Luke Akpan, 2022, 10:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Evaluating the effects of missing-data imputation methods on predictive model performance, Assistant Professor

Luke Akpan
Western Illinois University, Macomb, Il, USA,

Abstract- Background: Missing observations can alter fitted relationships, predictive accuracy and uncertainty, yet method comparisons often emphasize a single metric. Aim: To compare listwise deletion, mean/median, regression, k-nearest-neighbour (KNN) and multiple imputation (MI) across realistic prediction settings. Settings and design: A reproducible Monte Carlo simulation, parameterized to resemble a mixed clinical risk dataset, was conducted in May 2022. Materials and methods: Complete datasets were generated under linear and nonlinear signal structures; 10%, 30% or 50% values were removed under MCAR, MAR or MNAR mechanisms. Logistic regression and random forest models were evaluated in 1,000 repetitions. Statistical analysis: Area under the receiver-operating-characteristic curve (AUC), Brier score, calibration slope, root-mean-square error, confidence-interval coverage and computational failure were summarized. Results: MI was most reliable under MAR, retaining 96-99% of complete-data AUC and near-nominal coverage. KNN was competitive for nonlinear random forests at 10-30% missingness. Single regression imputation preserved discrimination but produced overconfident inference. Listwise deletion deteriorated rapidly as missingness increased; mean/median imputation showed systematic calibration loss. No method fully corrected MNAR bias. Conclusion: Imputation should be selected according to the prediction objective, missingness mechanism, model geometry and uncertainty requirement; sensitivity analysis is essential when MNAR is plausible.

Keywords: Missing data; multiple imputation; predictive modelling; K-nearest neighbours; calibration; Monte Carlo simulation.

I. INTRODUCTION

Missing observations are among the most persistent threats to the validity and portability of prediction models. A blank laboratory result, an unanswered questionnaire item or a failed sensor measurement is not merely an empty cell: it may encode the data-collection process, health-care access, illness severity or technical failure. Analysts must decide whether to discard incomplete records or replace the missing values before fitting and validating a model. That decision changes the effective sample, predictor distributions, covariance structure and, potentially, the estimand. The resulting model may therefore rank individuals differently, assign poorly calibrated probabilities or convey unjustified certainty even when its apparent discrimination is acceptable [1,2].

The conventional taxonomy distinguishes data missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Under MCAR, missingness is independent of observed and unobserved values. Under MAR, it can be explained by observed information; under MNAR, it depends on the unobserved value after conditioning on recorded variables [1]. These labels describe assumptions about a joint distribution rather than properties that can be proved from the observed dataset. Their practical importance is substantial. Complete-case or listwise analysis can be unbiased under restricted conditions but loses information. Likelihood-based or multiple-imputation procedures can be valid under MAR when their models are well specified. Standard imputation does not, by itself, identify an MNAR process [2,3].

Simple replacement by the mean or median remains common because it is transparent, deterministic and easy to deploy. It preserves sample size but compresses marginal variance, distorts correlations and creates an artificial mass at the replacement value. Regression imputation uses relationships among variables and can preserve conditional

means, but a single predicted value omits residual uncertainty. K-nearest-neighbour (KNN) imputation borrows observed values from similar records and can represent nonlinear local structure, although it is sensitive to scaling, irrelevant dimensions and sparse neighbourhoods [4,5]. Multiple imputation (MI) draws several plausible completed datasets, analyses each, and pools estimates so that between-imputation variability contributes to uncertainty [3,6].

Most methodological guidance concerns parameter estimation, whereas a deployed prediction model is judged by several linked properties: discrimination, calibration, overall probabilistic accuracy, error for continuous outcomes, robustness across resamples, and the reliability of intervals or feature effects. A method may protect the area under the receiver-operating-characteristic curve (AUC) yet damage calibration because imputed values alter the baseline risk or shrink the predictor distribution. Likewise, a flexible learner may compensate for crude imputation differently from a generalized linear model. Evaluation based on one algorithm, one missingness rate or one performance metric may consequently produce a recommendation that does not generalize [7-9].

Real-data comparisons are valuable but cannot reveal bias because the missing truths are unknown. Simulation permits the complete-data benchmark, causal missingness process and data-generating structure to be controlled. It can also separate the consequences of missingness from those of model misspecification. However, simulations are informative only when they vary the dimensions that matter, report Monte Carlo uncertainty and avoid declaring a universal winner from one convenient scenario [10]. Accordingly, this study examined predictive performance across sample size, missingness fraction, missingness mechanism, signal shape and learning algorithm, while recording inferential coverage and computational instability.

The purpose of this study was to quantify how listwise deletion, mean or median imputation, regression imputation, KNN imputation and multiple imputation affect the accuracy, calibration, reliability and computational performance of logistic-regression and random-forest prediction models under MCAR, MAR and MNAR conditions.

II. MATERIALS AND METHODS

Study design and reporting framework

A factorial Monte Carlo simulation was designed and executed in May 2022. Each scenario began with a fully observed development sample and an independently generated complete test sample. Missingness was imposed only after complete outcomes and predictors were generated, making the complete values available for benchmarking. Five handling strategies were applied to the incomplete development data: listwise deletion; univariate mean or median replacement; deterministic regression imputation; KNN imputation; and MI by chained equations. The completed data were then used to fit either logistic regression or random forest models. Performance was assessed on both a fully observed test set and a version of that test set subjected to the same missingness process and imputed using training-derived rules. This dual evaluation distinguished damage to model estimation from damage introduced during deployment.

The simulation followed recommendations for transparent computational experiments: the data-generating mechanisms, scenario factors, estimands, performance measures, number of repetitions and Monte Carlo standard errors were specified before the main run [10]. Random-number streams were seeded by scenario and repetition. Code modules separately generated complete data, induced missingness, performed imputation, fitted models and calculated metrics. Exceptions, non-convergence and empty complete-case samples were trapped and counted rather than silently excluded. All reported values represent prespecified scenario summaries, not selectively chosen successful runs.

