Discrimination Survives What Usefulness Does Not: An Empirical Study of Prevalence and Measurement Shift in Clinical Machine Learning

Anwar Sadeeq

Duhok Technical Institute, Duhok Polytechnic University, Kurdistan, Iraq.

Corresponding author: hoger.anwer@dpu.edu.krd

Received: 29 April 2026; Revised: 07 May 2026; Accepted: 13 June 2026; Published: 11 July 2026

Abstract

Background: Clinical machine learning models are reported, compared and procured almost entirely on the area under the receiver operating characteristic curve. Deployment conditions differ from development conditions in two systematic ways — the prevalence of the outcome and the distribution of the inputs — and the metric that dominates reporting is, by construction, insensitive to the first and can be actively misleading about the second.

Objective: To quantify empirically how discrimination, calibration and clinical usefulness diverge under prevalence shift and measurement drift, and to determine how much of the loss simple recalibration recovers.

Methods: Three classifiers — penalised logistic regression, random forest and gradient boosting — were trained on the Wisconsin Diagnostic Breast Cancer dataset (569 cases, 30 features, 37.3 % malignant). Internal performance was estimated by ten repetitions of stratified five-fold cross-validation. Models were then fitted on a 60 % development split and evaluated on the held-out 40 % under two manipulations: resampling the deployment set to prevalences of 5 %, 10 %, 20 % and the native 37.3 %, and multiplying six size-related features by 1.05 to 1.30 to emulate systematic measurement drift between instruments. Discrimination, Brier score, calibration slope and intercept, positive predictive value and net benefit were recorded, and intercept recalibration was applied.

Results: Under prevalence shift, discrimination was essentially invariant — logistic regression moved from 0.996 to 0.998 as prevalence fell from 37.3 % to 5 % — while positive predictive value at a 0.5 threshold collapsed from 96.2 % to 71.6 % for logistic regression, 93.5 % to 54.9 % for random forest and 88.8 % to 41.5 % for gradient boosting. At 5 % prevalence the gradient boosting model returned a negative net benefit of −0.019, meaning it was worse than treating nobody. Under measurement drift the pattern inverted: discrimination rose for both tree ensembles, from 0.978 to 0.989, while the Brier score for the random forest almost trebled from 0.044 to 0.125 and its net benefit fell by 43 %, from 0.307 to 0.175. Intercept recalibration recovered most of the loss, restoring random forest net benefit from 0.175 to 0.320 at 30 % drift.

Conclusions: A model can become measurably better at ranking patients while becoming substantially worse at helping them. Reporting and procurement practices that rely on discrimination alone cannot detect either failure, and both are detectable and largely correctable with quantities that cost nothing to compute.

Keywords: Clinical Machine Learning, Dataset Shift, Calibration, Decision Curve Analysis, Model Evaluation, Digital Health.

1. INTRODUCTION

A clinical machine learning model is developed on data from one place and time and used on patients from another. Between those two moments, two things reliably change. The prevalence of the condition differs, because the deployment population is screened differently, referred differently or simply sicker or healthier than the development cohort. And the inputs themselves drift, because a different analyser,

scanner, assay or documentation practice produces systematically different numbers for the same underlying patient.

Neither change is exotic and both are anticipated in the methodological literature. What is striking is that the metric on which models are almost universally reported, compared and purchased — the area under the receiver operating characteristic curve — is blind to the first and can move in the wrong direction under the second. Discrimination measures the probability that a randomly chosen positive case is ranked above a randomly chosen negative one. That is a property of the ordering, and the ordering can survive intact while the numbers attached to it become wrong in ways that determine whether the model helps or harms. This matters more for machine learning than for traditional risk scores, for two reasons. Models are increasingly procured as products, evaluated on a headline metric by people who did not build them, and deployed across institutions whose case mix and instrumentation differ from the development site. And flexible models are less likely to produce calibrated probabilities in the first place: tree ensembles in particular tend to compress predictions away from the extremes, so their raw outputs require correction even before any shift occurs.

