

A Comparative Investigation of Class Imbalance Mitigation Strategies in Clinical Machine Learning: Evaluating Data Resampling, Synthetic Sample Generation, Cost-Aware Algorithms, and Decision Threshold Optimization Under Five Levels of Class Imbalance

Julia López ¹, Mia Kamińska ²

¹University of Nairobi, KENYA

²University of Cape Town, SOUTH AFRICA

Keywords

Class Imbalance, SMOTE, Resampling, Calibration, Clinical Prediction Models, Precision–Recall Curve, Threshold Tuning, Cost-Sensitive Learning

ABSTRACT

Background. Clinical outcomes of interest are usually rare, and applying a class-imbalance correction — most often SMOTE — has become near-automatic in medical machine learning. Whether these corrections improve anything a clinician would value has been questioned but is rarely tested across imbalance regimes. **Objective.** To quantify the effect of ten imbalance strategies on discrimination, classification and probability calibration across systematically varied imbalance ratios. **Methods.** Three real clinical datasets (Pima diabetes, Haberman survival, Wisconsin Diagnostic Breast Cancer) were subsampled to induce majority-to-minority ratios of 1:2, 1:5, 1:10, 1:20 and 1:50. Ten strategies — no correction, random over- and under sampling, SMOTE, Borderline-SMOTE, ADASYN, SMOTE+Tomek, class weighting, decision-threshold tuning and balanced random forests — were combined with three classifiers under five-fold stratified cross-validation, giving 392 evaluated configurations. Every correction was applied inside the cross-validation loop to training folds only. **Results.** No resampling method improved ranking performance. Mean change in area under the precision–recall curve relative to no correction was -0.008 for SMOTE (winning 47.6% of 42 paired comparisons; paired t -test $p = 0.035$ in favour of no correction), -0.011 for random oversampling and -0.025 for random under sampling. Corrections did improve classification: F1 rose from 0.414 to 0.499 under SMOTE and G-mean from 0.451 to 0.641. However, simple threshold tuning achieved statistically indistinguishable F1 (0.494 vs 0.499, Wilcoxon $p = 0.229$) while leaving probabilities untouched. Calibration damage was severe and systematic: mean predicted risk rose from a true prevalence of 0.121 to 0.214 under SMOTE (+9.3 percentage points) and 0.343 under random under sampling (+22.2 points), with Brier scores 56% and 160% worse than baseline, respectively. Separately, AUC-ROC proved almost blind to imbalance, falling only from 0.906 to 0.819 as prevalence dropped from 33% to 3.6%, while AUC-PR halved from 0.841 to 0.457. **Conclusions.** Imbalance corrections buy a sensitivity shift that threshold tuning provides for free, and they pay for it with systematically overestimated risk. For clinical models whose output is a probability, the correction is not neutral — it is harmful.

Introduction

The clinical events worth predicting are, by their nature, uncommon. In-hospital mortality, post-operative complications, rare cancers and adverse drug reactions all occur in a small fraction of patients, and a model trained naively on such data can achieve high accuracy by predicting that nothing will happen. The standard response in the applied literature is to correct the imbalance before training by oversampling the minority

class, undersampling the majority, generating synthetic minority examples with SMOTE, or reweighting the loss function [1].

This has become close to a reflex. SMOTE, introduced in 2002 [1], now appears in a large share of clinical machine learning papers, frequently without justification and almost never with an uncorrected comparator reported alongside. The implicit assumption is that correction is at worst neutral and at best necessary.

That assumption has been challenged. Van den Goorbergh and colleagues showed by simulation and a case study in ovarian cancer diagnosis that random undersampling, oversampling and SMOTE produced poorly calibrated logistic regression models with strongly overestimated minority probabilities, no improvement in AUC, and a sensitivity–specificity shift that could be obtained equally well by moving the decision threshold [2]. Subsequent work has reported consistent findings across larger collections of clinical datasets [3], [4]. Nonetheless, the practice continues, in part because the evidence is scattered across simulation studies and single-dataset illustrations rather than a systematic comparison across imbalance regimes and model families.

This study provides that comparison. Rather than accepting each dataset’s natural prevalence, it induces imbalance systematically — from a mild 1:2 to a severe 1:50 — on three real clinical datasets, so that imbalance itself becomes the experimental variable. Contributions are:

- 1.a factorial design over three datasets, five imbalance ratios, ten strategies and three classifiers (392 evaluated configurations), with all corrections applied strictly inside the cross-validation loop;
- 2.evaluation on discrimination, classification and calibration simultaneously, since strategies that help on one axis are shown here to harm another;
- 3.a direct paired test of the practical question the literature leaves open — does SMOTE outperform simply moving the decision threshold?;
- 4.quantification of calibration damage as risk inflation in percentage points, a form directly interpretable by clinicians;
- 5.an empirical demonstration of how far AUC-ROC diverges from AUC-PR as prevalence falls, which bears on how the entire literature reports its result [2-7]s.

Background

Correction Strategies

Strategies divide into four families. Data-level methods change the training distribution: random oversampling duplicates minority cases, random undersampling discards majority cases, and SMOTE interpolates new synthetic minority points between a case and its nearest minority neighbours [1]. Borderline-SMOTE restricts generation to cases near the decision boundary and ADASYN concentrates generation where the local minority density is lowest; SMOTE+Tomek follows generation with removal of borderline-overlapping pairs. Algorithm-level methods leave the data alone and reweight the loss, either by class weights in logistic regression and random forests or by a positive-class scaling factor in gradient boosting. Ensemble methods such as balanced random forests resample within each bootstrap. Decision-level methods change neither data nor loss, adjusting only the probability threshold at which a positive prediction is issued.

The last of these is conceptually distinct and important for what follows. Threshold tuning cannot alter any ranking-based metric, because it does not change the ordering of predicted probabilities. It also cannot damage calibration, because it does not change the probabilities at all. Whatever it achieves on classification metrics is achieved for free.

