Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Does the Effect of Missing-Data Imputation on Classification Performance Depend on Feature Scaling? A Factorial Study Across Classifiers and Clinical Benchmark Datasets

Charles R. Haruna
University of Cape Coast
Collins Amoo
camoo@stu.ucc.edu.gh
University of Cape Coast

Research Article

Keywords: Missing data, Imputation, Feature scaling, Classifier, Cross-validation, Clinical prediction models, Biomedical

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11674176/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Does the Effect of Missing-Data Imputation on Classification Performance Depend on Feature Scaling? A Factorial Study Across Classifiers and Clinical Benchmark Datasets

***Charles R. Haruna*[1]*, Collins Amoo*[2]**

Department of Computer Science and Information Technology, University of Cape Coast, Cape Coast, Ghana[1]

Department of Data Science and Economic Policy, University of Cape Coast, Cape Coast, Ghana [2]

Corresponding author: Collins Amoo [email: camoo@stu.ucc.edu.gh]

Abstract

Background Missing data are common in clinical datasets, and imputation performance depends on the dataset and downstream classifier. Feature scaling is usually fixed rather than evaluated experimentally. We investigated whether imputation effects on classification performance depend on scaling and classifier.

Methods We used the Wisconsin Breast Cancer Diagnostic (WBCD) and Pima Indians Diabetes datasets, simulating missingness under a missing-at-random (MAR) mechanism at a target rate of 20%. Mean, median, and MICE-style iterative single imputation were crossed with three scaling strategies (none, standard, robust) and three classifiers (logistic regression, random forest, radial basis function support vector machine [SVM]). Iterative imputation produced one completed dataset per run without pooling using Rubin's rules. Imputers and scalers were fitted exclusively on training partitions. Five-fold stratified cross-validation repeated five times, with three missingness masks per split, yielded 75 matched split-mask units and 2,025 evaluations per dataset. ROC-AUC was the primary outcome. Analysis used sum-coded ordinary least-squares models with split-cluster-robust covariance and paired comparisons with cluster-bootstrap intervals. Inference was exploratory.

Results In WBCD, scaling and classifier choice dominated ROC-AUC variation. Scaling effects were largest for SVM (none versus standard: −0.0219; 95% CI −0.0267 to −0.0175), smaller for logistic regression (−0.0034), and absent for random forest. Marginal imputation differences were at most 0.0009; the largest conditional paired difference was approximately 0.0017. In Pima, SVM performance depended on the imputation-scaling combination: median imputation with robust scaling yielded mean ROC-AUC of 0.757, compared with 0.799 using standard scaling. Joint imputation × scaling and three-way interaction tests were significant in both datasets ($p<0.01$), although WBCD interactions were practically negligible.

Conclusions Imputation effects should be assessed jointly with scaling and classifier choice, particularly for distance-based models. Clinical prediction benchmarks should report scaling and evaluate it as an experimental factor. Findings remain conditional on the two datasets, simulated missingness, and fixed classifier settings; overlapping cross-validation training sets limit inference. Pima also contained pre-existing zero-coded missing values. The second dataset assessed consistency across prediction tasks rather than independent external clinical validation. These results do not establish a universally preferred pipeline or fully account for uncertainty across multiple imputations.

Keywords Missing data; Imputation; Feature scaling; Classifier; Cross-validation, Clinical prediction models; Biomedical

Background

Missing data are common in clinical and epidemiological datasets and can compromise statistical power, precision, and analytical validity (Schober & Vetter, 2020). Rubin (1976) distinguished missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Under MAR, missingness depends on observed variables. This is a common imputation assumption, although it cannot generally be verified from observed data alone (Schober & Vetter, 2020). Complete case analysis may introduce bias when data are not MCAR, making imputation important in clinical prediction workflows (Pan & Chen, 2023; Schober & Vetter, 2020). However, predictive performance also depends on subsequent preprocessing and modelling.

No imputation method performs best across all settings. Garciarena and Santana (2017) examined interactions among missing-data types, imputation methods, and classifiers, supporting pipeline-oriented benchmarking. Pan and Chen (2023) found that imputation did not consistently outperform listwise deletion and cautioned against generalizing beyond similar datasets. Systematic reviews recommend considering missingness mechanism, pattern, and proportion and call for standardized benchmarking (Afkanpour et al., 2024; Ren et al., 2024). Comparative studies also identify variable correlations and the amount of missingness as important influences on performance (Chen & Chang, 2024; Digitale et al., 2025).

Clinical studies illustrate this variability. Liu and Gopalakrishnan (2017) reported similar performance for imputation-augmented and unaugmented models in pediatric cardiac imaging. Li et al. (2024) found lower AUC values for SVM models trained on imputed cohort data (0.668–0.777) than on complete data (0.804). Joel et al. (2025) identified MissForest and multiple imputation by chained equations as the most accurate of seven techniques across three healthcare datasets at missingness rates of 10%–25%. Outside clinical prediction, Bilicioglu Gunes and Zeybekoglu Yesil (2026) crossed missingness mechanisms and rates with item parameter drift in test equating, demonstrating the relevance of jointly evaluating design factors.

Although imputation–classifier interactions have been studied, feature scaling is often fixed. Tougui et al. (2021), for example, used imputation, normalization, and standardization as fixed pipeline components. Scaling matters for distance-sensitive and gradient-based models, whereas tree-based algorithms generally do not require it (Koçak, 2022). Mean and median imputation concentrate replacement values at a single point, changing distributions from which subsequent scalers estimate their parameters. Imputation comparisons may therefore depend on scaling. To the best of our knowledge, the reviewed studies have not explicitly evaluated imputation × scaling as an experimental factor. Fixing the scaler could conceal conditional differences and restrict the applicability of benchmark conclusions to other pipelines.

This study aimed to evaluate the effects of imputation, scaling, and classifier choice, and their interactions, on ROC-AUC under simulated MAR conditions. A full factorial design crossed mean, median, and MICE-style iterative single imputation with no scaling, standard scaling, and robust scaling, and with logistic regression, random forest, and an RBF-kernel SVM. The Wisconsin Breast Cancer Diagnostic and Pima Indians Diabetes datasets were analysed separately to examine consistency across prediction tasks, not external clinical validation. The research questions were: (1) Does the effect of imputation on ROC-AUC depend on scaling? (2) Does this dependence differ among linear, tree-based, and kernel-based classifiers? (3) Are findings consistent across breast cancer classification and diabetes

prediction? The contribution is to treat scaling as an experimental factor, identifying conditional preprocessing effects rather than an overall ranking of imputation methods.