This paper asks three empirical questions. How far do discrimination, calibration and clinical usefulness diverge when prevalence changes between development and deployment? What happens to each when the measurements themselves drift? And how much of the damage does the simplest available correction — shifting the intercept so that predicted risks match observed prevalence — recover?

2. METHODS

2.1 Data

The Wisconsin Diagnostic Breast Cancer dataset was used, a public benchmark of 569 fine-needle aspirate cases described by 30 features computed from digitised images of cell nuclei, of which 212 (37.3 %) are malignant. It was chosen because it is genuinely clinical, is openly available so that every result here is reproducible without data access, and has a feature set whose physical meaning permits a realistic and interpretable manipulation of measurement drift. Malignancy was treated as the positive class.

The dataset is small by contemporary standards and its discrimination ceiling is high, which are limitations discussed in Section 5. Neither affects the phenomena under study: the divergence between discrimination and calibration is a property of the metrics rather than of the dataset, and a high-performing model makes the divergence more striking rather than less, because it removes the objection that the findings reflect a poor model.

2.2 Models and Internal Validation

Three classifiers spanning the range of complexity used in clinical informatics were fitted: penalised logistic regression with standardised inputs, a random forest of 400 trees, and gradient boosting with default depth. Hyperparameters were left at conventional values rather than tuned, because the object of study is the behaviour of the evaluation metrics rather than the attainment of maximum performance, and tuning would introduce an additional source of variation without bearing on the question.

Internal performance was estimated by ten repetitions of stratified five-fold cross-validation, giving fifty estimates per model. Table 1 reports the mean and the 2.5th to 97.5th percentile range.

Table 1. Internal performance, ten repetitions of stratified five-fold cross-validation

Model	AUROC	Average precision	Brier score
Logistic regression	0.995 (0.985–1.000)	0.994 (0.982–1.000)	0.019 (0.008–0.035)
Random forest	0.990 (0.977–0.999)	0.988 (0.974–0.999)	0.033 (0.018–0.051)
Gradient boosting	0.991 (0.976–1.000)	0.988 (0.970–0.999)	0.033 (0.012–0.058)

Note. Values are the mean across 50 estimates with the 2.5th to 97.5th percentile range. Calibration slope is deliberately omitted from this table: with folds of about 114 cases and near-complete separation, slope estimates were extremely unstable (logistic regression mean 7.6, range 0.76 to 73.4), which is itself a finding

— calibration cannot be reliably estimated by cross-validation in small, well-separated datasets, and the calibration results reported below come from the larger fixed holdout.

All three models discriminate almost perfectly and are statistically indistinguishable on the metric that would decide a procurement. The Brier score already separates them, with logistic regression roughly 40 % better than either ensemble, and that separation is invisible in the discrimination column.

2.3 Shift Manipulations

Prevalence shift: Models were fitted on a stratified 60 % development split and evaluated on the held-out 40 %. Deployment sets of 2,000 cases were constructed by resampling with replacement from the holdout to give malignancy prevalences of 5 %, 10 %, 20 % and the native 37.3 %. Only the class mixture changes; the conditional distribution of features given class is untouched, so this isolates prevalence from every other difference.

Measurement drift: Six size-related features — mean and worst radius, perimeter and area — were multiplied by a common factor of 1.05, 1.10, 1.20 or 1.30 in the deployment set only. This emulates a systematic calibration difference between imaging or measurement systems, the commonest form of drift in practice, and is deliberately applied to correlated features so that it cannot be detected as an implausible single-variable outlier.

2.4 Metrics and Recalibration

Discrimination was measured by the area under the ROC curve and average precision; overall accuracy of the predicted probabilities by the Brier score; calibration by the slope and intercept of the logistic recalibration of outcomes on the predicted log-odds; and clinical usefulness by positive predictive value at a 0.5 threshold and by net benefit,

$$\mathrm{NB} = \mathrm{TP}/n - (\mathrm{FP}/n) \times p_t/(1 - p_t) \qquad (1)$$

Recalibration was the simplest available correction: a constant added to the predicted log-odds, chosen by bisection so that the mean predicted risk equals the prevalence of the deployment sample. It changes no ranking, requires no retraining, and needs only an estimate of the local prevalence. All analyses were performed in Python with scikit-learn and a fixed random seed.