Why Calibration Matters Clinically

A clinical prediction model usually outputs a probability that is acted on directly: a 30% risk of complication and a 3% risk lead to different conversations, even when both fall on the same side of a classification threshold. Calibration is the agreement between predicted probabilities and observed frequencies [5], and a model that systematically overestimates risk will lead to over-treatment, unnecessary anxiety and misallocated resources regardless of how well it ranks patients.

Resampling breaks calibration for a structural reason. A model trained on data rebalanced to 1:1 learns the base rate of the rebalanced sample, not of the population. If it is then applied to patients whose true event rate is 5%, its probability estimates carry a prior of 50%. The distortion is not an implementation flaw; it is what the method does, and it is only removable by an explicit recalibration step that the applied literature seldom performs [2].

Materials and Methods

Datasets

Three publicly available clinical datasets with genuinely different characteristics were used (Table 1). The Pima Indians diabetes dataset contains 768 records with eight predictors and a 34.9% prevalence of diabetes onset. The Haberman survival dataset contains 306 records with only three predictors and 26.5% five-year mortality after breast-cancer surgery; it is deliberately included as a hard, low-signal problem. The Wisconsin Diagnostic Breast Cancer dataset contains 569 records with 30 predictors and 37.3% malignancy, and is an easy, high-signal problem. Together they span the range of signal strength encountered in practice.

Table 1: Source Datasets Before Imbalance Induction

Dataset	Clinical outcome	n	Predictors	Positives	Prevalence
Pima Indians diabetes	Type 2 diabetes onset	768	8	268	34.9%
Haberman survival	Death within 5 years of surgery	306	3	81	26.5%
Wisconsin Diagnostic (WDBC)	Malignant breast tumour	569	30	212	37.3%

Inducing Imbalance

For each dataset and each target ratio $r \in \{2, 5, 10, 20, 50\}$, the majority class was retained entirely and the minority class randomly subsampled to n_{majority} / r cases, with a floor of 12 minority cases to keep cross-validation feasible. Configurations that could not reach the target ratio were skipped rather than approximated. This design holds dataset, features and majority class fixed while varying only imbalance, so any change in performance is attributable to imbalance alone [8-13].

Strategies, Classifiers and Protocol

Ten strategies were evaluated: no correction; random oversampling; random undersampling; SMOTE; Borderline-SMOTE; ADASYN; SMOTE+Tomek; class weighting; decision-threshold tuning; and balanced random forests. Each was paired with logistic regression, a 250-tree random forest and gradient boosting (250 rounds, depth 3, learning rate 0.06), except balanced random forests which is intrinsically tree-based.

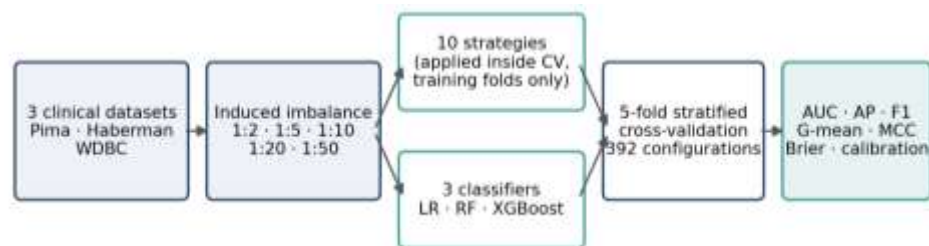
The methodological point that most often goes wrong in this literature is where the correction is applied. Resampling the full dataset before splitting leaks synthetic points derived from test cases into the training set and inflates every metric. Here the protocol was:

```
for each of 5 stratified folds:
    fit scaler on TRAIN fold only
    apply resampling to TRAIN fold only
    fit model; tune threshold on TRAIN fold if applicable
    evaluate on the untouched TEST fold
```

No synthetic sample ever reached a test fold, and thresholds were selected on training data only. Nine metrics were recorded [6]: AUC-ROC, average precision (AUC-PR) [7], F1, sensitivity, specificity, G-mean, Matthews correlation coefficient [8], Brier score [9], and mean predicted probability. Calibration slope was estimated by logistic recalibration of the predicted logit against the observed outcome [5]. Figure 1 summarises the design.

Figure 1: Experimental Design

Figure 1 *Experimental design. Three datasets × five induced imbalance ratios × ten strategies × three classifiers, evaluated by five-fold stratified cross-validation, giving 392 completed configurations.*



Statistical Analysis

Strategies were compared to the uncorrected baseline within matched configurations — same dataset, same ratio, same classifier — so that comparisons are paired rather than marginal. Differences are reported as mean change with a 95% confidence interval and a win rate. Paired comparisons used the Wilcoxon signed-rank test with the paired t-test reported alongside, since the two diverge in one instructive case. All computation was performed in Python (scikit-learn, imbalanced-learn [10], XGBoost, SciPy) at seed 42.

