المجلة الليبية للدراسات الأكاديمية المعاصرة (Libyan Journal of Contemporary Academic Studies)
LJCAS, eISSN: 5970-3005
Volume 4, Issue 2, 2026, Pages: 77-94

# When Preprocessing Changes the Winner: Sensitivity of Medical Prediction Model Rankings to Missing Data Handling

Albahlool Abood *
Department of Software Engineering, Faculty of Information Technology, University of Gharyan, Gharyan, Libya
*Corresponding author: albahlool.Abood@gu.edu.ly

# عندما تغيّر المعالجة المسبقة النموذج الفائز: حساسية ترتيب نماذج التنبؤ الطبي لأساليب معالجة البيانات المفقودة

البهلول عبود *
قسم هندسة البرمجيات، كلية تقنية المعلومات، جامعة غريان، غريان، ليبيا

**Received:** 27-06-2026; **Accepted:** 28-08-2026; **Published:** 13-09-2026

**Abstract:**
Missing data are common in clinical prediction studies, yet their handling may affect not only predictive performance but also which model is judged best. This study examined the stability of classifier selection when missing data handling was changed under controlled, paired evaluation conditions. Using SUPPORT2 data from 9,105 patients, Logistic Regression, Random Forest, and XGBoost were evaluated with median, K nearest neighbor, and iterative imputation. The analysis retained natural missingness, added nested missing completely at random (MCAR) perturbations of 5%, 10%, 20%, and 30%, and included a separate 20% missing at random (MAR) condition. Repeated stratified five fold cross validation used the same patient partitions, model seeds, and artificial missingness masks across corresponding comparisons. Ranking stability was assessed through condition level winner changes, paired rank reversals, and agreement across receiver operating characteristic area under the curve (ROC AUC), Average Precision, and Brier Score. The condition level winner remained stable under natural and mild additional missingness, but became dependent on imputation at higher missingness. Winner changes occurred in 6 of 12 imputation comparisons and 9 of 15 missingness comparisons, while pairwise rank reversals occurred in about 44% of matched repeat comparisons. All three metrics selected the same winner in 8 of 18 conditions. Because competing winners were separated by small ROC AUC margins, the results indicate sensitivity of model selection rather than large performance advantages. Reporting ranking stability alongside conventional performance estimates may therefore provide a more cautious basis for comparative clinical prediction studies.

**Keywords:** Missing data; clinical prediction; model ranking stability; imputation; machine learning.

**الملخص :**
تُعد البيانات المفقودة شائعة في دراسات التنبؤ السريري، وقد يؤثر أسلوب التعامل معها في الأداء التنبؤي وفي تحديد النموذج الذي يُعد الأفضل. هدفت هذه الدراسة إلى تقييم استقرار اختيار المصنف عند تغيير معالجة البيانات المفقودة ضمن شروط تقييم مضبوطة ومقترنة. باستخدام بيانات SUPPORT2 التي تضم 9,105 مرضى، جرى تقييم الانحدار اللوجستي والغابة العشوائية وXGBoost مع ثلاث استراتيجيات للتعويض: الوسيط، وأقرب الجيران، والتعويض التكراري. احتفظ التحليل بالفقد الطبيعي، وأضاف مستويات متداخلة من الفقد العشوائي الكامل MCAR بنسبة 5% و10% و20% و30%، إلى جانب حالة مستقلة من الفقد العشوائي MAR بنسبة 20%. واستُخدم تحقق متقاطع طبقي مكرر بخمسة طيات وخمسة تكرارات، مع تثبيت تقسيمات المرضى، وبذور النماذج، وأقنعة الفقد الاصطناعي في المقارنات المناظرة. قيس استقرار الترتيب من خلال تغير النموذج الفائز، وانعكاسات الترتيب المقترنة، والاتفاق بين ROC AUC ومتوسط الدقة ودرجة Brier. ظل النموذج الفائز مستقرًا في البيانات الطبيعية وعند المستويات المنخفضة من الفقد الإضافي، لكنه أصبح معتمدًا على طريقة التعويض عند ارتفاع الفقد. تغير النموذج الفائز في 6 من 12 مقارنة بين طرق التعويض وفي 9 من 15 مقارنة بين حالات الفقد، وظهرت انعكاسات زوجية في الترتيب في نحو 44% من المقارنات المتطابقة عبر التكرارات. كما اختارت المقاييس الثلاثة النموذج نفسه في 8 من 18 حالة. ونظرًا إلى صغر فروق ROC AUC بين النماذج الفائزة، تشير النتائج إلى حساسية اختيار النموذج أكثر من تفوق كبير في الأداء. وقد يوفر الإبلاغ عن استقرار الترتيب إلى جانب مقاييس الأداء التقليدية أساسًا أكثر حذرًا للمقارنة بين نماذج التنبؤ السريري.

**الكلمات المفتاحية:** البيانات المفقودة؛ التنبؤ السريري؛ استقرار ترتيب النماذج؛ التعويض عن القيم المفقودة؛ تعلم الآلة.

## 1. INTRODUCTION

Clinical prediction increasingly depends on routinely collected data, and in such data missing values are often part of the data collection process rather than isolated exceptions. Laboratory tests may not be requested for every patient, physiological measurements may be unavailable at particular times, and some clinical variables may be recorded selectively according to patient condition or workflow. As a result, missingness can change the information available to a predictive model and influence comparisons between competing algorithms. This becomes especially important when a study identifies one classifier as superior while the preceding missing data strategy is treated as a fixed preprocessing decision. A review of 152 clinical prediction studies using machine learning found that missing data were frequently handled and reported inadequately, with deletion and simple imputation remaining common despite their potential influence on predictive performance [1].

Research on this problem has expanded considerably. Hasan et al. reviewed 191 studies and documented wide variation in both statistical and machine learning approaches to missing value imputation, as well as in the way imputation quality and downstream prediction were assessed [2]. Afkanpour et al. reached a related conclusion in a systematic review of structured clinical datasets, showing that the suitability of an imputation method depends on factors such as the mechanism, pattern, and extent of missingness rather than on a universally preferable technique [3]. This shifts the question from how missing values should be filled to a broader issue: whether conclusions drawn after preprocessing remain stable when that preprocessing choice changes.

That distinction matters because successful reconstruction of missing values does not necessarily imply better downstream prediction. Perez-Lebel et al. showed across several health databases that alternative missing data strategies can change predictive performance as well as computational behavior [4]. Shadbahr et al. further demonstrated that an imputation method with stronger reconstruction performance does not necessarily produce the strongest classifier and reported that classifier rankings could remain stable in some settings while the preferred model differed across datasets [5]. These findings suggest that the complete prediction pipeline, rather than the imputation method in isolation, should be considered when comparing predictive models.

A recent review of electronic health record research reinforces this context dependence. Ren et al. reported substantial heterogeneity in missing data assumptions, analytical methods, and evaluation criteria, and concluded that no single strategy can be expected to perform uniformly across different data settings [6]. This has a direct consequence for comparative model evaluation. When candidate classifiers achieve similar predictive performance, even modest changes introduced during preprocessing may be enough to alter their ordering. In such situations, reporting only the highest score can give a stronger impression of model superiority than the evidence actually supports.

Most existing studies still focus primarily on identifying the most effective imputation method or the imputation and classifier combination that produces the highest predictive performance. A different question arises when the choice of classifier is itself the scientific conclusion: Would the same classifier still be selected if the missing data handling decision changed while the patients, folds, and missingness perturbations remained controlled? Absolute performance sensitivity asks how much a predictive score changes. Ranking sensitivity asks whether the conclusion about which model is preferred changes as well.

The present study examines this question using the SUPPORT2 clinical dataset for in-hospital mortality prediction. Logistic Regression, Random Forest, and XGBoost are evaluated with median, K nearest neighbor, and iterative imputation. The analysis retains the naturally incomplete data, introduces nested additional missing completely at random conditions of 5%, 10%, 20%, and 30%, and includes a separate missing at random condition at 20%. Corresponding comparisons use the same patient partitions, missingness masks, and prespecified random seeds, while all preprocessing steps are estimated from training data only.

