
Target-set and all-feature missingness stress tests for two tabular clinical benchmarks

Research report | Clinical missing-data robustness benchmark

Sohan Mohanty

Abstract

Comparisons of simulated missing-data mechanisms are difficult to interpret when each mechanism can erase a different portion of the feature matrix. I studied logistic regression, random forest, and histogram-based gradient boosting on the Wisconsin Diagnostic Breast Cancer (WDBC; 569 observations, 30 features) and Statlog Heart (270 observations, 13 features) datasets. In the primary analysis, I restricted missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) to the same prespecified target features. I evaluated a separate MCAR-all condition as an all-feature stress test. I used two repeats of stratified five-fold cross-validation, three-fold inner tuning, four requested missingness rates, and five mask seeds for each outer split. I averaged mask seeds within split and estimated descriptive 95% intervals by bootstrapping the 10 outer repeat-fold splits. At a requested 50% target-feature rate, the achieved whole-matrix missingness was about 8.3% for WDBC and 11.5% for Statlog Heart. Mean ROC-AUC changes in this scope-matched analysis ranged from -0.0033 to -0.0016 for WDBC and from -0.0057 to +0.0005 for Statlog Heart. Differences among mechanisms were small compared with the all-feature stress test. At approximately 50% whole-matrix MCAR, ROC-AUC changes ranged from -0.0275 to -0.0115 for WDBC and from -0.1015 to -0.0847 for Statlog Heart. The corresponding Brier-score increases ranged from +0.0412 to +0.1024 and from +0.0510 to +0.0659. In exploratory comparisons of imputers, corrupted training, and missingness indicators, I found no strategy that was consistently superior. In these experiments, the scope of missingness had a larger effect than the nominal mechanism label. I recommend reporting both target-set and whole-matrix rates and examining probability error alongside discrimination. My descriptive benchmark does not establish clinical utility, deployment safety, or the causal structure of real-world missingness.

Keywords: missing data; calibration; robustness; clinical machine learning; benchmark; MCAR; MAR; MNAR

Research status and review

I based this report on the paper_* artifacts in the repository, with all numeric results reconciled against the saved prediction probabilities before drafting. My analysis is descriptive and exploratory. I did not conduct a confirmatory hypothesis test, power calculation, or clinical validation.

I am the sole author. I have given final approval of the paper and references for sharing on GitHub. I record reporting coverage, source-by-source verification, and computational reproducibility separately; approval does not certify those checks. I am not preparing a journal submission.

1. Introduction

Clinical prediction models may encounter tests that were not performed, measurements that are delayed, or records that are only partly observed. A model developed and evaluated on complete inputs can therefore be used under conditions that differ from its training data. The consequences are not captured by a single robustness measure: missing inputs may affect classification, ranking, probability error, and calibration differently.

MCAR, MAR, and MNAR describe relationships between missingness and data [1]. Those labels do not specify which variables are eligible for masking or how much of the full feature matrix is removed. If one condition can erase every feature while another acts only on a small target set, a difference caused by feature scope may be mistaken for a difference between mechanisms. I therefore treat the scope-matched mechanism comparison and the all-feature stress test as separate analyses.

ROC-AUC measures ranking, not the accuracy of predicted probabilities. Brier score is a squared probability-error measure [2]. Reliability diagrams and expected calibration error (ECE) summarize agreement between predicted and observed event rates [3]. Calibration intercept and slope can also be useful, but in small evaluation folds their estimates may not exist because predicted logits perfectly separate the outcomes. For that reason, I use Brier score and ECE for the main probability-quality results. I retain calibration regression as an exploratory diagnostic only when it yields a finite estimate.

I asked three questions:

Question 1. What happens to discrimination and probability error when MCAR, MAR, and MNAR are applied to the same target features?

Question 2. How do those results compare with MCAR applied across the full feature matrix?

Question 3. Do alternative imputers, training on corrupted inputs, or explicit missingness indicators consistently improve robustness in exploratory extensions?

I designed this project as a reproducible benchmark and an example of precise reporting for simulated missingness. I do not present it as a diagnostic model, patient-level risk tool, or evaluation of clinical effectiveness.

2. Methods

2.1 Study design

I ran a controlled, descriptive benchmark on two complete public tabular datasets. Checked-in YAML files record the analysis plan used for the reported runs. The core experiment crossed two datasets, three model families, four missingness constructions, four requested rates, five mask seeds, and 10 outer repeat-fold splits. In three smaller experiment grids, I evaluated imputers, clean or corrupted training, and missingness indicators. I excluded no observations after loading, and neither raw feature matrix contained native missing cells.

The dataset sizes were fixed by the selected benchmarks, and I performed no formal sample-size or power calculation. My intervals describe variation across the prespecified outer splits. They are not population-level confidence intervals and should not be read as evidence of clinical generalizability.