Methods

Study aim, design and setting

The aim was to determine whether the effect of missing-data imputation on classification performance depended on feature scaling and classifier choice, and whether these patterns were consistent across two clinical prediction tasks. This was a computational benchmark experiment using the publicly available Wisconsin Breast Cancer Diagnostic (WBCD) and Pima Indians Diabetes datasets, rather than a prospective clinical intervention study. The datasets were analysed separately; the second dataset provided a consistency check across prediction tasks, not external validation in an independent clinical population.

A full factorial design crossed three imputation methods (mean, median, and MICE-style iterative single imputation), three feature-scaling strategies (none, standard, and robust), and three classifiers (logistic regression, random forest, and an RBF-kernel support vector machine). All 27 conditions were evaluated on the same cross-validation splits and missingness masks within each dataset, allowing matched comparisons. The iterative condition used scikit-learn's IterativeImputer to produce one completed dataset per run, without pooling using Rubin's rules. ROC-AUC was the primary outcome, and F1-score was secondary. The computational analyses were performed in Python using pandas, NumPy, scikit-learn (Pedregosa et al., 2011), and statsmodels (Seabold & Perktold, 2010).

Experimental processes and comparisons

Within each dataset, the workflow comprised stratified splitting, separate missingness generation in the training and validation partitions, imputation, feature scaling where specified, classifier fitting, and validation scoring. Each imputer was fitted using training data only, and the fitted imputer transformed both partitions. Each scaler was then fitted on the imputed training data and applied to the imputed validation data. Classifiers were trained on the resulting training predictors and evaluated on the corresponding validation predictors. The same split-mask unit was used for all conditions, so comparisons changed the imputation method, scaler, or classifier while retaining the data partition and missingness configuration.

Simulation of missing data

Missing data were simulated under a MAR mechanism at a target missingness rate of 20%, so that the preprocessing methods could be compared in a controlled incomplete-data setting. Because the true mechanism cannot be established from the observed data alone (Schober & Vetter, 2020), the simulated mechanism was defined explicitly. Missingness depended on an observed predictor through a coefficient of 0.8. In WBCD, radius_mean was the driving variable, and missingness was introduced into ten other predictors that were selected reproducibly with a predefined random seed. In Pima, Age was the driving variable, and missingness was introduced into the other seven predictors; the existing zero-coded missing values were handled as missing data during preprocessing. For each cross-validation split, three missingness masks (seeds 3001, 3002, and 3003) were generated, and missingness was generated

separately for the training and validation partitions with split-dependent seeds. The realized missingness rates were recorded.

Cross-validation and leakage control

Performance was estimated with five-fold stratified cross-validation repeated five times (seed 2028). Stratification preserved the outcome class distribution in the training and validation partitions, and repetition was used to obtain a more stable estimate than a single run (Tougui et al., 2021). Both datasets contain one record per person, so record-wise splitting did not place records of the same subject in different partitions; the subject-wise problem described by Tougui et al. (2021) therefore did not apply to the present data.

The imputers and the scalers were fitted on the training partition of each split and then applied to the corresponding validation partition. Performing preprocessing steps such as imputation on the complete dataset before the split is a recognized form of data leakage that produces overoptimistic performance estimates (Kapoor & Narayanan, 2023), and scaling on the whole dataset is a recognized pitfall (Koçak, 2022). Grouping the preprocessing steps and the classifier into a single pipeline is a way to avoid this problem (Tougui et al., 2021).

The three masks were evaluated for each of the 25 splits, which yielded 75 split-mask units. Each unit was evaluated under all 27 combinations, which resulted in 2,025 model evaluations per dataset. Comparisons between conditions were therefore made within the same split-mask unit, which reduced the influence of the data partition and of the simulated missingness on the estimated differences. The random seeds were: master seed 2026, missingness generation 2027, cross-validation 2028, imputation 2029, and model initialization 2030.

Imputation methods

Mean and median imputation replaced each missing value with the mean or median of the corresponding predictor, calculated exclusively from the training observations. Both are single-imputation methods that are straightforward to implement but may distort the variance and correlation structure of the data (Li et al., 2024; Schober & Vetter, 2020).

The third imputation condition used a MICE-style iterative single imputation procedure based on chained conditional updates, implemented using scikit-learn's IterativeImputer (Pedregosa et al., 2011). The algorithm iteratively estimates missing values in each incomplete predictor using the remaining predictors. The implementation used the default Bayesian ridge regression estimator, a maximum of 40 iterations, a convergence tolerance of 10^{-3}, and random seed 2029. Posterior sampling was not enabled. One completed dataset was generated per experimental run; therefore, this procedure did not constitute full multiple imputation, and no pooling of estimates using Rubin's rules was performed.

For the Pima dataset, the initial imputation strategy was set to the median and complete columns were also processed (skip_complete=False), whereas the WBCD implementation used the default initial mean imputation strategy. The fitted imputer was trained exclusively on the training partition and subsequently applied to the corresponding validation partition to prevent information leakage. MICE-style iterative single imputation was used to evaluate predictive performance within the factorial classification framework, not to perform statistical inference that accounts for uncertainty across multiple imputations.

Feature-scaling strategies

In the no-scaling condition, the imputed predictors remained on their original scales. Standard scaling centered each predictor with its training mean and divided it by its training standard deviation. Robust scaling used the training median and the interquartile range, which makes it less sensitive to extreme values.

Classification algorithms

Logistic regression used the liblinear solver, a maximum of 10,000 iterations, and seed 2030. Random forest used 100 trees and seed 2030. The SVM used an RBF kernel with C=1 and gamma="scale". SVM probability estimation was not requested; its ROC-AUC was calculated from decision-function values. These classifiers represented linear, tree-based, and kernel-based models with differing sensitivity to predictor scale (Koçak, 2022). Hyperparameters were fixed rather than tuned across conditions. Comparisons therefore evaluated preprocessing effects under these specified model settings, rather than the performance of separately optimized pipelines.

