SAFWAN ALMODHAYAN



OPINION AND ANALYSIS

Editorial date: November 1, 2026

The Prediction That Changed the Future It Predicted

> An alert used to tell us what might happen. Then we decided to act on it. When the warning did not come true, we asked the alert to defend its honesty—as if putting a firefighter on trial because the house he entered did not burn down.

The Hospital Wards That Heard an Alert Others Did Not

Between October 2019 and July 2021, five hospitals in Saudi Arabia's Ministry of National Guard Health Affairs conducted an experiment whose significance cannot be reduced to “the technology worked.” Forty-five medical and surgical wards were assigned across nine transition waves between usual care and an electronic screening system for sepsis, the dangerous bodily response that can accompany infection. This was not a robot diagnosing disease in secret, nor a language model composing treatment advice. It was simpler than that—and therefore more useful for the question at hand.

The electronic health record calculated an alert from three components of the quick Sequential Organ Failure Assessment, or qSOFA: systolic blood pressure below a specified threshold, an elevated respiratory rate, or altered consciousness. When two signs appeared within a defined window, the system generated an alert. In wards that had crossed into the intervention phase, those alerts became visible to the care team and connected to a response pathway; in the other wards, the system ran in an unannounced mode for comparison. Researchers could therefore ask a question more important than whether the calculation was accurate: what changed once people had an alert on which they could act?1

The study included 60,055 patients: 29,442 in the electronic screening group and 30,613 in the comparison group. After an alert, patients in the intervention group were more likely to receive lactate testing and orders for intravenous fluids. Yet a decontextualized reading of the mortality table is puzzling. Crude in-hospital mortality within ninety days was **3.2%** in the intervention group and **3.1%** in the comparison group, with no clear crude difference favoring screening. In the study's prespecified analysis adjusting for time periods, the clustering of patients within wards and hospitals, and COVID status, the adjusted relative risk was estimated at **0.85**, with a 95% confidence interval from 0.77 to 0.93. That is an adjusted estimate of relative risk reduction. It is not a claim that every survivor was saved by the program, or that the crude comparison alone is enough to interpret the result.1

Why did the crude and adjusted results differ? The trial used a stepped-wedge design: wards changed conditions at different times, and then the COVID pandemic changed patient profiles and risks across the study periods. The researchers noted that estimating the effect required adjustment for those factors. This is central to any inquiry about predictions. Even a randomized trial does not hand us a ready-made causal answer in a single row of figures. We need to know how and when groups were exposed to the intervention, and how the circumstances around them changed.1

This is not evidence that an AI model predicted danger and then lost accuracy because patients were saved. The Saudi system used an electronic qSOFA rule, and the paper did not test a claim about machine-learning performance drifting after deployment. It is, however, experimental evidence of the loop under discussion: **a risk signal changes staff actions, and staff actions can change the event we are trying to predict**. That is where a problem quite different from classification quality begins.

What Does It Mean for a Successful Alert to Be Wrong?

Imagine a patient showing signs of deterioration. A risk system assigns a high probability to a serious event within twenty-four hours. The alert reaches the team; the patient is assessed quickly and receives the appropriate intervention; the event does not occur. A month later, the data office returns to the record: “The system predicted danger, and it did not happen. Was this not a false alarm?”

Perhaps it was. Perhaps it was not.

If the model's probability means “the chance of the event **under continued usual care without adding this alert**,” it cannot properly be tested as though it meant “the chance of the event **after we responded to the alert**.” The new information changed the path whose destination the

model was supposed to predict. If instead the model was designed to predict what would occur after the staff's expected response, then comparing it with the observed outcome under that policy is a legitimate question—but a different one.

In statistical notation, we might write the event risk under the usual policy as $R_0(x) = P(Y(0) = 1 \mid X = x)$, and the risk under the alert-and-response program as $R_1(x) = P(Y(1) = 1 \mid X = x)$. Here X represents information known before action is taken. $Y(0)$ is the potential outcome under **usual care without the new program**, not in a world without medicine or treatment, while $Y(1)$ is the potential outcome under the alert program. Their difference is what the policy might change for people of the same kind, under specified design assumptions.