2.2 Datasets and outcomes

WDBC contains 569 observations and 30 continuous features derived from digitized breast-mass images. Its binary outcome has 357 observations with target value 1 and 212 with target value 0 [5]. Statlog Heart contains 270 observations, 13 mixed numeric, binary, and categorical features, and a binary heart-disease outcome. Of these observations, 120 are labeled present and 150 absent [6]. My loaders encode both outcomes as integers 0 and 1. In the cached scikit-learn WDBC data, target 1 is the benign class. I study classification behavior and do not interpret either encoded value as a clinical decision threshold.

Dataset	Rows	Features	Target 1, n (%)	Target 0, n (%)	Native missing cells
WDBC	569	30	357 (62.7%)	212 (37.3%)	0
Statlog Heart	270	13	120 (44.4%)	150 (55.6%)	0

2.3 Prespecified feature scope

I gave MCAR, MAR, and MNAR the same eligible target features in the primary comparison. Five of 30 WDBC features were eligible: mean texture, mean perimeter, mean area, worst texture, and worst perimeter. Three of 13 Statlog Heart features were eligible: resting blood pressure, serum cholesterol, and maximum heart rate achieved. I declared these feature sets in configs/datasets.yaml.

A requested rate is the expected fraction erased within the eligible target set. For every run, I saved both the achieved target-set rate and the achieved rate across the complete feature matrix. Since the target sets contain only 5/30 of WDBC features and 3/13 of Statlog Heart features, 50% target-feature missingness is not equivalent to 50% whole-matrix missingness.

In the separate `mcAR_all` condition, every feature was sampled independently at the requested probability. I treated this as an all-feature stress test and did not pool it with the scope-matched mechanism comparison.

2.4 Missingness constructions

Scope-matched MCAR erased each eligible cell independently with the requested probability. Under MAR, the masking probability for a target feature followed a logistic function of a standardized observed anchor: mean radius for WDBC and age for Statlog Heart. Under MNAR, that probability followed a logistic function of the standardized value being erased. MAR and MNAR used a log-odds coefficient of 2.0. A numerical search found an intercept that set the mean masking probability to the requested rate within the current training or test partition. Random sampling produced small differences between requested and achieved rates.

I use these rules to define controlled simulations, not to estimate missingness in either source dataset. In particular, the MNAR rule is a value-dependent stress condition and does not identify the data-generating mechanism of WDBC or Statlog Heart.

2.5 Models, preprocessing, and tuning

I evaluated logistic regression, random forest, and scikit-learn's `HistGradientBoostingClassifier` [4] in the core analysis. I used median imputation for numeric features and most-frequent imputation for binary and categorical features. The logistic-regression pipeline also scaled its numeric features. I fitted imputation and scaling within each training pipeline to prevent information from the test fold from entering preprocessing.

I selected hyperparameters by three-fold inner stratified cross-validation using ROC-AUC. The grids in `configs/models.yaml` included logistic-regression C values of 0.01, 0.1, 1, and 10; random forests with 200 or 500 trees, unrestricted depth or depth 4 or 8, minimum leaf sizes of 1, 3, or 5, and square-root or log2 feature sampling; and histogram-gradient-boosting models with learning rates of 0.03, 0.05, or 0.1, unrestricted depth or depth 3 or 5, and 100 or 200 iterations.

For the primary regime, I tuned and fitted each model on the complete outer-training fold, then evaluated the selected pipeline on complete and masked versions of the outer-test fold. Each corrupted result therefore had a clean reference from the same split.

2.6 Cross-validation and randomization

Two repeats of stratified five-fold outer cross-validation, using split seed 42, produced 10 repeat-fold splits for each dataset. The primary analysis used five independent mask seeds per split, and the extension grids used three. Mask generation did not depend on model or imputer choice. As a result, paired comparisons used identical masked test inputs for the same dataset, mechanism, requested rate, seed, repeat, and fold.

I used the outer repeat-fold split as the unit of descriptive interval estimation and first averaged mask seeds within split. I then calculated the mean of the 10 split estimates and obtained a two-sided descriptive 95% interval from the 2.5th and 97.5th percentiles of 4,000 bootstrap resamples of those estimates. Neither repeated folds nor mask seeds are independent patient samples. The intervals quantify resampling variation under this benchmark design.

2.7 Outcomes

Accuracy and ROC-AUC summarized classification and discrimination. Brier score and 10-bin fixed-width ECE summarized probability error and calibration discrepancy. Higher accuracy and ROC-AUC are better; lower Brier score and ECE are better. Results for corrupted conditions were reported as absolute values and as paired changes from the corresponding clean outer-split reference.

I estimated calibration intercept and slope with unpenalized logistic calibration regression when a finite maximum-likelihood estimate existed. When evaluation logits perfectly separated the outcomes, I recorded the calibration estimates as not estimable. After this correction, 2,590 of 4,800 run-level calibration regressions in the core grid were not estimable. My main conclusions therefore do not rely on calibration intercept or slope.

