AI on Friday

What to Know, Why It Matters

Issue 04, 11 September 2026

Pan American Health Organization
World Health Organization Americas Region
Information Systems and Digital Health Unit
Department of Evidence and Intelligence for Action in Health

This weekly email shares key developments, use cases and practical AI-in-health concepts in plain English, helping us apply AI in our technical cooperation and discuss emerging opportunities with Member States and partners.

DEVELOPMENT TO WATCH

AI scribes face growing patient-safety scrutiny

What happened: Healthwatch England reported patient accounts of errors in medical records produced with AI scribes, sometimes identified by patients rather than health professionals. The UK's Health Services Safety Investigations Body has now launched a national investigation into this technology.

Why it matters: AI scribes may reduce documentation burdens, but inaccurate records can affect future clinical decisions and patient safety.

Read more: *Healthwatch England* • *HSSIB investigation*

AI explanations can help, but they can also influence judgment

What happened: A 2026 study in Nature Medicine examined how AI-supported dermatological diagnosis affected 623 members of the public and 153 primary care physicians. AI support improved overall accuracy and reduced performance differences across skin tones.

What it also found: When AI was incorrect, explanations written in convincing language could increase overreliance. Showing the AI recommendation before asking for a human judgment also increased anchoring, including among clinicians.

What to watch: The sequence matters. Asking users to form an initial judgment before seeing the AI recommendation may help preserve independent clinical reasoning.

Read more: *Nature Medicine*

AI IN PLAIN ENGLISH

Hallucinations

An AI hallucination occurs when an AI system generates information that is incorrect, unsupported or entirely invented, while presenting it as if it were reliable.

Why it happens: Generative AI predicts plausible responses based on patterns in data. It does not automatically verify whether each statement, statistic or reference is true.

Health example: An AI tool may invent a scientific article, misquote a clinical guideline or provide an inaccurate recommendation using convincing medical language. **Essential safeguard:** Treat AI-generated information as a draft. Verify important claims against original and authoritative sources, especially before making clinical, policy or technical decisions.

PRACTICAL IDEAS

Ground the answer in a trusted document

When using AI to analyze a report, guideline or policy, provide the original document and add:

"Answer using only the attached document. For each main claim, identify the corresponding page or section. If the requested information is not included, say 'Not found in the provided document.' Do not fill gaps using assumptions or external information."

Before using the response, open two or three of the cited passages and confirm that the AI represented them accurately.

Important reminder: AI-generated content may contain errors, outdated information or incorrect references. Always verify original sources, publication dates and links before using or sharing it, particularly for clinical, policy or technical decisions.

FROM THE AMERICAS

Bogotá, Colombia, uses AI to strengthen suicide prevention

What happened: Bogotá's District Health Secretariat began implementing Project STOP, an initiative that uses AI to identify patterns in social media associated with suicidal ideation, depression and other mental health risks.

How it works: The initiative analyzes anonymized information and uses targeted messages to connect people who may need support with Bogotá's Línea 106 mental health service. It complements the work of health professionals and existing pathways of care.

What to watch: This is a promising public health application, but its implementation requires careful attention to privacy, potential bias, false alerts and human review. Measuring whether digital outreach results in timely and effective support will also be essential.

Read more: *Official information from the City of Bogotá*

WORTH READING 1

Addressing benchmarking gaps in large language models for health and medicine with dynamic red-teaming

Published in 2026 in *Nature Health*, this peer-reviewed, open-access study evaluates 15 leading language models across robustness, privacy, bias and hallucinations. It identifies an important gap between high scores on standard medical benchmarks and reliability under changing or adversarial conditions. In these tests, hallucinations or factual inaccuracies occurred in more than 74% of cases.
The study shows why benchmark results alone are insufficient and why continuous, context-specific evaluation is essential for health applications. Read the open-access article

WORTH READING 2

Mapping the susceptibility of large language models to medical misinformation across clinical notes and social media

Published in 2026 in The Lancet Digital Health, this study tested 20 language models using more than 3.4 million prompts containing fabricated health information. The models accepted false information in 46.1% of prompts based on realistic hospital notes, compared with 8.9% of prompts based on social media. This suggests that authoritative language and credible-looking documents can make misinformation more difficult for AI to recognize.

The findings reinforce the need to verify the content of source documents, not only whether a source appears trustworthy. Read the article

Additional resources (Co-authored by PAHO's staff)



Link



Link

Important reminder: AI-generated content may contain errors, outdated information or incorrect references. Always verify original sources, publication dates and links before using or sharing it, particularly for clinical, policy or technical decisions.