Performance metrics

ROC-AUC measures the ability of a classifier to rank positive observations above negative observations across all thresholds and was selected as the primary metric because it does not depend on a threshold (Koçak, 2022). The F1-score, the harmonic mean of precision and recall, was calculated at a probability threshold of 0.50 for logistic regression and random forest and at a decision-function threshold of 0 for the SVM. It was interpreted as a secondary metric because it depends on the threshold, and the thresholds are not equivalent across classifiers.

Statistical analysis

A full factorial ordinary least-squares (OLS) model of ROC-AUC (classifier × imputation × scaling, with all factors treated as categorical and sum-coded) was fitted using cluster-robust covariance estimation (CRV1), with the cross-validation split as the clustering unit. Joint Wald tests were used to assess the main effects and interaction terms. The matched experimental structure was also used for paired comparisons of classifiers, imputation methods, and scaling strategies, including comparisons of scaling strategies within each classifier. The effect size was the mean paired difference in ROC-AUC. For each contrast, ROC-AUC differences were first averaged over the other experimental factors within each matched split-mask unit and then averaged across the three missingness masks within each cross-validation split, yielding 25 split-level differences. Two-sided 95% percentile confidence intervals were obtained from 5,000 bootstrap resamples of these split-level differences, using a NumPy random generator initialized with seed 20261021. Exploratory p-values were obtained from one-sample t-tests of the same 25 split-level differences against zero. Holm-adjusted p-values (Holm, 1979) were calculated separately within each factor's pairwise comparison family for the main-effect analyses and across the corresponding sets of subgroup contrasts for the stratified analyses. All inferential results were treated as exploratory.

The 2,025 evaluations per dataset were not independent observations: conditions shared split-mask units, and training sets overlapped across cross-validation splits. Clustering by split accounted for part of this dependence, but not for overlapping training sets, and only 25 clusters were available. Split-level bootstrap intervals and paired tests likewise did not fully resolve this dependence. All inferential results

were therefore exploratory; interpretation emphasized paired effect sizes, uncertainty intervals, and consistency rather than p-values alone (Pan & Chen, 2023). Robustness was assessed by repeating the key summaries and contrasts within each of the five cross-validation repeats and each of the three missingness-mask seeds.

Results

Realized missingness

The simulated missingness was close to the 20% target. In WBCD, the mean realized rates were 0.2003 in the training partitions (range 0.1881 to 0.2108) and 0.1999 in the validation partitions (range 0.1640 to 0.2482). In Pima, the corresponding values were 0.1993 (range 0.1883 to 0.2129) and 0.2001 (range 0.1681 to 0.2484). The full design was complete in both datasets (75 units, 25 splits, and 2,025 evaluations).

Wisconsin Breast Cancer Diagnostic dataset

Overall performance. Mean ROC-AUC across conditions was high. Joint tests of the factorial effects are reported in Table 1. The mean ROC-AUC was 0.9926 for logistic regression, 0.9892 for random forest, and 0.9871 for the SVM (Fig. 1). Averaged over the other factors, the mean ROC-AUC was 0.9841 without scaling, 0.9923 with robust scaling, and 0.9925 with standard scaling, and it was 0.9894, 0.9893, and 0.9902 for mean, median, and MICE-style iterative single imputation, respectively. The corresponding F1-scores were 0.918, 0.954, and 0.957 for no, robust, and standard scaling. The sum-coded model explained 37.4% of the variance in ROC-AUC.

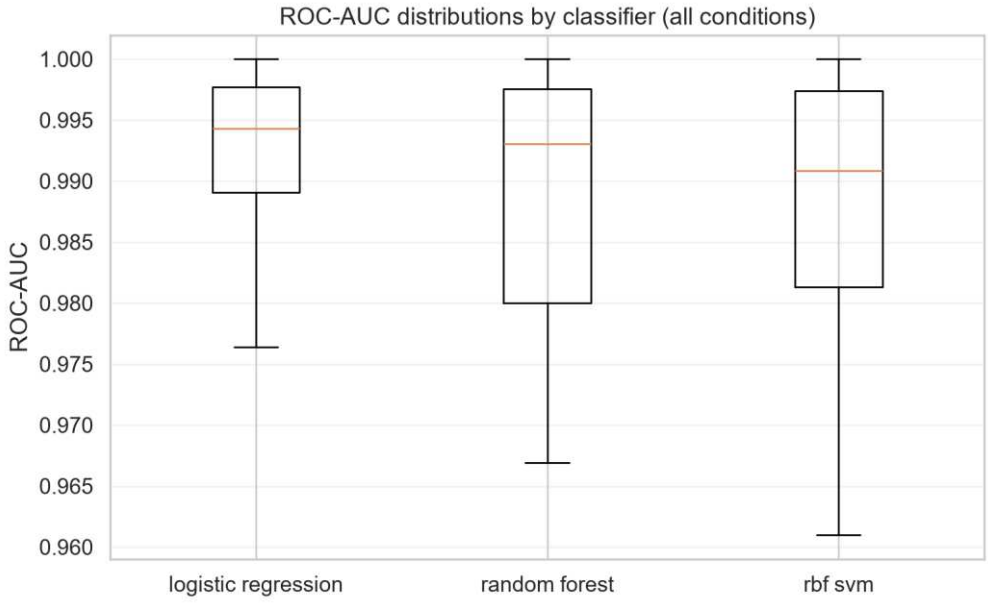


Fig. 1. ROC-AUC distributions by classifier in WBCD. Box plots show ROC-AUC for logistic regression, random forest, and RBF SVM across all conditions. WBCD = Wisconsin Breast Cancer Diagnostic dataset; RBF SVM = radial basis function support vector machine.