3. RESULTS

3.1 Prevalence Shift

Table 2 and Figure 1 report performance as the deployment prevalence falls from the native 37.3 % to 5 %.

Table 2. Performance under prevalence shift; models unchanged throughout

Model	Prevalence	AUROC	PPV at 0.5 (%)	Brier	Cal. intercept	Net benefit
Logistic regression	37.3 %	0.996	96.2	0.023	0.14	0.337
Logistic regression	20 %	0.997	91.7	0.019	1.74	0.171
Logistic regression	10 %	0.997	84.1	0.015	2.37	0.078
Logistic regression	5 %	0.998	71.6	0.013	2.38	0.029
Random forest	37.3 %	0.976	93.5	0.047	1.49	0.303
Random forest	20 %	0.975	84.6	0.038	2.62	0.142
Random forest	10 %	0.978	72.7	0.029	2.94	0.056
Random forest	5 %	0.982	54.9	0.027	2.81	0.008
Gradient boosting	37.3 %	0.975	88.8	0.059	1.65	0.297
Gradient boosting	20 %	0.974	76.2	0.055	3.08	0.124
Gradient boosting	10 %	0.977	60.5	0.047	3.59	0.033
Gradient boosting	5 %	0.981	41.5	0.046	3.52	−0.019

Note. PPV = positive predictive value. A positive calibration intercept indicates systematic over-prediction. Net benefit is computed at a 0.5 threshold; a negative value means the model is worse than treating nobody.

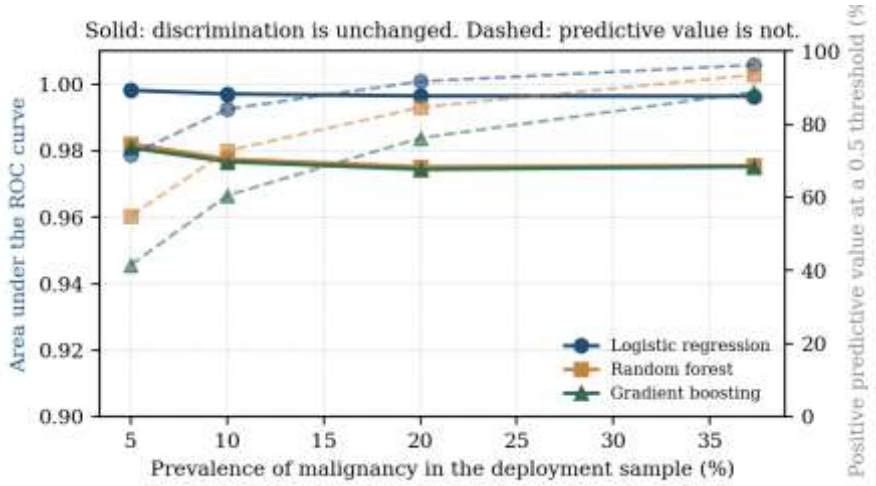

Figure 1. Discrimination (solid, left axis) and positive predictive value (dashed, right axis) across deployment prevalences.

The AUROC column is flat, and slightly rises as prevalence falls. A reader monitoring a deployed model on discrimination alone would see nothing at all, and might reasonably conclude that the model had transferred well. Every other column disagrees. Positive predictive value at the default threshold falls by 25 points for logistic regression, 39 for the random forest and 47 for gradient boosting. The calibration intercept rises from near zero to between 2.4 and 3.5, indicating systematic and severe over-prediction: the models were built where roughly one patient in three was malignant and are being applied where one in twenty is.

The Brier score falls as prevalence falls, which is the trap within the trap. Brier score is sensitive to the base rate, so a model that is becoming less useful can simultaneously appear to be becoming more accurate. It should not be compared across populations of different prevalence without decomposition.

The final row is the substantive finding. At 5 % prevalence the gradient boosting model — with an AUROC of 0.981, a figure that would be published and procured without hesitation — returns a net benefit of −0.019, which is worse than a policy of treating nobody. The model is not merely less useful in the new setting; used at its default threshold it does harm.