The contribution of the study is not a new classifier or a new imputation algorithm. Instead, classifier ranking stability is defined before model performance is inspected and treated as the primary analytical target. The analysis considers condition-level winner changes and repeat-matched pairwise rank reversals, with ROC AUC as the primary performance criterion and Average Precision, Brier Score, and calibration used as complementary views of predictive behavior. This makes it possible to distinguish between two related but different questions: whether predictive performance declines as missingness increases, and whether the identity of the preferred classifier remains stable while that decline occurs.

The remainder of this paper is organized as follows. Section 2 reviews the closest prior work on missing data handling, downstream prediction, and model ranking behavior. Section 3 describes the dataset, missingness perturbation design, preprocessing procedures, classifiers, cross-validation protocol, and ranking stability analysis. Section 4 presents the predictive performance and ranking stability results, including winner changes, pairwise rank flips, cross-metric consistency, and the iterative imputation robustness analysis. Section 5 discusses the

findings in relation to previous studies, possible explanations, methodological implications, and study limitations. Section 6 concludes the paper and outlines directions for future work.

## 2. RELATED WORK

Prior research on missing data in machine learning has progressively moved from evaluating imputation fidelity toward examining downstream prediction and, more recently, complete pipeline selection. The present study builds on this progression but focuses specifically on whether the relative ordering of classifiers remains stable when missing data handling is changed under paired experimental conditions.

Abassi and Msengwa [7] simulated MCAR and MAR missingness in breast cancer recurrence data and compared multiple imputers with Binary Logistic Regression, Linear Discriminant Analysis, and Quadratic Discriminant Analysis. Their results showed that preferred classifier and imputation combinations could change across missingness conditions, but ranking changes were outcomes of a performance comparison rather than a prespecified stability estimand. Memon et al. [8] also examined consistency across missingness conditions using several imputers and classifiers. Although they employed Kendall's coefficient of concordance, the ranked objects were imputation methods rather than classifier identities.

The interaction among missingness, imputation, and classifier behavior has been demonstrated in broader benchmarking studies. Gabr et al. [9] evaluated several missing data patterns and imputation methods with multiple supervised classifiers and showed that predictive consequences depend jointly on these factors. Tiwaskar et al. [10] extended this factorial comparison in medical datasets using several imputers, classifiers, and missingness levels. Their work illustrates that crossing multiple imputers, classifiers, and missingness rates is no longer sufficient by itself as a methodological contribution. Shadbahr et al. [5] moved closer to the present question by explicitly discussing classifier ranking consistency and showing that superior reconstruction does not necessarily correspond to superior downstream classification. Ranking stability, however, remained a secondary observation within a broader evaluation of imputation quality and prediction.

Evidence from real clinical records further demonstrates that preprocessing choices can alter preferred prediction pipelines. Ma et al. [11] evaluated six imputation methods with four classifiers for gestational diabetes prediction and found that preferred combinations varied with the degree of missingness. Digitale et al. [12], using extensive simulation grounded in pediatric intensive care data, likewise showed that the predictive consequences of missing data handling depend on both missingness mechanism and proportion. Neither study treated paired classifier rank reversal as the primary analytical endpoint.

Recent work has made conventional imputation benchmarking even less distinctive. Zhang et al. [13] compared five imputation strategies across three clinical datasets under MCAR, MAR, and MNAR mechanisms and missingness rates ranging from 10% to 90%, using Random Forest for downstream prediction. Their analysis provides extensive evidence about imputation behavior but does not compare rankings across multiple classifiers. Nyakundi et al. [14] evaluated nine imputers with Logistic Regression, Random Forest, and LightGBM under several missingness mechanisms and rates. Although their design reports the strongest pipeline combinations, it does not formalize paired classifier rank stability across preprocessing alternatives.

The framework proposed by Li et al. [15] represents a further methodological step by treating incomplete-data pipeline selection as a decision problem and using missingness fingerprints, winner fractions, and selection regret to recommend prediction pipelines across datasets. The present question is different. Rather than learning which pipeline should be selected for a new dataset, the analysis asks whether the model selection conclusion within one fixed prediction task remains stable when a defensible preprocessing choice is changed. Roy et al. [16] similarly demonstrated that missing data processing can affect both predictive performance and the substantive interpretation of clinical machine learning analyses, but classifier ranking stability was not isolated as a prespecified outcome.

Taken together, previous studies establish that missingness severity and mechanism can affect prediction, that strong reconstruction does not necessarily imply strong downstream classification, and that preferred imputation and classifier combinations may vary across settings. What remains less directly examined is the **stability of the classifier selection conclusion itself** when observational units, resampling structure, and artificial missingness are paired across preprocessing alternatives.

Accordingly, the present study does not claim novelty from the selected imputers or classifiers. Its contribution lies in the experimental structure: the ranking question is specified before outcome inspection, corresponding pipelines use the same folds and missingness masks, condition-level winner changes and repeat-matched pairwise rank flips are quantified explicitly, and discrimination is considered alongside complementary probabilistic

criteria. Within the literature reviewed for this study, we did not identify a clinical tabular prediction study combining these elements while treating preprocessing-conditioned classifier ranking stability as its primary estimand.

Table I summarizes the studies most directly relevant to the present experimental question. It is intended as a comparative narrative synthesis rather than a systematic review.

**TABLE I. Comparison of the Closest Prior Studies and the Present Study**

| Ref. | Setting / Missingness | Imputation / Models | Ranking Focus | Main Distinction |
|---|---|---|---|---|
| [7] | Breast cancer recurrence; MCAR/MAR 15% to 60% | Six imputers; BLR, LDA, QDA | Best combinations compared | No prespecified paired ranking stability |
| [8] | Clinical categorical data; MCAR 5% to 50% | Five imputers; four classifiers | Kendall W for imputer consistency | Classifiers were not the ranked target |
| [5] | Two synthetic and three clinical datasets; natural/induced missingness | Multiple imputers; five classifiers | Classifier ranking consistency discussed | Ranking stability was descriptive, not the paired primary estimand |
| [9] | Benchmark datasets; MCAR/MNAR, multiple rates | Several imputers; five classifiers | Performance changes compared | No dedicated classifier ranking endpoint |
| [10] | Medical datasets; simulated missingness 10% to 50% | KNNI, MICE, MissForest; RF, SVM, AdaBoost, XGB | Best combinations compared | Focus on performance, not ranking stability |
| [11] | 5,066 pregnancies; natural and varying missingness | Six imputers; four classifiers | Best combinations and robustness | No paired winner change or rank flip analysis |
| [12] | Pediatric intensive care EHR; multiple mechanisms/proportions | Several missing data strategies and prediction models | Predictive consequences compared | Ranking stability not isolated as an outcome |
| [13] | Three clinical datasets; MCAR/MAR/MNAR 10% to 90% | Five imputers; RF downstream classifier | Imputation methods compared | Does not compare rankings across multiple classifiers |
| [14] | Two breast cancer datasets; MCAR/MAR/MNAR 5%, 10%, 20% | Nine imputers; LR, RF, LightGBM | Best pipeline combinations | No formal paired classifier rank stability framework |
| [15] | Multiple EHR datasets; multiple missingness regimes | 21 prediction pipelines | Win rate and selection regret | Targets pipeline recommendation across datasets |
| [16] | Infectious disease prediction; natural/scenario based missingness | Six missing data strategies; six classifiers | Performance and clinical conclusions | No prespecified classifier rank stability estimand |
| Present study | SUPPORT2; natural, nested MCAR +5% to +30%, MAR +20% | Median, KNN, Iterative; LR, RF, XGB | Winner changes, paired flips, cross metric agreement | Classifier ranking stability is the primary prespecified estimand |

*Abbreviations: BLR, Binary Logistic Regression; LDA, Linear Discriminant Analysis; QDA, Quadratic Discriminant Analysis; LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; XGB, XGBoost; KNNI, K nearest neighbor imputation; MICE, multiple imputation by chained equations; EHR, electronic health record; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random.*

## 3. MATERIALS AND METHODS

### 3.1 Study Design

The study was designed to test whether the relative ordering of clinical prediction models remains stable when missing data handling is changed under controlled conditions.

Before the main classifier results were inspected, predictor eligibility, missingness conditions, imputation procedures, model configurations, cross validation folds, random seeds, evaluation metrics, and ranking definitions were specified in advance to separate analytical design from outcome inspection.