Table 1. Joint cluster-robust Wald tests for ROC-AUC in WBCD and Pima

Term	df	WBCD χ^2	WBCD p	Pima χ^2	Pima p
Classifier	2	63.15	<0.001	60.59	<0.001
Imputation	2	13.78	0.001	66.99	<0.001
Scaling	2	74.79	<0.001	41.04	<0.001
Classifier × imputation	4	1.04	0.904	48.24	<0.001
Classifier × scaling	4	79.35	<0.001	53.78	<0.001
Imputation × scaling	4	25.07	<0.001	87.41	<0.001
Classifier × imputation × scaling	8	25.02	0.002	110.59	<0.001

Note. WBCD = Wisconsin Breast Cancer Diagnostic dataset; df = degrees of freedom; χ^2 = Wald test statistic. Bold text identifies column headings. Tests are exploratory, with 25 split clusters and overlapping training sets.

Scaling comparisons within classifiers. The classifier × scaling interaction was significant in the exploratory joint test (χ^2=79.35, df=4, p<0.001; Table 1). Differences between scaling strategies for random forest were 0.00001 or less. For logistic regression, standard scaling increased ROC-AUC relative to no scaling by 0.0034 (95% CI 0.0018 to 0.0049), and robust scaling increased it by 0.0032 (95% CI 0.0014 to 0.0048). For the SVM, mean ROC-AUC was 0.9726 without scaling, 0.9945 with standard scaling, and 0.9941 with robust scaling. The paired differences were −0.0219 for none − standard (95% CI −0.0267 to −0.0175) and −0.0216 for none − robust (95% CI −0.0264 to −0.0169; Fig. 2). Standard–robust comparison intervals included zero for each classifier. Selected paired contrasts and Holm-adjusted p-values are presented in Table 2.

Table 2. Paired ROC-AUC differences between scaling strategies within classifiers

Dataset	Classifier	Contrast	Mean difference	95% CI	Holm p
WBCD	Logistic regression	None − standard	−0.0034	−0.0049 to −0.0018	0.002
WBCD	Random forest	None − standard	−0.0000	−0.0000 to 0.0000	1.000
WBCD	RBF SVM	None − standard	−0.0219	−0.0267 to −0.0175	<0.001
WBCD	RBF SVM	Robust − standard	−0.0004	−0.0009 to 0.0002	0.936
Pima	Logistic regression	None − standard	−0.0113	−0.0157 to −0.0067	0.001
Pima	Random forest	None − standard	0.0000	−0.0003 to 0.0003	1.000
Pima	RBF SVM	None − standard	−0.0129	−0.0192 to −0.0063	0.005
Pima	RBF SVM	None − robust	0.0026	−0.0047 to 0.0103	1.000
Pima	RBF SVM	Robust − standard	−0.0155	−0.0227 to −0.0086	0.002

Note. WBCD = Wisconsin Breast Cancer Diagnostic dataset; RBF SVM = radial basis function support vector machine; CI = confidence interval. Bold text identifies column headings. Negative differences favor the second strategy listed. Intervals are two-sided 95% percentile bootstrap intervals using split-level resampling. Holm p denotes the adjusted exploratory p-value for the corresponding set of stratified contrasts. Values displayed as 0.0000 or −0.0000 are rounded.

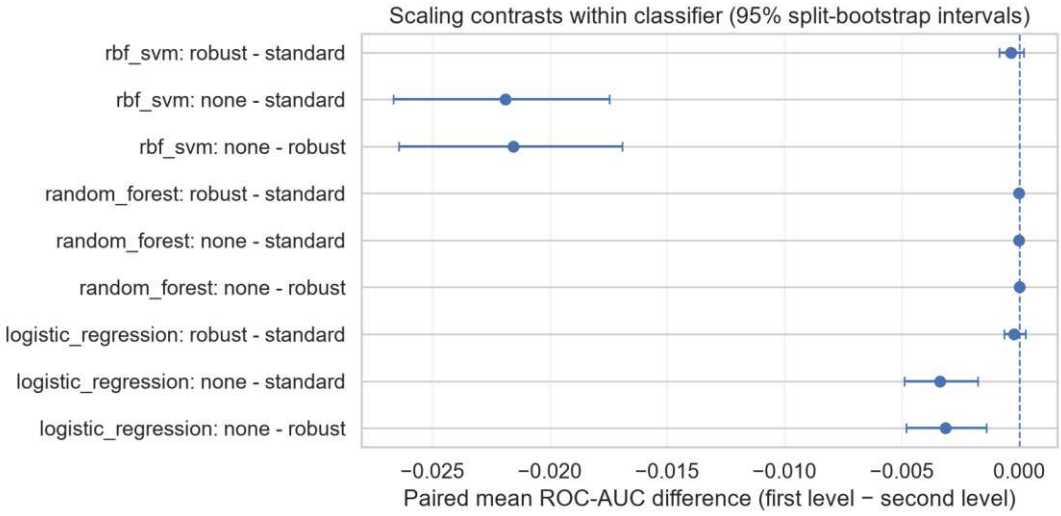


Fig. 2. Paired scaling contrasts within each classifier in WBCD. Points show paired mean ROC-AUC differences (first level minus second level) between scaling strategies (none, standard, robust) for logistic regression, random forest, and RBF SVM. Horizontal bars show 95% split-level bootstrap intervals, and the dashed vertical line marks zero difference. WBCD = Wisconsin Breast Cancer Diagnostic dataset; RBF SVM = radial basis function support vector machine.

Imputation effects and interactions. The imputation main effect was significant in the exploratory joint test (χ^2=13.78, df=2, p=0.001; Table 1). MICE-style iterative single imputation exceeded mean imputation by 0.0008 (95% CI 0.0003 to 0.0014) and median imputation by 0.0009 (95% CI 0.0005 to 0.0015). The mean − median difference was 0.0001 (95% CI −0.0001 to 0.0004). The classifier × imputation interaction was not significant (p=0.904), whereas imputation × scaling (χ^2=25.07, df=4, p<0.001) and the three-way interaction (χ^2=25.02, df=8, p=0.002; Fig. 3) were significant. Conditional paired imputation differences were 0.0017 or less. The largest occurred for the SVM with robust scaling: MICE-style iterative single imputation exceeded median imputation by 0.0017 (95% CI 0.0012 to 0.0023; Holm p<0.001). In a sensitivity analysis treating the 2,025 evaluations as independent, the imputation main effect was F=2.50 (p=0.082), and all interaction p-values exceeded 0.98; the corresponding clustered imputation test gave p=0.001.