Results

AUC-ROC Is Nearly Blind to Imbalance

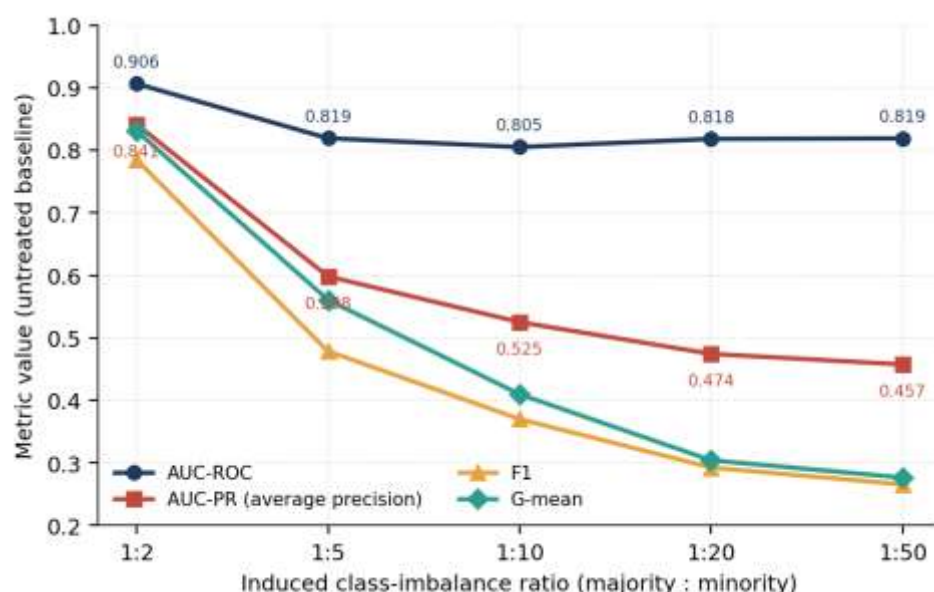
Before evaluating any correction, the uncorrected baseline shows something that bears on how the whole literature reports results (Figure 2). As the induced ratio moves from 1:2 to 1:50 — prevalence falling from 33.3% to 3.6% — AUC-ROC declines only from 0.906 to 0.819, and is essentially flat from 1:5 onward (0.819, 0.805, 0.818, 0.819). Over the same range, average precision falls from 0.841 to 0.457, F1 from 0.785 to 0.266 and G-mean from 0.831 to 0.277.

A reader given only AUC-ROC would conclude that severe imbalance costs about nine points of performance. A reader given AUC-PR would conclude it halves it. The second reader is closer to the clinical truth, because AUC-ROC's specificity term is computed over a majority class so large that even many false positives barely move it, whereas precision — the probability that a flagged patient really has the condition

— collapses [7]. Since most clinical machine learning papers report AUC-ROC as the headline metric, this alone accounts for a portion of the gap between published performance and clinical experience.

Figure 2: Metric Behaviour as Imbalance Increases

Figure 2 Behaviour of four metrics on the uncorrected baseline as induced imbalance increases. AUC-ROC is nearly flat while AUC-PR, F1 and G-mean fall sharply.



No Correction Improved Ranking Performance

Against the matched uncorrected baseline, the mean change in average precision was negative for every data-level and algorithm-level correction (Figure 3, Table 2). Random undersampling was worst at -0.025 (winning only 11.9% of 42 paired comparisons), followed by random oversampling at -0.011 and class weighting at -0.011 . SMOTE and ADASYN both lost 0.008, and SMOTE+Tomek 0.007. Only balanced random forests improved on baseline, at $+0.008$ across its 14 comparisons, and Borderline-SMOTE was effectively neutral at -0.001 .

The SMOTE comparison deserves care because the two paired tests disagree. SMOTE won 47.6% of its 42 comparisons — almost exactly a coin flip — with a mean deficit of 0.008. The Wilcoxon signed-rank test gives $p = 0.180$ and the paired t-test $p = 0.035$. The honest reading is that SMOTE's losses are slightly larger in magnitude than its wins, which the t-test detects and the rank-based test does not. Either way, the direction is not favourable, and nothing in these data supports the proposition that SMOTE improves the model's ability to rank patients by risk [14-17].

Figure 3: Change in AUC-PR Relative to Baseline

Figure 3 Change in AUC-PR relative to the matched uncorrected baseline, with 95% confidence intervals and win rates. Only balanced random forests improved on no correction.

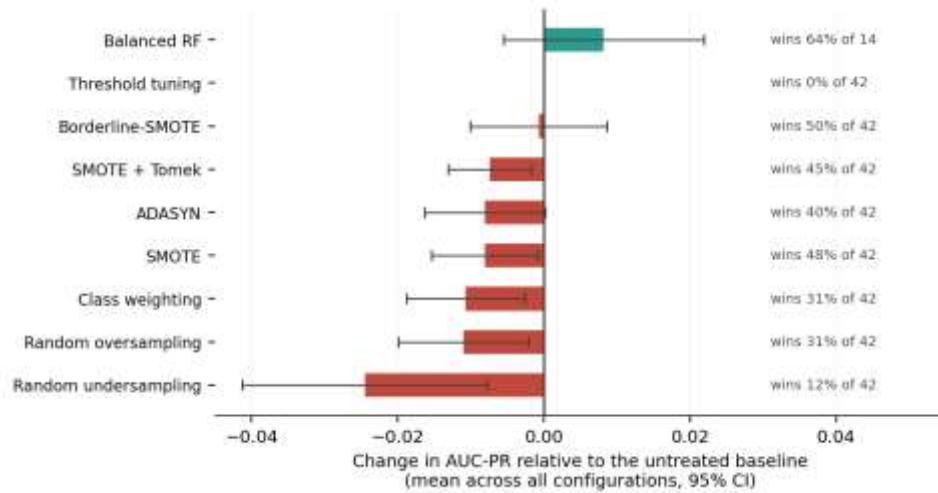


Table 2: Change in AUC-PR Relative to No Correction (Paired, All Configurations)

Strategy	Mean Δ AUC-PR	SD	Win rate	n pairs
Balanced random forest	+0.0082	0.0262	64.3%	14
Borderline-SMOTE	-0.0007	0.0308	50.0%	42
SMOTE + Tomek	-0.0074	0.0189	45.2%	42
SMOTE	-0.0081	0.0240	47.6%	42
ADASYN	-0.0081	0.0273	40.5%	42
Class weighting	-0.0107	0.0268	31.0%	42
Random oversampling	-0.0110	0.0296	31.0%	42
Random undersampling	-0.0245	0.0553	11.9%	42
Threshold tuning	0.0000	0.0000	—	42

Note: Threshold tuning cannot change any ranking metric by construction, since it does not alter the ordering of predicted probabilities; its Δ is exactly zero and it is shown here for completeness.

Figure 4: Change in AUC-PR by Strategy and Imbalance Ratio

Figure 4 Change in AUC-PR relative to baseline by strategy and imbalance ratio, in percentage points. Losses concentrate at severe imbalance, where corrections are most often recommended.