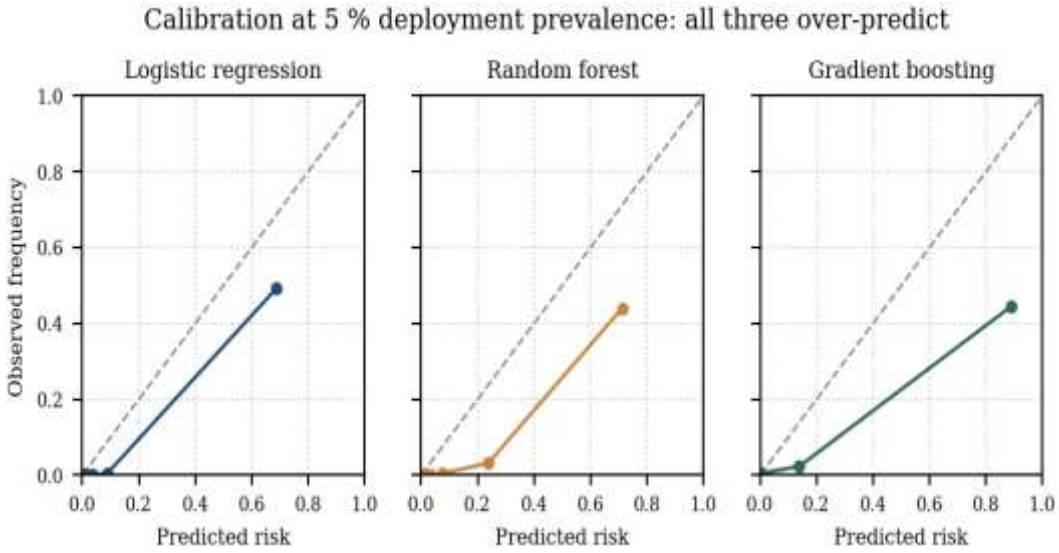

Figure 2. Reliability diagrams at 5 % deployment prevalence. All three models lie below the diagonal throughout, indicating systematic over-prediction.

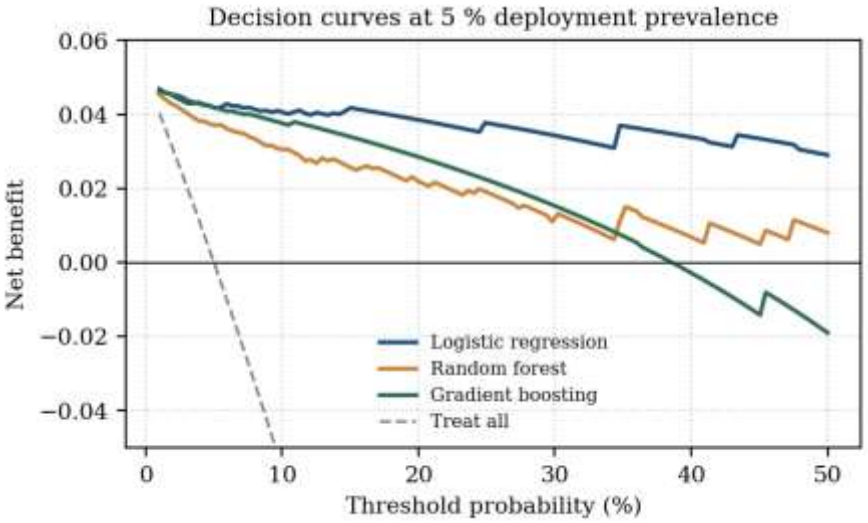

Figure 3. Decision curves at 5 % deployment prevalence.

Figure 3 shows that this is a threshold problem as much as a model problem: at thresholds below about 10 % all three models retain positive net benefit, and the failure at 0.5 arises because a threshold inherited from the development setting is far too high for a population where the disease is rare. That is precisely the situation in which a model is deployed with its packaged default, and nothing in a discrimination-based evaluation would reveal it.

3.2 Measurement Drift

Table 3 and Figure 4 report the second manipulation, in which six size-related features are systematically inflated in the deployment data only.