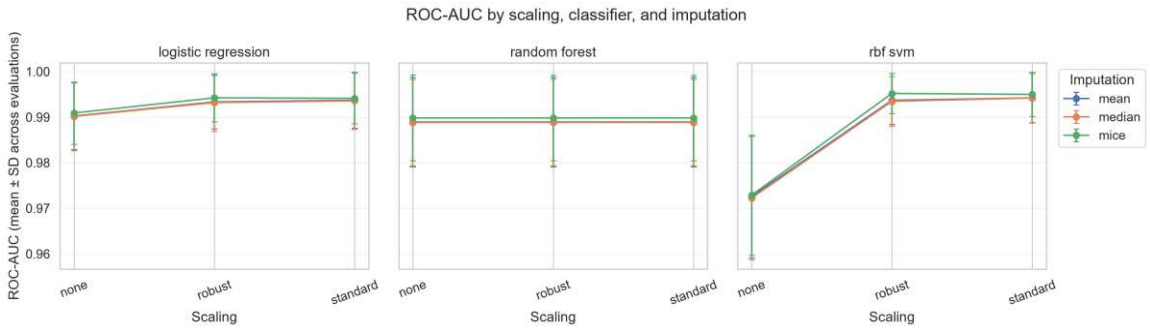


Fig. 3. ROC-AUC by scaling strategy, classifier, and imputation method in WBCD. Panels show logistic regression, random forest, and RBF SVM, and lines show mean, median, and MICE-style iterative single imputation (labelled mice in the plot). Points are mean ROC-AUC and error bars show the standard deviation across evaluations. WBCD = Wisconsin Breast Cancer Diagnostic dataset; RBF SVM = radial basis function support vector machine.

Highest and lowest observed condition means. The highest mean ROC-AUC was obtained with the SVM, MICE-style iterative single imputation, and robust scaling (0.9952, SD 0.0044; Fig. 4), followed by the SVM with iterative imputation and standard scaling (0.9950) and logistic regression with iterative imputation and robust scaling (0.9943). The lowest means occurred with the unscaled SVM across the three imputation methods (0.9723 to 0.9729). Averaged over imputation and scaling, logistic regression exceeded random forest by 0.0034 (95% CI 0.0011 to 0.0061) and the SVM by 0.0056 (95% CI 0.0043 to 0.0070). Random forest exceeded the SVM by 0.0022 (95% CI 0.0001 to 0.0040; Holm p=0.047). The SVM had the lowest classifier-level mean when averaged across scaling strategies, but the highest condition mean when scaling was applied.

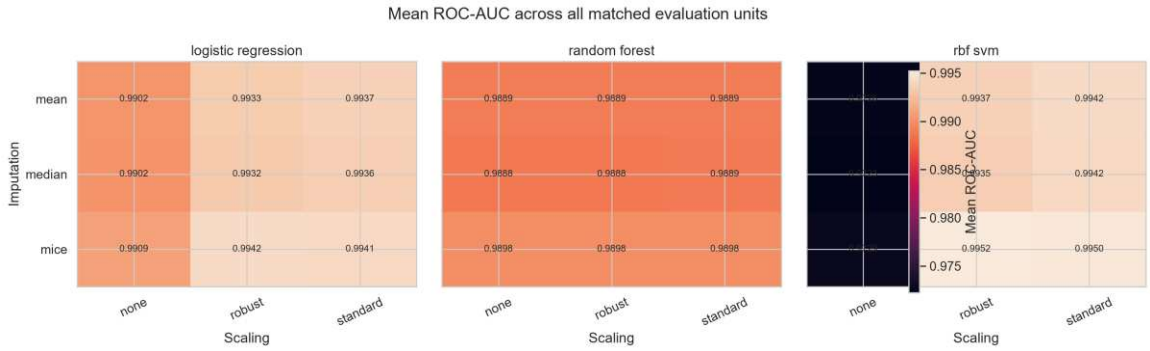


Fig. 4. Mean ROC-AUC by classifier, imputation method, and scaling strategy in WBCD. Heatmaps show mean ROC-AUC across all matched evaluation units for each combination of imputation method (rows) and scaling strategy (columns), with one panel per classifier. MICE-style iterative single imputation is labelled mice in the plot. WBCD = Wisconsin Breast Cancer Diagnostic dataset; RBF SVM = radial basis function support vector machine.

Robustness. The best condition was SVM with MICE-style iterative single imputation and robust scaling under all three mask seeds, and in four of the five repeats (logistic regression with MICE-style iterative single imputation and robust scaling was the best in repeat 1). The difference between the best and the worst condition ranged from 0.0228 to 0.0231 across mask seeds and from 0.0218 to 0.0251 across

repeats. The gain from standard scaling for the SVM was 0.0216 to 0.0221 across mask seeds and 0.0209 to 0.0241 across repeats, and for logistic regression it was between 0.0020 and 0.0046 in repeats 1 to 4.

Pima Indians Diabetes dataset

Overall performance. Mean ROC-AUC was 0.8074 for logistic regression, 0.7994 for random forest, and 0.7916 for the SVM (Fig. 5). Averaged over the other factors, the means were 0.7958 without scaling, 0.7987 with robust scaling, and 0.8039 with standard scaling. Imputation-level means were 0.8009 for mean, 0.7956 for median, and 0.8019 for MICE-style iterative single imputation. Within-condition standard deviations across split-mask units were approximately 0.03 to 0.05. The sum-coded model explained 8.5% of the variance in ROC-AUC. All main effects and interactions were significant in the exploratory joint tests (all $p<0.001$; Table 1). Condition-specific means and standard deviations are reported in Table 3 and Fig. 6.