Reliability diagrams used 10 fixed-width probability bins. They contain repeated outer-cross-validation predictions and, for corrupted conditions, multiple mask seeds. Counts refer to prediction records, not unique patients. Descriptive uncertainty was calculated by bootstrapping outer repeat-fold clusters.

2.8 Exploratory extensions

In the imputer experiment, I fixed the model to logistic regression and compared median or most-frequent simple imputation, uniform-weight KNN imputation, and iterative imputation. The KNN search included 3, 5, and 7 neighbors. Because the training and inner-validation inputs were complete, the neighbor settings tied; the first candidate, 3 neighbors, was selected in all 20 clean outer-split KNN fits. I trained models on complete inputs and tested them on masked inputs. Numeric imputation preceded scaling, so KNN distances used the original feature units.

In the training-regime experiment, I also used logistic regression, with iterative imputation. I compared complete training followed by masked testing with training and testing under the same mechanism and requested rate. In the corrupted-training condition, I refitted imputation and hyperparameter tuning within each masked training fold.

In the indicator experiment, I compared simple imputation with and without explicit missingness indicators under corrupted training and testing. All extensions used the scope-matched MCAR, MAR, and MNAR target sets. I treated them as exploratory, reported them regardless of direction, and made no adjustment for multiple comparisons.

2.9 Reproducibility and integrity checks

I used the commands in `scripts/run_paper.ps1` to produce run-level metrics, predictions, clean references, masks, figures, and manifests. I pinned the Python interpreter and analysis dependencies in `requirements.txt` and `environment.yml`; the run manifests record the environment used for each experiment.

An independent reconciliation script recomputed accuracy, ROC-AUC, Brier score, and ECE from all 825,576 saved corrupted prediction records. It found no duplicate run keys or prediction rows. Expected and observed prediction counts agreed, and the largest absolute metric discrepancy was approximately 3.1×10^{-16} . Correcting the calibration regression changed only intercept and slope fields; predictions and the four principal metrics did not change. Run, reconciliation, and analysis manifests contain SHA-256 hashes for their inputs and outputs.

3. Results

3.1 Clean-data reference performance

All three models discriminated well on complete WDBC inputs. Mean ROC-AUC was 0.9950 for logistic regression, 0.9908 for random forest, and 0.9918 for gradient boosting. Their Brier scores were 0.0216, 0.0327, and 0.0316. Statlog Heart was less separable: mean ROC-AUC was 0.9038, 0.9050, and 0.8874, and Brier score was 0.1300, 0.1302, and 0.1368. These values set dataset-specific clean references. They do not rank the models for clinical use.

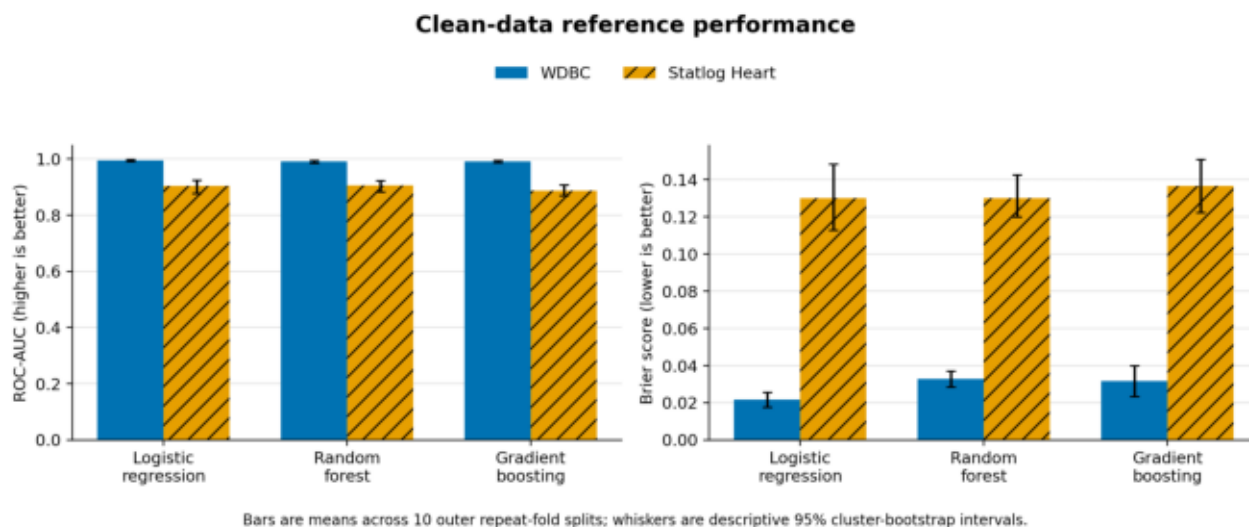


Figure 1. Clean-data reference performance. Bars show means over 10 outer repeat-fold splits; whiskers are descriptive 95% outer-split cluster-bootstrap intervals.