We never observe both outcomes for the same patient. We see what happened along the path taken, not a second version of the world in which the same patient existed but the team did not hear the alert. This is the potential-outcomes problem, and it makes “nothing happened” less informative than it sounds. Nothing happened **after what**? After which intervention? And what would probably have happened had the other path continued?

In 2020, Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt introduced the framework of **performative prediction**. The central idea is that when a model informs a decision, it may affect the future data distribution on which it will later be judged. The world it predicted is no longer the same world it inhabits after its predictions have been deployed. The paper studied mathematical properties of stability under repeated training and the interaction between prediction and data. It did not certify that every alert whose accuracy declines has successfully prevented harm.[2]

The difference between a forecast that changes our behavior and one that changes nothing is not philosophical trivia. A weather forecast does not ordinarily change the clouds. A prediction of patient deterioration may change a clinician's decision; a prediction that a student will fall behind may change the support that student receives; and a prediction of an industrial incident may change the maintenance schedule. Here AI is not a mirror held up to the future. It is a new variable in the making of that future.

A Hospital That Found Risk Signals in Nurses' Actions

We need a case that actually used machine learning, not merely an alert rule. In April 2025, *Nature Medicine* published a randomized trial of the **CONCERN** early-warning system for patient deterioration. The system does not principally rely on laboratory results alone. It learns from temporal patterns in nursing documentation and surveillance: when did a nurse

increase the frequency of measurements, enter an unusual note, or conduct monitoring at a time not dictated by routine? A change in the team's own behavior becomes a signal that may precede some of what is deteriorating in the patient's body.3

The trial covered **74 clinical units** randomized between the intervention and usual care at two US health systems, and included **60,893 hospital stays**. Displaying CONCERN alerts within the workflow was associated with a lower adjusted instantaneous risk of in-hospital mortality, with a hazard ratio of **0.64** and a 95% confidence interval from 0.53 to 0.78. A ratio of 0.64 does not mean that “64% of patients recovered because of AI.” It is the ratio of comparative instantaneous hazard rates in the study's survival analysis. Length of stay was also reduced in the analysis reported by the paper.3

Yet the same table showed an increase in the adjusted instantaneous risk of **unexpected transfer to intensive care**, with a hazard ratio of **1.25** and a confidence interval from 1.09 to 1.43. How can mortality fall while an indicator that appears to signal deterioration rises? One possibility is that earlier surveillance prompted teams to move more patients to a higher level of care, but the study does not allow us to classify every transfer as a successful rescue. In other cases, an unexpected transfer may reflect deterioration that should have been prevented. We need the timing, reason, clinical appropriateness, and change in outcomes—not merely a plus or minus sign beside the number.3

The management report now has two faces. One department can say that the system improved a dangerous outcome; another can note the rise in unexpected transfers and ask for an explanation. Both are reading a real result from the study. The difference between their readings exposes a larger truth: **a successful alert may change the map of events, and no new event should be treated as proof of success or failure until we understand the intervention pathway that produced it**.

The researchers also stated the study's limitations. It took place in two US health systems during a period affected by COVID, with differences in intensive-care unit mix, documentation practices, and the possibility that some results were influenced by unmeasured factors. Its implications are therefore stronger than a marketing story and narrower than a general guarantee for every hospital and every alert system.3

At School, Too: Risk Signals Fell but the Final Measure Did Not Move

To keep the paradox from being confined to intensive-care rooms, consider US high schools. In the 2014–2015 academic year, a study later released by the US Department of Education tested the **Early Warning Intervention and Monitoring System**, or EWIMS, with students at risk of falling off track for graduation. **Seventy-three high schools** were randomized: 37 implemented

the program, while 36 continued their usual practices during the year of evaluation. The study included **37,671 students** in the first two years of high school. This was not a language model or a modern machine-learning experiment, but a system of early indicators, follow-up, and intervention.[4]

After one year, chronic absence stood at **10%** in program schools, compared with **14%** in comparison schools. The proportion of students who failed at least one course was **21%**, compared with **26%**. Both differences were statistically significant. But the proportion of students who had not earned enough credits to remain on track for graduation within four years was **14% in both groups**; after one year, the study detected no effect on that measure of academic progress. It also reported difficulties in implementing the program's seven steps and generally weak implementation in many schools.[4]

What is the lesson? The behavior that produces risk may change before the final outcome we care about changes. Chronic absence is a signal; course failure is a signal; progress toward graduation is an outcome closer to the long-term goal. If we train a new system on records created after interventions prevented some absences or failures, then treat every student who improved as a “wrong risk classification,” we conflate two questions: who needed support, and who fell behind despite receiving it? Conversely, if attendance indicators improve but progress toward graduation does not, we cannot claim that prevention achieved its ultimate objective.