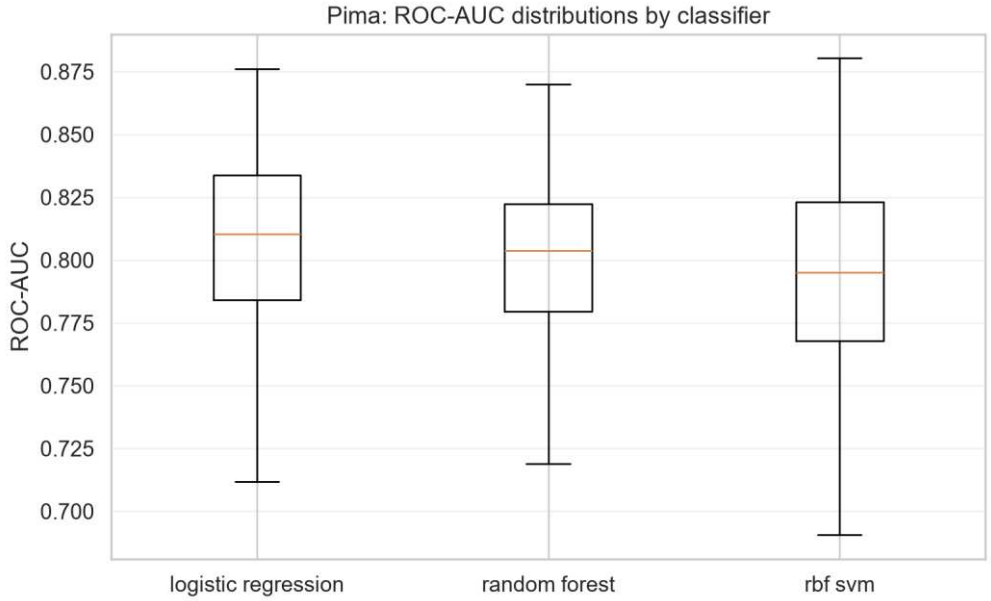


Fig. 5. ROC-AUC distributions by classifier in Pima. Box plots show ROC-AUC for logistic regression, random forest, and RBF SVM. RBF SVM = radial basis function support vector machine.

Table 3. Pima ROC-AUC by classifier, imputation method, and scaling strategy

Classifier	Imputation	No scaling	Standard	Robust
Logistic regression	Mean	0.8005 (0.0325)	0.8121 (0.0334)	0.8120 (0.0334)
Logistic regression	Median	0.7995 (0.0332)	0.8110 (0.0338)	0.8109 (0.0340)
Logistic regression	MICE-style iterative single imputation	0.7997 (0.0358)	0.8106 (0.0336)	0.8107 (0.0339)
Random forest	Mean	0.7989 (0.0325)	0.7988 (0.0326)	0.7991 (0.0324)
Random forest	Median	0.7988 (0.0327)	0.7989 (0.0327)	0.7983 (0.0327)
Random forest	MICE-style iterative single imputation	0.8007 (0.0351)	0.8007 (0.0350)	0.8006 (0.0350)

RBF SVM	Mean	0.7913 (0.0400)	0.7999 (0.0320)	0.7957 (0.0445)
RBF SVM	Median	0.7872 (0.0397)	0.7987 (0.0333)	0.7572 (0.0505)
RBF SVM	MICE-style iterative single imputation	0.7859 (0.0371)	0.8045 (0.0342)	0.8038 (0.0349)

Note. Values are mean ROC-AUC with standard deviation (SD) in parentheses, based on 75 split-mask units per cell. RBF SVM = radial basis function support vector machine. Bold text identifies column headings. MICE-style iterative single imputation produced one completed dataset per run without pooling using Rubin's rules.

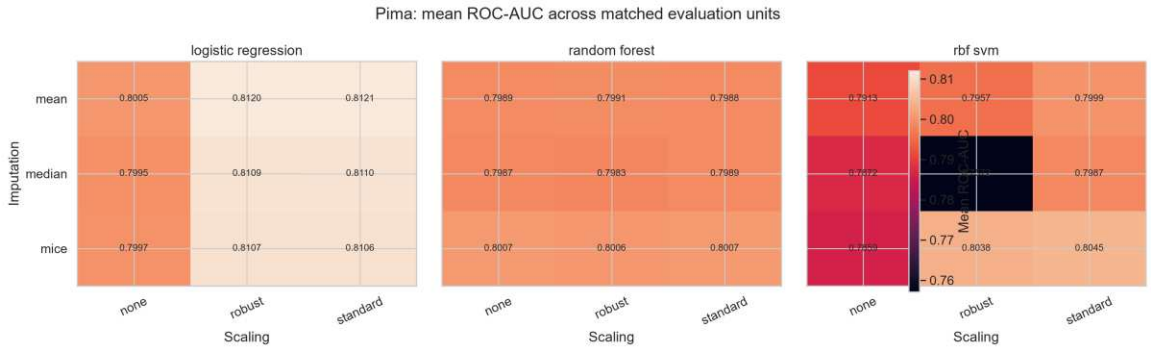


Fig. 6. Mean ROC-AUC by classifier, imputation method, and scaling strategy in Pima. Heatmaps show mean ROC-AUC across matched evaluation units for each combination of imputation method (rows) and scaling strategy (columns), with one panel per classifier. MICE-style iterative single imputation is labelled mice in the plot. RBF SVM = radial basis function support vector machine.

Scaling comparisons within classifiers. For random forest, the none − standard paired difference was 0.0000 (95% CI −0.0003 to 0.0003; Holm p=1.000; Table 2). For logistic regression, standard scaling increased ROC-AUC relative to no scaling by 0.0113 (95% CI 0.0067 to 0.0157), and robust scaling increased it by 0.0113 (95% CI 0.0067 to 0.0156); the standard–robust comparison did not differ significantly. For the SVM, standard scaling increased ROC-AUC relative to no scaling by 0.0129 (95% CI 0.0063 to 0.0192). The none − robust difference was 0.0026 (95% CI −0.0047 to 0.0103), and the robust − standard difference was −0.0155 (95% CI −0.0227 to −0.0086; Holm p=0.002).

Imputation effects and interactions. Averaged over the other factors, MICE-style iterative single imputation exceeded median imputation by 0.0063 (95% CI 0.0037 to 0.0093), and mean imputation exceeded median imputation by 0.0053 (95% CI 0.0041 to 0.0065). The mean − iterative imputation difference was −0.0010 (95% CI −0.0035 to 0.0012). The imputation × scaling joint test gave χ^2=87.41 (df=4, p<0.001), and the three-way interaction gave χ^2=110.59 (df=8, p<0.001; Table 1; Fig. 7).

