

The Problem of False Positives in AI Health Screening

ABSURD Health Institute · Volume Two · Research Essay 56

Research series: 08 · Signals, measurement and responsible medical AI

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 arithmetic of low-prevalence disease

False positives are a central biological and statistical obstacle to population-scale AI health screening, not a cosmetic limitation in model presentation.

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.

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.

False alarms and cascades of investigation

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.

False positives become especially damaging when a screening system is applied repeatedly to large populations with low disease prevalence. Even an apparently impressive specificity can generate many erroneous alerts when true cases are uncommon. Sequential testing makes the problem more

complicated because dozens of correlated measurements may trigger repeated warnings, and users can mistakenly interpret each as new independent evidence. The correct denominator is the number of meaningful detection opportunities and patients affected, not just the number of sensor samples.

A responsible system should report sensitivity, specificity, positive predictive value, false-alarm burden per person-month and performance in clinically relevant subgroups. A model can achieve high area under the receiver-operating curve yet fail to generate a useful threshold in practice. Calibration drift after device upgrades, changes in disease prevalence and differences among user groups all need ongoing monitoring. Suppressing all alerts may reduce annoyance while missing urgent deterioration, so the decision threshold must consider consequences rather than appearance of accuracy alone.

When true disease is rare, even a test with attractive sensitivity and specificity may flag many more unaffected than affected people. Correlated repeated measurements can compound alarm rates, and confirmation pathways may be costly or invasive. The controversial implication is that a seemingly advanced detection system may worsen care if it fails to incorporate prevalence and the consequences of investigation. A high accuracy percentage is not enough to establish clinical usefulness.

Evaluation in the intended screening population

The scientific endpoint is net benefit under a defined response protocol. Investigators should compare the number of useful early actions against unnecessary consultations, testing costs and anxiety. The app should expose uncertainty plainly and create pathways for escalating severe symptoms independently of its classifier. For high-consequence conditions such as stroke or severe infection, a user must not be reassured solely because an automated score is low.

Base rates dominate interpretation. If only a small fraction of users will have a particular emergency during the surveillance period, a classifier with high sensitivity but imperfect specificity may still issue mostly false-positive warnings. Repeated daily use can compound the burden. This does not make screening inherently useless; it means alerts must be evaluated at the intended prevalence and decision threshold. Headline accuracy from a balanced case-control test dataset cannot substitute for real-world positive predictive value.

Calibration and subgroup harms. A model may overpredict risk in one age group and underpredict it in another even if its overall discrimination looks good. People with chronic conditions or irregular wear patterns can be disproportionately exposed to false alerts or missed events. Model recalibration and independent monitoring should be part of deployment, not a one-time quality check.

A safer output system. Alerts should indicate their underlying observation, uncertainty and suggested level of response, with appropriate clinical review for concerning cases. The platform should never suppress urgent symptom-based instructions because a sensor-based model assigns low risk. An explicit false-positive budget can help evaluate thresholds, but it must be weighed against the seriousness of missed disease. The ethical measure of success is improved decisions, not simply fewer notifications.

Report calibration, positive predictive value and downstream harms in realistic low-prevalence populations, with the unit of analysis as alerts per patient-year as well as per test. Compare against simple rules and clinical judgment. A model should not be deployed to recommend invasive evaluation until benefits outweigh false-positive burdens.

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.

A numerical example shows why prevalence cannot be ignored

Suppose a screening model has 90 percent sensitivity and 90 percent specificity for a condition affecting 1 percent of the screened population. Among 10,000 people, approximately 90 of the 100 affected people would be flagged, but about 990 of the 9,900 unaffected people would also be flagged. Thus only about 8 percent of positive alerts would correspond to the condition despite the model's seemingly strong performance. Real data may violate the simple assumptions, but the example exposes why impressive headline accuracy can mislead. Repeating such tests daily can multiply false alarms if alerts are not consolidated. An operational health system should report positive predictive value, number needed to evaluate, false alerts per user-year and the outcomes produced by acting on the alarms. The provocative conclusion is that a complex AI tool may perform worse for patients than a simpler, better-calibrated rule, even when its classification score is statistically higher.

A model can be accurate and still cause harm

One way to communicate false-positive risk is to show expected outcomes for a realistic cohort rather than only percentages. A model used among mostly healthy people may generate far more false alerts than true detections even when its test-set accuracy appears impressive. Repeated surveillance, unequal disease prevalence and overlapping sensor features further complicate the calculation. Decision thresholds should therefore be selected according to the seriousness of the missed condition and the burden of unnecessary evaluation. An alert with no meaningful next action may be counterproductive even if statistically associated with later illness. Calibration should be reassessed when populations or hardware change, and subgroup harms should be reported separately. A clinically responsible app must allow severe symptom-based escalation regardless of its automated prediction. The engineering achievement is controlled, understandable uncertainty, not maximal alert volume or a claim that AI can diagnose all diseases from ordinary fluctuations.

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.

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.

Sources and research leads

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

Editorial status: Third-pass conceptual research manuscript. The argument has been individually strengthened and shared boilerplate removed. Original reference leads were retained; a complete independent source-by-source scholarly verification has not been performed. No new clinical trial data are presented.