Silent Failure: a Simulation Study of Why Discrimination Metrics Do Not Detect the Deployment Failures That Matter In Clinical Prediction Models

Makbul Hussain Choudhury

Assistant Professor, Rahman Institute of Pharmaceutical Sciences and Research, Guwahati, Assam, India.

Corresponding author: makbulchy200@gmail.com

Received: 01 June 2026; Revised: 11 July 2026; Accepted: 17 August 2026; Published: 03 September 2026

Abstract

Background: Deployed clinical prediction models are most often monitored by the area under the receiver operating characteristic curve (AUROC). AUROC is a rank statistic and is therefore insensitive to changes that alter the absolute risk a model reports. We quantified which monitoring metrics detect which deployment failures, and what each failure costs in clinical utility.

Methods: We developed a logistic prediction model on a simulated cohort of 40,000 encounters with 10 predictors and an outcome prevalence of 9.0 per cent, then froze it and evaluated it on five simulated deployment cohorts of 20,000 each: no shift, prevalence shift, covariate shift, concept drift, and case-mix narrowing. We measured discrimination, Cox calibration intercept and slope, operating characteristics at a fixed alert threshold of 0.10, and decision-curve net benefit, then tested intercept-only and intercept-plus-slope recalibration.

Results: Under prevalence shift, AUROC was unchanged (0.828 against 0.820 with no shift) while positive predictive value fell from 0.236 to 0.122, number needed to evaluate rose from 4.23 to 8.19, and net benefit fell by 83 per cent. Calibration intercept moved to -0.86 and detected the failure. Under covariate shift, discrimination and calibration were both preserved while alert burden rose from 277 to 483 per 1,000 encounters. Case-mix narrowing lowered AUROC to 0.729 without any change in the predictor–outcome relationship. Intercept-only recalibration of the prevalence-shifted cohort restored calibration, raised positive predictive value from 0.119 to 0.186, reduced alerts from 284 to 123 per 1,000, and nearly doubled net benefit, with AUROC identical throughout.

Conclusion: Discrimination was silent for the two failures with the largest operational consequences. Monitoring that reports AUROC alone will miss them. We recommend a minimum monitoring set of calibration intercept, calibration slope, alert rate and net benefit, reported alongside discrimination.

Keywords: Clinical Prediction Models, Model Monitoring, Dataset Shift, Calibration, Decision Curve Analysis, Clinical Decision Support.

1. INTRODUCTION

Clinical prediction models are now embedded in electronic health record systems at scale, firing alerts for sepsis, deterioration, readmission and a widening range of other outcomes. The literature describing their development is large and methodologically mature. The literature describing what happens to them afterwards is neither.

In practice, a deployed model is usually monitored by periodic recomputation of the area under the receiver operating characteristic curve. AUROC is familiar, it is reported in every development paper, it requires no decision about a threshold, and it produces a single number that can be tracked on a dashboard. These are real virtues, and they explain its dominance.

They also conceal a structural limitation. AUROC is a rank statistic: it measures the probability that a randomly chosen case receives a higher predicted score than a randomly chosen non-case [1,2]. It is therefore invariant to any monotone transformation of the predicted probabilities, and it is insensitive to the absolute level of risk the model reports. A model whose predictions are all twice as large as they should be has exactly the same AUROC as a correctly calibrated one, and will alert at twice the rate.

This property is well known to methodologists and is repeatedly noted in the prediction-model literature [3–6]. What is less often done is to quantify it in terms that a deployment team can act on: for each realistic mode of dataset shift, which monitoring metric moves, by how much, and what does the failure cost in clinical utility? That is the question this study addresses.

We use simulation rather than a retrospective clinical dataset deliberately. In a real deployment the true shift mechanism is unobserved, several mechanisms operate simultaneously, and the counterfactual — what the model would have done under no shift — is unavailable. Simulation allows each mechanism to be isolated, applied at a known magnitude, and evaluated against a known truth, which is what makes the attribution of a metric's silence to a specific failure mode possible.

1.1 Objectives

We set out to quantify the response of discrimination, calibration and utility metrics to four isolated mechanisms of dataset shift; to determine which failures are invisible to discrimination; to measure the clinical utility lost under each; to test whether simple recalibration recovers it; and to derive from these results a minimum monitoring specification for deployed models.

2. METHODS

2.1 Overall Design

The design is shown in Figure 1. A logistic regression model was developed on a simulated development cohort, then frozen — no coefficient was updated thereafter — and evaluated on five independently simulated deployment cohorts, each representing a distinct shift mechanism. This mirrors the usual real-world situation, in which a model validated at one site and period continues to run unchanged as its environment moves.