Conditional imputation comparisons. The SVM with median imputation and robust scaling had mean ROC-AUC of 0.7572 (SD 0.0505; minimum 0.6204), with 11 of 75 evaluations below 0.70 (Table 3). Under this classifier–scaler combination, median imputation was lower than mean imputation by 0.0385 (95% CI 0.0285 to 0.0493; Holm p<0.001) and lower than MICE-style iterative single imputation by 0.0467 (95% CI 0.0344 to 0.0592; Holm p<0.001). In the other SVM conditions and in logistic regression and random forest, absolute paired imputation differences were 0.0057 or less, and most confidence intervals included zero.

Highest observed condition mean and robustness. Logistic regression with mean imputation and standard scaling had the highest mean ROC-AUC (0.8121). Logistic regression had the highest classifier-level mean under every mask seed and repeat. The highest-performing imputation–scaling combination varied across mask seeds and repeats, involving mean or MICE-style iterative single imputation and standard or robust scaling. The difference between the highest and lowest condition means ranged from 0.0514 to 0.0595 across mask seeds and from 0.0471 to 0.0609 across repeats.

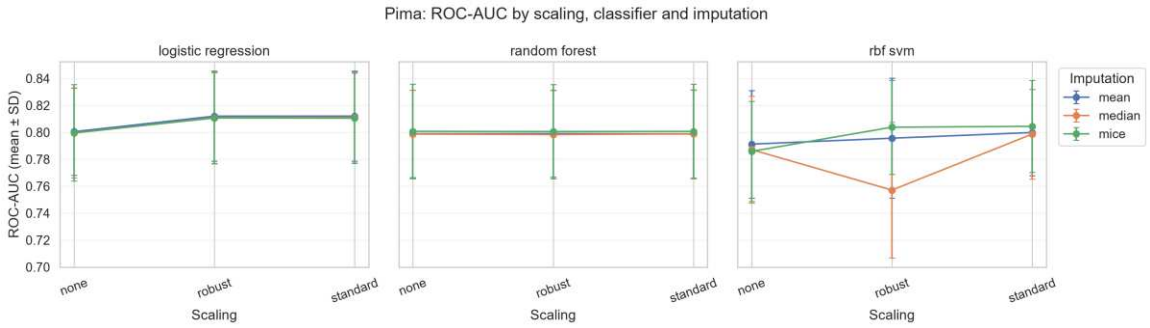


Fig. 7. ROC-AUC by scaling strategy, classifier, and imputation method in Pima. Panels show logistic regression, random forest, and RBF SVM, and lines show mean, median, and MICE-style iterative single imputation (labelled mice in the plot). Points are mean ROC-AUC and error bars show the standard deviation. RBF SVM = radial basis function support vector machine.

Secondary outcome: F1-score. In WBCD, F1-scores followed the same scaling ordering as ROC-AUC: means were 0.918 without scaling, 0.954 with robust scaling, and 0.957 with standard scaling. In Pima, the SVM mean F1-score was 0.562 with standard scaling, 0.465 without scaling, and 0.378 with robust scaling. These results were obtained using the prespecified thresholds of 0.50 for logistic regression and random forest probabilities and 0 for the SVM decision function.

Discussion

Imputation performance depended on the scaling strategy and classifier, but the practical importance of these interactions differed between datasets. Conditional imputation differences were small in WBCD. In Pima, the SVM with median imputation and robust scaling had substantially lower ROC-AUC than the same classifier–scaler combination with mean or iterative imputation. These findings favor evaluating preprocessing combinations rather than ranking imputation methods in isolation.

The scaling patterns are consistent with research on scale-sensitive models (Koçak, 2022). Random forest showed near-zero differences, whereas standard scaling improved logistic regression and SVM performance. Predictor ranges may affect the distances used by the RBF kernel, while regularization and optimization may contribute to the logistic regression pattern. These mechanisms were not tested, and the findings are conditional on the fixed hyperparameters. Nonsignificant standard–robust differences in WBCD do not establish scaler equivalence.

The small WBCD imputation effects align with reports of limited downstream benefit in some clinical settings (Liu & Gopalakrishnan, 2017) and evidence that dataset structure and missingness influence comparative performance (Chen & Chang, 2024; Digitale et al., 2025). The slight marginal advantage of iterative imputation does not establish general superiority. Comparisons with favorable chained-equations results (Joel et al., 2025) must account for implementation differences: this study used single,

not multiple, imputation. Li et al. (2024) reported larger AUC differences among imputed SVM models, but the present design cannot explain the difference in magnitude.

In Pima, median imputation may have concentrated values at the median and altered the interquartile ranges used by robust scaling, affecting RBF-SVM distances. This hypothesis requires direct examination of imputed distributions, scaler parameters, and feature distances. Pre-existing missingness of unknown origin also limits attribution to the simulated MAR mechanism.

The study extends context-dependent imputation benchmarking (Afkanpour et al., 2024; Garciarena & Santana, 2017; Pan & Chen, 2023; Ren et al., 2024) by varying scaling rather than fixing preprocessing as in Tougui et al. (2021). The Pima SVM comparison illustrates the relevance: median imputation was only slightly lower than the alternatives with standard scaling but markedly lower with robust scaling. A fixed-scaler benchmark could obscure this distinction.

Benchmarks should report the imputer, scaler, preprocessing order, and classifier settings, and test whether imputation comparisons persist across alternative scalers for scale-sensitive models. The Pima finding warrants checking imputed distributions and robust-scaling parameters. These recommendations concern benchmark design, not clinical deployment, because calibration and clinical utility were not evaluated.

Matched comparisons and training-only preprocessing strengthen the design and reduce preprocessing leakage (Kapoor & Narayanan, 2023; Koçak, 2022), but do not eliminate cross-validation dependence. Differences between independence-assuming and clustered analyses show sensitivity to its treatment. Interpretation should emphasize effect sizes, uncertainty, and consistency across repeats and masks: statistical detectability does not make the small WBCD gains practically important, and the larger Pima contrast requires independent replication.

Limitations

First, only two small clinical benchmark datasets were evaluated. The second provided a consistency check, not external clinical validation, limiting generalizability to other populations.