Logistic Regression, Random Forest, and XGBoost were each evaluated with median, K nearest neighbor, and iterative imputation under six data conditions: the naturally incomplete data, nested additional MCAR perturbations of 5%, 10%, 20%, and 30%, and a separate MAR perturbation targeting 20% additional missingness. The evaluation was paired within each repeat and fold: corresponding pipelines used the same patient partition, artificial missingness mask, and model random state. RQ1 assessed ranking changes across imputation strategies while holding missingness fixed, whereas RQ2 assessed ranking changes across missingness conditions while holding imputation fixed. Table II summarizes the study design.

**TABLE II. Study Design Summary**

| Item | Specification |
|---|---|
| Dataset | SUPPORT2 |
| Patients | 9,105 |
| Outcome | In hospital mortality (hospdead) |
| Positive outcome | 2,360 / 9,105 (25.92%) |
| Predictors | 20 (17 numeric, 3 categorical) |
| Classifiers | Logistic Regression; Random Forest; XGBoost |
| Imputation | Median; KNN; Iterative |
| Missingness | Natural; nested MCAR +5%, +10%, +20%, +30%; MAR +20% |
| Cross validation | Repeated stratified 5 fold × 5 repeats |
| Primary unit | Repeat level pooled out of fold predictions |
| Primary metric | ROC AUC |
| Ranking endpoints | Winner; winner change; pairwise rank flip |
| Hyperparameter tuning | None |
| Decision threshold | 0.50 |

Figure 1 summarizes the complete experimental workflow and the paired structure used to evaluate classifier ranking stability.

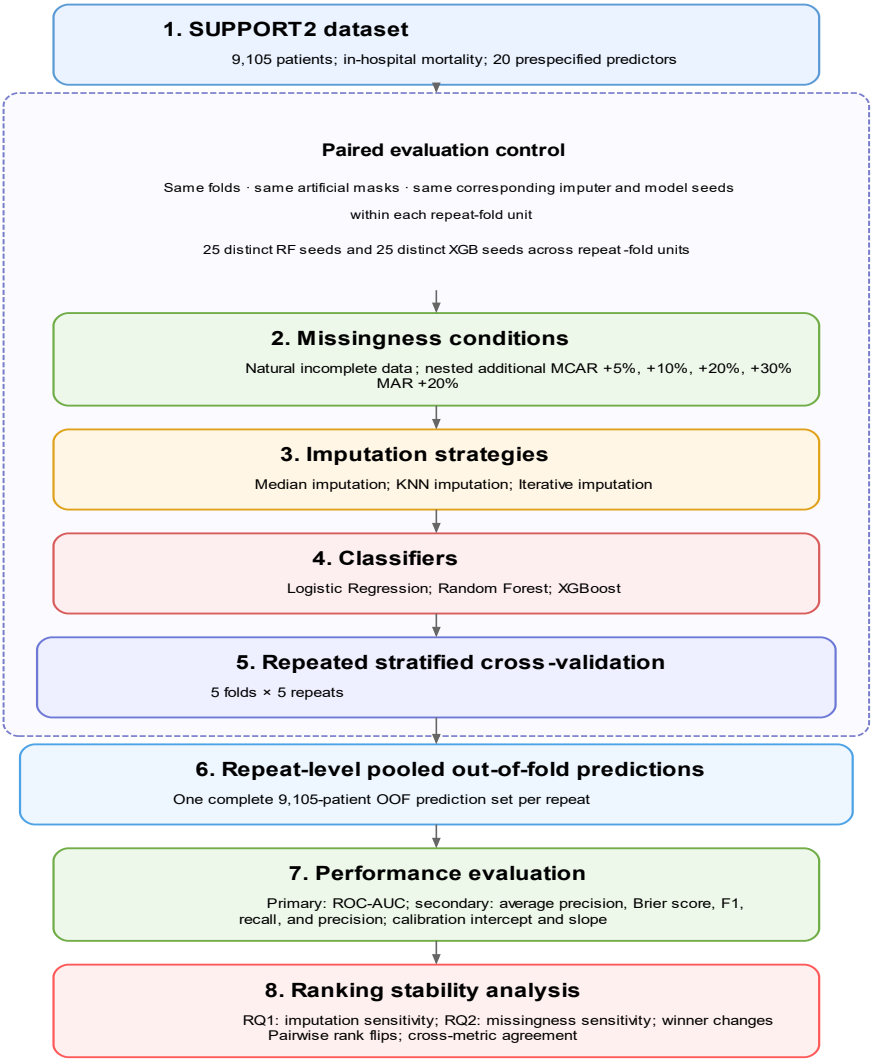


**Figure 1:** Experimental workflow and paired evaluation framework for assessing classifier ranking stability across alternative missing data handling conditions.

### 3.2 Data Source, Outcome, and Predictors

The study used the SUPPORT2 dataset from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments, which enrolled 9,105 seriously ill hospitalized adults at five U.S. teaching hospitals [17]. The original SUPPORT prognostic framework centered clinical assessment on information available around study day 3 [18]. The outcome was hospdead, indicating in hospital mortality. Of the 9,105 patients, 2,360 died during hospitalization and 6,745 survived to discharge, giving an event rate of 25.92%. The analysis was framed as a day 3 landmark prediction study and did not make claims about prospective clinical deployment.

Predictors were selected before model training. Variables representing identifiers, direct or future outcome information, previous prognostic scores, resource utilization, post landmark information, or potential timing and circularity concerns were excluded. The final feature set contained 20 predictors. Seventeen were numeric: age, number of comorbidities (num.co), coma score (scoma), study hospital day (hday), diabetes, dementia, mean blood pressure (meanbp), white blood cell count (wblc), heart rate (hrt), respiratory rate (resp), temperature (temp), PaO2/FiO2 ratio (pafi), albumin (alb), bilirubin (bili), creatinine (crea), sodium (sod), and arterial pH (ph). Sex, diagnostic group (dzgroup), and cancer status (ca) were treated as categorical predictors. The original data were preserved unchanged.

### 3.3 Missingness Perturbation Design

The natural condition retained the missing values already present in SUPPORT2. Among the selected predictors, natural missingness ranged from 0% to 37.03%, with the highest rates for albumin (37.03%), bilirubin

(28.57%), PaO2/FiO2 ratio (25.54%), and arterial pH (25.09%). The complete predictor level audit is reported in Supplementary Table S2.

Artificial masking was permitted for 15 predictors: age, num.co, scoma, hday, meanbp, wblc, hrt, resp, temp, pafi, alb, bili, crea, sod, and ph. Sex, dzgroup, ca, diabetes, and dementia were never artificially masked. Only originally observed cells were eligible for additional missingness.

For MCAR, additional cell level missingness was introduced at nominal rates of 5%, 10%, 20%, and 30%. Masks were nested through a single frozen random field, so the 5% mask was a subset of the 10% mask, which was itself contained within the 20% and 30% masks. Across the 50 train and test split realizations, the mean achieved rates were 4.99%, 9.98%, 19.97%, and 29.98%, respectively.

The MAR condition targeted 20% additional cell level missingness using a fixed synthetic logistic propensity model based only on five fully observed, never masked variables: sex, diagnostic group, cancer status, diabetes, and dementia. The outcome was not used. Coefficients were fixed before the main experiment and represented simulation parameters rather than estimated clinical associations. The intercept was calibrated within each split to an expected additional rate of 20%, yielding a mean achieved rate of 20.07%. Full coefficients and audit ranges are reported in Supplementary Table S3. Training and test masks were generated separately with frozen seeds, and corresponding models and imputation strategies used the same masks.

### 3.4 Preprocessing and Imputation

All preprocessing parameters were estimated from the training portion of each fold. Held out data were not used to estimate scaling parameters, imputation values, categorical encodings, or model parameters.

The 17 numeric predictors were standardized using StandardScaler from scikit learn [19]. Scaling parameters were estimated from available training values, while missing entries remained missing until imputation. Scaling preceded imputation for all three strategies, providing a common numeric representation and an appropriate distance scale for KNN.

Median imputation used the training fold median for each numeric variable. KNN imputation used five neighbors, uniform weighting, and the nan_euclidean distance metric. Iterative imputation used IterativeImputer with the default Bayesian Ridge estimator, median initialization, max_iter = 10, tolerance $10^{-3}$, sample_posterior = False, skip_complete = True, and a frozen random state. The three categorical predictors had no missing values in the verified data and were never artificially masked; they were one hot encoded with handle_unknown = ignore and no dropped reference category. No missingness indicators were added.