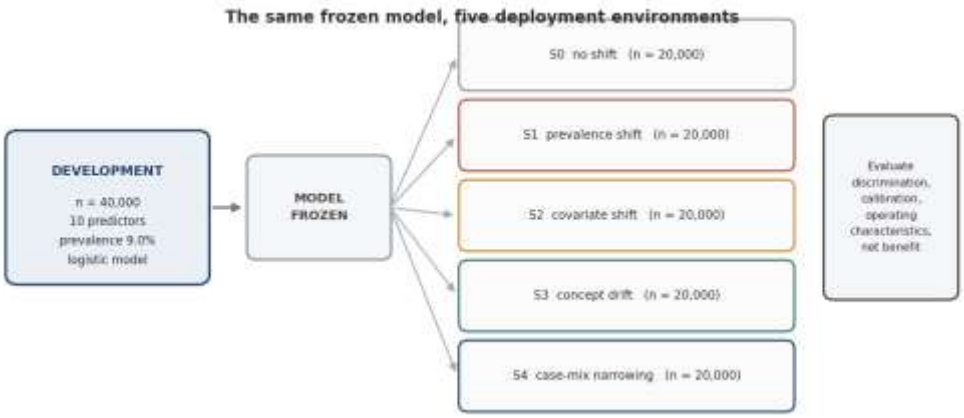

Figure 1. Study design. A single frozen model is evaluated across five simulated deployment environments differing in the mechanism of dataset shift.

2.2 Data Generation

Each encounter had 10 independent standard normal predictors. The outcome was Bernoulli with probability given by the logistic function of a linear predictor with fixed coefficients and an intercept chosen to give an outcome prevalence of approximately 9 per cent, which is typical of deterioration and readmission targets. The development cohort contained 40,000 encounters and each deployment cohort 20,000. All random draws used a fixed seed.

```
lp = intercept + X @ beta y  ~ Bernoulli(1 / (1 + exp(-lp))) beta = [0.85, -0.70, 0.55, 0.45, -0.40, 0.30, 0.25, -0.20, 0.15, 0.10]
```

2.3 Shift Mechanisms

Four mechanisms were implemented, each in isolation, and defined in Table 2. Prevalence shift (S1) lowered the intercept, reducing outcome frequency while leaving every predictor–outcome relationship

intact. Covariate shift (S2) moved the means of three predictors, changing the distribution of inputs while leaving the conditional relationship unchanged. Concept drift (S3) weakened the coefficient of the strongest predictor, changing the relationship itself. Case-mix narrowing (S4) compressed the linear predictor about its mean, producing a more homogeneous population without altering any coefficient.

The fourth deserves emphasis because it is frequently misread in practice. A model moved from a heterogeneous referral population to a narrower one will show falling AUROC even though nothing about its validity has changed, because discrimination depends on the spread of risk in the population being ranked [4].

2.4 Metrics

Discrimination Was Measured By AUROC And By Area Under The Precision–Recall Curve. Calibration Was assessed by the Cox method: the calibration slope from a logistic regression of outcome on the linear predictor, and calibration-in-the-large as the intercept of that regression with slope constrained to unity [3,5]. Zero and one respectively indicate perfect calibration.

Operating characteristics were computed at a fixed alert threshold of 0.10, held constant across all cohorts to represent a deployed system whose threshold is configured once. We report sensitivity, specificity, positive predictive value, alerts per 1,000 encounters, and number needed to evaluate, defined as the count of alerts per true positive identified.

Clinical utility was quantified by decision-curve net benefit [7,8], computed as the true positive rate minus the false positive rate weighted by the odds of the threshold probability. Net benefit is reported at the deployment threshold and across a range of thresholds.

2.5 Recalibration

For the two cohorts in which calibration deteriorated, we tested recovery by intercept-only recalibration and by intercept-plus-slope recalibration. Each shifted cohort was split in half; recalibration parameters were estimated on the first half and all metrics were computed on the held-out second half, so that no recalibrated result is reported on the data used to fit it.

2.6 Software and Reproducibility

Analyses used Python with scikit-learn for model fitting, statsmodels for calibration regressions and SciPy for supporting computation. The complete generating script, including all seeds, is provided so that every number in this paper can be regenerated exactly. No human participants, human tissue or identifiable records were involved; all data are synthetic. Institutional review board approval and informed consent were therefore not applicable.

Table 1. Simulation parameters.

Parameter	Value	Rationale
Predictors	10, independent standard normal	Tractable; avoids collinearity as a confounder
Coefficients	0.85 to 0.10, mixed sign	Produces realistic discrimination near 0.82
Development cohort	40,000 encounters	Large enough that estimation error is negligible
Deployment cohorts	20,000 encounters each	Precision adequate for the comparisons made
Outcome prevalence (development)	9.0%	Typical of deterioration and readmission targets
Model	Logistic regression, negligible regularisation	Correctly specified; isolates shift from misspecification
Alert threshold	0.10, fixed across all cohorts	Represents a deployed threshold configured once

Table 2. Shift mechanisms implemented.