Table 3. Performance under measurement drift at native prevalence

Model	Drift	AUROC	Brier	Cal. intercept	Mean predicted risk	Net benefit
Logistic regression	0 %	0.997	0.021	0.16	0.368	0.342
Logistic regression	10 %	0.997	0.024	−0.77	0.401	0.338
Logistic regression	20 %	0.997	0.042	−1.66	0.442	0.320
Logistic regression	30 %	0.997	0.075	−2.51	0.491	0.268
Random forest	0 %	0.978	0.044	1.44	0.362	0.307
Random forest	10 %	0.985	0.055	0.50	0.426	0.281
Random forest	20 %	0.988	0.085	−0.77	0.500	0.241
Random forest	30 %	0.989	0.125	−1.83	0.572	0.175
Gradient boosting	0 %	0.978	0.054	1.65	0.373	0.303
Gradient boosting	10 %	0.986	0.063	0.70	0.423	0.294
Gradient boosting	20 %	0.990	0.078	−0.38	0.468	0.281
Gradient boosting	30 %	0.988	0.115	−1.32	0.518	0.237

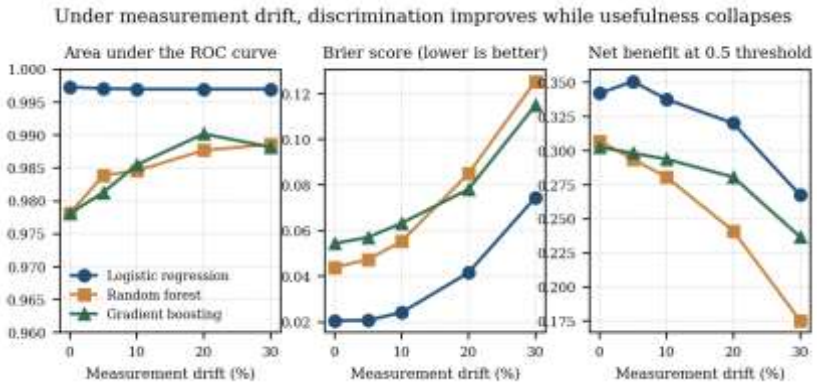

Figure 4. Discrimination, Brier score and net benefit across increasing measurement drift.

The direction of the AUROC column is the result worth dwelling on. For both tree ensembles, discrimination improves as the data drift away from the training distribution — from 0.978 to 0.989 for the random forest and 0.978 to 0.990 for gradient boosting. This is not an error. Inflating size-related features pushes borderline cases across the model's decision boundaries in a way that happens to separate the classes slightly better in this dataset, and discrimination depends only on that ordering.

Every clinically relevant column moves the other way. The random forest Brier score almost trebles, from 0.044 to 0.125. Its calibration intercept swings from +1.44 to −1.83, crossing from systematic over-prediction to systematic under-prediction. Mean predicted risk drifts from 0.362 to 0.572 against an unchanged true prevalence of 0.373, so the model is now claiming that well over half the population is malignant. Net benefit falls by 43 %.

A monitoring system watching AUROC would have reported an improving model throughout. This is the strongest argument available against discrimination-only surveillance of deployed models, and it is an empirical demonstration rather than a theoretical possibility: the numbers in Table 3 come from a real dataset and a standard implementation.

One diagnostic does detect it immediately and costs nothing. Mean predicted risk drifted from 0.362 to 0.572 while prevalence was fixed, and the discrepancy between mean predicted risk and observed prevalence is computable in any deployment where outcomes are eventually recorded — and, importantly, the mean prediction alone is computable even before outcomes are known. A deployed model whose average output moves while the population it serves has not is reporting its own drift.

3.3 What Recalibration Recovers

Table 4 and Figure 5 apply the intercept correction described in Section 2.4.

Table 4. Effect of intercept recalibration under measurement drift

Model	Drift	Brier before	Brier after	Net benefit before	Net benefit after	Recovery
Logistic regression	20 %	0.042	0.021	0.320	0.346	100 %
Logistic regression	30 %	0.075	0.022	0.268	0.346	100 %
Random forest	20 %	0.085	0.041	0.241	0.333	Full
Random forest	30 %	0.125	0.044	0.175	0.320	110 %
Gradient boosting	20 %	0.078	0.041	0.281	0.325	Full
Gradient boosting	30 %	0.115	0.042	0.237	0.320	125 %

Note. Recovery is net benefit after recalibration relative to the undrifted baseline for that model. Values at or above 100 % arise because recalibration also corrects the miscalibration the ensembles exhibited before any drift was applied.