### 3.5 Prediction Models

Logistic Regression used L2 regularization with $C = 1.0$, the lbfgs solver, and a maximum of 2,000 iterations.

Random Forest followed the ensemble formulation of Breiman [20] and used 300 trees, the Gini criterion, bootstrap sampling, square root feature selection at each split, a minimum leaf size of one, a minimum split size of two, and unrestricted tree depth.

XGBoost followed the gradient boosting framework of Chen and Guestrin [21]. The model used a binary logistic objective, 300 boosting trees, learning rate 0.05, maximum depth 4, minimum child weight 1, full row and feature sampling, gamma 0, L1 regularization 0, L2 regularization 1, and histogram based tree construction.

The experiment compared fixed analytical configurations rather than the maximum attainable performance of each algorithm. No hyperparameter search, class weighting, resampling, early stopping, or threshold tuning was performed. Stochastic model states followed the frozen seed registry. Probability based metrics used predicted probabilities of in hospital death, while threshold dependent metrics used a fixed threshold of 0.50.

### 3.6 Cross Validation and Performance Evaluation

Predictive performance was evaluated using stratified five fold cross validation repeated five times. The resulting 25 held out folds used identical patient assignments for every model, imputation strategy, and missingness condition.

The 25 fold scores were not treated as 25 independent observations. Within each repeat, predictions from the five held out folds were concatenated to form one complete out of fold prediction vector containing all 9,105 patients exactly once. Each model, imputation strategy, and missingness condition therefore produced five repeat level performance estimates. This avoids interpreting correlated cross validation folds as independent replicates [22].

ROC AUC was the primary performance metric. Secondary measures were PR AUC, operationalized using Average Precision, Brier Score, F1 score, recall, and precision. F1, recall, and precision were evaluated at the fixed threshold of 0.50. Full secondary metric summaries are provided in Supplementary Table S4.

Probabilistic performance and calibration were examined separately. Brier Score summarized overall probabilistic prediction quality, while calibration intercept and calibration slope assessed calibration more specifically [23]. Probabilities were clipped to $[10^{-6}, 1 - 10^{-6}]$, transformed to the logit scale, and entered as the sole predictor in an unpenalized logistic calibration model. Ideal calibration corresponds to an intercept of 0 and a slope of 1.

For each experimental combination, means, standard deviations, and ranges across the five repeat level estimates were reported descriptively. No confidence interval or hypothesis test was constructed by treating the 25 folds as independent replicates.

### 3.7 Classifier Ranking Stability

The primary scientific analysis concerned the stability of classifier ordering rather than the identification of the largest absolute performance value.

For a given experimental condition c, the condition level winner was the classifier with the largest mean repeat level ROC AUC. Let μAUC(m,c) denote the mean repeat level ROC AUC of classifier m under condition c. The winner was defined as:

$$W(c) = arg\ max_m\ \mu AUC(m,c)$$

The three classifiers were ranked in descending order of this quantity. A numerical tolerance of $10^{-12}$ was used only for computational ties; near equal values were not forced to share a rank.

For RQ1, median imputation was the prespecified reference. Within each missingness condition, rankings under KNN and iterative imputation were compared with the ranking under median imputation. For RQ2, the natural data condition was the prespecified reference. Within each imputation method, rankings under MCAR 5%, MCAR 10%, MCAR 20%, MCAR 30%, and MAR 20% were compared with the corresponding natural condition.

A condition level winner change occurred when the top ranked classifier differed from the appropriate reference winner. Winner identities were also compared within the same repeat to preserve pairing.

Pairwise rank flips were evaluated for Logistic Regression versus Random Forest, Logistic Regression versus XGBoost, and Random Forest versus XGBoost. For classifiers a and b, repeat r, and condition c, the paired difference was:

$$\Delta(a,b,r,c) = AUC(a,r,c) - AUC(b,r,c)$$

A rank flip was recorded when the sign of this difference under a comparison condition was opposite to its sign under the prespecified reference condition for the same repeat. Differences within the $10^{-12}$ tolerance were treated as ties rather than flips.

Winner change and rank flip proportions summarize the prespecified comparison grid. They are descriptive point estimate stability measures and are not interpreted as population probabilities or as inference from independent experimental replicates. The numerical magnitude of performance differences and repeat consistency were retained to avoid interpreting small reversals as large practical advantages.

Mean rank, rank range, Kendall tau, and Spearman rho were retained as secondary descriptive summaries. Cross metric consistency was examined by comparing the winning classifier under ROC AUC, Average Precision, and Brier Score; higher values were preferred for ROC AUC and Average Precision, whereas lower Brier Score values were preferred.

### 3.8 Iterative Imputation Robustness Analysis

The primary iterative imputation configuration used max_iter = 10, as specified before model training. During execution, convergence warnings were observed for a subset of iterative imputation fits. Before any classifier ranking results were inspected, a secondary robustness analysis was declared using max_iter = 30.

No other component of the experiment was changed. The same patient folds, missingness masks, seeds, preprocessing rules, classifier specifications, thresholds, and evaluation metrics were retained. The original max_iter = 10 analysis remained the primary analysis and was not replaced.

Robustness was assessed by comparing condition level winners, repeat level winner agreement, RQ1 and RQ2 conclusions, pairwise flip counts, cross metric winner identities, and numerical differences in performance metrics.

### 3.9 Reproducibility

The experiments were executed in Google Colab using Python 3.13.15, NumPy 2.1.3, pandas 2.2.3, scikit learn 1.6.1, and XGBoost 3.4.1.

Patient fold assignments, artificial missingness generation, iterative imputation random states, and stochastic classifier random states were governed by frozen registries. Random Forest and XGBoost each used 25 distinct model seeds across the 25 repeat fold units. Within a given repeat fold unit, the same model seed was retained across the corresponding missingness and imputation comparisons, preserving the paired design. The 25 unit level seed assignments are reported in Supplementary Table S5.

The raw SUPPORT2 file remained unchanged. SHA 256 digests were retained for the source dataset, design specifications, principal result files, and final analysis packages to preserve computational provenance.

## 4. RESULTS

### 4.1 Overall Predictive Performance and Condition Level Winners

Predictive discrimination declined progressively as additional missingness increased, but the decline was accompanied by changes in the identity of the highest ranked classifier. Under the natural data condition, Random Forest achieved the highest mean repeat level ROC AUC under all three imputation strategies: 0.8096 with median imputation, 0.8084 with KNN imputation, and 0.8090 with iterative imputation. The condition level winner also remained Random Forest under MCAR 5% and MCAR 10%.

At MCAR 20%, XGBoost became the highest ranked classifier under median imputation with a mean ROC AUC of 0.7972, whereas Logistic Regression became the highest ranked classifier under KNN and iterative imputation, with means of 0.7961 and 0.7965. The same division persisted at MCAR 30%, where XGBoost led under median imputation (0.7911) and Logistic Regression led under KNN (0.7889) and iterative imputation (0.7864).

Under MAR 20%, median imputation again favored XGBoost (0.7957), whereas KNN and iterative imputation favored Logistic Regression (0.7949 and 0.7957). Table III reports the condition level results; Supplementary Table S1 contains the full 54 model level ROC AUC summaries.

Winner margins were small across the 18 missingness and imputation combinations, approximately 0.0003 to 0.0028 ROC AUC. These differences therefore indicate sensitivity of point estimate model selection rather than large absolute performance advantages.