Code	Mechanism	Implementation	Real-world analogue
S0	No shift	Same generating process as development	Internal validation
S1	Prevalence shift	Intercept lowered by 0.80; coefficients unchanged	Successful prevention programme; seasonal variation; change in case ascertainment
S2	Covariate shift	Means of three predictors moved; conditional relationship unchanged	New catchment; change in admission policy; new care setting
S3	Concept drift	Coefficient of strongest predictor reduced from 0.85 to 0.30	Treatment change breaks a previously predictive relationship
S4	Case-mix narrowing	Linear predictor compressed by factor 0.62 about its mean	Deployment to a more homogeneous unit or triaged population

3. RESULTS

3.1 Development and Internal Performance

The development cohort contained 40,000 encounters with an outcome prevalence of 9.0 per cent. On an independent cohort drawn from the same distribution, the frozen model achieved an AUROC of 0.830, calibration intercept 0.026 and calibration slope 1.050 — performance that would be reported as good internal validation in any development paper.

3.2 The Central Dissociation

Figure 2 places discrimination beside clinical utility across the five deployment environments, and the dissociation is the principal finding of this study. Under prevalence shift the model's AUROC was 0.828, marginally higher than the 0.820 observed with no shift. Net benefit over the same comparison fell from 0.0420 to 0.0070, a reduction of 83 per cent.

Under covariate shift the divergence ran the other way. AUROC was 0.819, essentially unchanged, and calibration remained good, yet the alert rate rose from 277 to 483 per 1,000 encounters — a 74 per cent increase in operational load that no discrimination or calibration metric registered as a problem.

Only two of the four shift mechanisms moved AUROC appreciably: concept drift, to 0.780, and case-mix narrowing, to 0.729. The second of these is the awkward case, because nothing about the model's validity changed — the predictor–outcome relationships were identical to development. A monitoring system reporting AUROC alone would have flagged a model that was working correctly, while remaining silent on a model whose alerts had become half as likely to identify a true case.

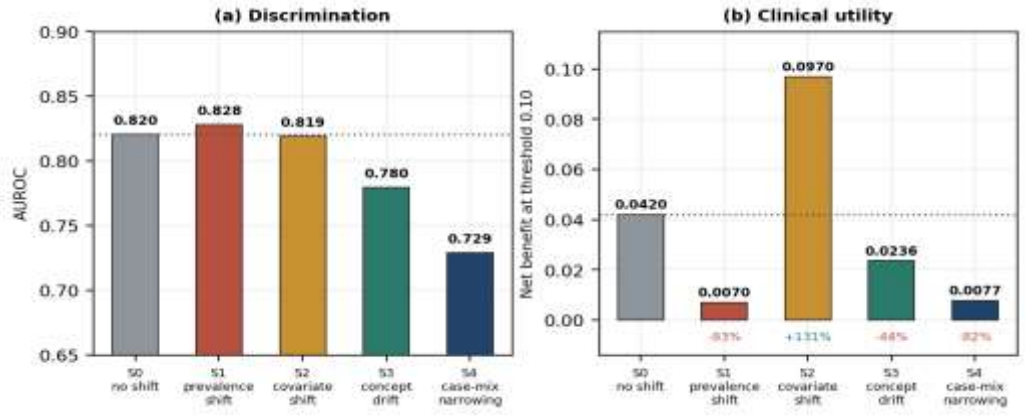

Figure 2. Discrimination and clinical utility across the five deployment environments. Dotted lines mark the no-shift reference. Percentages give the change in net benefit relative to no shift.

3.3 Discrimination and Calibration in Detail

Table 3 reports the discrimination metrics and Table 4 the calibration metrics. Figure 3 shows the corresponding ROC and calibration curves, and the contrast between the two panels is the visual form of the argument: ROC curves for the no-shift and prevalence-shift cohorts are nearly superimposed, while their calibration curves diverge sharply.

Area under the precision–recall curve behaved differently from AUROC, falling from 0.358 to 0.221 under prevalence shift and rising to 0.496 under covariate shift. This is expected, since precision–recall summaries depend on prevalence, and it makes AUPRC a more informative monitoring statistic than AUROC for this purpose — though it conflates a change in prevalence with a change in model quality, which calibration metrics do not.

Calibration intercept detected prevalence shift decisively, moving from 0.034 to -0.861. Calibration slope was near unity in that cohort (0.982), correctly indicating that the model's ranking remained sound and only its level was wrong. Concept drift moved the slope to 0.816 and case-mix narrowing to 0.612, in both cases signalling that predictions were too extreme for the new population.