The school and hospital cases do not share a technology or a definition of success. They share a logic. When a risk marker is used to change human conduct, the outcomes that follow become part of the **institution's response**, not a neutral test of the marker's predictive power.

Selecting the Person at Risk Is Not Enough

This distinction matters because “invisible prevention success” must not become a convenient excuse for any weak model. In Camden, New Jersey, researchers tested an intensive care-coordination program for patients known to have frequent hospital admissions and complex needs. Its operating premise was compelling: we know which people return repeatedly, so let us provide coordinated medical and social support and reduce their readmissions.

But a randomized trial of **800 patients** found no statistically significant reduction in readmission within 180 days. The rate was **62.3%** among program recipients and **61.7%** under usual care; the adjusted difference was **0.82 percentage points**, with a confidence interval ranging from a decrease of 5.97 points to an increase of 7.61 points. More importantly, comparing the program group with itself before and after the intervention would have suggested a large reduction of about **38 percentage points**. Patients entered the trial after an exceptionally high-use period, and visits subsequently declined in both groups. That fall cannot automatically be credited to the program.[5]

The Camden program was not an AI experiment. It nevertheless provides a critical empirical objection to conflating **finding the people most likely to experience an event** with **finding the people for whom the available intervention can prevent it**. The person at highest risk does not necessarily have the greatest expected benefit from a given program. The causes of that person's risk may not respond to the program, may require resources the program does not provide, or may be temporary by their nature.

On paper, two models could rank patients by the chance of hospital readmission and agree on who is most likely to return. But if we ask which patients would benefit most from a specific intervention, we may need an entirely different ranking. One question concerns **risk level**; the other concerns **how changeable that risk is**. Conflating them makes high predictive accuracy look like excellent preventive policy without evidence that it is.

The same problem appears in a study published in February 2026 on passive displays of AI predictions of deterioration in a cardiac unit. It included **10,422 admissions** and tested showing risk trajectories without requiring a defined clinical response. The study found no improvement in its primary outcome. Clinicians moved about **11%** of patients between beds with and without displays, weakening some features of the practical randomization. The result is not a verdict against every alert model. It is a reminder that **knowing the risk does not automatically produce an action capable of reducing it**.[6]

We are left with two possibilities that cannot safely be separated. The alert may have been right and the intervention may have saved someone; or the alert may have been inaccurate, or the intervention ineffective. What distinguishes these possibilities is causal design, not goodwill toward the algorithm.

The Outcome We Need Is Not the Outcome We Have

In a paper published in 2020, Nan van Geloven and colleagues emphasized the need to define exactly what a model predicts when treatment begins after risk is measured. They called this target a **predictimand**, by analogy with specifying an estimand in a clinical trial. Are we predicting what will happen if no additional treatment begins? What will happen if treatment practices continue as they did in the training data? Or whether the event will occur before treatment starts? These are different questions, with different answers, estimation methods, and assumptions.[7]

In October 2025, Christopher B. Boyer and colleagues addressed the more direct problem of **estimating and evaluating counterfactual prediction models**. They explained why a model that predicts an outcome under a hypothetical treatment policy cannot be evaluated simply with data observed under a different policy. They proposed methods for estimating properties such as loss, calibration, and area under the ROC curve in the counterfactual world, while stating

the statistical identification conditions and assumptions. Their illustration concerned cardiovascular-disease risk among people who did not initiate statin therapy. These are conditional methodological results, not a recipe that makes unobserved outcomes visible.8

We therefore need to distinguish three quantities:

1. **Baseline risk:** the probability of harm under the reference policy used for comparison, such as usual care without the new alert system.
2. **Post-intervention risk:** the probability of harm when the alert is used under a known policy and people make the responses they actually follow.
3. **Policy benefit:** the amount of harm reduced by moving from the first policy to the second, after accounting for side effects, cost, resources, and other consequences.