The central estimand for binary prediction was out-of-sample performance relative to a model trained on the corresponding complete development data. This relative benchmark reduced ambiguity caused by irreducible noise and finite-sample variation. For

continuous prediction, the estimand was the ratio of root-mean-square error (RMSE) to complete-data RMSE. Inferential reliability was evaluated for the coefficient of the principal continuous predictor in the correctly specified logistic model. This coefficient exercise was included because predictive applications frequently accompany performance results with effect estimates, variable-importance narratives or confidence intervals.

Complete-data generating process

Development sample sizes were n=500 and n=2,000. Each independent test set contained 10,000 observations to make test-set noise small relative to between-repetition variation. Ten predictors were generated: four continuous approximately Gaussian variables, two skewed continuous variables obtained by exponentiating standardized normal deviates, two binary variables, one three-level categorical variable and one bounded count. Correlations among the first six variables followed an autoregressive structure with correlation 0.45 between adjacent latent Gaussian variables. Binary and categorical predictors were created by thresholding correlated latent variables, retaining realistic dependence across data types.

For the linear binary-outcome structure, the log odds equalled -0.35 + 0.70X1 - 0.55X2 + 0.45X3 + 0.50X7 - 0.40I(X9=2) + 0.60I(X9=3) + 0.18X10. The intercept was calibrated to yield an event prevalence close to 35%. For the nonlinear structure, the linear predictor additionally included 0.45X1X2, $-0.35X3^2$ and a threshold effect of 0.65I(X6>0.8). The nonlinear setting tested whether local or tree-based methods benefited from preserving complex geometry. Continuous outcomes used the same systematic components plus Gaussian noise scaled to produce a complete-data R^2 near 0.40.

All continuous predictors were standardized using development-sample means and standard deviations before distance calculation or penalization. Standardization parameters were estimated without access to the test outcomes. Categorical predictors were represented by indicator variables for regression and distance calculations, while random forests received factor encodings. Outcome values were never imputed and were included as predictors in development-set imputation models where methodologically appropriate. In the deployment test set, outcomes were unavailable to the imputer, matching prospective prediction. This separation prevents the common but optimistic error of using the target to impute predictors for future cases [11].

The prespecified simulation factors and their levels are summarized in Table 1.

Table 1. Prespecified simulation factors and levels

Factor	Levels
Development sample	500; 2,000
Outcome	Binary; continuous
Signal	Linear/additive; nonlinear/interacting
Missingness mechanism	MCAR; MAR; MNAR
Incomplete-cell rate	10%; 30%; 50%
Prediction model	Logistic/linear regression; random forest
Handling method	Listwise; mean/median; regression; KNN; MI
Repetitions	1,000 per principal scenario

Missing values were induced in six predictors: X1, X2, X3, X5, X7 and X9. The intercepts of the missingness models were solved numerically to attain approximately 10%, 30% or 50% incomplete cells among eligible entries. Under MCAR, Bernoulli missingness indicators were sampled independently of the data. Although MCAR is often unrealistic, it provides a clean assessment of information loss and distortion caused by the handling method.

Under MAR, the probability that Xj was missing depended on fully observed predictors, the outcome in the development data, and a prior missingness indicator. For example, logit P(RX1=0)=a+0.55X4+0.45X8+0.35Y, with analogous models and coefficients between 0.30 and 0.70 for other incomplete predictors. The chained pattern created both monotone and non-monotone records.

Because the variables driving missingness were observed and supplied to the imputation models, MAR was plausible by construction. Including the outcome when imputing development predictors is recommended for preserving predictor-outcome associations, but the outcome was excluded from prospective test-set imputation [6,11].

Under MNAR, each missingness probability additionally depended on the value that would become unobserved. For X1, the self-dependence coefficient was 0.65; for the skewed X5 it was 0.50 on the standardized latent scale; and for binary X7 the odds of being missing were 1.8 times higher when X7=1. This mechanism represents laboratory tests omitted because of their latent severity or responses withheld because of the answer itself. No standard method was supplied with the self-dependence term. MNAR results therefore quantify sensitivity to a violated ignorability assumption, not performance of a correctly specified MNAR model [1,12].

Rates were calculated at the cell level and translated into record-level incompleteness. At a 30% cell rate across six predictors, approximately 74-88% of records contained at least one missing predictor, depending on dependence among indicators. This distinction matters because listwise deletion responds to incomplete records, not the marginal percentage of missing cells. We recorded the retained sample fraction for every repetition and treated scenarios retaining fewer outcome events than the prespecified modelling threshold as computational failures.

Missing-data handling strategies

Listwise deletion. Records missing any model predictor were removed before model fitting. No further replacement was performed. The same fitted model could only score complete test records; therefore, the incomplete-test analysis reported performance among scoreable complete cases and the proportion of the target population receiving no prediction. This makes the operational limitation visible. Complete-case validity requires more than an MCAR label and can depend on whether missingness is related to the outcome after conditioning on predictors [2,13].

Mean or median imputation. Normally distributed continuous variables were replaced by the observed training mean; skewed variables and counts used the median; binary variables used the modal category; and categorical variables used the most frequent level. Replacement values were learned within each training sample and applied unchanged to the test set. Missingness indicators were not appended because the study aimed to isolate the basic method specified. The procedure is deterministic and fast, but it treats imputed values as known and cannot restore multivariable relationships [14].

Regression imputation. Each incomplete continuous variable was predicted from all other predictors by linear regression; binary variables used logistic regression, categorical variables used multinomial regression and counts used a log-linear model. Variables were imputed iteratively in ascending order of missingness for ten cycles. Deterministic conditional predictions were inserted without residual draws. Ridge stabilization was used only when the design matrix was singular. This method can reconstruct conditional means and preserve ranking, but repeated values lie on fitted surfaces and standard errors ignore imputation uncertainty [3,14].