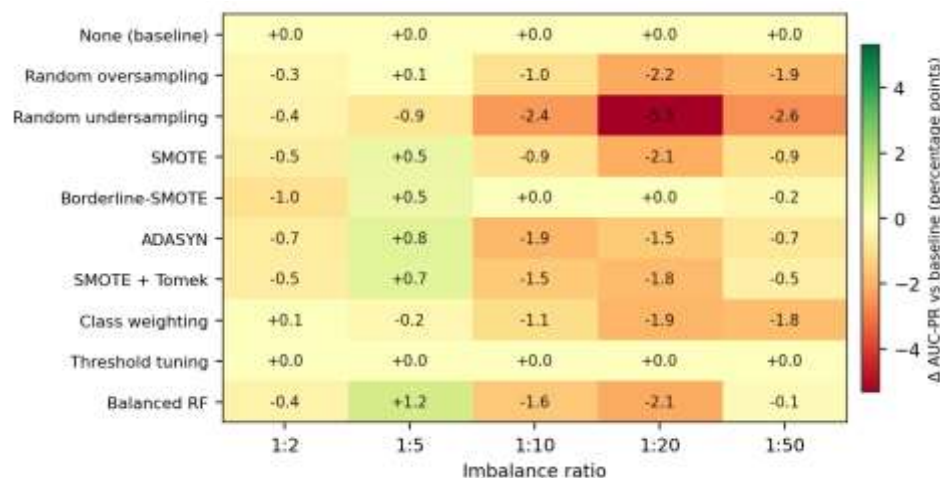


Figure 4 shows the pattern by ratio, and it is the opposite of the usual justification for these methods. At mild imbalance (1:2) all strategies are within 1 percentage point of baseline. Losses grow as imbalance worsens: at 1:20, random undersampling loses 5.3 points of AUC-PR and SMOTE 2.1. The corrections perform worst precisely in the regime for which they are prescribed, because with few minority cases the synthetic points are interpolations among a handful of real examples and encode little genuine information.

Corrections Do Improve Classification — and So Does Moving the Threshold

On classification metrics the picture reverses (Figure 5, Table 3). The uncorrected model is a poor classifier at the default threshold, with mean sensitivity 0.380 against specificity 0.979 — it is close to predicting the majority class throughout. Every correction shifts this. SMOTE raises sensitivity to 0.568 and G-mean from 0.451 to 0.641; random undersampling pushes sensitivity to 0.774 and achieves the best G-mean of any strategy at 0.758.

But threshold tuning achieves much of the same shift while changing nothing about the model. It raises sensitivity from 0.380 to 0.519, F1 from 0.414 to 0.494 and G-mean from 0.451 to 0.601. Compared head-to-head with SMOTE across 42 matched configurations, the F1 difference is 0.494 versus 0.499 — a gap of 0.005 in SMOTE's favour that is not statistically significant (Wilcoxon $p = 0.229$), and one that vanishes entirely at severe imbalance where threshold tuning is marginally ahead (Figure 6) [18-20].

This is the central practical result. The benefit that resampling is credited with — a better sensitivity–specificity balance — is a threshold effect, and it is available without touching the training data. It replicates, on different datasets and with a wider set of methods, the conclusion reached by van den Goorbergh and colleagues [2].

Figure 5: Classification Metrics by Strategy

Figure 5 Classification metrics by strategy, ordered by G-mean. The uncorrected baseline sits at the far left with the highest specificity and lowest sensitivity of any strategy.