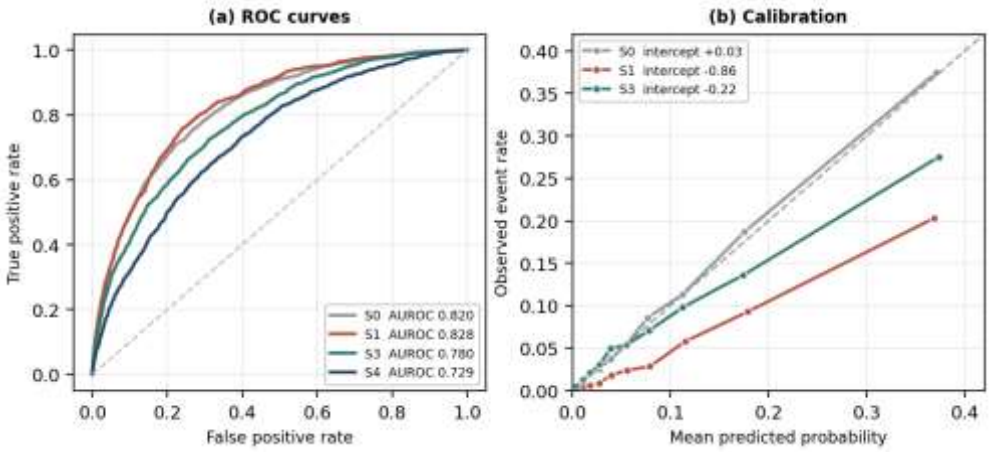

Figure 3. (a) ROC curves under four deployment conditions. (b) Calibration curves by decile of predicted risk, with the diagonal indicating perfect calibration.

Table 3. Discrimination by deployment environment.

Cohort	Prevalence	AUROC	AUPRC	Brier score
Internal validation	0.0916	0.8301	0.3881	0.0683
S0 no shift	0.0914	0.8202	0.3577	0.0699
S1 prevalence shift	0.0447	0.8277	0.2209	0.0430
S2 covariate shift	0.1569	0.8190	0.4965	0.1039
S3 concept drift	0.0753	0.7795	0.2641	0.0637
S4 case-mix narrowing	0.0596	0.7290	0.1598	0.0586

Each deployment cohort contains 20,000 simulated encounters. AUPRC = area under the precision–recall curve.

Table 4. Calibration by deployment environment.

Cohort	Calibration intercept	Calibration slope	Mean predicted risk	Observed / expected
Internal validation	0.026	1.050	0.0898	1.020
S0 no shift	0.034	0.995	0.0891	1.026
S1 prevalence shift	-0.861	0.982	0.0898	0.498
S2 covariate shift	0.027	1.040	0.1541	1.018
S3 concept drift	-0.220	0.816	0.0893	0.843

S4 case-mix narrowing	-0.519	0.612	0.0896	0.665

Perfect calibration corresponds to an intercept of 0 and a slope of 1. Observed/expected is the ratio of the observed event rate to the mean predicted risk.

3.4 What the Alerting System Actually Did

Table 5 and Figure 4 report the consequences at the deployed threshold, which is where a clinician meets the model. Under prevalence shift, positive predictive value fell from 0.236 to 0.122 and the number needed to evaluate rose from 4.23 to 8.19: a clinician would work through roughly twice as many alerts to find one true case, with no change whatever in the metric being monitored.

Alert volume behaved differently again. It was near-constant under prevalence shift, concept drift and case-mix narrowing (285, 279 and 275 per 1,000 respectively, against 277 with no shift) and rose sharply only under covariate shift, to 483. Alert rate is therefore a genuinely complementary monitoring signal: it detects the one failure that both discrimination and calibration miss, and it is the cheapest quantity in this paper to compute, requiring no outcome labels at all.

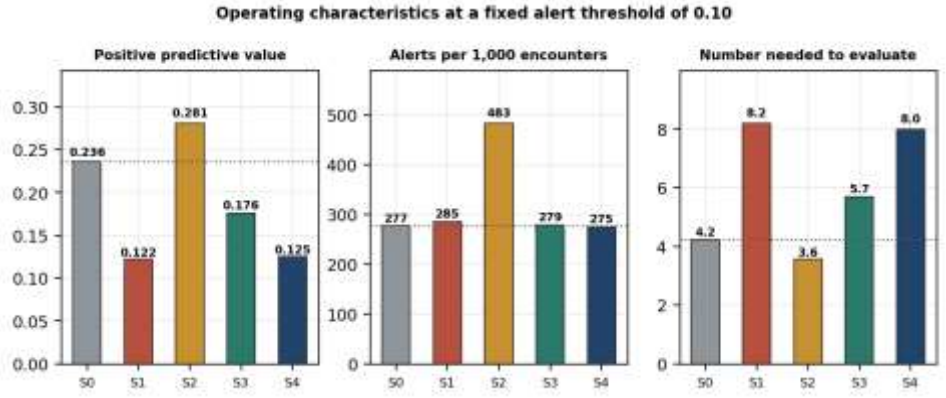

Figure 4. Operating characteristics at a fixed alert threshold of 0.10. Dotted lines mark the no-shift reference.

Table 5. Operating characteristics at the deployed alert threshold of 0.10.