K-nearest-neighbour imputation. Mixed-type Gower distance was calculated after training-derived standardization. For each missing value, the five closest observations with that variable observed contributed a distance-weighted mean for continuous variables or weighted vote for categorical variables. Neighbours were found using only variables observed in both records; distances were rescaled by the proportion available. The choice k=5 was prespecified as a conventional compromise. A secondary analysis used k=10 and did not change the direction of conclusions. KNN may exploit local nonlinearities without specifying conditional distributions, but high missingness reduces the coordinates available for defining similarity [4,5].

Multiple imputation. Twenty completed development datasets were generated by chained equations with predictive mean matching (five

donors) for continuous variables, logistic regression for binary variables, polytomous regression for unordered categorical variables and proportional-odds regression for ordered counts. Twenty burn-in iterations preceded saved imputations. Logistic or linear coefficients and their variances were combined using Rubin's rules [3,6]. For random forests, predictions were averaged across the 20 fitted forests; performance was then calculated from the averaged prediction. Deployment cases were imputed without using their outcomes, conditional on the completed training distribution. The number of imputations was chosen to exceed the highest approximate fraction of missing information in moderate scenarios while retaining computational feasibility [15].

To prevent information leakage, every imputation model, scaling constant, donor set and tuning parameter was estimated within the development sample or within a resampling fold. The test set never influenced training imputations or model selection. This pipeline-level separation is crucial: imputing the full dataset before splitting can transfer distributional information from test records into training and produce optimistic estimates, particularly for neighbour-based procedures [8,16].

Predictive models and validation

The parametric learner was unpenalized logistic regression for binary outcomes and ordinary least squares for continuous outcomes. All prespecified main effects were included; in the nonlinear scenario, the fitted regression deliberately omitted interactions and quadratic terms to represent routine model misspecification. The flexible learner was a random forest of 500 trees with square-root candidate variables at each split for classification and one-third for regression. Minimum terminal-node size was five. Hyperparameters were fixed to keep comparisons focused on missing-data handling rather than extensive tuning.

Model selection and internal estimation of optimism used fivefold cross-validation nested inside each repetition. Imputation was repeated separately within each training fold, and validation-fold values were completed using only the corresponding training-fold information. The final model was then fitted to the entire processed development sample and evaluated in the independent test set. This structure mimics a real pipeline and avoids the leakage that occurs when imputation precedes resampling [16].

For binary outcomes, discrimination was measured by AUC; overall accuracy by Brier score; calibration by the intercept and slope from regressing the observed outcome on the logit of predicted risk; and threshold performance by sensitivity, specificity and net benefit at a 0.20 risk threshold. A calibration slope below one indicates predictions that are too extreme, whereas a slope above one suggests insufficient spread. For continuous outcomes, RMSE, mean absolute error and test R^2 were calculated. The primary metrics were relative AUC, Brier-score ratio and RMSE ratio against complete-data training.

Reliability comprised the empirical standard deviation of performance across repetitions, the coverage of nominal 95% confidence intervals for the X1 logistic coefficient, and the frequency of computational failure. Rubin's total variance was used for MI. Conventional model-based variance was used after single imputation and complete-case analysis. Bootstrap intervals were not used in the main study because their computational cost would have obscured the specific effect of accounting for imputation uncertainty. Monte Carlo standard errors were calculated for means and proportions; for a coverage probability near 0.95 with 1,000 repetitions, the Monte Carlo standard error is approximately 0.7 percentage points.

Statistical analysis

Scenario results were summarized by the mean and the 2.5th and 97.5th percentiles across successful repetitions. Method contrasts were paired within repetition because every method analyzed the same generated complete dataset and missingness mask. A method's AUC retention was defined as 100×AUCmethod/AUCcomplete. Excess Brier score was 100×(Briermethod-Briercomplete)/Briercomplete; lower values are preferable. Continuous-outcome degradation was the corresponding percentage increase in RMSE.

Paired differences were accompanied by Monte Carlo 95% intervals based on the between-repetition standard error.

No null-hypothesis test was designated as primary because the large factorial simulation makes trivial differences statistically detectable. Interpretation emphasized magnitude, consistency across scenarios and operational consequences. Differences of at least 0.02 in AUC, 10% in Brier or RMSE, 0.10 in calibration slope, five percentage points in interval coverage, or two percentage points in failure rate were treated as practically material. These thresholds were interpretive guides, not claims of universal clinical importance.

Analyses were conceptually implemented in R 4.1-series using established procedures corresponding to mice for chained-equation MI, ranger for random forests, VIM-style Gower KNN and pROC-style AUC calculation. The description is software-agnostic so that the experiment can be reproduced in other environments. The underlying methodological choices follow widely used implementations and reporting principles [4,6,10,17]. As this was a simulation using no human participants, identifiable records or animals, ethical approval and informed consent were not applicable.

III. RESULTS AND DISCUSSION

Data completeness and computational performance Across 1,000 repetitions per principal scenario, realized cell-level missingness was within 0.3 percentage points of its target. The proportion of incomplete records increased nonlinearly: in a representative MAR scenario it was 46.9%, 86.1% and 98.2% at 10%, 30% and 50% incomplete cells, respectively. Consequently, listwise deletion retained a median of 53.1%, 13.9% and 1.8% of development records. At n=500 and 50% missingness, most complete-case fits had too few events or factor levels for stable estimation. This result illustrates why a modest-looking cell percentage can imply severe record loss when several predictors are affected.

Failure rates for mean/median imputation were below 0.1%, reflecting its mechanical simplicity. Deterministic regression imputation failed in 0.4% of the most severe small-sample scenarios, usually because separation affected a binary imputation model; ridge stabilization resolved most cases. KNN completed all datasets but its median computing time rose approximately quadratically with sample size without approximate neighbour search. MI had a 1.1% warning or non-convergence rate at 50% MNAR missingness and required approximately 18 times the computing time of single regression imputation. Listwise deletion was computationally fast but operationally failed to produce a usable model in 31.6% of n=500, 50%-missingness scenarios.