3.2 Requested and whole-matrix rates

Within the target features, the masking algorithms closely matched the requested rates. At the requested 50% rate, the mean achieved whole-matrix rate was approximately 0.083 for WDBC and 0.115 for Statlog Heart under the scope-matched mechanisms. These values follow from targeting 5/30 and 3/13 of the respective feature matrices. MCAR-all, in contrast, reached approximately 0.500 of the complete matrix.

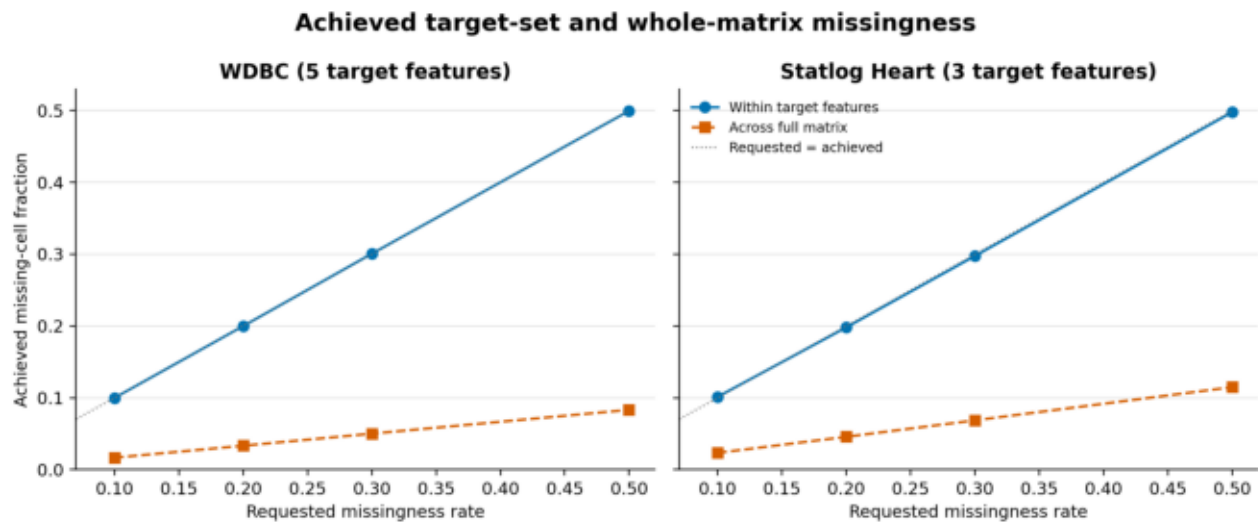


Figure 2. Achieved missingness within target features and across the complete feature matrix. Scope-matched MCAR, MAR, and MNAR use the same target sets.

3.3 Scope-matched mechanism comparison

Discrimination changed little at the requested 50% target-set rate. Across the nine model and mechanism combinations, mean ROC-AUC change ranged from -0.0033 to -0.0016 on WDBC. The Statlog Heart range was -0.0057 to +0.0005, and most descriptive intervals included zero. Once feature scope was matched, the data did not support a general claim that MCAR was uniformly the harshest mechanism.

The probability-error results differed by dataset. Brier score increased in all nine 50% WDBC conditions, from +0.0070 for logistic regression under MCAR to +0.0157 for random forest under MNAR. Mean Statlog Heart changes were smaller, ranging from -0.0013 to +0.0023. Targeted removal of the selected WDBC variables therefore changed probability quality despite little loss of discrimination. The selected Statlog Heart target set had little average effect under this protocol.