**TABLE III. Primary ROC AUC Ranking Matrix**

| Missingness | Median | KNN | Iterative |
|---|---|---|---|
| Natural | RF<br>0.8096 ± 0.0007<br>Δ=0.0008 | RF<br>0.8084 ± 0.0006<br>Δ=0.0009 | RF<br>0.8090 ± 0.0009<br>Δ=0.0013 |
| MCAR +5% | RF<br>0.8064 ± 0.0012<br>Δ=0.0003 | RF<br>0.8046 ± 0.0006<br>Δ=0.0007 | RF<br>0.8055 ± 0.0010<br>Δ=0.0012 |
| MCAR +10% | RF<br>0.8038 ± 0.0013<br>Δ=0.0008 | RF<br>0.8017 ± 0.0006<br>Δ=0.0007 | RF<br>0.8029 ± 0.0012<br>Δ=0.0011 |
| MCAR +20% | XGB<br>0.7972 ± 0.0015<br>Δ=0.0017 | LR<br>0.7961 ± 0.0012<br>Δ=0.0024 | LR<br>0.7965 ± 0.0006<br>Δ=0.0010 |
| MCAR +30% | XGB<br>0.7911 ± 0.0023<br>Δ=0.0028 | LR<br>0.7889 ± 0.0011<br>Δ=0.0024 | LR<br>0.7864 ± 0.0011<br>Δ=0.0017 |
| MAR +20% | XGB<br>0.7957 ± 0.0032<br>Δ=0.0023 | LR<br>0.7949 ± 0.0015<br>Δ=0.0019 | LR<br>0.7957 ± 0.0015<br>Δ=0.0026 |

*Cells show the condition level winner, mean ROC AUC ± SD across five repeat level pooled out of fold estimates, and winner margin Δ relative to the second ranked classifier. RF, Random Forest; LR, Logistic Regression; XGB, XGBoost.*

Figure 2 visualizes the transition from Random Forest dominance under natural and low additional missingness to an imputation dependent winner pattern at higher missingness.

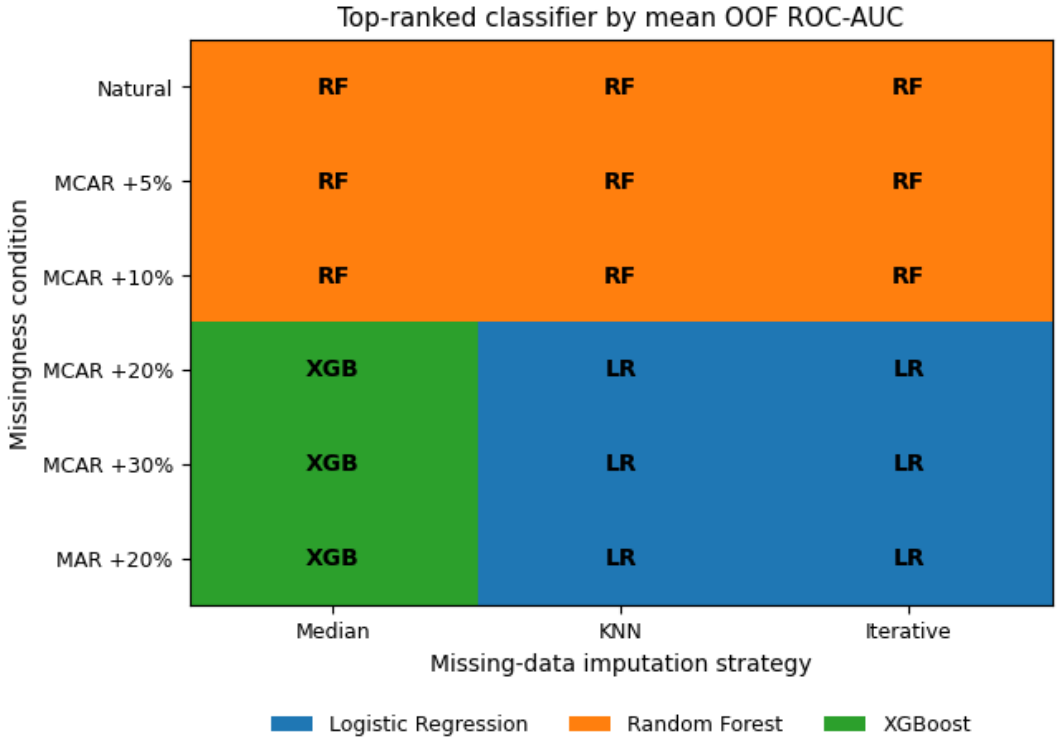


**Figure 2:** Top ranked classifier across missingness conditions and imputation strategies based on mean repeat level pooled out of fold ROC AUC.

### 4.2 Performance Trajectories Under Increasing MCAR Missingness

Figure 3 follows the three classifiers from the natural condition through the nested MCAR perturbations. Because the masks were nested, increasing missingness represented progressively stronger perturbation of the same underlying masking structure rather than unrelated random scenarios.

All classifiers exhibited decreasing ROC AUC as additional MCAR missingness increased, but their trajectories were not parallel. Random Forest retained a small advantage under natural and low missingness. By MCAR 20%, XGBoost had overtaken Random Forest under median imputation, whereas Logistic Regression became the highest ranked classifier under KNN and iterative imputation. The ranking changes therefore arose from different rates of deterioration rather than improvement in absolute performance.

The error bars in Figure 3 represent one standard deviation across the five repeat level pooled out of fold estimates and are descriptive measures of resampling variation, not confidence intervals.

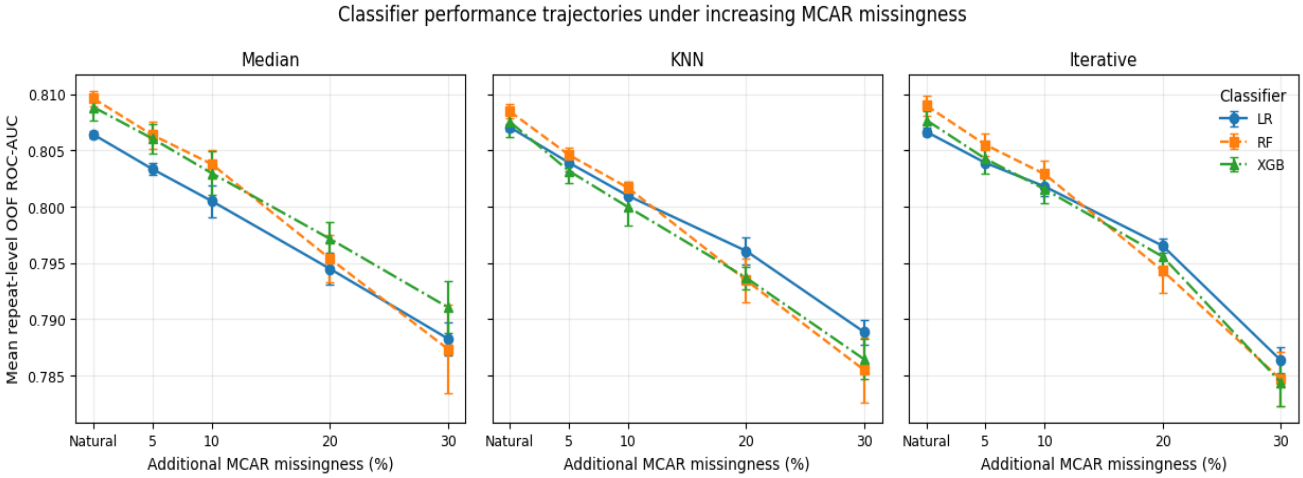


**Figure 3:** Mean ROC AUC trajectories under increasing nested MCAR missingness. Error bars represent ±1 standard deviation across the five repeat level pooled out of fold estimates.

### 4.3 RQ1: Sensitivity to Imputation Strategy

RQ1 compared KNN and iterative imputation with the prespecified median reference while holding the missingness condition fixed. Under the natural condition, MCAR 5%, and MCAR 10%, the condition level winner remained Random Forest in all six comparisons.

At MCAR 20%, MCAR 30%, and MAR 20%, median imputation selected XGBoost, whereas both KNN and iterative imputation selected Logistic Regression. Thus, 6 of 12 RQ1 comparisons (50.0%) changed the condition level winner.

At the matched repeat level, 80 of 180 classifier pair comparisons (44.44%) produced a rank flip relative to the median imputation reference. These counts summarize the prespecified grid and should not be interpreted as independent replicate probabilities.

### 4.4 RQ2: Sensitivity to Missingness Severity and Mechanism

RQ2 held the imputation method fixed and compared each perturbed condition with the natural data reference. Random Forest was the natural condition winner under all three imputers and remained the condition level winner at MCAR 5% and MCAR 10%.