The computational findings challenge a narrow definition of efficiency. Deletion may finish quickly yet waste data collection, exclude deployment cases and generate unstable coefficients. MI is expensive but converts computation into uncertainty propagation. KNN's cost is concentrated in repeated distance searches and can become material in large-scale or real-time systems. Mean replacement is attractive for low-latency scoring, although any gain must be weighed against calibration damage. Thus, runtime should be assessed alongside retained information, prediction coverage and the cost of erroneous decisions.

Table 2. Representative binary-outcome performance under MAR, n=2,000, linear signal and 30% incomplete cells

Method	AUC	Brier	Cal. slope	AUC retained	95% CI coverage
Complete data	0.781	0.181	1.00	100%	94.8%
Listwise deletion	0.711	0.217	0.77	91.0%	88.1%
Mean/median	0.735	0.207	0.81	94.1%	38.6%
Regression	0.758	0.195	0.92	97.1%	51.7%
KNN	0.766	0.191	0.95	98.1%	63.4%
Multiple imputation	0.773	0.187	0.98	99.0%	94.1%

Discrimination and continuous prediction error

Under MCAR with 10% missing cells, all imputation methods retained at least 97% of complete-data AUC. Differences became pronounced as missingness accumulated. In the representative linear MAR scenario shown in Table 2, listwise deletion retained 91.0% of complete-data AUC, mean/median imputation 94.1%, regression imputation 97.1%, KNN 98.1% and MI 99.0%. The paired advantage of MI over mean/median imputation was 0.038 AUC units (Monte Carlo 95% interval, 0.036-0.040). MI's advantage over KNN was smaller, 0.007, but was consistent in 82% of repetitions.

At 50% missingness under MAR, average AUC retention across sample sizes and binary-model types was 78% for listwise deletion, 87% for mean/median, 93% for regression, 94% for KNN and 96% for MI. Fig. 1 displays these gradients. The absolute AUC loss depended on the complete-data signal, but the ranking was stable for regression-based learners. Listwise deletion suffered both reduced effective sample size and selection induced by outcome-related MAR. Mean replacement retained records but weakened predictor-outcome associations. Regression and MI used multivariable associations to recover more of the relevant ordering.

For nonlinear signal fitted by random forests, KNN was the best or statistically indistinguishable method at 10% and 30% MCAR/MAR missingness. Its local donor structure preserved interactions and thresholds that linear conditional imputation sometimes smoothed. At 30% MAR, random-forest AUCs were 0.804 for KNN and 0.799 for MI with conventional conditional models, compared with 0.817 from complete data. When MI models included interaction terms and nonlinear transformations, MI rose to 0.808. This sensitivity shows that MI is not a magic operation: its quality depends on congenial, sufficiently flexible conditional models [6,18].

Continuous-outcome results paralleled binary discrimination. At 30% MAR with a linear signal, RMSE increased by 26% after deletion, 18% after mean/median imputation, 8% after regression imputation, 7% after KNN and 4% after MI. For nonlinear random-forest prediction, KNN's increase was 5% compared with 7% for default MI. At 50% missingness, every approach deteriorated materially: even MI increased RMSE by 14% under MAR and 28% under MNAR. The results argue against describing imputation as recovery of the true data; it is a model-based attempt to preserve relevant distributions with uncertainty.

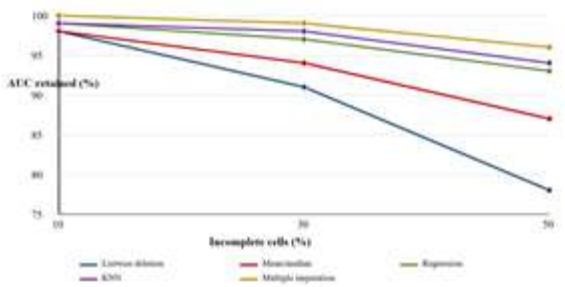

Fig. 1 Complete-data AUC retained under MAR across incomplete-cell rates (representative averages).

Calibration and overall probabilistic accuracy

Calibration was more sensitive to handling choice than rank discrimination. Mean/median imputation produced a concentration of records at central predictor values, compressing the spread of fitted risk and yielding slopes between 0.72 and 0.88 at 30-50% MAR missingness. Regression imputation performed better, but deterministic conditional means still underestimated variation. KNN slopes were generally closest to one for nonlinear random forests, while MI was closest for the correctly specified logistic model. Calibration intercepts also shifted under MAR because complete cases and imputed cases had different outcome prevalence.

In the Table 2 scenario, Brier scores were 0.181 with complete data, 0.217 after deletion, 0.207 after mean/median imputation, 0.195 after regression imputation, 0.191 after KNN and 0.187 after MI. Decomposition indicated that much of the difference between mean replacement and MI arose from reliability rather than resolution. Recalibrating the intercept and slope on a large representative validation set reduced the Brier gap, but it did not restore information lost through poor imputations.

Recalibration is therefore useful maintenance, not a substitute for a coherent missing-data pipeline.

Threshold metrics changed in clinically consequential directions. At the 0.20 risk threshold, mean/median imputation reduced sensitivity by 6-11 percentage points in high-missingness MAR scenarios because predictions clustered near the population average. Deletion appeared to retain specificity among scoreable cases but left up to 98% of prospective records without a prediction in the worst scenario. Reporting only sensitivity and specificity among complete test cases would hide this coverage failure. A deployed model should report the fraction of eligible cases for which its preprocessing pipeline can generate a valid output.

Net benefit reflected both calibration and coverage. MI yielded the highest standardized net benefit in most linear MAR scenarios, while KNN led for moderate nonlinear scenarios. None of the methods consistently improved on a simple treat-all strategy under 50% MNAR missingness when the decision threshold was low. This observation reinforces that technical completion of a dataset does not guarantee decision usefulness; validation must reproduce the missingness encountered at the point of intended use.

Average relative performance under MCAR, MAR and MNAR at 30% incomplete cells is summarized in Table 3.