Cohort	Sensitivity	Specificity	PPV	Alerts per 1,000	NNE
S0 no shift	0.717	0.767	0.236	277	4.23
S1 prevalence shift	**0.778**	**0.738**	**0.122**	**285**	**8.19**
S2 covariate shift	**0.864**	**0.588**	**0.281**	**483**	**3.56**
S3 concept drift	0.652	0.752	0.176	279	5.68
S4 case-mix narrowing	0.577	0.744	0.125	275	7.99

PPV = positive predictive value; NNE = number needed to evaluate, the number of alerts generated per true positive identified.

3.5 Decision Curves

Figure 5 shows net benefit across a range of threshold probabilities. The no-shift model provides benefit over treating all and treating none across the plotted range. Under prevalence shift the model's curve collapses toward the horizontal axis and the region over which it provides material benefit narrows substantially. Under concept drift the loss is intermediate.

Decision-curve analysis has the advantage of making the threshold dependence explicit, which matters because a deployed threshold is a policy choice that outlives the conditions it was chosen under. A threshold optimal at development prevalence is not optimal at half that prevalence, and the decision curve shows this directly where a single summary statistic cannot.

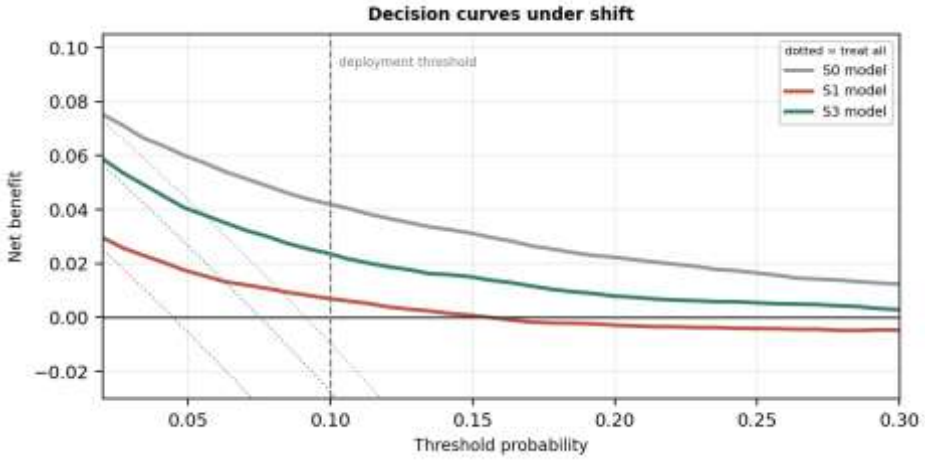

Figure 5. Decision curves under three deployment conditions. Solid lines show the model; dotted lines show a treat-all strategy at the corresponding prevalence.

3.6 Recovery by Recalibration

Table 6 and Figure 6 report recalibration, and the result is the practical counterpart of the paper's main finding. Intercept-only recalibration of the prevalence-shifted cohort moved the calibration intercept from -0.886 to -0.050, raised positive predictive value from 0.119 to 0.186, reduced alerts from 284 to 123 per 1,000, and increased net benefit from 0.0061 to 0.0118.

AUROC was 0.8248 before recalibration and 0.8248 after — identical to four decimal places, as it must be, since intercept recalibration is a monotone transformation of the predicted probabilities. The intervention that nearly doubled the model's clinical value was completely invisible to the metric most services monitor. Adding slope recalibration made no further difference in that cohort, because the slope was already near unity (0.964). In the concept-drift cohort, where the slope had fallen to 0.824, full recalibration corrected it to 1.019 and produced a modest further gain over intercept-only correction. Recalibration is not a substitute for redevelopment where the predictor–outcome relationship has genuinely changed, but it recovers a substantial part of the loss at trivial cost.

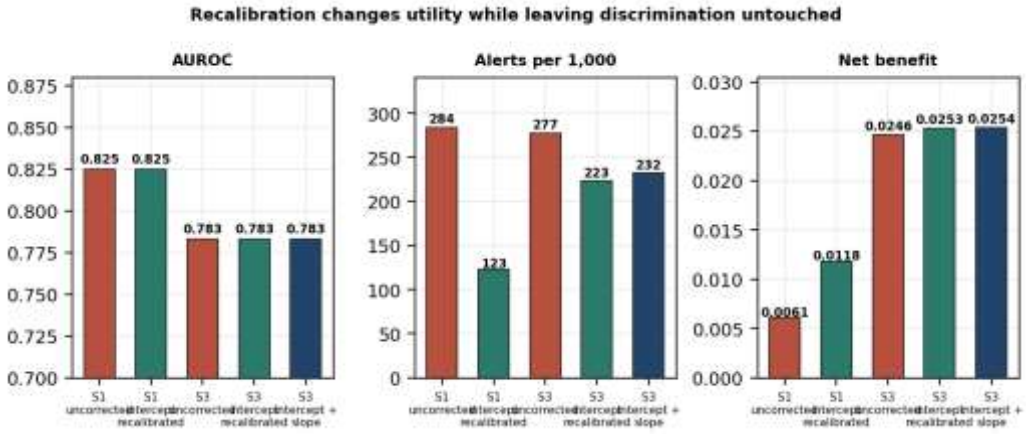

Figure 6. Effect of recalibration, evaluated on a held-out half of each shifted cohort. Discrimination is unchanged by construction; utility is not.