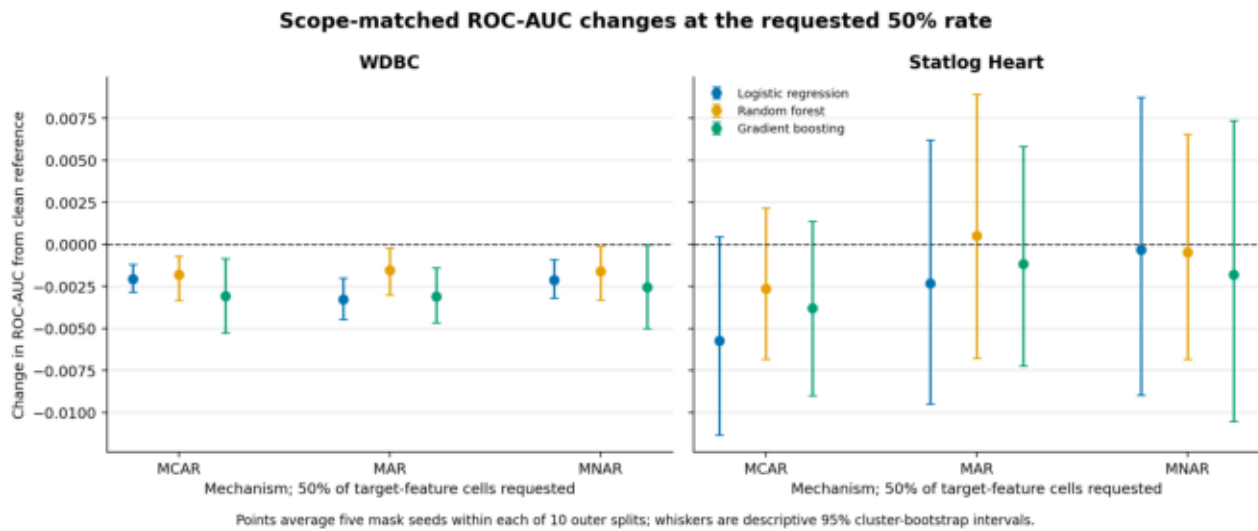


Figure 3. ROC-AUC change at a requested 50% target-feature rate. Mask seeds were averaged within each outer split; whiskers are descriptive 95% outer-split cluster-bootstrap intervals.

3.4 All-feature MCAR stress test

Applying MCAR to every feature produced larger changes. At approximately 50% whole-matrix missingness, mean WDBC ROC-AUC changed by -0.0124 for logistic regression, -0.0115 for random forest, and -0.0275 for gradient boosting. Brier score increased by +0.0412, +0.0721, and +0.1024.

Statlog Heart showed greater discrimination loss under the same scope. ROC-AUC changed by -0.0847 for logistic regression, -0.0927 for random forest, and -0.1015 for gradient boosting; Brier score increased by +0.0510, +0.0577, and +0.0659. The descriptive interval excluded zero for each of these six 50% contrasts. Feature scope separated the mild target-set perturbation from this severe information-loss condition more clearly than mechanism labels separated the scope-matched conditions.

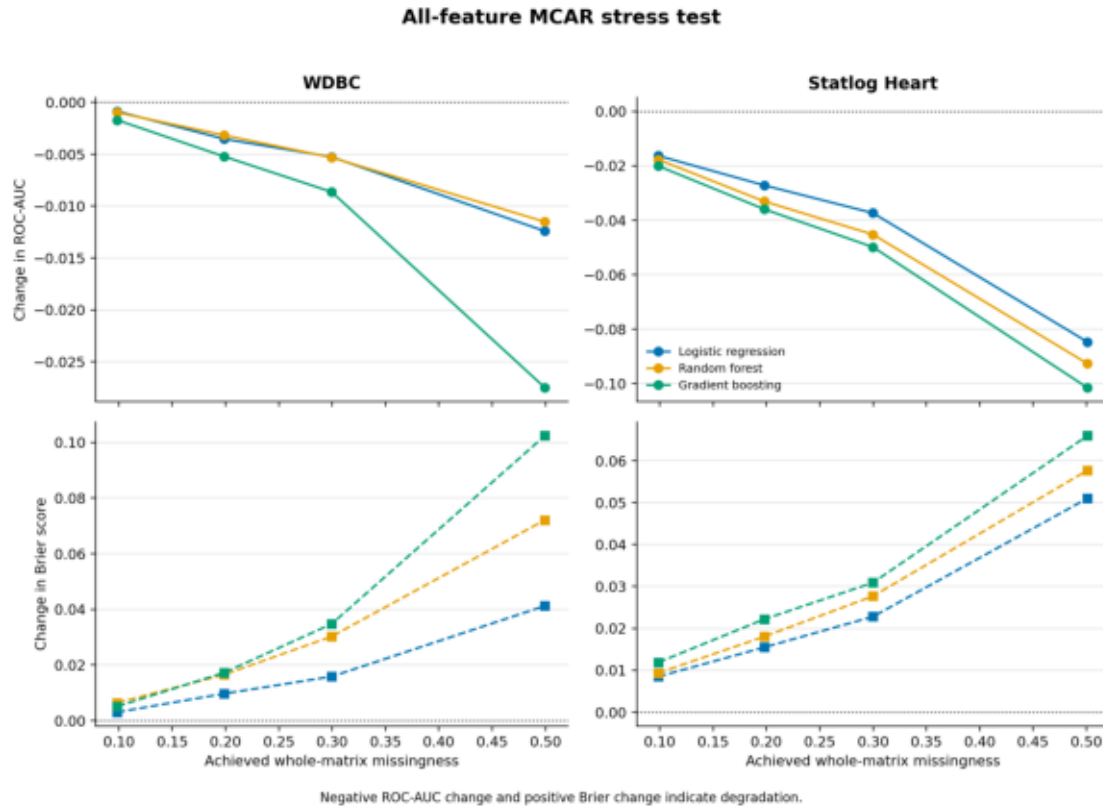


Figure 4. ROC-AUC and Brier-score changes across achieved whole-matrix MCAR rates. Negative ROC-AUC change and positive Brier change indicate degradation.