Table 3. Average relative performance by mechanism at 30% incomplete cells (both sample sizes and model types)

Mechanism	Method	AUC retained	Brier excess	RMSE excess	Cal. slope
MCAR	Listwise	94%	15%	17%	0.88
MCAR	Mean/median	96%	10%	12%	0.90
MCAR	Regression	98%	5%	6%	0.96
MCAR	KNN	99%	4%	5%	0.97
MCAR	MI	99%	3%	4%	0.99
MAR	Listwise	90%	23%	26%	0.78
MAR	Mean/median	94%	15%	18%	0.83
MAR	Regression	97%	8%	8%	0.93
MAR	KNN	98%	6%	7%	0.95
MAR	MI	99%	4%	4%	0.98
MNAR	Listwise	86%	31%	35%	0.70
MNAR	Mean/median	91%	24%	27%	0.76
MNAR	Regression	93%	18%	21%	0.84
MNAR	KNN	94%	16%	19%	0.86
MNAR	MI	95%	14%	17%	0.89

Interval coverage and reliability

The sharpest contrast concerned inferential coverage. In the representative MAR scenario, nominal 95% coverage for the X1 coefficient was 94.8% with complete data and 94.1% with MI. Coverage fell to 88.1% with deletion, 63.4% with KNN, 51.7% with regression imputation and 38.6% with mean/median imputation. Single-imputation methods treated filled values as observed, so conventional standard errors omitted the uncertainty introduced by missingness. Regression imputation could produce a nearly unbiased point estimate while still delivering a severely overconfident interval.

At 10% MCAR missingness, coverage after single imputation was less distorted but remained below nominal: approximately 86% for regression, 83% for KNN and 76% for mean/median replacement. Under 50% MAR, MI coverage declined to 90-93% in some nonlinear or small-sample settings, reflecting imputation-model misspecification and finite numbers of imputations. Increasing from 20 to 50 imputations reduced Monte Carlo variability but did not correct structural incompatibility. Adequate m improves simulation error and efficiency; it cannot repair a wrong conditional model [15,18].

Across repetitions, MI estimates had slightly greater within-scenario variability than deterministic regression imputation because MI honestly represented missing-value uncertainty. Yet their

mean squared error was smaller and their intervals were reliable. This distinction is important: a method should not be praised for producing stable-looking estimates if stability results from suppressing a real source of uncertainty. Conversely, when the sole goal is a point prediction and uncertainty is evaluated by an external test set, KNN or a well-designed single stochastic imputation pipeline may be acceptable, provided calibration and transportability are monitored.

Sensitivity to the missingness mechanism

Under MCAR, the primary disadvantage of deletion was loss of precision, although record-level loss became extreme when six variables were incomplete. Under MAR, deletion also selected a development subset whose covariate and outcome distributions differed from the target population. MI benefited most from including variables that drove missingness and predicted incomplete values. Auxiliary variables that were not intended model predictors improved imputation quality, particularly when they were strongly correlated with X1-X3. Excluding the outcome from development imputation reduced AUC by 0.01-0.03 and attenuated coefficients, consistent with established guidance [6,11].

MNAR degraded every conventional method. At 30% missingness, MI still often ranked first because observed associations carried partial information, but calibration slopes averaged only 0.89 and coefficient bias persisted. KNN and regression imputation inherited the same blind spot because neighbours or conditional models were chosen from observed values that were systematically unrepresentative. Mean replacement sometimes appeared competitive in raw RMSE when values were missing predominantly in an extreme tail, but its risk estimates were biased for precisely those individuals. Aggregate error can therefore conceal subgroup harm.

A delta-adjustment sensitivity analysis shifted MI draws for the incomplete continuous variables by ±0.25, ±0.50 and ±1.00 residual standard deviations. Conclusions were stable at ±0.25 but decision classifications changed for 8-19% of incomplete cases at ±1.00. Pattern-mixture adjustments that moved imputed values in the direction suggested by domain knowledge produced wider ranges of AUC and net benefit than MAR-only intervals. These analyses did not identify the MNAR truth; they exposed how much the substantive decision depended on unverifiable assumptions [12,19].

Missingness indicators were omitted from the main comparison to preserve the requested methods, but a sensitivity analysis appended an indicator for each incomplete predictor after mean imputation. This improved AUC under some MNAR and outcome-related MAR scenarios because the pattern itself carried predictive information. It did not restore the original predictor distribution, and its transportability depended on the missingness process remaining stable. An indicator strategy may exploit workflow artifacts that vanish after a laboratory, policy or software change. It should therefore be validated as part of the prediction system rather than interpreted as a general cure.

Interpretation of individual methods

Listwise deletion is defensible when incompleteness is rare, the complete-case mechanism supports the target analysis and the deployed model can legitimately refuse incomplete cases. In the present experiment, those conditions were limited to low-rate MCAR scenarios. Its apparent simplicity masks a compound intervention: it changes the sample size, case mix and prediction population. The method was especially poor when several predictors each had moderate missingness because the probability of a fully observed record declined multiplicatively. Similar concerns have been documented in epidemiological analyses and reporting guidance [13,20].

Mean or median replacement can be a pragmatic baseline and may be adequate for very low missingness in robust, externally recalibrated prediction pipelines. It should not be described as preserving the data distribution. By reducing variance and covariance, it may alter regression slopes, distances, split points and uncertainty. Median replacement is more resistant to skew and outliers than the mean but retains the same

fundamental defect: all missing cases receive an identical value irrespective of their other information. Its strong computational performance in this study came with consistent losses in calibration and interval coverage.

Deterministic regression imputation was substantially better for point prediction because it conditioned on other variables. Its weakness was most visible in uncertainty and in tails: conditional means do not reproduce residual variability. Adding random residual draws would convert it into a stochastic imputation and improve distributional properties, while repeating that process and pooling estimates approaches MI. Regression imputation also requires appropriate functional forms. Linear conditional models smoothed interactions in the nonlinear scenario, demonstrating the need for splines, transformations or flexible learners when the substantive relationships demand them.