Table 6. Recalibration of the shifted cohorts, evaluated on held-out data.

Cohort and correction	AUROC	Cal. intercept	PPV	Alerts per 1,000	Net benefit
S1 uncorrected	0.8248	-0.886	0.119	284	0.0061
S1 intercept recalibrated	0.8248	-0.050	0.186	123	0.0118

S1 intercept + slope	0.8248	-0.050	0.186	123	0.0118
S3 uncorrected	0.7829	-0.201	0.180	277	0.0246
S3 intercept recalibrated	0.7829	0.038	0.202	223	0.0253
S3 intercept + slope	0.7829	0.033	0.198	233	0.0254

Recalibration parameters were estimated on one half of each shifted cohort and all metrics computed on the other half. AUROC is unchanged by recalibration because it is invariant to monotone transformation.

4. DISCUSSION

4.1 Principal Findings

Three findings follow from these results. Discrimination was silent for the two failures with the largest operational consequences: prevalence shift, which halved the yield of every alert, and covariate shift, which increased alert volume by 74 per cent. Calibration intercept detected the first and nothing in our metric set except alert rate detected the second. And a correction that took one parameter and no new model development recovered most of the lost utility, again without moving the monitored metric at all.

The case-mix result deserves separate emphasis because it is the inverse error. A monitoring system tracking AUROC would have raised an alarm about the model deployed to a narrower population, where no relationship had changed and the correct response would have been to do nothing to the model. False alarms of this kind consume the credibility that a monitoring system needs when a real failure arrives.

4.2 Relation to Existing Work

That AUROC is insensitive to calibration is not a new observation; it is stated in standard methodological references and in the TRIPOD reporting guidance, which requires calibration to be reported alongside discrimination [3,9,10]. Decision-curve analysis was introduced precisely to bring threshold-dependent utility into model evaluation [7,8], and dataset shift has been characterised in the machine learning literature for two decades [11,12].

The contribution here is not the phenomenon but its quantification under controlled conditions, expressed in units a deployment team uses: alerts per 1,000, number needed to evaluate, net benefit. Reporting guidance tells developers what to publish; it says comparatively little about what an operations team should watch on a Tuesday morning three years after go-live, which is where most deployed models now are.

4.3 A Minimum Monitoring Specification

Figure 7 and Table 7 set out which metric detects which failure and a corresponding minimum set. Four quantities are proposed, and the argument for each is empirical rather than theoretical.

Calibration intercept, because it detected the prevalence shift that discrimination missed entirely and that cost 83 per cent of net benefit. Calibration slope, because it detected concept drift and case-mix narrowing and distinguishes a model whose level is wrong from one whose ranking has degraded. Alert rate, because it was the only metric that moved under covariate shift, and because it requires no outcome labels and can therefore be computed continuously rather than waiting for outcome ascertainment. And net benefit at the deployed threshold, because it is the only quantity in the set that expresses what the model is worth rather than how it behaves.

We would retain discrimination in the set, but demote it. It answers a question about the model's ranking ability that remains worth knowing, and it is the metric most comparable with the development literature. It should not be the trigger for action.

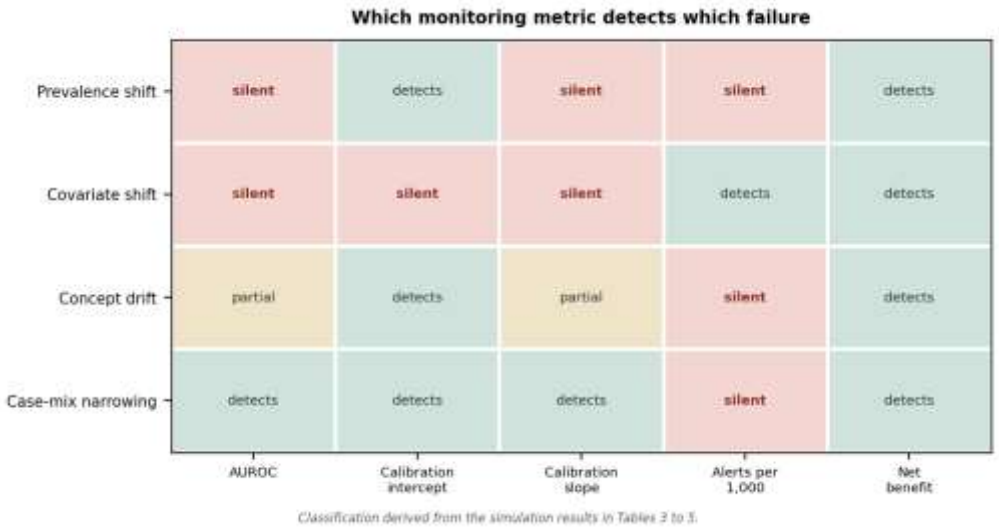

Figure 7. Which monitoring metric detects which failure mode, derived from the results in Tables 3 to 5.