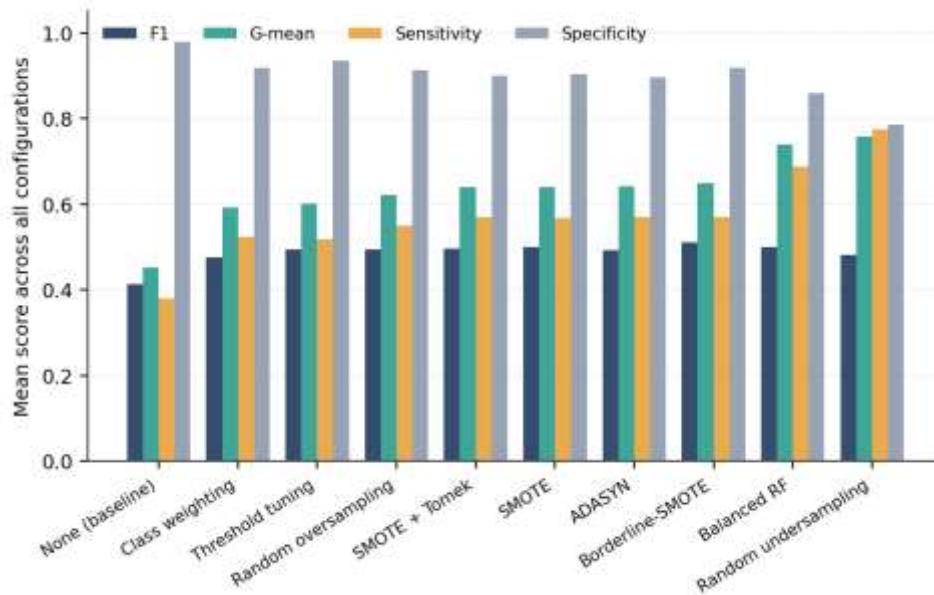


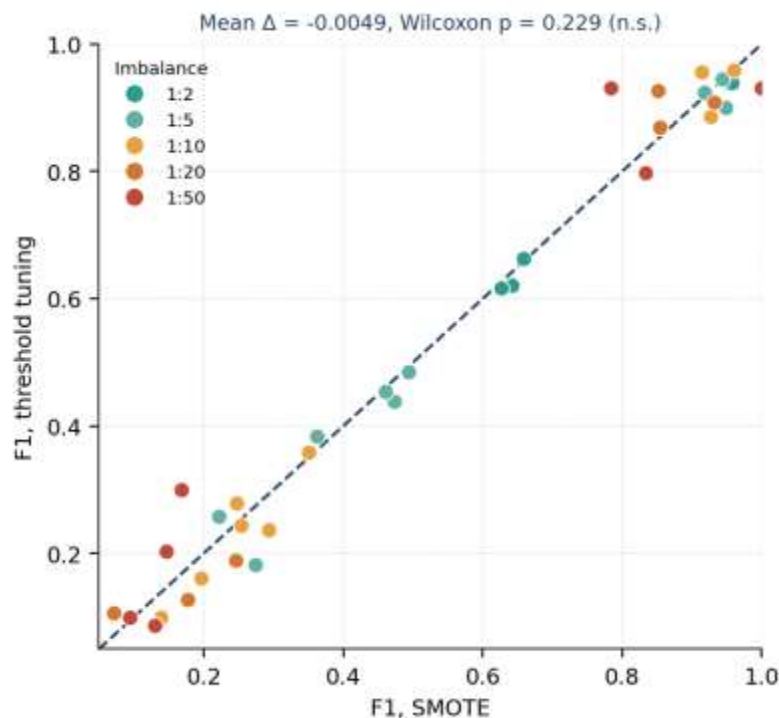
Table 3: Mean Performance by Strategy Across All 392 Configurations

Strategy	AUC-PR	AUC-ROC	F1	G-mean	MCC	Sens.	Spec.	Brier
Threshold tuning	0.5606	0.8283	0.4944	0.6007	0.4450	0.519	0.935	0.0589
None (baseline)	0.5606	0.8283	0.4136	0.4511	0.3931	0.380	0.979	0.0589
Borderline-SMOTE	0.5599	0.8291	0.5108	0.6490	0.4608	0.570	0.918	0.0844
Balanced random forest	0.5542	0.8331	0.5006	0.7389	0.4524	0.687	0.860	0.1106
SMOTE + Tomek	0.5532	0.8275	0.4956	0.6390	0.4410	0.569	0.899	0.0933
ADASYN	0.5525	0.8249	0.4916	0.6419	0.4368	0.570	0.896	0.0948
SMOTE	0.5525	0.8266	0.4993	0.6408	0.4461	0.568	0.903	0.0921
Class weighting	0.5499	0.8217	0.4754	0.5916	0.4307	0.525	0.918	0.0867
Random oversampling	0.5496	0.8221	0.4944	0.6214	0.4434	0.549	0.913	0.0891
Random undersampling	0.5361	0.8288	0.4810	0.7582	0.4379	0.774	0.785	0.1530

Note: Ordered by AUC-PR descending. Threshold tuning and the baseline share identical ranking metrics and Brier scores by construction, differing only in the operating point.

Figure 6: Paired F1, SMOTE Versus Threshold Tuning

Figure 6 Paired comparison of F1 between SMOTE and threshold tuning across 42 matched configurations, coloured by imbalance ratio. Points cluster on the diagonal.



Calibration Damage Is Severe and Systematic

The cost of correction appears in the probabilities (Figure 7, Table 4). Across all configurations the true mean prevalence was 0.121. The uncorrected model predicted a mean probability of 0.117 — an error of 0.4 percentage points, essentially perfect on average. Every correction inflated it: Borderline-SMOTE to 0.196 (+7.6 points), class weighting to 0.197 (+7.7), random oversampling to 0.203 (+8.2), SMOTE to 0.214 (+9.3), ADASYN to 0.220 (+10.0), balanced random forests to 0.292 (+17.1) and random undersampling to 0.343 (+22.2).

In clinical terms, a model corrected by random undersampling tells a clinician that a patient with a true 12% risk has a 34% risk. Brier scores follow directly (Figure 8): against a baseline of 0.0589, SMOTE reaches 0.0921 (56% worse), balanced random forests 0.1106 (88% worse) and random undersampling 0.1530 (160% worse). Threshold tuning alone leaves the Brier score exactly unchanged, because it never touches a probability.

The ordering of the damage is itself informative: the more aggressively a method alters the training prevalence, the worse the inflation. Random undersampling, which discards the most majority data, is the most destructive; Borderline-SMOTE, which generates the fewest synthetic points, is the least. This is consistent with the mechanism described under Why Calibration Matters Clinically and with published simulation work reporting calibration intercepts of -4.5 or worse at 1% event fractions [2].

Figure 7: Mean Predicted Probability Against True Prevalence

Figure 7 Mean predicted probability by strategy against the true mean prevalence (dashed line). Every correction overestimates risk; only the baseline and threshold tuning sit on the true prevalence.

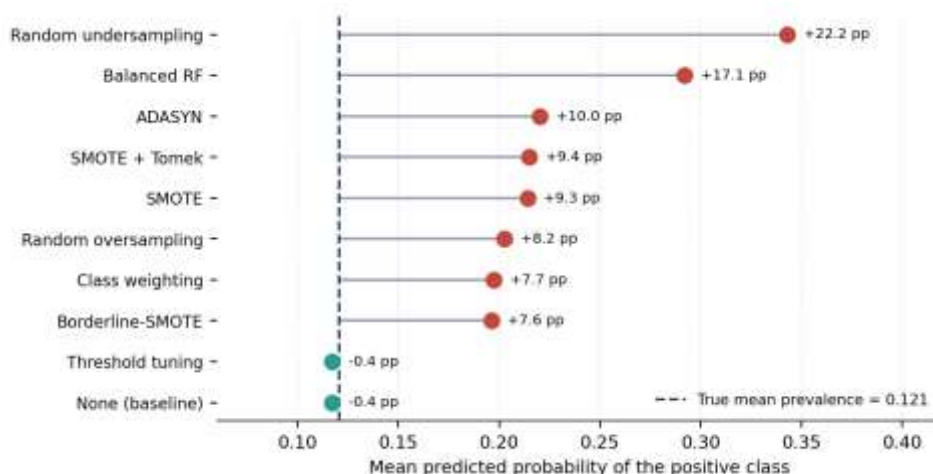


Figure 8: Brier Scores by Strategy

Figure 8 Brier scores by strategy with percentage change from baseline. Lower is better.