At MCAR 20%, MCAR 30%, and MAR 20%, median imputation shifted from Random Forest to XGBoost, whereas KNN and iterative imputation shifted from Random Forest to Logistic Regression. Overall, 9 of 15 RQ2 comparisons (60.0%) changed the condition level winner, and 99 of 225 matched repeat classifier pair comparisons (44.0%) produced a rank flip.

**TABLE IV. Ranking Stability Endpoints**

| Analysis | Endpoint | Changed / Total | Percent |
|---|---|---|---|
| RQ1 | Condition level winner change | 6 / 12 | 50.00% |
| RQ1 | Repeat level pairwise rank flip | 80 / 180 | 44.44% |
| RQ2 | Condition level winner change | 9 / 15 | 60.00% |
| RQ2 | Repeat level pairwise rank flip | 99 / 225 | 44.00% |
| Cross metric | Winner disagreement | 10 / 18 | 55.56% |

Figure 4 summarizes the main instability endpoints and shows that the pattern was not confined to a single comparison family.

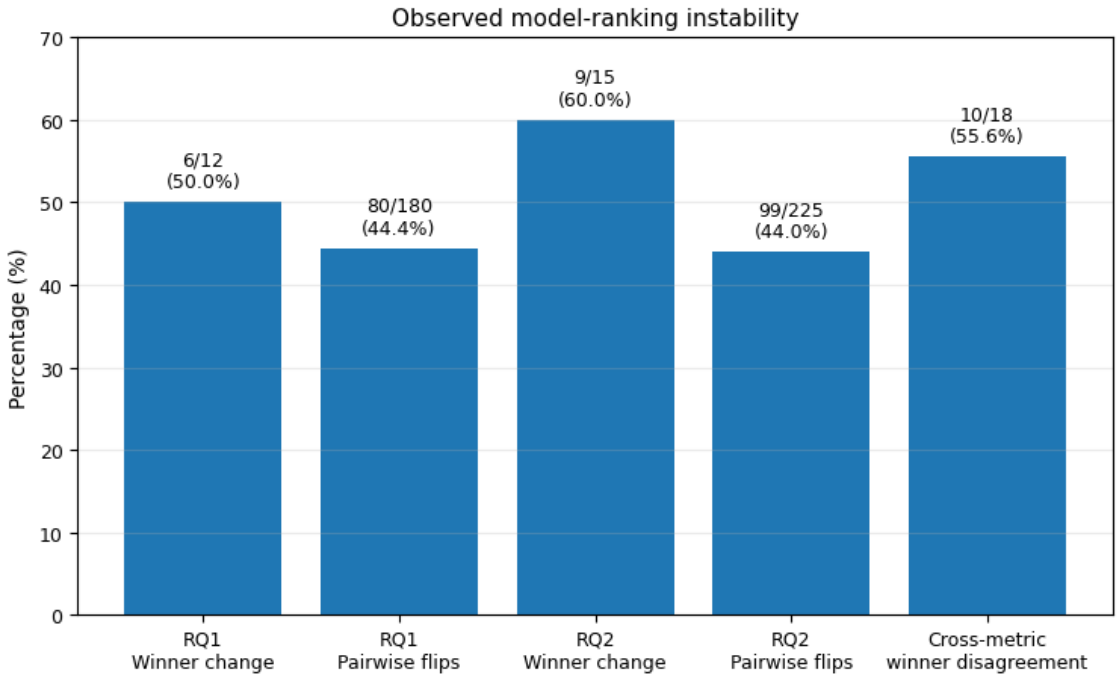


**Figure 4:** Summary of observed classifier ranking instability across the prespecified comparison grid.

### 4.5 Pairwise Rank Flip Patterns

Aggregate flip rates conceal differences among classifier pairs. Figure 5 therefore separates Logistic Regression versus Random Forest, Logistic Regression versus XGBoost, and Random Forest versus XGBoost for RQ1 and RQ2.

Pairwise reversals were present in some natural and low missingness comparisons even when the condition level winner was unchanged, showing that a stable winner does not imply a stable complete ordering. Under several MCAR 20%, MCAR 30%, and MAR 20% comparisons, flip rates reached 80% or 100%, meaning that the sign reversal occurred in four or all five matched repeats.

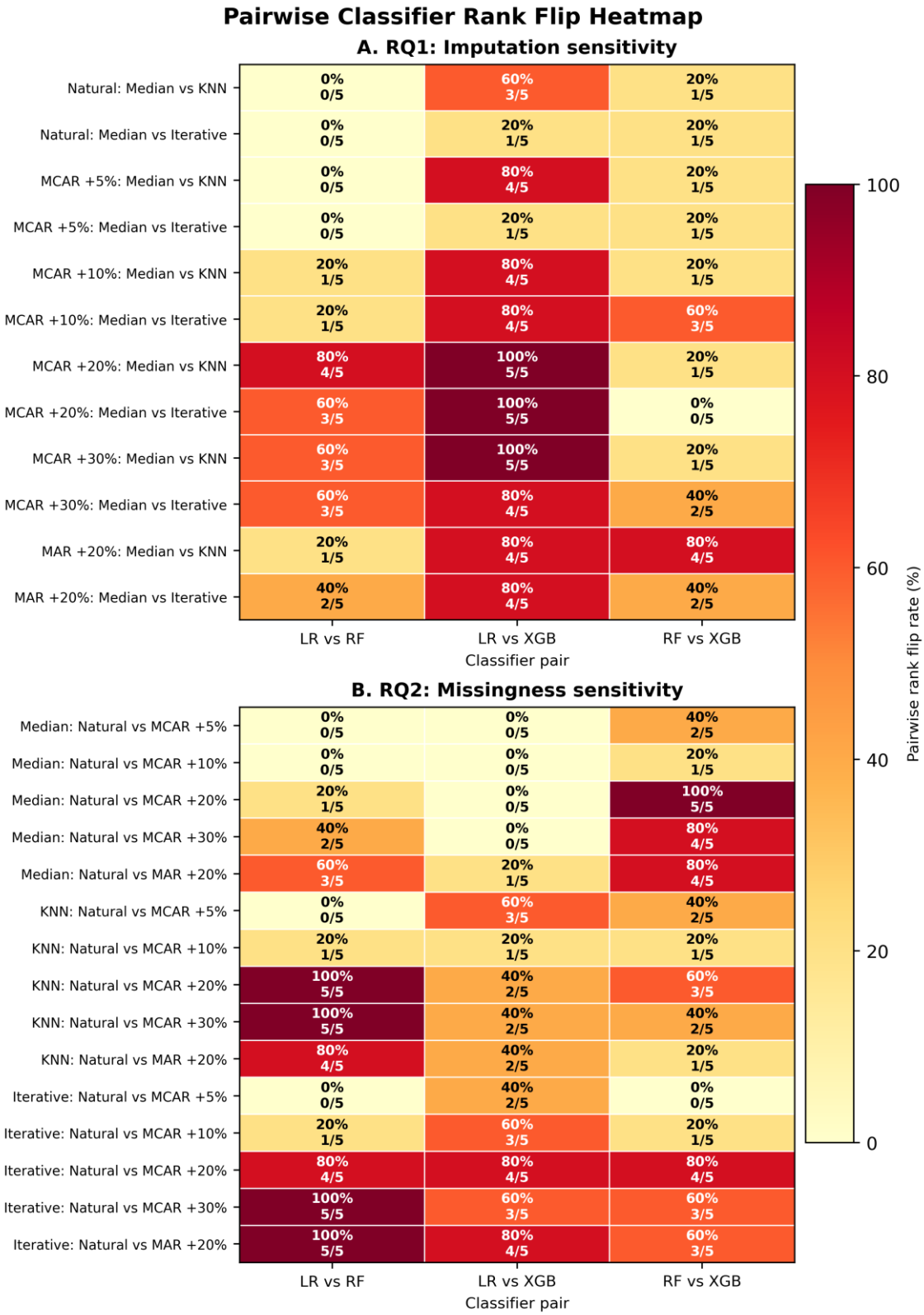


**Figure 5:** Pairwise classifier rank flip rates for RQ1 and RQ2. Each cell reports the percentage and count of the five matched repeats in which the sign of the ROC AUC difference reversed relative to the prespecified reference.