The joint ROC-AUC and Brier display also shows that probability error can rise while ranking changes remain modest, especially on WDBC. This pattern appeared across model families in the all-feature condition and supports reporting a proper scoring rule with ROC-AUC.

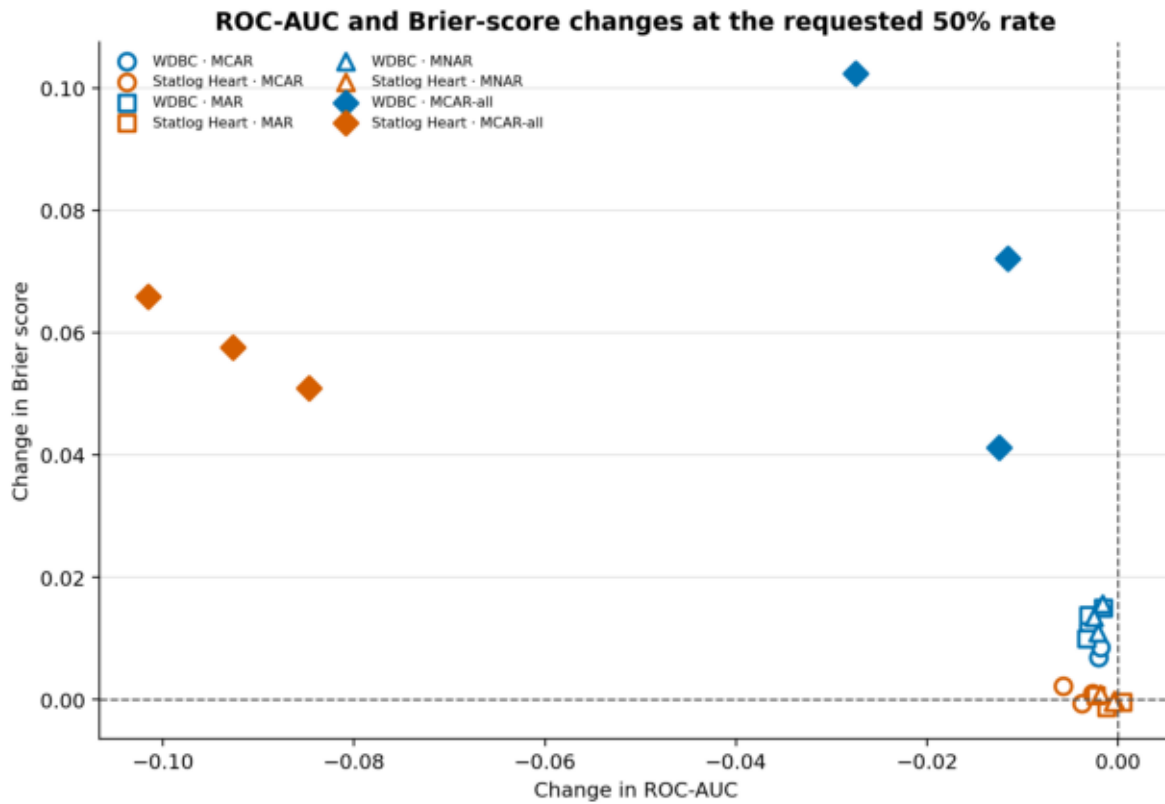


Figure 5. Joint changes in discrimination and probability error at the requested 50% rate. Open markers denote target-set mechanisms; filled diamonds denote MCAR-all.

Reliability diagrams give another view of the probability changes. For WDBC logistic regression, the MCAR-all curve fell below the ideal line across much of the middle probability range, consistent with the Brier-score increase. Statlog Heart had wider split-level uncertainty and irregular clean-data reliability, which is consistent with its smaller sample and lower separability. [results/tables/paper/reliability_bins.csv](#) contains the bin-level prediction-record counts and intervals.

