same week supplied the reasons those details matter. Google's AMIE chatbot was used, under physician supervision, by patients before urgent primary-care visits. It was a single-centre, single-arm feasibility study. Zero of 98 completed chats triggered a pre-set safety stop, and supervisors noted one hallucination. That is useful information about whether the approach can be tried safely. It is not evidence that the chatbot improves diagnoses, reduces missed problems, or helps patients. A randomised comparison with patient outcomes is still required.

At the other end of the evidence ladder sits an already-cleared tool. Epic's Deterioration Index Version 2 received 510(k) clearance. The FDA summary shows, at the medium-risk threshold, markedly lower sensitivity for Black and Hispanic patients than for White patients. Hospitals choose their own alert thresholds. A headline AUC of 0.83 can therefore hide uneven performance that only

becomes visible once the model is switched on in a particular hospital's patients. Local subgroup validation and published monitoring are not optional extras; they are the conditions under which the clearance remains meaningful.

Even documentation tools that are not classified as medical devices show the same pattern. An outpatient-note assistant scored higher on integrity than physicians but lower on factuality. Time saved in writing can reappear as time spent checking. The savings are real only if the checking is done.

Before any organisation buys, deploys, or expands an AI tool, three questions remain the practical filter:

How was it tested on patients who look like the ones we actually see—by age, comorbidity, language, ethnicity, and care setting?

How will performance be monitored once it is live, including by subgroup, and what triggers a pause or a new review?

What happens when the model is retrained or the workflow changes? Is there a pre-agreed change-control plan, or does every update require a fresh assessment?

Developers and vendors who can answer these questions with data, not slides, will earn the trust that funding announcements cannot buy. Regulators who insist on the answers—before and after launch—will do more for safe adoption than any single high-profile clearance. The UK's shift toward lifecycle oversight is a necessary experiment. The rest of the world, including India as it tightens rules on AI claims and device monitoring, will be watching whether the experiment produces reusable measures or only further promises.

A good week for AI in health was a week of questions. The useful answers came from tools that were watched, not merely launched. There is still no finish line for what is good enough.

The end of the one-off clearance

The UK has taken up a blueprint in which an AI medical device is judged for as long as it is used. Most of the hard parts are still ahead.

WHO SAID WHAT

The National Commission into the Regulation of AI in Healthcare was independent and led by NHS doctors. It drew on evidence from more than 12,000 patients, clinicians and members of the public, and reported in September. Its central conclusion was about time, not technology (see Mechanism of Action, below).

WHAT IS CONFIRMED

The government started at once. The Airlock sandbox has three more years of funding, and its next phase is aimed at what happens after deployment. A Programme Board now owns nine priority areas. Patients are promised to be told when AI is used in their care, and to have routes to redress when standards fall short.

WHAT IS ONLY PROMISED

Read closely, much of the response is a direction rather than a rule. Staged authorisation pathways, which would let promising tools into the NHS earlier under close supervision, are being explored. How to define and classify an AI medical device goes to consultation next year. Timelines for most recommendations arrive only

with the roadmap in spring 2027. The release does not say what level of performance would trigger action, who pays for monitoring, or how results by patient group will be reported.

WHY IT MATTERS NOW

The tension is the interesting part. Earlier access under supervision depends on monitoring that works, and monitoring depends on measures chosen before launch. This week's evidence shows why. A feasibility study can show a chatbot is safe enough to try, but not that it improves care. A cleared model can carry uneven results across patient groups into every hospital that switches it on. A lifecycle regime should, in principle, catch both. Whether it does depends on details not yet written: which outcomes, which groups, reported to whom.

THE VIEW FROM INDIA

The contrast is useful. Reports this week describe limits on AI advertising and wider device-safety monitoring, but the notices themselves were not seen. The UK is trying to build monitoring into approval itself. Neither approach has yet been tested on a model that changes after launch.