### 4.6 Cross Metric Consistency

The preferred classifier also depended on the performance criterion. ROC AUC winners were compared with winners based on Average Precision and Brier Score. All three metrics selected the same classifier in 8 of 18 missingness and imputation combinations (44.44%); in the remaining 10 conditions (55.56%), at least one metric selected a different winner.

Disagreement was most common under the natural and lower MCAR conditions, while several more severe missingness conditions produced agreement across all three metrics. This is not a contradiction among metrics: ROC AUC measures discrimination across thresholds, Average Precision emphasizes precision recall behavior, and Brier Score summarizes probabilistic prediction quality. Table V reports the winner comparison, and Supplementary Table S4 provides the full secondary metric and calibration summaries.

**TABLE V. Cross Metric Winner Comparison**

| Missingness | Imputation | ROC AUC | Average Precision | Brier | Agreement |
|---|---|---|---|---|---|
| Natural | Median | RF | XGB | XGB | No |
| Natural | KNN | RF | XGB | XGB | No |
| Natural | Iterative | RF | XGB | XGB | No |
| MCAR +5% | Median | RF | XGB | XGB | No |
| MCAR +5% | KNN | RF | LR | LR | No |
| MCAR +5% | Iterative | RF | XGB | LR | No |
| MCAR +10% | Median | RF | XGB | XGB | No |
| MCAR +10% | KNN | RF | XGB | LR | No |
| MCAR +10% | Iterative | RF | XGB | LR | No |
| MCAR +20% | Median | XGB | XGB | XGB | Yes |
| MCAR +20% | KNN | LR | LR | LR | Yes |
| MCAR +20% | Iterative | LR | XGB | LR | No |
| MCAR +30% | Median | XGB | XGB | XGB | Yes |
| MCAR +30% | KNN | LR | LR | LR | Yes |
| MCAR +30% | Iterative | LR | LR | LR | Yes |
| MAR +20% | Median | XGB | XGB | XGB | Yes |
| MAR +20% | KNN | LR | LR | LR | Yes |
| MAR +20% | Iterative | LR | LR | LR | Yes |

*RF, Random Forest; LR, Logistic Regression; XGB, XGBoost. Agreement = all three metrics selected the same classifier.*

### 4.7 Iterative Imputation Robustness Analysis

The primary iterative imputation analysis used the prespecified max_iter = 10 setting and produced 55 convergence warnings. The predeclared robustness branch repeated iterative imputation with max_iter = 30 while retaining all other experimental elements.

Increasing the iteration limit reduced convergence warnings from 55 to 26. Condition level winners were identical in all six missingness conditions, repeat level winner identity agreed in 29 of 30 comparisons, all RQ1 and RQ2 condition level conclusions were preserved, and cross metric winner identities agreed in all 18 comparisons.

The mean absolute ROC AUC difference between the max_iter = 30 and max_iter = 10 analyses was approximately 0.00021, with a maximum absolute difference of approximately 0.00210. Thus, the primary ranking conclusions were not explained solely by incomplete convergence of iterative imputation.

**TABLE VI. Iterative Imputation Robustness**

| Condition | Winner, 10 iter. | Winner, 30 iter. | Agreement | Margin, 10 | Margin, 30 |
|---|---|---|---|---|---|
| Natural | RF | RF | Yes | 0.0013 | 0.0013 |
| MCAR +5% | RF | RF | Yes | 0.0012 | 0.0012 |
| MCAR +10% | RF | RF | Yes | 0.0011 | 0.0011 |
| MCAR +20% | LR | LR | Yes | 0.0010 | 0.0016 |
| MCAR +30% | LR | LR | Yes | 0.0017 | 0.0020 |
| MAR +20% | LR | LR | Yes | 0.0026 | 0.0024 |

## 5. DISCUSSION

### 5.1 Principal Findings

The results show that the answer to the study's central question is conditional. Random Forest remained the condition-level ROC AUC winner under the natural data and mild MCAR perturbations, but the preferred classifier changed once additional missingness reached 20% or more.

These changes should not be interpreted as evidence of substantial algorithmic superiority. Discrimination declined for all three classifiers, and winner margins were generally below 0.003 ROC AUC. The more important finding is that model selection became sensitive when competing models were close. About 44% of matched pairwise comparisons reversed order, and ROC AUC, Average Precision, and Brier Score selected different winners in 10 of 18 conditions. Thus, the identity of the “best” model depended not only on the classifier but also on the preprocessing path and the criterion used to define best.

### 5.2 Possible Explanations for the Observed Ranking Changes

The observed changes may reflect interactions among missingness severity, imputation behavior, and model structure. Random Forest captures nonlinearities and interactions through an ensemble of decision trees [20], whereas Logistic Regression imposes a smoother linear decision function. As information was progressively removed and reconstructed, these modeling families did not lose discrimination at the same rate.

One plausible interpretation is that KNN and iterative imputation produced representations in which the regularized linear model became relatively more competitive as complex local structure became harder to recover. Under median imputation, XGBoost may have retained a relative advantage by exploiting nonlinear thresholds in the remaining information [21]. These explanations were not tested directly, however, and should be regarded as hypotheses rather than established mechanisms.

### 5.3 Relation to Previous Studies

The findings are consistent with earlier evidence that missing data handling can influence downstream prediction. Abassi and Msengwa [7] reported changes in preferred classifier-imputation combinations across missingness conditions, while Memon et al. [8] showed that imputation rankings were not constant as missingness increased. Shadbahr et al. [5] are particularly relevant because they discussed classifier ranking consistency and showed that better reconstruction did not necessarily lead to better downstream classification.

More recent studies have reinforced the same broader pattern. Ma et al. [11], Digitale et al. [12], Zhang et al. [13], and Nyakundi et al. [14] showed that imputation and downstream prediction depend on missingness severity, mechanism, or data context. Li et al. [15] extended this line of work toward pipeline recommendation using winner fractions and selection regret, while Roy et al. [16] showed that missing data processing can alter both predictive performance and substantive interpretation.

The present study differs mainly in analytical focus. Rather than asking which pipeline performs best across datasets or conditions, it treats the stability of the classifier selection conclusion itself as the primary outcome under paired folds and missingness masks. The resulting claim is therefore narrower than universal algorithm superiority: when competing models perform similarly, preprocessing choices can become part of the evidential basis on which a winner is declared.

### 5.4 Paired Evaluation, Metric Dependence, and Small Margins

The paired design strengthens this interpretation because corresponding comparisons used the same patients, fold assignments, artificial missingness masks, and within-unit model seeds. Rank reversals therefore reflect controlled analytical perturbations rather than different random splits or independently generated masks.

The small winner margins nevertheless require caution. The study quantifies instability in point-estimate rankings, not the population probability that one classifier is truly superior. The repeated cross-validation folds were not treated as independent replicates, and no significance threshold was required to define a rank flip. A future extension could examine winner uncertainty using paired patient-level bootstrap procedures.

Metric disagreement provides a related warning. ROC AUC, Average Precision, and Brier Score measure different aspects of prediction, so they need not identify the same model when performance is close. Claims about the "best" model should therefore state the criterion on which that claim is based.

### 5.5 Robustness and Model Stochasticity

Increasing iterative imputation from max_iter = 10 to max_iter = 30 reduced convergence warnings while preserving all condition-level winners and 29 of 30 repeat-level winners. This makes it unlikely that the main ranking pattern was driven by the original convergence setting.

Random Forest and XGBoost each used 25 distinct seeds across the 25 repeat-fold units, with the corresponding seed retained across paired preprocessing comparisons. This controlled random-state differences within those comparisons but did not isolate model stochasticity through a dedicated same-fold multi-seed experiment. Such an analysis would be useful when winner margins are very small, but it was outside the prespecified primary design.

### 5.6 Strengths and Limitations

Several design features strengthen the study. Ranking definitions were specified before the main results were inspected, folds and missingness masks were paired, MCAR perturbations were nested, preprocessing was fitted only on training data, and repeat-level pooled out-of-fold predictions were used rather than treating individual folds as independent replicates. The analysis also separates absolute performance changes from changes in model selection.

The main limitation is external validity. The study used one historical clinical dataset and one outcome, so the frequency and direction of ranking changes should not be generalized to other populations or prediction tasks. Only three classifiers and three imputation strategies were evaluated, and the MAR mechanism was synthetic rather than estimated from an observed clinical missingness process. MNAR was not examined because doing so would require additional assumptions about unobserved values.