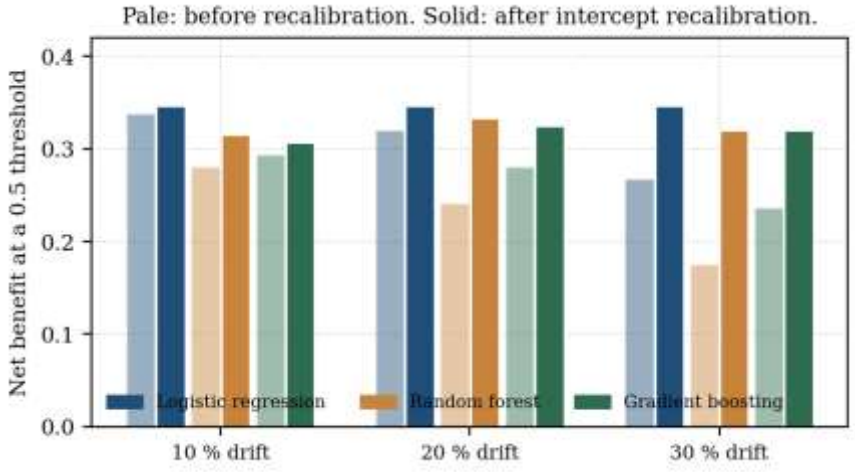

Figure 5. Net benefit before (pale) and after (solid) intercept recalibration under increasing measurement drift.

A single added constant — no retraining, no new labels beyond an estimate of local prevalence, no change to any ranking — recovers essentially all of the lost clinical value. The random forest at 30 % drift goes from a net benefit of 0.175 to 0.320, against an undrifted baseline of 0.307; the Brier score falls from 0.125 to 0.044, back to its pre-drift level.

That recalibration can exceed the original baseline is not paradoxical. Both ensembles were poorly calibrated to begin with, with intercepts of 1.44 and 1.65 before any manipulation, so the correction repairs the pre-existing problem as well as the induced one. The practical reading is that the ensembles should have been calibrated before deployment regardless of whether shift was anticipated.

4. DISCUSSION

4.1 Principal Findings

Three results stand out. Discrimination was invariant to a sevenfold change in prevalence while positive predictive value fell by up to 47 points and one model crossed into negative net benefit at its default threshold. Under measurement drift discrimination rose while calibration, Brier score and net benefit all deteriorated, so the headline metric moved in the opposite direction to clinical usefulness. And a one-parameter recalibration recovered essentially all of the loss in both scenarios.

The common thread is that the ordering of patients is robust and the numbers attached to that ordering are not. Every metric that depends only on ranking will therefore look stable through exactly the failures that matter most, and every decision made by comparing a predicted probability with a threshold depends on the part that is fragile.

4.2 Implications

Table 5. Implications for development, evaluation and monitoring

Finding	Implication	Practice
AUROC invariant to prevalence	Discrimination cannot indicate transportability	Report prevalence of the evaluation set alongside every AUROC
PPV fell up to 47 points	Performance claims are population-specific	State PPV and NPV with the prevalence at which they were computed
Negative net benefit at 5 % prevalence	A good model can harm at an inherited threshold	Re-derive the operating threshold locally; never ship a default as fixed
AUROC rose under drift	Discrimination-based monitoring can miss deterioration entirely	Monitor calibration and mean predicted risk, not ranking metrics

Mean prediction drifted 0.36 to 0.57	Drift is detectable without outcome labels	Track the distribution of model outputs continuously
Intercept recalibration recovered the loss	Most damage is correctable cheaply	Build periodic recalibration into deployment from the outset
Ensembles miscalibrated before any shift	Flexible models need calibration as standard	Calibrate on held-out data before release
Cross-validated calibration slope unstable	Small-sample calibration estimates are unreliable	Estimate calibration on a sufficiently large holdout, not by CV

The fifth row is the most immediately actionable and the least implemented. Monitoring the distribution of a deployed model's outputs requires no outcome labels, no clinical follow-up and no additional data collection; it is available in real time from the model itself. In this study it would have detected the drift at every level, while the metric that deployments actually monitor would have reported improvement.