Second, the simulation used one MAR mechanism, driven by one predictor with a coefficient of 0.8, and a 20% target missingness rate. MCAR, MNAR, other rates, and other patterns were not examined (Schober & Vetter, 2020). Pima also contained pre-existing zero-coded missingness of unknown origin, so its results do not represent an exclusively simulated MAR setting.

Third, iterative imputation generated one completed dataset per run without posterior sampling or Rubin's rules pooling. Findings therefore concern this single-imputation implementation, not full multiple imputation or inference accounting for imputation uncertainty (Vuong et al., 2025). K-nearest neighbors, random forest-based imputation, and native missing-value handling were not evaluated (Digitale et al., 2025; Joel et al., 2025; Li et al., 2024).

Fourth, fixed classifier hyperparameters limit the findings to the specified settings; tuning could alter scaling sensitivity. Fifth, F1-scores used fixed thresholds (0.50 for probability outputs and 0 for SVM decision values) that were neither equivalent across classifiers nor optimized for these tasks.

Sixth, evaluations shared split-mask units and overlapping training sets. Split clustering, bootstrap intervals, and paired tests addressed only part of this dependence, with 25 clusters. Intervals may understate uncertainty, so inference remains exploratory and requires independent replication.

Finally, the proposed mechanism for the Pima median-imputation–robust-scaling result was not tested. Only discrimination and threshold-dependent F1 were assessed; calibration and clinical utility were not evaluated (Koçak, 2022). The findings establish neither the interaction's cause nor clinical benefit.

Conclusions

Under the conditions evaluated, the effect of missing-data imputation on classification performance depended on feature scaling and classifier choice. Imputation differences were small in WBCD, whereas the Pima dataset showed a substantial conditional reduction in SVM performance when median imputation was followed by robust scaling. Scaling effects were classifier-specific, with little change in random forest discrimination. These findings do not identify a universally preferred imputation method or scaler; they demonstrate why preprocessing combinations should be evaluated rather than imputation methods ranked in isolation.

The study's importance lies in treating feature scaling as an experimental factor in imputation benchmarking. This approach makes conditional performance differences visible that a fixed-scaler comparison may obscure. For clinical prediction research, the findings support reporting the complete preprocessing pipeline and assessing whether imputation comparisons remain consistent across alternative scaling strategies, particularly for scale-sensitive classifiers. The contribution is methodological rather than evidence of clinical benefit. Conclusions remain conditional on the two benchmark datasets, the simulated missingness and pre-existing Pima missingness, and the fixed classifier settings. Because cross-validation dependence limits inference, independent replication and evaluation in other missingness scenarios and clinical populations are needed.

Declarations

Ethics approval and consent to participate

Not applicable. The study used publicly available, de-identified benchmark datasets.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are available in the imputation_x_scaler repository on GitHub, https://github.com/Kolinx25/imputation_x_scaler.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no funding.

Authors' contributions

Charles R. Haruna (CRH) wrote the manuscript and reviewed the documents. Collins Amoo (CA) performed the analysis and wrote the Results and Discussion sections.

Acknowledgements

Thanks Prof. Stephen Moore for his encouragement to pursue the research idea that led to this study.

References

Afkanpour, M., Hosseinzadeh, E., & Tabesh, H. (2024). Identify the most appropriate imputation method for handling missing values in clinical structured datasets: A systematic review. *BMC Medical Research Methodology, 24*, Article 188. https://doi.org/10.1186/s12874-024-02310-6

Bilicioglu Gunes, A., & Zeybekoglu Yesil, S. (2026). Examining the joint impact of missing data mechanisms and item parameter drift on the accuracy of item response theory-based test equating. *PLOS ONE, 21*(7), Article e0353665. https://doi.org/10.1371/journal.pone.0353665

Chen, C.-Y., & Chang, Y.-W. (2024). Missing data imputation using classification and regression trees. *PeerJ Computer Science, 10*, Article e2119. https://doi.org/10.7717/peerj-cs.2119

Digitale, J., Franzon, D., Pletcher, M. J., McCulloch, C. E., & Gennatas, E. D. (2025). Methods for addressing missingness in electronic health record data for clinical prediction models: Comparative evaluation. *JMIR Medical Informatics, 13*, Article e79307. https://doi.org/10.2196/79307

Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications, 89*, 52–65. https://doi.org/10.1016/j.eswa.2017.07.026

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Joel, L. O., Doorsamy, W., & Paul, B. S. (2025). A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-025-00825-9

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), Article 100804. https://doi.org/10.1016/j.patter.2023.100804

Koçak, B. (2022). Key concepts, common pitfalls, and best practices in artificial intelligence and machine learning: Focus on radiomics. *Diagnostic and Interventional Radiology, 28*(5), 450–462. https://doi.org/10.5152/dir.2022.211297

Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., Ding, Y., Li, Y., Jian, L., Cheng, J., & Guo, H. (2024). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research Methodology, 24*, Article 41. https://doi.org/10.1186/s12874-024-02173-x

Liu, Y., & Gopalakrishnan, V. (2017). An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. *Data, 2*(1), Article 8. https://doi.org/10.3390/data2010008

Pan, S., & Chen, S. (2023). Empirical comparison of imputation methods for multivariate missing data in public health. *International Journal of Environmental Research and Public Health, 20*(2), Article 1524. https://doi.org/10.3390/ijerph20021524

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., &

Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Ren, W., Liu, Z., Wu, Y., Zhang, Z., Hong, S., Liu, H., & MINDER Group. (2024). Moving beyond medical statistics: A systematic review on missing data handling in electronic health records. *Health Data Science, 4*, Article 0176. https://doi.org/10.34133/hds.0176

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Schober, P., & Vetter, T. R. (2020). Missing data and imputation methods. *Anesthesia & Analgesia, 131*(5), 1419–1420.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 92–96).

Tougui, I., Jilbab, A., & El Mhamdi, J. (2021). Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthcare Informatics Research, 27*(3), 189–199. https://doi.org/10.4258/hir.2021.27.3.189

Vuong, M.-L., Verbeke, G., & Dreesen, E. (2025). A multiple imputation workflow for handling missing covariate data in pharmacometrics modeling. *CPT: Pharmacometrics & Systems Pharmacology*. https://doi.org/10.1002/psp4.70039