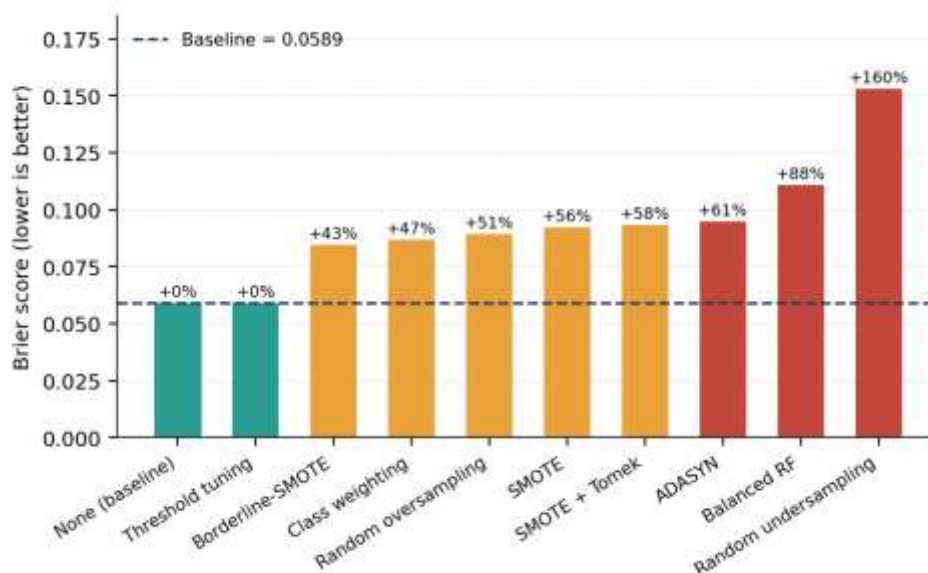


Table 4: Calibration by Strategy (True Mean Prevalence = 0.121)

Strategy	Mean predicted risk	Inflation	Brier	vs baseline	Cal. slope
None (baseline)	0.117	−0.4 pp	0.0589	—	0.775
Threshold tuning	0.117	−0.4 pp	0.0589	0%	0.775
Borderline-SMOTE	0.196	+7.6 pp	0.0844	+43%	0.664
Class weighting	0.197	+7.7 pp	0.0867	+47%	0.712
Random oversampling	0.203	+8.2 pp	0.0891	+51%	0.682

Strategy	Mean predicted risk	Inflation	Brier	vs baseline	Cal. slope
SMOTE	0.214	+9.3 pp	0.0921	+56%	0.682
SMOTE + Tomek	0.215	+9.4 pp	0.0933	+58%	0.678
ADASYN	0.220	+10.0 pp	0.0948	+61%	0.683
Balanced random forest	0.292	+17.1 pp	0.1106	+88%	0.990
Random undersampling	0.343	+22.2 pp	0.1530	+160%	0.826

Note: pp = percentage points. A calibration slope below 1 indicates predictions that are too extreme.

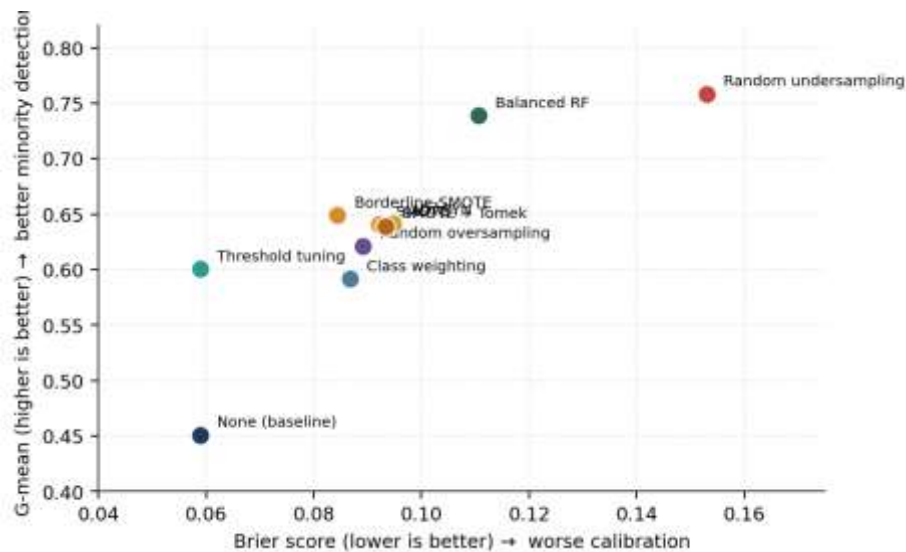
The Trade-Off Surface

Plotting minority detection against calibration quality (Figure 9) makes the choice explicit. Random undersampling occupies the top-right: the best G-mean of any strategy (0.758) and the worst Brier score (0.153). The baseline and threshold tuning occupy the left edge with the best calibration; threshold tuning sits above the baseline at identical calibration, which is the only strictly dominant move available. Balanced random forests offer an intermediate position, with G-mean 0.739 at Brier 0.111.

No strategy achieves both high G-mean and good calibration, and this is not an artefact of these datasets: shifting the operating point toward the minority class and preserving population-level probability estimates are different objectives, and only threshold tuning separates them cleanly by leaving the probabilities intact and moving only the decision rule [21-24].

Figure 9: Trade-Off Between Detection and Calibration

Figure 9 Trade-off between minority-class detection (G-mean) and calibration (Brier). No strategy occupies the desirable upper-left region.