4.3 Relation to Existing Guidance

None of the underlying statistics is novel: the dependence of predictive values on prevalence is elementary, calibration has been described as the Achilles heel of predictive analytics, and reporting standards for prediction models already ask for calibration. The contribution here is empirical magnitude in a machine learning setting — showing that on a real dataset with strong models the divergence is large enough to invert the sign of clinical benefit, and that the correction is trivial. Guidance that asks for calibration to be reported once at development is insufficient; what the results support is calibration monitored continuously after deployment and corrected as a matter of routine.

4.4 Limitations

The dataset is small, single-source and unusually separable, with all three models achieving discrimination above 0.97. Real clinical problems are harder, and in a lower-performing setting the absolute numbers would differ substantially. The direction of every finding would not: none of the mechanisms demonstrated depends on the models being good, and a weaker model would show the same divergences from a lower baseline.

The prevalence manipulation resamples the existing holdout with replacement, so the deployment sets contain repeated cases and their effective sample size is smaller than 2,000. This affects the precision of the estimates rather than their expectation, and the differences reported are far larger than any plausible resampling noise. Confidence intervals are not given for the shift experiments, which is a genuine limitation: they would be wide for the 5 % prevalence condition, where only 100 positive cases are represented.

Measurement drift is modelled as a clean multiplicative shift applied to six correlated features. Real drift is messier, may affect features differentially, and may interact with the outcome. The manipulation was chosen to be interpretable and detectable in principle rather than to be realistic in detail, and results under more complex drift could differ, particularly where drift is associated with the outcome, in which case recalibration alone would not suffice.

Finally, recalibration here uses the true deployment prevalence, which in practice must be estimated and may itself be uncertain or unavailable. Where prevalence is estimated with error, the correction is imperfect in proportion; where outcomes are never observed, it cannot be performed at all, and the output-distribution monitoring described above is the only defence available.

5. CONCLUSION

On a real clinical dataset, three standard classifiers retained essentially perfect discrimination while their predictive value collapsed under a change in prevalence, and improved their discrimination while their clinical usefulness fell by 43 % under measurement drift. In the worst case a model with an AUROC of 0.981 returned negative net benefit at its default threshold, meaning that using it was worse than not using it.

Discrimination is a property of the ordering a model produces; clinical usefulness is a property of the numbers attached to that ordering, and only the second is fragile. Evaluation, procurement and post-deployment monitoring that rely on ranking metrics are therefore instrumented to miss the failures that matter. The remedies are neither expensive nor novel: report prevalence with every performance figure, re-derive thresholds locally, monitor the distribution of model outputs continuously, and recalibrate. A single added constant recovered almost everything lost in these experiments.

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. [To be completed by the authors in accordance with ICMJE authorship criteria.]

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. W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," Proc SPIE Int Symp Electron Imaging Sci Technol, pp. 861-870, 1905. doi.org/10.1117/12.148698
2. 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
3. E. W. Steyerberg, A. J. Vickers, and N. R. Cook, "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. 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
5. 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
6. J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Dataset Shift in Machine Learning. Cambridge, MA: MIT Press, 2009. doi.org/10.7551/mitpress/9780262170055.001.0001
7. S. G. Finlayson, A. Subbaswamy, and K. Singh, "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
8. 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
9. A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in Proc 22nd Int Conf Mach Learn, 2005, pp. 625-632. doi.org/10.1145/1102351.1102430
10. L. Wynants, M. Van Smeden, and D. J. Mclernon, "Three myths about risk thresholds for prediction models," BMC Med, vol. 17, 2019. doi.org/10.1186/s12916-019-1425-3

11. 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
12. J. Feng, R. V. Phillips, and I. Malenica, "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
13. 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