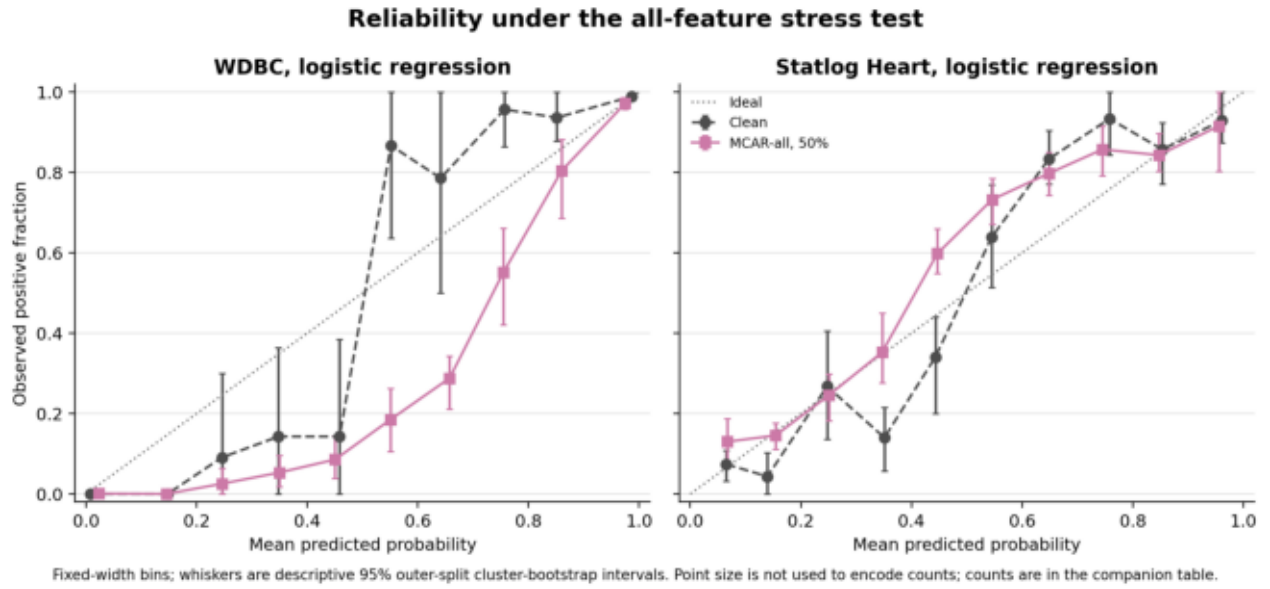


Figure 6. Logistic-regression reliability under clean inputs and approximately 50% whole-matrix MCAR. Fixed-width bin counts are reported in the companion table; whiskers are descriptive outer-split cluster-bootstrap intervals.

3.5 Imputer comparison

I found no imputer that was superior on both datasets. At the 50% target-set rate on WDBC, iterative imputation improved ROC-AUC by +0.0011 to +0.0018 relative to simple imputation across mechanisms and reduced Brier score by 0.0038 to 0.0058. Descriptive intervals for those paired contrasts excluded zero. KNN imputation also reduced WDBC Brier score, although its ROC-AUC intervals included zero. On Statlog Heart, every 50% KNN-versus-simple and iterative-versus-simple interval for ROC-AUC and Brier score included zero. I interpret this as a WDBC-specific benefit, not evidence that iterative imputation is generally best.

3.6 Corrupted-training regime

Training on corrupted inputs offered no consistent advantage. At the 50% target-set rate, paired ROC-AUC contrasts between corrupted and complete training ranged from approximately -0.0016 to +0.0066 across datasets and mechanisms. Every descriptive interval included zero, as did the Brier-score intervals. This null and heterogeneous result does not support a general claim that training exposure to simulated missingness improves robustness.

3.7 Missingness indicators

Indicator effects changed by dataset and mechanism. At 50% on Statlog Heart, indicators improved MAR ROC-AUC by +0.0056 (descriptive 95% interval +0.0016 to +0.0100) but reduced MCAR ROC-AUC by -0.0050 (-0.0112 to -0.0009). On WDBC, indicators increased MCAR Brier score by +0.0016 (+0.0007 to +0.0027) and reduced MNAR Brier score by -0.0024 (-0.0040 to -0.0005). The direction changes do not support describing indicators as uniformly protective.

4. Discussion

4.1 Main findings

Comparisons between missingness mechanisms depend on both the eligible features and the denominator used for the reported rate. Once MCAR was restricted to the same target sets as MAR and MNAR, the mechanism-specific differences were small. The larger degradation under MCAR-all came from erasing every feature rather than from randomness alone.

The all-feature stress test also affected the datasets differently. At approximately 50% MCAR across the full matrix, Statlog Heart lost about 0.085 to 0.102 ROC-AUC, compared with about 0.012 to 0.028 for WDBC. Redundancy, class structure, feature informativeness, and clean separability could contribute to this difference, but the benchmark was not designed to distinguish among those explanations.

I found that discrimination and probability error did not always move together. In the scope-matched WDBC conditions, ROC-AUC changed little while Brier score consistently deteriorated. A model may therefore preserve its ordering of cases while producing less accurate probabilities. That distinction matters when model outputs are presented as risk estimates, although I validated neither a decision threshold nor a clinical consequence.

My exploratory extensions did not yield a common remedy. Iterative imputation had a modest WDBC-specific advantage, corrupted training did not consistently help, and indicator effects reversed direction across mechanisms and datasets. I retain the null and adverse results to avoid overstating any single preprocessing choice.