Source: GOV.UK, government response and AI Airlock phase 3

WHAT TO WATCH

- Does the draft guidance say when a retrained model needs new review?
- Does phase 3 publish the monitoring measures it tests, so others can reuse them?
- Will staged access require a monitoring plan before supervised use begins?
- Will results by patient group be reported as a matter of routine?

MECHANISM OF ACTION

Lifecycle regulation

Approval judges a product once, before sale. AI can drift or be retrained, so the commission argued for monitoring it through its working life. The AI Airlock is the MHRA sandbox testing how; phase 3 covers post-market surveillance.

What a 510(k) means

Substantial equivalence to a device already on the market, not proof of clinical benefit. A reviewer of Epic's summary reports a locked model with no predetermined change-control plan, which would mean a retrained version needs a new submission.

Feasibility versus trial

A single-arm feasibility study can show that supervised use was safe and acceptable, not that outcomes improved. A 'safety stop' is a pre-set trigger for a physician to step in. A randomised trial adds a control arm.

Subgroup sensitivity

Sensitivity is the share of truly deteriorating patients flagged at a chosen threshold. A good headline score can hide uneven results by group, and hospitals set their own thresholds, so validate locally.

What we follow, and where it stands

Updated every Sunday. Each item carries a stage and a change flag: **NEW** **ADVANCED** **UNCHANGED** **CLOSED**. This is the first issue, so every item is new. Dates for what happens next are in Follow-up.

ITEM	WHAT IT IS	WHERE IT STANDS	WHAT DECIDES IT
STORY UK AI Airlock, phase 3 NEW	MHRA sandbox testing how AI devices are monitored once in use.	OPEN Applications open; first cohort not yet chosen. Funded for three more years. Source: GOV.UK	Whether it publishes methods and measures that regulators then adopt in guidance.
STORY MHRA guidance on changing AI devices NEW	Planned draft rules on how an AI device may adapt after clearance.	PROMISED Draft not yet published.	When a retrained model needs fresh review, and whether change-control plans become expected.
STORY UK classification consultation NEW	First Commission recommendation: how to define and classify AI-enabled devices.	COMMITTED Set for next year; not yet open.	Which tools fall inside regulation at all.
STORY Epic Deterioration Index v2 NEW	FDA-cleared inpatient deterioration model, in use in hospitals running Epic.	CLEARED FDA summary shows lower sensitivity for Black and Hispanic patients. A reviewer reports it is locked. Source: FDA 510(k) K260008	Subgroup validation by hospitals, published monitoring, and any retrained submission. Would change our view: the locked, no-change-plan claim is a reviewer's; an FDA decision document or Epic statement would confirm or correct it.
STORY FDA generative-AI device paper NEW	FDA discussion paper on risk, premarket evaluation and postmarket monitoring of generative-AI devices.	COMMENTING Public docket open for feedback. Source: FDA bulletin	What FDA does with the comments.
STORY AI scribe regulation NEW	MHRA guidance (July): tools that only transcribe, summarise or draft letters for clinician review are not medical devices, per a press report.	IN FORCE Press report of the guidance; GOV.UK refers to it as recent.	Whether findings on note accuracy prompt a second look. Would change our view: the MHRA guidance text itself.
STORY India: AI claims and device rules NEW	NMC curbs on AI testimonials in medical ads; device-safety guidance extended to AI devices.	UNVERIFIED Press-reported; notices not seen.	Hospital compliance and enforcement. Would change our view: the NMC notice and the revised device-safety guidelines themselves, with effective dates.
INNOVATION Staged authorisation NEW	UK idea: let promising AI tools into the NHS early, under close supervision.	EXPLORING No pathway defined.	Entry criteria, and who monitors.
INNOVATION Supervised patient-facing AI NEW	Chatbots taking histories before urgent primary-care visits.	FEASIBILITY One single-arm, single-centre study.	A randomised trial with a comparison arm and patient outcomes.
INNOVATION Hospital-operations AI NEW	Staffing, bed-capacity, authorisation and billing agents, where the week's capital went.	COMPANY Company-reported results only.	Whether savings hold up. Would change our view: an independent or peer-reviewed evaluation of the savings.