Hyperparameter optimization was also not performed because the aim was to compare fixed representative configurations rather than maximum attainable performance. Different tuning policies could alter relative rankings and deserve future investigation. Finally, the study did not assess deployment utility or prospective clinical consequences, so a ranking change should not be interpreted as evidence that clinical decisions would necessarily change.

### 5.7 Implications for Comparative Clinical Prediction Studies

When candidate models perform similarly in datasets with substantial missingness, reporting only the largest performance value may imply more certainty than the evidence supports. A winner that remains stable across several defensible preprocessing choices provides stronger support for model preference than one that appears only under a single analytical path.

This does not mean that every prediction study requires an extensive missingness simulation. It does suggest that preprocessing sensitivity should be considered when model differences are small and missing data are substantial. In such settings, the relevant question is not only how well a model predicts, but also whether its apparent advantage survives plausible changes in the analytical decisions that precede prediction.

## 6. CONCLUSION

Overall, classifier selection in SUPPORT2 was stable under natural and mild additional missingness but became dependent on the imputation strategy as missingness increased. At higher missingness levels, median imputation favored XGBoost, whereas KNN and iterative imputation favored Logistic Regression.

Across the prespecified comparison grid, the condition-level winner changed in 50% of the imputation comparisons and 60% of the missingness comparisons, while matched pairwise rank reversals occurred in about 44% of comparisons. ROC AUC, Average Precision, and Brier Score selected the same classifier in only 8 of 18 conditions. Because the competing winners were separated by small ROC AUC margins, these findings should not be interpreted as evidence of clinically important superiority. Rather, they show that the identity of the reported winner can be sensitive to analytical choices made before model evaluation.

The iterative imputation robustness analysis preserved the main conclusions, while the paired design controlled patient folds, missingness masks, and corresponding model seeds across compared pipelines. These findings highlight the distinction between estimating predictive performance and assessing the stability of model selection. When competing models perform similarly and missing data are substantial, reporting whether their ranking persists across defensible missing data treatments may provide a more informative basis for comparison than presenting a single pipeline winner as unequivocal. Future work should evaluate this framework across additional datasets, outcomes, model classes, tuning strategies, and missingness mechanisms.

## Compliance with ethical standards

### *Ethics statement*

This work is a secondary methodological analysis of a publicly available historical dataset. No new participants were enrolled, no intervention or participant contact occurred, and no direct identifiers were accessed. The original SUPPORT study and its clinical protocol are described in [17]. No new clinical data were collected for the present analysis.

### *Funding*

This research received no external funding.

### *Data availability*

The SUPPORT2 dataset analyzed in this study is publicly available from the Vanderbilt University Department of Biostatistics HBiostat data repository [24]. The raw file used for the experiments was preserved unchanged, and its cryptographic identity was recorded in the reproducibility materials.

### *Reproducibility materials*

The study generated frozen fold and seed registries, missingness specifications and audits, result manifests, and SHA 256 provenance records. Supplementary material accompanying this manuscript reports the full ROC AUC summaries, natural missingness audit, synthetic MAR parameters, secondary performance and calibration summaries, and the 25 repeat fold seed assignments.

## Compliance with ethical standards

*Disclosure of conflict of interest*

The authors declare that they have no conflict of interest.

## References

1. Nijman, S.W.J.; et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. *J. Clin. Epidemiol.* **2022**, *142*, 218–229. doi: 10.1016/j.jclinepi.2021.11.023.
2. Hasan, M.K.; Alam, M.A.; Roy, S.; Dutta, A.; Jawad, M.T.; Das, S. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Inform. Med. Unlocked* **2021**, *27*, 100799. doi: 10.1016/j.imu.2021.100799.
3. Afkanpour, M.; Hosseinzadeh, E.; Tabesh, H. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. *BMC Med. Res. Methodol.* **2024**, *24*, 188. doi: 10.1186/s12874-024-02310-6.
4. Perez-Lebel, A.; Varoquaux, G.; Le Morvan, M.; Josse, J.; Poline, J.B. Benchmarking missing-values approaches for predictive models on health databases. *GigaScience* **2022**, *11*, giac013. doi: 10.1093/gigascience/giac013.
5. Shadbahr, T.; et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. *Commun. Med.* **2023**, *3*, 139. doi: 10.1038/s43856-023-00356-z.

6. Ren, W.; Liu, Z.; Wu, Y.; Zhang, Z.; Hong, S.; Liu, H. Moving Beyond Medical Statistics: A Systematic Review on Missing Data Handling in Electronic Health Records. *Health Data Sci.* **2024**, *4*, 0176. doi: 10.34133/hds.0176.

7. Abassi, R.A.; Msengwa, A.S. Classification of breast cancer recurrence based on imputed data: a simulation study. *BioData Min.* **2022**, *15*, 30. doi: 10.1186/s13040-022-00316-8.

8. Memon, S.M.Z.; Wamala, R.; Kabano, I.H. A comparison of imputation methods for categorical data. *Inform. Med. Unlocked* **2023**, *42*, 101382. doi: 10.1016/j.imu.2023.101382.

9. Gabr, M.I.; Helmy, Y.M.; Elzanfaly, D.S. Effect of Missing Data Types and Imputation Methods on Supervised Classifiers: An Evaluation Study. *Big Data Cogn. Comput.* **2023**, *7*, 55. doi: 10.3390/bdcc7010055.

10. Tiwaskar, S.; Rashid, M.; Gokhale, P. Impact of machine learning-based imputation techniques on medical datasets: a comparative analysis. *Multimed. Tools Appl.* **2025**, *84*, 5905–5925. doi: 10.1007/s11042-024-19103-0.

11. Ma, L.; et al. Enhancing early gestational diabetes mellitus prediction with imputation-based machine learning framework: A comparative study on real-world clinical records. *Digit. Health* **2025**, *11*, 20552076251352436. doi: 10.1177/20552076251352436.

12. Digitale, J.; Franzon, D.; Pletcher, M.J.; McCulloch, C.E.; Gennatas, E.D. Methods for Addressing Missingness in Electronic Health Record Data for Clinical Prediction Models: Comparative Evaluation. *JMIR Med. Inform.* **2025**, *13*, e79307. doi: 10.2196/79307.

13. Zhang, S.; Zhang, Z.; Zhou, Y.; Hong, S.; Liu, H. Comparison of Imputation Strategies for Incomplete Electronic Health Data. *Health Data Sci.* **2026**, Article 0493. doi: 10.34133/hds.0493.

14. Nyakundi, N.G.; Ndiritu, J.; Mwaniki, I.J.; Kamanu, T.K. Machine Learning-Based Imputation for Breast Cancer Prediction: Evaluating Performance Under Complex Missing Data Mechanisms. *AppliedMath* **2026**, *6*, 134. doi: 10.3390/appliedmath6080134.

15. Li, R.; Shen, Z.; Wu, C.; Li, J.; Tian, Y. Prediction Pipeline Selection for Incomplete Clinical Data via Missingness Fingerprints and Instance Augmentation. *Bioengineering* **2026**, *13*, 497. doi: 10.3390/bioengineering13050497.

16. Roy, S.S.; et al. Mission imputable: Effects of missing data processing on infectious disease detection and prognosis. *PLoS ONE* **2026**, *21*, e0320105. doi: 10.1371/journal.pone.0320105.

17. Connors, A.F., Jr.; et al. A controlled trial to improve care for seriously ill hospitalized patients: The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). *JAMA* **1995**, *274*, 1591–1598. doi: 10.1001/jama.1995.03530200027032.

18. Knaus, W.A.; et al. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Ann. Intern. Med.* **1995**, *122*, 191–203. doi: 10.7326/0003-4819-122-3-199502010-00007.

19. Pedregosa, F.; et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

20. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. doi: 10.1023/A:1010933404324.

21. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016; pp. 785–794. doi: 10.1145/2939672.2939785.

22. Bengio, Y.; Grandvalet, Y. No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.* **2004**, *5*, 1089–1105.

23. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. doi: 10.1186/s12916-019-1466-7.

24. Vanderbilt University Department of Biostatistics. Datasets: SUPPORT study datasets. HBiostat. Available online: https://hbiostat.org/data/ (accessed on 7 September 2026).