Table 7. Proposed minimum monitoring set for a deployed prediction model.

Metric	Detects	Label requirement	Suggested cadence
Alert rate per 1,000	Covariate shift; upstream data pipeline failures	None	Continuous
Mean predicted risk	Covariate shift; input distribution change	None	Continuous
Calibration intercept	Prevalence shift; calibration drift	Outcomes required	Monthly or quarterly
Calibration slope	Concept drift; case-mix change	Outcomes required	Quarterly
Positive predictive value and NNE	The clinician's experience of the alert	Outcomes required	Monthly
Net benefit at the deployed threshold	Loss of clinical value from any cause	Outcomes required	Quarterly
AUROC	Concept drift; case-mix change (with false alarms)	Outcomes required	Quarterly, as context

4.4 Limitations

This is a simulation, and its external validity rests on whether the mechanisms modelled resemble those encountered in practice. Real deployments experience several mechanisms at once, at varying magnitudes, often gradually rather than as a step change. Our shifts were isolated and abrupt by design, because that is what permits attribution of a metric's behaviour to a mechanism, but the resulting estimates of metric sensitivity should be read as demonstrations of direction and order of magnitude rather than as expected values for any particular system.

Predictors were independent, normally distributed and correctly specified, and the outcome model was logistic. Real clinical data are correlated, non-normal, heavily missing and generated by processes no model specifies correctly. Model misspecification and missingness were deliberately excluded so that shift could be isolated; both would plausibly amplify rather than attenuate the dissociations reported.

The magnitudes of shift were chosen by us and are not calibrated to any observed distribution of real-world drift. A larger prevalence shift would produce a larger loss of net benefit, and a smaller one less; the qualitative claim — that AUROC does not move while utility does — holds for any magnitude, but the specific percentages do not generalise.

We evaluated a single model class at a single threshold on a single outcome. More flexible learners may respond differently to covariate shift in particular, since they can extrapolate less safely outside the training distribution, and this would be a worthwhile extension.

Finally, net benefit and number needed to evaluate quantify the informational value of an alert. They do not capture alert fatigue, workflow disruption, or the downstream consequences of acting on a prediction, all of which shape whether a deployed model helps patients.

5. CONCLUSION

A frozen prediction model moved into a population with half the outcome prevalence retained its discrimination exactly — AUROC 0.828 against 0.820 — while the clinician acting on its alerts had to evaluate 8.2 patients rather than 4.2 to find one true case, and the model's net benefit fell by 83 per cent. A one-parameter correction recovered much of that loss and changed AUROC not at all.

A monitoring practice built on discrimination will therefore miss the failures that matter most operationally and will raise false alarms about a model that has merely moved to a more homogeneous population. The remedy requires no new methodology: calibration intercept, calibration slope, alert rate and net benefit are all established quantities, and two of the four need no outcome labels.

As deployed models age, and the environments they were built in move away from them, what a monitoring dashboard chooses to display becomes a clinical safety decision rather than a technical preference. The case for displaying more than one number is, on these results, straightforward.

Declarations

Ethical approval: Not applicable.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of interes: The authors declare that there are no conflicts of interest regarding the publication of this paper.

Data availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author contributions:

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Bashar Sabah Sahib	✓		✓			✓	✓					✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

6. REFERENCES

1. J. A. Hanley and B. J. Mcneil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29-36, 1982. doi.org/10.1148/radiology.143.1.7063747
2. M. J. Pencina, D. R. Sr, D. R. Jr, and R. S. Vasan, "Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond," Stat Med, vol. 27, no. 2, pp. 157-172, 2008. doi.org/10.1002/sim.2929
3. E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, and N. Obuchowski, "Assessing the performance of prediction models: a framework for traditional and novel measures," Epidemiology, vol. 21, no. 1, pp. 128-138, 2010. doi.org/10.1097/EDE.0b013e3181c30fb2
4. B. Van Calster, D. J. Mclernon, M. Van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," BMC Med, vol. 17, 2019. doi.org/10.1186/s12916-019-1466-7