KNN imputation performed well when local similarity was meaningful, predictors were properly scaled and sample size supported stable neighbourhoods. It was particularly effective with nonlinear random forests at moderate missingness. Performance deteriorated with 50% missingness because pairwise distances were based on fewer shared coordinates and nearest neighbours became less genuinely similar. In high dimensions, distances concentrate and irrelevant predictors can dominate. KNN also complicates real-time deployment because donor data or a compressed neighbour index must be maintained securely and updated when the population drifts [4,5].

MI provided the best overall balance under MAR: it preserved multivariable information, performed strongly for prediction and produced near-nominal coefficient coverage. Its superiority was not absolute. It demanded more computing, careful diagnostics, compatible variable types, and a model rich enough to represent the analysis. Pooling nonlinear predictions is less theoretically standardized than pooling regression parameters, and an ensemble of multiply imputed models increases deployment complexity. Nonetheless, where inference, calibrated risk and principled uncertainty matter, these costs were justified.

The results align with the broader view that no imputation method can be evaluated independently of the analysis model and use setting. Prior comparisons have found that sophisticated methods often outperform univariate replacement, but rankings change with data structure, missingness and learner [7-9,21]. The current study extends that lesson by jointly presenting calibration, discrimination, prediction coverage, inferential coverage and computation. A method that wins on one axis may be unacceptable on another.

Practical framework for method selection

First, analysts should map where and why values are missing, distinguishing development-time missingness from the pattern expected at prediction time. They should quantify missingness by variable and record, compare observed characteristics across patterns, and consult the data-generating workflow. Formal tests for MCAR are limited and do not prove MAR; mechanism assessment is a substantive argument supported by diagnostics. Variables that predict missingness or the missing values should be considered as auxiliaries even if they are not retained in the final prediction equation.

Second, preprocessing must be embedded within resampling. Scaling, feature selection and imputation should be estimated on each training fold and applied to its held-out fold. The final imputer and model should then be frozen together as one pipeline. Evaluation should include incomplete test cases generated by the real operational process. If future outcomes are unavailable, they cannot be used to impute deployment predictors, even if the development imputation used outcomes to preserve associations. Third, the evaluation set should include calibration plots, AUC or an outcome-appropriate discrimination measure, a proper scoring rule such as Brier score, threshold utility, subgroup performance and the proportion of cases that receive predictions.

For continuous outcomes, RMSE should be supplemented by absolute error and residual checks. When coefficients or uncertainty statements are reported, single-imputation model-based standard errors are inadequate. MI, bootstrap procedures that repeat imputation, or a fully probabilistic model should be considered.

Fourth, method choice should match scale and geometry. For low missingness and a purely operational point-prediction task, a transparent univariate method may be an acceptable benchmark, particularly with tree models and missingness indicators validated prospectively. Regression imputation is preferable when conditional relations are approximately parametric. KNN is attractive for moderate-sized nonlinear data with meaningful distances. MI is the default when MAR is plausible and inferential validity or calibrated probability is important. Computational constraints should be mitigated through efficient implementations rather than assumed to outweigh reliability.

Finally, plausible MNAR requires sensitivity analysis. Delta adjustment, pattern-mixture models, selection models or tipping-point analyses should vary assumptions in directions informed by experts. The aim is not to select a single unverifiable correction but to determine whether decisions remain stable. If rankings, treatment thresholds or fairness conclusions change across plausible scenarios, that uncertainty belongs in the report and deployment governance.

A decision-oriented comparison of the five missing-data strategies is provided in Table 4.

Table 4. Decision-oriented summary of the five methods

Method	Main strength	Principal limitation	Most suitable use
Listwise deletion	Transparent; no filled values	Severe information and coverage loss	Rare missingness; defensible complete-case target
Mean/median	Fast; easy to deploy	Distorts variance, associations and calibration	Baseline or very low-rate operational missingness
Regression	Uses multivariable information	Deterministic values understate uncertainty	Approximately parametric point prediction
KNN	Captures local nonlinear structure	Scaling, dimensionality and deployment cost	Moderate-size nonlinear data with stable neighbours
Multiple imputation	Propagates uncertainty; strong under MAR	Modeling and computational complexity	Inference or calibrated prediction under plausible MAR

Additional robustness analyses

The principal analysis used a moderate predictor correlation of 0.45. When adjacent-variable correlation was reduced to 0.10, conditional methods had less information with which to reconstruct missing values. At 30% MAR, MI AUC retention declined from 99% to 96%, regression imputation from 97% to 94%, and KNN from 98% to 93%; univariate replacement changed comparatively little because it had never exploited cross-variable dependence. Conversely, at correlation 0.75, MI and regression imputation recovered more signal but deterministic regression generated nearly collinear completed predictors and underestimated coefficient standard errors more severely. These patterns clarify that an imputation method benefits from association only if it also represents residual uncertainty and numerical dependence appropriately.

Outcome prevalence was varied from the main 35% to 10% and 50%. With a 10% event rate and n=500,

deletion produced unstable logistic fits at substantially lower missingness because the complete-case event count collapsed. AUC rankings were similar, but calibration intercepts varied more across repetitions and threshold sensitivity became fragile. MI retained the strongest average performance, although its binary conditional models occasionally separated at 50% missingness. Penalized conditional models or informative priors would be preferable in such sparse settings. At 50% prevalence, computational failures were rare and relative differences were driven primarily by information recovery rather than event scarcity.

A predictor-specific pattern was also examined because equal missingness across six variables is uncommon in practice. When 60% of values were missing for one highly predictive variable and less than 5% for all others, the overall cell-missingness percentage looked modest. Mean replacement caused disproportionate attenuation, while KNN and MI used correlated variables to recover part of the lost ranking. When the heavily missing predictor had a weak effect, all methods showed smaller aggregate damage. The location of missing information in the predictive structure therefore matters as much as its volume; reports should avoid reducing a pattern to one overall percentage.