Baseline risk may be high, post-intervention risk low, and policy benefit large. Or baseline risk may be high while the intervention does little, producing almost no benefit. The intervention may even harm some groups. A single “risk score” cannot tell us which of these is true.

For a numerical illustration only, imagine that person A has a 30% event risk under usual policy and a 10% risk under a particular intervention. Person B has a 20% risk that falls to 19%. The estimated difference in benefit is 20 percentage points for the first person and one point for the second. These are **hypothetical figures for explanation, not study results**. Nor can an individual's realized benefit be known by observing that person only once. What we can attempt to estimate are average effects for groups under suitable assumptions and study designs.

The problem deepens when a system periodically retrains itself on data produced through its own use. When the model sees that people it scored as high risk did not experience the event after they were treated, transferred, or supported, it may learn a new relation between their characteristics and the observed outcome. It may become better at predicting what happens **under the current policy** and less suitable for estimating what would have happened **without that policy**. Retraining is not always a mistake. But the question being predicted must remain stable. We cannot silently change the target and call the result merely a new version.

Why Accuracy Can Mislead Us

If a company announces that its model achieved a high area under the ROC curve, that may demonstrate discrimination in a specified test set. By itself, it does not establish the effects of deployment on patients, students, or facilities. If a predictive measure falls after a successful intervention is introduced, the decline cannot be interpreted without knowing the original

target and the nature of the policy. The distribution of events may have changed because of the intervention, the population may have changed, or the model may genuinely have deteriorated.

There are three different trials, and none should swallow the other two:

First, the trial of the prediction: were the risk estimates sound in the specified reference world? This includes calibration, discrimination, data integrity, and fitness for the period and population.

Second, the trial of the response: did the message change the decision-maker's behavior? How long elapsed between alert and examination or action? Was the action appropriate, and did the group concerned respond to it?

Third, the trial of the outcome: what was the policy's effect relative to usual care or an alternative policy? Did harm decline? Did other harms, costs, unnecessary transfers, or drains on staff and resources appear?

These trials cannot be collapsed into one percentage. In the Saudi experiment, increased testing and fluid orders were part of the staff response, while mortality was a separate outcome. In CONCERN, one hazard declined while intensive-care transfers increased. In the EWIMS schools, some indicators improved without a detected effect on graduation progress after a year. In Camden, the program did not demonstrate benefit on its primary target despite selecting a group with high service use. The mechanisms differ, but scientific judgment in every case requires us to separate prediction, action, and effect.

This problem also differs from cases in which labels are missing because some people never had a chance to manifest the outcome. In preventive prediction, the outcome may be **fully observed**, but it is an outcome that occurred **after the policy intervened**. The absence of the event in the record does not reveal whether it would have arisen without the intervention. This is not a blank field in a table. It is a missing alternative path through the world.

How Do We Test a Model That Prevents Some of What It Predicts?

When ethically and institutionally possible, there is a direct experimental route. Preserve usual care or service for everyone, randomize the **addition of the new program** among comparable units or introduce it in carefully designed stages, and record the predictions before disclosure, the moment they are disclosed, staff actions, and subsequent outcomes. Such a comparison does not mean withholding required treatment or exposing someone to a

known danger. The object of the test is an added feature whose benefit has not yet been established. In sensitive settings, relevant professionals and ethical governance bodies must approve it.

Some systems can continue calculating a risk score in a concealed mode during the comparison, provided information is protected and the boundaries of usual care are respected. The effect of **making the alert available** is measured across groups or wards, rather than by claiming to know which survivors would otherwise have died. Stepped-wedge and cluster-randomized trials can be useful when knowledge spills among staff within the same unit, but they bring challenges involving periods and external events, as the COVID-era Saudi trial demonstrated.

When a randomized trial is impossible, researchers can try to emulate a target trial with observational data, or use methods that adjust for treatment interference with the prediction and estimate counterfactual effects. These methods depend on assumptions that should not be hidden: knowing the confounders that influence both the decision and the outcome; being able to compare people who had plausible opportunities to receive different options; defining the intervention consistently; and not corrupting the comparison through untreated interactions among units. The assumptions will not hold in every setting.

If intervening for one person consumes an intensive-care bed another needs, the question extends beyond an isolated individual effect. It becomes a problem of resource allocation, congestion, and **interference**. A policy that helps dozens of patients may produce a different result when deployed across every ward in a country. Testing model accuracy is therefore not enough to justify the assumption that doubling alerts doubles the number of rescues. Response capacity varies, excessive alerts have costs, and team behavior itself may change over time.