5. B. Van Calster, D. Nieboer, Y. Vergouwe, D. Cock, B. Pencina, and M. J. Steyerberg, "A calibration hierarchy for risk models was defined: from utopia to empirical data," J Clin Epidemiol, vol. 74, pp. 167-176, 2016. doi.org/10.1016/j.jclinepi.2015.12.005
6. R. D. Riley, L. Archer, K. Snell, J. Ensor, P. Dhiman, and G. P. Martin, "Evaluation of clinical prediction models (part 2): how to undertake an external validation study," BMJ, vol. 384, 2024. doi.org/10.1136/bmj-2023-074820
7. A. J. Vickers and E. B. Elkin, "Decision curve analysis: a novel method for evaluating prediction models," Med Decis Making, vol. 26, no. 6, pp. 565-574, 2006. doi.org/10.1177/0272989X06295361
8. A. J. Vickers, B. Van Calster, and E. W. Steyerberg, "Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests," BMJ, vol. 352, 2016. doi.org/10.1136/bmj.i6
9. G. S. Collins, J. B. Reitsma, D. G. Altman, and K. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement," BMJ, vol. 350, 2015. doi.org/10.1136/bmj.g7594
10. J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Dataset shift in machine learning. Cambridge: MIT Press, 2009. doi.org/10.7551/mitpress/9780262170055.001.0001
11. S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, and J. Zittrain, "The clinician and dataset shift in artificial intelligence," N Engl J Med, vol. 385, no. 3, pp. 283-286, 2021. doi.org/10.1056/NEJMc2104626
12. S. E. Davis, T. A. Lasko, G. Chen, E. D. Siew, and M. E. Matheny, "Calibration drift in regression and machine learning models for acute kidney injury," J Am Med Inform Assoc, vol. 24, no. 6, pp. 1052-1061, 2017. doi.org/10.1093/jamia/ocx030
13. S. E. Davis, R. A. Greevy, T. A. Lasko, C. G. Walsh, and M. E. Matheny, "Detection of calibration drift in clinical prediction models to inform model updating," J Biomed Inform, vol. 112, 2020. doi.org/10.1016/j.jbi.2020.103611
14. A. Wong, E. Otles, J. P. Donnelly, A. Krumm, J. Mccullough, and O. Detroyer-Cooley, "External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients," JAMA Intern Med, vol. 181, no. 8, pp. 1065-1070, 2021. doi.org/10.1001/jamainternmed.2021.2626
15. A. R. Habib, A. L. Lin, and R. W. Grant, "The epic sepsis model falls short - the importance of external validation," JAMA Intern Med, vol. 181, no. 8, pp. 1040-1041, 2021. doi.org/10.1001/jamainternmed.2021.3333
16. A. Youssef, M. Pencina, A. Thakur, T. Zhu, D. Clifton, and N. H. Shah, "External validation of AI models in health should be replaced with recurring local validation," Nat Med, vol. 29, no. 11, pp. 2686-2687, 2023. doi.org/10.1038/s41591-023-02540-z
17. J. Feng, R. V. Phillips, I. Malenica, A. Bishara, A. E. Hubbard, and L. A. Celi, "Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare," NPJ Digit Med, vol. 5, 2022. doi.org/10.1038/s41746-022-00611-y
18. A. Subbaswamy and S. Saria, "From development to deployment: dataset shift, causality, and shift-stable models in health AI," Biostatistics, vol. 21, no. 2, pp. 345-352, 2020. doi.org/10.1093/biostatistics/kxz041
19. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLoS One, vol. 10, no. 3, 2015. doi.org/10.1371/journal.pone.0118432
20. N. R. Cook, "Use and misuse of the receiver operating characteristic curve in risk prediction," Circulation, vol. 115, no. 7, pp. 928-935, 2007. doi.org/10.1161/CIRCULATIONAHA.106.672402
21. Y. Vergouwe, K. Moons, and E. W. Steyerberg, "External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients," Am J Epidemiol, vol. 172, no. 8, pp. 971-980, 2010. doi.org/10.1093/aje/kwq223
22. K. Janssen, K. Moons, C. J. Kalkman, D. E. Grobbee, and Y. Vergouwe, "Updating methods improved the performance of a clinical prediction model in new patients," J Clin Epidemiol, vol. 61, no. 1, pp. 76-86, 2008. doi.org/10.1016/j.jclinepi.2007.04.018

23. E. W. Steyerberg, G. Borsboom, H. C. Van Houwelingen, M. Eijkemans, and J. Habbema, "Validation and updating of predictive logistic regression models: a study on sample size and shrinkage," Stat Med, vol. 23, no. 16, pp. 2567-2586, 2004. doi.org/10.1002/sim.1844
24. T. L. Su, T. Jaki, G. L. Hickey, I. Buchan, and M. Sperrin, "A review of statistical updating methods for clinical prediction models," Stat Methods Med Res, vol. 27, no. 1, pp. 185-197, 2018. doi.org/10.1177/0962280215626466
25. L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, and E. Schuit, "Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal," BMJ, vol. 369, 2020. doi.org/10.1136/bmj.m1328
26. J. S. Ancker, A. Edwards, S. Nosal, D. Hauser, E. Mauer, and R. Kaushal, "Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system," BMC Med Inform Decis Mak, vol. 17, 2017. doi.org/10.1186/s12911-017-0430-8
27. P. J. Embi and A. C. Leonard, "Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study," J Am Med Inform Assoc, vol. 19, no. e1, pp. e145-148, 2012. doi.org/10.1136/amiajnl-2011-000743