RAPID ROUNDS, BY DOMAIN

REGULATION & POLICY

- REPORT** **Utah sandbox:** an AI may prescribe acne creams alone after two physicians approve each of the first 100 patients.
- REPORT** **Device scope:** one cardiac device announcement excluded its AI algorithm from the clearance. Read the scope, not the headline.
- OPINION** **Lancet correspondence:** authors argue an independent multidisciplinary body should examine a medical AI's intended use and limits before development starts.
- REPORT** **Anthropic and HHS:** a closed-door pitch of a Claude healthcare connector to HHS officials is reported; details are secondhand.
- EVIDENCE & TRIALS**
- TRIAL** **LungDiag and eye-history trials:** two randomised studies in China showed LLM gains over web search and residents; one used simulated cases.
- NEGATIVE** **Parkinson's telerehabilitation:** exploratory trial missed its primary endpoint; hypothesis-generating only.
- STUDY** **Reproducibility:** few prediction-model papers share their code, which limits re-running them.
- STUDY** **Where models look:** an IIIT Hyderabad audit found medical vision-language models may not look where radiologists look on chest X-rays; press-reported.
- OPINION** **Harm with no record:** a Nature Health perspective argues bad chatbot health advice stays on the company platform, so users have nowhere to report it.
- OPINION** **Six steps:** an npj Digital Medicine opinion piece sets out six steps to separate generative AI's promise from its dangers; abstract only.

MARKET & FUNDING

- The pattern: money went to workflow, not diagnosis. Staffing, screening, billing and calls drew the most, then biomanufacturing and drug discovery. Most figures are company-reported.*
- COMPANY** **Vitalize:** \$31M Series A to link hospital staffing with bed capacity; savings are customer-reported.
- COMPANY** **Healthleap:** \$38M seed and Series A for LLM screening of clinician notes; says it runs in 50-plus hospitals.
- COMPANY** **GenHealth.ai:** \$16.5M for prior-authorisation and billing agents.
- COMPANY** **Vocca and General Medicine:** \$20M for a patient phone assistant; \$120M for a consumer healthcare storefront.
- COMPANY** **Abridge:** expands across eight University of Texas health institutions; terms not seen.
- COMPANY** **Multiply Labs:** \$75M Series B for robotic cell and gene therapy manufacturing.
- COMPANY** **Illumina and IQVIA:** SpliceAI2 splice-variant prediction and trial-outcome models; company announcements.
- COMPANY** **insitro:** models added to Lilly TuneLab; needs prospective validation.
- REPORT** **Blackford Analysis:** changes owner, from Bayer to Sopris Capital; newsletter summary only, terms not seen.
- INDIA**
- REPORT** **PGI and IIT Delhi:** MoU on an AI breast-cancer tool trained on Indian data.
- REPORT** **Arunachal Pradesh:** rolls out a newborn-measurement app.
- REPORT** **IISc TANUH:** its lead describes AI screening for oral and breast cancer; an interview, not a study.

VITAL SIGNS

0 of 98 AMIE chats needing a physician stop	161 real cases for ten GPT-5 setups	12.2% of 3,967 prediction papers shared code	\$1.8B biology-data commitment: intent, not delivered	0.83 Epic AUC on its combined outcome	3 sites in Epic's validation
---	---	--	---	---	--

FOLLOW-UP: THE WEEK AHEAD

19 OCT	22 OCT	NOV 2026	DEC 2026
FDA comments due on its generative-AI device discussion paper (docket FDA-2026-N-7874).	MHRA webinar for AI Airlock phase 3 applicants (10:00).	AI Airlock phase 3 applications close; first cohort selected.	MHRA draft guidance on managing changes to AI-enabled devices.

AI Health Weekly is synthesised from the week's daily editions of 6-11 October 2026. Offers, targets and company-reported figures are not audited outcomes. Regulatory status as of this edition.