The Forecast–Intervention Ledger: A Falsifiable Proposal

I propose that an institution maintain a **Forecast–Intervention Ledger**. This is an editorial proposal, not an approved standard or a system with proven performance. Its purpose is not to multiply dashboard fields. It is to separate three things that tend to be conflated: what was known before the intervention, what was done because of it, and what happened afterward.

For every eligible case, the ledger would record the time at which prior information was collected, the model version, the definition of the risk represented by the score, the prediction horizon, when the alert was issued, whether a decision-maker saw it, the response that followed and its timing, the case outcome, and possible harms. A separate analytical layer would estimate causal effects across comparable groups, preserving uncertainty and the

assumptions behind the estimate. The field “what would have happened to this patient without an alert” would not be populated with a fabricated individual number, because that future was never observed.

The ledger could become three separate panels for institutional review:

- **Prediction panel:** model calibration under the reference policy, alert sensitivity, and the suitability of the risk score for the target group.
- **Intervention panel:** the proportion of alerts followed by an appropriate response, time intervals, burdens, referrals, and harms.
- **Impact panel:** the difference in outcomes against a specified comparison policy, with confidence intervals, cost, and resource-distribution effects—not merely what happened to people who received alerts.

The evidence that would falsify the system's value should be explicit. If a sound comparison shows no meaningful difference in the important outcome, if the harms of the alerts exceed their benefits, or if identifying the highest-risk group does not change the usefulness of the intervention, rhetoric about “events we invisibly prevented” cannot rescue the system. The model must be allowed to have failed to improve the world, just as it was allowed to have succeeded before the results arrived.

The ledger's most important rule is that the answer can never be: “Cases fell after installation, so the system saved them.” Cases may fall because the population, season, care practices, or a previous surge changed. Camden's lesson is unforgiving on this point: a before-and-after comparison suggested a large improvement, while the randomized comparison group followed a similar path.

The test that matters to a hospital board, an education ministry, or a public-safety authority is not how many alerts the system generated, nor simply how many days passed without disaster. It is this: **how much harm changed because of this particular policy, for whom, at what cost, and with what confidence in the inference?**

When the Second Future Becomes Part of Scientific Accountability

AI is not required to know the unseen, nor is a researcher expected to display an unrealized future on a screen. What is required is recognition that preventive decisions create a paradox traditional classification tests cannot solve by themselves. Some favorable outcomes may result from effective intervention; others may result from excessive alerts, luck, or natural change. The algorithm cannot distinguish them by choosing the kindest story about itself.

We should not reward a model for fearing everything, or punish it merely because what it feared did not happen. We should not confuse a patient's risk with that patient's responsiveness to treatment, or a student's difficulty with the chance of improvement when a teacher provides help. Nor should every decline after a program be attributed to the program.

The deepest institutional shift demanded by performative prediction is a move from the apparently scientific question “Did you predict what happened?” to one that is harder and fairer: “What would have happened if we had not acted on your prediction, and how can we test the difference without pretending we saw both futures?”

At that point, success changes meaning. The best alert system will not be the one whose warnings come true most often, because the realization of danger is not prevention's purpose. Nor will it necessarily be the one after which every event disappears, because disappearance may not be its doing. The best will be the system we can show—through an experiment or disciplined causal inference—helped people make the bad outcome less likely without hiding costs and alternative harms.

We have learned how to test whether a computer knows something about the future. We still need to learn how to test whether **its knowledge made the future better**.

References

1. Arabi et al., JAMA, Electronic Sepsis Screening (2025 print; 2024 online).
2. Perdomo et al., Performative Prediction, ICML 2020.
3. Rossetti et al., CONCERN RCT, Nature Medicine 2025.
4. Institute of Education Sciences, EWIMS Impact Study, April 2017.
5. Finkelstein et al., Health Care Hotspotting RCT, NEJM 2020.
6. Keim-Malpass et al., AI-based Analytics RCT, Scientific Reports 2026.
7. van Geloven et al., Predictimands, European Journal of Epidemiology 2020.
8. Boyer et al., Counterfactual Prediction Evaluation, Statistics in Medicine 2025.