Consistency Across Classifiers

The pattern is not an artefact of a single model family (Figures 10 and 11). Logistic regression achieves the highest mean AUC-PR of the three classifiers on these problems (0.588 uncorrected), ahead of random forest (0.546) and gradient boosting (0.548), consistent with systematic evidence that machine learning offers no general performance benefit over logistic regression for clinical prediction [11]. Within each classifier the

ordering is the same: no correction and threshold tuning tie for best, SMOTE and Borderline-SMOTE follow closely, and random undersampling is consistently worst — 0.561, 0.535 and 0.512 respectively. The magnitude of the loss is largest for gradient boosting, which is also the family most often paired with SMOTE in the applied literature.

Figure 10: Paired AUC-PR, Baseline Versus SMOTE

Figure 10 Paired AUC-PR, uncorrected baseline against SMOTE, across 42 matched configurations. Points below the diagonal are configurations where SMOTE performed worse.

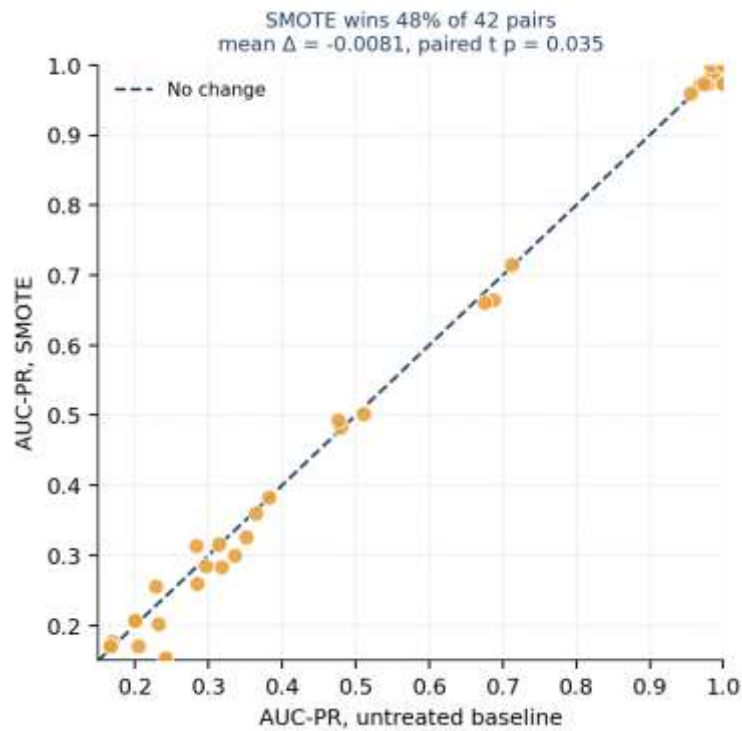
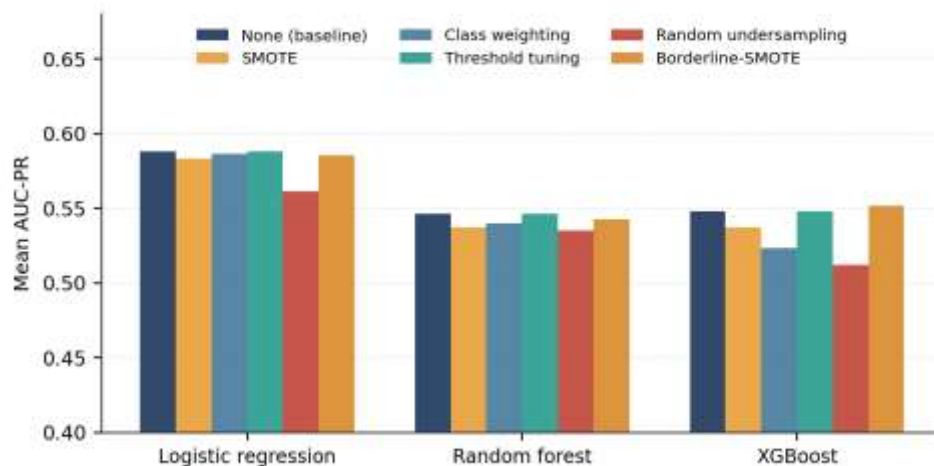


Figure 11: Mean AUC-PR by Classifier and Strategy

Figure 11 Mean AUC-PR by classifier and strategy. The ordering of strategies is stable across all three model families.



Discussion

What the Corrections Actually Do

Across 392 configurations, imbalance correction did three things. It did not improve the model's ability to rank patients by risk — every data-level and algorithm-level method lost ground on AUC-PR, most heavily at the severe imbalance where they are most often prescribed. It did improve classification at a fixed threshold, raising sensitivity substantially. And it systematically inflated predicted risk, by between 7.6 and 22.2 percentage points depending on method.

The first and third of these are costs; only the second is a benefit, and threshold tuning delivers the same benefit at zero cost. That makes the practical recommendation unusually clean: train on the data as it is, then choose an operating point that reflects the clinical costs of false negatives and false positives. This is not a novel recommendation — it is what van den Goorbergh and colleagues concluded from simulation and an ovarian-cancer case study [2], and what a recent evaluation across ten clinical datasets and 605,842 patients found again, reporting AUROC changes of -0.002 to -0.010 alongside Brier increases of 0.029 to 0.080 [4]. The present study reproduces it independently on different data, with a wider set of methods and across a controlled imbalance gradient.

The Metric Problem Underneath

The finding that AUC-ROC is nearly blind to imbalance suggests why the practice persists despite the evidence. If a study reports AUC-ROC only, imbalance appears to be a mild problem, and a correction that leaves AUC-ROC roughly unchanged while improving accuracy, F1 and sensitivity at a fixed threshold will appear to have helped. The damage is invisible unless calibration is measured, and calibration is reported in a minority of clinical machine learning papers [12].

The findings therefore carry a reporting recommendation as much as a methodological one. Any study on imbalanced clinical data should report AUC-PR alongside AUC-ROC, since the two diverge by more than 0.35 at 1:50 imbalance in these experiments; should report a calibration measure whenever the model output

is a probability; and should include the uncorrected model as a comparator so that readers can see what the correction bought, in line with current reporting guidance for clinical prediction models [13].