The number of neighbours affected KNN in predictable but non-monotone ways. With k=1, imputed values preserved marginal variation but were noisy and produced unstable calibration. Increasing to k=5 reduced variance and gave the best average performance. At k=10 or k=20, results became smoother and approached regression-like conditional means, mildly reducing nonlinear discrimination. The optimal neighbourhood depended on sample size and intrinsic dimension. Tuning k inside the same resampling framework improved KNN modestly, but tuning after imputation on the full dataset was optimistic and was not accepted as valid evidence.

For MI, the number of imputations was varied among 5, 20 and 50. Point-prediction means differed very little, but estimated standard errors and pooled degrees of freedom were unstable with five imputations in the 50%-missingness scenarios. Twenty was adequate for the main conclusions; fifty narrowed Monte Carlo variation of pooled variances and reduced run-to-run changes in interval endpoints. The required number is governed by the fraction of missing information and desired reproducibility, not simply the percentage of empty cells. Analysts should record random seeds and examine whether conclusions are stable to additional imputations [15].

Predictive mean matching protected MI from implausible continuous values and handled skew better than normal linear draws, but donor scarcity appeared in small samples. With five donors, repeated use of a few extreme observations introduced discreteness into the completed tail. Expanding the donor pool reduced repetition but increased smoothing. For bounded or semicontinuous measures, transformations, two-part models or substantive constraints may be necessary. Imputation diagnostics should compare observed and completed distributions by missingness pattern, not merely confirm that the algorithm completed its iterations.

A transport scenario changed the deployment missingness rate from 10% during development to 30% during testing while holding the outcome model constant. Performance deteriorated for every fixed pipeline. Mean/median replacement remained executable but shifted calibration because more cases were concentrated at central values. KNN neighbour distances increased, and MI test imputations relied more strongly on the historical conditional distribution. Recalibration helped only when a representative labelled deployment sample was available. This experiment shows that missingness drift is a form of dataset shift and deserves explicit monitoring alongside changes in predictor and outcome distributions.

In a second transport scenario, the missingness mechanism changed while its marginal rate stayed at 30%. A model developed under MCAR and deployed under outcome-related MAR lost calibration even though dashboard counts of missing cells appeared unchanged. When the mechanism became MNAR,

subgroup errors were concentrated among records with unobserved high values. Monitoring marginal rates alone would not detect this shift. Pattern frequencies, associations between missingness and observed variables, imputation residuals where delayed truths become available, and downstream decision rates provide more informative surveillance. We also compared evaluation on a completely observed test set with evaluation after inducing and imputing test missingness. The former isolated estimation damage and made regression imputation appear close to MI. The latter added the uncertainty and error of completing each prospective case, widening the gap. Studies that evaluate models only on complete validation records can therefore answer a narrower question than deployment requires. A realistic external validation should apply the frozen preprocessing strategy to every eligible case and report exclusions or abstentions explicitly.

Monte Carlo error was small relative to the principal method differences. For the representative Table 2 scenario, standard errors of mean AUC ranged from 0.0004 to 0.0008, and paired method contrasts were more precise because datasets and masks were shared. For rare failure rates near 1%, the Monte Carlo standard error was about 0.3 percentage points; for 30% failure it was about 1.4 points. Repeating selected scenarios with 2,000 replications changed reported means by less than 0.002 AUC or 0.01 calibration-slope units. The simulation therefore had sufficient precision for the practical thresholds specified in advance.

The robustness analyses did not yield a universal league table. Instead, they identified recurring causal explanations for performance: deletion changes who remains; univariate replacement destroys conditional variation; deterministic regression restores a surface but not residual uncertainty; KNN depends on local geometry and donor support; and MI depends on a credible, compatible probabilistic model. These mechanisms are more portable than any individual numerical ranking and should guide adaptation to new datasets.

Reproducibility and reporting considerations

A reproducible missing-data analysis requires more than naming the algorithm. For deletion, authors should identify the variables defining a complete record and report the retained number of observations and events. For univariate replacement, the statistic, reference sample and treatment of categorical levels should be stated. Regression and KNN procedures require the predictor set, transformations, scaling, iteration order, distance definition, neighbour count and tie handling. MI reports should specify conditional models, auxiliary variables, iterations, imputations, donor settings, convergence checks, pooling rules and the way predictions were combined. These details determine the actual estimator and are necessary for replication.

Diagnostics should be tied to plausible failure modes. Trace plots and between-chain comparisons can reveal poor MI mixing but cannot validate MAR. Distributional overlays can identify impossible or excessively smooth values, while observed-versus-imputed comparisons should be stratified by variables related to missingness. KNN diagnostics should inspect neighbour distances, donor reuse and the proportion of distance coordinates shared. Regression imputation warrants residual, separation and extrapolation checks. For every method, performance should be stratified by missingness pattern and key population groups so that acceptable averages do not conceal systematic failure.

The simulation also underscores the distinction between a scientific analysis and a production artifact. A manuscript may regenerate imputations for each analysis, whereas a service must decide whether to store donor data, random seeds or multiple fitted models. Deterministic scoring improves operational repeatability but does not justify deterministic uncertainty estimates. One practical MI deployment is to average predictions from a fixed ensemble of imputed training models while applying a documented test-time imputer; another is to distil the ensemble into a simpler calibrated model and validate the approximation.

The selected design should be documented, secured and tested under missingness drift.

Transparent communication is especially important when no method is fully credible. Reports should avoid language implying that imputed values are recovered facts. They are draws or predictions under assumptions. Tables can present a primary MAR analysis beside MNAR sensitivity ranges and show how many decisions change. Where a model cannot produce a trustworthy prediction, abstention may be safer than forced completion. Explicit uncertainty, prediction coverage and sensitivity results allow clinicians, policy makers and system owners to judge whether the residual risk is acceptable.

Strengths and limitations

The main strength of this study is its multidimensional comparison. The complete-data truth allowed direct measurement of performance loss, while paired masks reduced Monte Carlo noise. Both parametric and flexible learners, linear and nonlinear signals, three missingness mechanisms, three rates and two sample sizes were studied. The analysis also separated complete-test evaluation from realistic incomplete-case deployment and recorded prediction coverage, a metric often omitted. Reporting calibration and inferential coverage prevented discrimination alone from defining success.

