

The Problem of False Positives in AI Health Screening

ABSURD HEALTH INSTITUTE · 08 · Signals, measurement and responsible medical AI · 15 OCTOBER 2026

Abstract

A health-screening tool can have respectable sensitivity and specificity and still generate mostly false alarms when applied to rare conditions. This paper makes base rates, calibration and abstention central to medical AI. Its thesis is that restraint is a core performance feature: the ability to say “insufficient evidence” can be more protective than attaching a diagnosis to every image.

The Clinical Question and Its Context

Positive predictive value depends on disease prevalence. Screening for hundreds of uncommon conditions multiplies the opportunity for incidental flags. Even if each model were independently well behaved, correlated outputs and repeated testing can create misleading cascades. Model confidence scores are not automatically calibrated probabilities, and external deployment can shift prevalence, camera conditions and patient mix. A ranking of 1,300 possible diagnoses may look impressive while being clinically unusable.

Proposed Biological or Explanatory Model

A trustworthy design should disclose which targets have validation evidence, reserve diagnostic labels for justified use and separate symptom education from screening. Low-confidence states should trigger abstention and recommendations for ordinary clinical evaluation. Calibration curves, precision-recall analysis, performance by demographic subgroup and prespecified referral thresholds matter more than a headline accuracy statistic from a balanced case-control dataset.

Mechanism-Specific Interpretation

The mathematics deserves to be visible to the reader. At 1% disease prevalence, a test with 90% sensitivity and 90% specificity yields about 9 true positives and 99 false positives per 1,000 screened people; the positive predictive value is therefore only about 8%. That result is not evidence that the test is poorly engineered, but that screening rare disease is intrinsically difficult. Add repeated testing and hundreds of candidate diagnoses, and the number of incidental alerts can increase rapidly. A system must therefore prioritize carefully selected clinical targets and transparent thresholds.

Competing Explanations and Limits of Inference

The counterargument is that overcaution may miss early disease. That is a real tradeoff, but the solution is explicit threshold setting around known clinical harms, not universal maximal sensitivity. Emergency warnings such as sudden focal weakness should follow clinical safety rules regardless of model confidence, while nonspecific visual hints require different handling.

Discriminating Observations and Evidence Thresholds

The model succeeds when it knows when not to claim. Calibration against real-world prevalence, explicit reporting of false alarms and the ability to abstain are indispensable. Independent external cohorts are necessary because impressive performance in a balanced research dataset may fail in an unselected population. This is an ethical and engineering requirement, not a cosmetic disclaimer.

Implications for Interpretation and Future Inquiry

For Absurd Health this paper is strategically important because it sets an ethical standard for Signals. AI-supported attention to health is valuable when it reduces uncertainty and improves care pathways; an unvalidated disease count is not an achievement. The system's ability to avoid unwarranted certainty should be regarded as a clinical strength.

Selected Evidence and Further Reading

1. [Quer et al. Wearable sensor data and self-reported symptoms for COVID-19 detection](#)
2. [Esteve et al. Dermatologist-level classification of skin cancer with deep neural networks](#)

Source lineage: Signals predictive diagnostic claims and app design

Editorial note: Conceptual scientific commentary; not an original clinical trial, systematic review, peer-reviewed publication, or medical advice. Hypotheses are not established treatment effects.