When Correction May Still Be Reasonable

The results do not support a blanket prohibition. Balanced random forests were the only method to improve mean AUC-PR (+0.008) and reached a G-mean of 0.739, which suggests that resampling inside an ensemble where each learner sees a different balanced bootstrap and predictions are averaged — is less damaging than resampling the training set once. Random undersampling, despite the worst calibration of any method, achieved the highest G-mean (0.758) and might be defensible for a pure screening triage where the output is a binary flag, no probability is shown to anyone, and missing a case is far costlier than a false alarm.

The condition in both cases is that the model's output is a decision rather than a risk estimate. Once a number is shown to a clinician or enters a downstream calculation, the calibration damage documented in the Results becomes a direct clinical harm. And if correction is used regardless, an explicit recalibration step on uncorrected data should follow — a step that is available, well described, and almost never performed.

Conclusion

Across three clinical datasets, five induced imbalance ratios, ten strategies and three classifiers — 392 configurations in total — no data-level or algorithm-level imbalance correction improved the ranking of patients by risk. SMOTE lost 0.008 AUC-PR on average and won fewer than half its paired comparisons; random undersampling lost 0.025 and won 11.9%. Corrections did improve classification at a fixed threshold, but simple threshold tuning matched SMOTE's F1 (0.494 vs 0.499, $p = 0.229$) without altering a single predicted probability. The cost of correction was systematic risk inflation of 7.6 to 22.2 percentage points and Brier scores 43% to 160% worse than baseline. A model told to expect a balanced world reports a balanced world's probabilities, and clinicians act on those numbers. For clinical prediction models, class imbalance is best handled at the decision threshold, not in the training data — and the near-automatic application of SMOTE in this literature should be treated as a methodological choice requiring justification rather than a default.

References

- [1] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
- [2] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.
- [3] van den Goorbergh, R., van Smeden, M., Timmerman, D., & Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, 29(9), 1525–1534.
- [4] Karim, N. B., Pentyala, R. K., & Simko, R. (2024b). Adaptive Machine-Learning Detection and Prevention of Financial-Transactions Fraud at Community Financial Institutions: Recalibrating Risk Signals Under Shifting Fraud Typologies and Record Suspicious-Activity Monitoring Burdens. *International Journal of Acta Informatica*, 3(1), 259-278.

- [5] Rahman, M. S., Islam, C. S., & Tirupathi, D. (2024). Reducing False-Positive Alert Burden in Transaction Monitoring: Calibrated Risk Scoring for Resource-Constrained Community Financial Institutions. *International Journal of Modern Computing*, 7(1), 266-282.
- [6] Rashid, S. U., & Cheng, X. (2024). Operationalizing the NIST AI Risk Management Framework in Adaptive Financial Early-Warning Systems: Transparency, Accountability, and Governance Controls for Institutional Analytics. *The Metascience*, 2(4), 142-154.
- [7] Potharaju, B., & Sajja, R. R. (2024). Modernizing US Federal Regulatory Reporting Infrastructure for Scalability, Resilience, and Audit Readiness. *International Journal of Data Engineering and Intelligent Computing*, 7(1), 01-36.
- [8] Piccininni, M., Wechsung, M., Van Calster, B., Rohmann, J. L., Kurth, T., & van Smeden, M. (2024). Understanding random resampling techniques for class imbalance correction and their consequences on calibration and discrimination of clinical risk prediction models. *Journal of Biomedical Informatics*, Article 104666.
- [9] Rahman, M. M., Computer, M. M. R., Das, N., Uddin, M. S., Shammah, R. S., & Rahman, M. M. (2025, May). An Investigation of Face Detection and Feature Extraction Techniques in Deepfake Detection Using a Multi-Model Machine Learning Approach. In *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)* (pp. 1-8). IEEE.
- [10] Andersen, A. K., Mehdizavareh, H., Khan, A., Becher, T., Britsch, S., Britsch, M., Böttcher, M., Winther, S., Rohde, P. D., Jensen, M. H., & Cichosz, S. L. (2026). *Tipping the balance: Impact of class imbalance correction on the performance of clinical risk prediction models* (arXiv:2603.00208). arXiv.
- [11] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17, Article 230.
- [12] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 21(1), 128–138.
- [13] Ahmed, K. R., Ahmmed, M. R., Kowser, S., Faruq, O., Shammah, R. S., & Sufian, M. A. (2025, July). The Future of AI-Powered SaaS: Market Trends and Investment Opportunities. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.
- [14] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- [15] Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110, 12–22.
- [16] Andaur Navarro, C. L., Damen, J. A. A., van Smeden, M., Takada, T., Nijman, S. W. J., Dhiman, P., ... Moons, K. G. M. (2022). Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *Journal of Clinical Epidemiology*, 154, 8–22.

- [17] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, Article e078378.
- [18] Karim, N. B., Vrettos, K., & Seitz, J. (2024a). Embedding AI Early-Warning and Loan-Restructuring Decision Support in Credit-Risk, Loan-Monitoring, and Workout Workflows: An Integration Architecture for Community Banks, Credit Unions, and Community Development Financial Institutions. *The Computertech*, 25-40.
- [19] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), Article e0118432.
- [20] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, Article 6.
- [21] Rahman, M. S., & Kim, E. Y. (2024). Embedding Adaptive Machine-Learning Decision Support in Suspicious Activity Report Investigation Workflows: An Enterprise Information-Systems Framework. *International Journal of Acta Informatica*, 3(1), 245-258.
- [22] Rashid, S. U., Junck, L., & Albous, M. R. (2024). Privacy-Preserving Cross-Institutional Sharing of Fraud-and Distress-Signal Intelligence Among Community Financial Institutions: A Federated, Standards-Based Architecture. *International Journal of Modern Computing*, 7(1), 283-297.
- [23] Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17), 1–5.
- [24] Potharaju, B., & Ravipudi, J. (2023). High-Availability and Disaster-Recovery Architecture for Containerized Banking Workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 01-32.