4.2 Experimental and reporting implications

A simulated-missingness study should state the eligible feature set, the requested conditional rate, the achieved rate within that set, and the achieved rate across the full matrix. A mechanism label without those quantities is incomplete.

All-feature stress testing can still be informative, provided it is named and interpreted separately from a target-feature comparison.

Multiple mask seeds improve Monte Carlo stability, but they are not independent datasets. I average seeds within each outer split before calculating descriptive uncertainty. Repeated cross-validation predictions likewise represent repeated observations of the same finite sample, so the reliability-bin counts are not unique-patient counts.

Small evaluation folds require care when fitting calibration regressions. Under separation, very large finite coefficients create false precision. Recording the estimates as not estimable, reporting how often that occurs, and relying on Brier score, ECE, and explicit reliability bins is more defensible for these data.

4.3 Limitations

I studied only two small, complete public benchmarks. Neither dataset represents a prospective cohort, hospital workflow, temporal drift, treatment selection, or deployed clinical system. I use the terms clinical and medical to describe the source domains, not validated use.

The MAR and MNAR mechanisms are stylized logistic rules with one coefficient value and prespecified variables. They do not show that either process generated missingness in the source populations. Because both original datasets are complete, the analysis cannot validate imputation against a native missing-data process.

I selected target features for a controlled benchmark rather than a study of clinical measurement practice. Different target sets, anchor variables, coefficients, requested rates, outcome encodings, model grids, or calibration-bin definitions could change the results.

I did not select neighbor counts using missing-input validation in the clean-training imputer comparison. KNN also used unscaled numeric distances, which can favor variables with larger units. These results therefore describe my pipeline and selected setting, not an optimized KNN imputation strategy.

My uncertainty intervals resample only 10 outer splits and are descriptive. They do not capture every dependency introduced by repeated cross-validation, dataset selection, hyperparameter search, or multiple exploratory comparisons. I performed no external validation, decision-curve analysis, subgroup analysis, fairness analysis, threshold analysis, or clinical-utility evaluation.

Independent replication has not yet been completed. My approval of the paper and references does not establish completed source-by-source verification or confirm any unrecorded declaration.

5. Conclusion

In these two benchmarks, feature scope mattered more than the differences among scope-matched MCAR, MAR, and MNAR simulations. A requested 50% target-feature rate removed only about 8% to 12% of the full matrices and produced small discrimination changes. MCAR applied to approximately 50% of the whole matrix caused substantially larger deterioration in ROC-AUC and Brier score. Alternative imputers, corrupted training, and missingness indicators did not provide a universal solution.

I recommend distinguishing target-set rates from whole-matrix rates, treating mask seeds as Monte Carlo repetitions, and pairing discrimination with a measure of probability error. My findings inform benchmark design and reporting. They do not establish clinical safety or effectiveness.

6. Data, code, and artifact availability

The source repository is

<https://github.com/sohanmohanty/what-medical-ai-doesnt-know>. Experiment configurations are in `configs/paper_*/`, and `scripts/run_paper.ps1` runs the combined workflow. The `results/` directory contains run-level metrics, summaries, predictions, masks, and manifests. Report tables are in `results/tables/paper/`, and report figures are in `figures/paper/`.

`results/manifests/paper_reconciliation.json` records the prediction-to-metric reconciliation. `results/manifests/paper_analysis_manifest.json` records hashes of analysis inputs and outputs.

I release my original code, website, and documentation, including this report, under the MIT License in the repository's LICENSE file. I retain the separate license terms for the public benchmark datasets and third-party software.

7. Authorship and reporting records

I am the sole author. I record authorship in `report/research_records/authorship.json`, reproducibility status and commands in `report/research_records/reproducibility.json`, and project-specific reporting coverage in `report/research_records/reporting_coverage.json`. The coverage record is an audit, not certification against a named journal checklist.

References

- [1] Little RJA, Rubin DB. Statistical Analysis with Missing Data. 3rd ed. Wiley; 2019.
<https://www.wiley.com/en-us/Statistical+Analysis+with+Missing+Data%2C+3rd+Edition-p-9780470526798>
- [2] Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950;78(1):1-3.
[https://doi.org/10.1175/1520-0493\(1950\)078%3C0001:VOFEIT%3E2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2)
- [3] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning. 2017;70:1321-1330. <https://proceedings.mlr.press/v70/guo17a.html>
- [4] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;12:2825-2830. <https://www.jmlr.org/papers/v12/pedregosa11a.html>
- [5] UCI Machine Learning Repository. Breast Cancer Wisconsin (Diagnostic). <https://doi.org/10.24432/C5DW2B>
- [6] UCI Machine Learning Repository. Statlog (Heart). <https://doi.org/10.24432/C57303>