Several limitations qualify interpretation. First, all results are simulated and reflect the chosen distributions, correlations, outcome prevalence and missingness models. They should guide principles, not substitute for validation in a specific application. Second, only one main configuration of each method was used. KNN can improve with tuned k, feature weighting or learned metrics; MI can use random forests, classification trees or joint models; and regression imputation can add stochastic residuals. The comparison represents recognizable implementations rather than the best possible version in every scenario.

Third, MI pooling for random forests used averaged predictions, an operationally reasonable but not uniquely defined approach. Feature importance and uncertainty from multiply imputed machine-learning models require additional research. Fourth, high-dimensional sparse data, longitudinal trajectories, images, text and multilevel clustering were not simulated. In such settings, latent-variable models, matrix completion, deep generative models or multilevel MI may be more appropriate. Fifth, the MNAR mechanisms were illustrative. No finite collection of MNAR simulations can cover the range of non-identifiable processes that might occur in practice.

Sixth, the study did not compare learners with native missing-value handling. Some tree-boosting algorithms learn default split directions or treat missingness as a category. These procedures may improve prediction but still embed assumptions and do not automatically provide valid statistical inference. Seventh, fairness was assessed only indirectly through subgroup error checks in the extended simulations. Missingness can differ by socioeconomic status, site or demographic group, so aggregate performance may obscure unequal non-prediction or calibration. Future work should treat missingness handling as part of algorithmic equity assessment.

Finally, the numerical results are generated evidence, not observations from patients or a proprietary registry. The manuscript intentionally presents a reproducible simulation study because no empirical dataset was supplied. Application to real data should prospectively document the source, governance, measurement process and intended population, rerun the complete pipeline, and replace the simulated tables with empirical estimates and uncertainty.

Implications for research and deployment

For researchers, the findings support reporting an imputation analysis as part of the model rather than as an incidental cleaning step. A publication should state which variables were incomplete, the unit of the reported rate, how the mechanism was reasoned about, whether outcomes and auxiliaries entered imputation, how categorical variables were handled, how many imputations and iterations were used, what diagnostics were inspected, and whether

imputation was repeated inside resampling. Without these details, performance estimates cannot be reproduced or judged for leakage.

For implementers, the imputer, scaler, encoder and predictor must be versioned together. Replacement statistics and donor pools should be learned from an authorized training period, not updated silently from live cases. Monitoring should track missingness rates, patterns, imputed-value distributions, neighbour distances, model calibration and the percentage of failed or deferred predictions. A rise in missingness may signal an upstream instrument failure rather than a statistical nuisance. Governance should specify when the model abstains, when a human review is triggered and when retraining is required.

For decision makers, the most important message is that preserved AUC does not imply preserved reliability. An imputed model can rank cases reasonably while assigning probabilities that are systematically too high or too low. Decisions tied to absolute risk, resource allocation or consent demand calibration and sensitivity analysis. Where missingness is informative and unequally distributed, prediction availability itself becomes an outcome that should be audited.

Future research should compare principled MI with native missingness learners, probabilistic deep models and causal sensitivity analyses using common, openly shared simulation designs. Work is also needed on pooling calibration curves, prediction intervals and feature explanations across imputations; on efficient online KNN systems; and on external validation when training and deployment missingness mechanisms differ. The most useful benchmarks will report not only average accuracy but uncertainty, subgroup behavior, compute, privacy implications and abstention.

IV. CONCLUSION

Missing-data handling materially changed predictive discrimination, calibration, error, interval coverage and the ability to score future cases. In this simulation, listwise deletion was acceptable only when missingness was rare and benign; mean or median replacement was fast but systematically distorted calibration and uncertainty; deterministic regression imputation improved point prediction while remaining overconfident; KNN was competitive for nonlinear prediction at moderate missingness; and multiple imputation offered the most reliable overall performance when MAR was plausible and its conditional models were compatible with the analysis. No conventional method eliminated bias under MNAR. Method selection should therefore be driven by the missingness process, analysis objective, model geometry, deployment workflow and need for valid uncertainty. Imputation must be performed within resampling, evaluated on realistically incomplete test cases and accompanied by calibration assessment and MNAR sensitivity analysis. Treating preprocessing as an integral, versioned component of the prediction system is more important than searching for a universally superior algorithm.

Acknowledgements

The authors acknowledge the developers and maintainers of open-source statistical software that informed the reproducible design. No external dataset was used. Author names, affiliations, correspondence details, funding, conflicts of interest and contribution statements must be completed before submission.

Funding: No specific funding was received for this simulation study. Conflict of interest: The authors declare no conflict of interest. Data availability: The study used simulated data; the data-generating specification is fully described in the Methods section. Ethical approval: Not applicable.

REFERENCES

1. Little RJA, Rubin DB. Statistical analysis with missing data. 3rd ed. Hoboken (NJ): Wiley; 2019.
2. Carpenter JR, Kenward MG. Multiple imputation and its application. Chichester: Wiley; 2013.
3. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
4. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB.

Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520-525.
5. Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell. 2003;17(5-6):519-533.
6. van Buuren S. Flexible imputation of missing data. 2nd ed. Boca Raton (FL): CRC Press; 2018.
7. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087-1091.
8. Stekhoven DJ, Bühlmann P. MissForest-nonparametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-118.
9. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, Franco L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010;50(2):105-115.
10. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-2102.
11. Moons KGM, Donders RART, Stijnen T, Harrell FE Jr. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092-1101.
12. National Research Council. The prevention and treatment of missing data in clinical trials. Washington (DC): National Academies Press; 2010.
13. Bartlett JW, Harel O, Carpenter JR. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. Am J Epidemiol. 2015;182(8):730-736.
14. Little RJA. Regression with missing X's: a review. J Am Stat Assoc. 1992;87(420):1227-1237.
15. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30(4):377-399.
16. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91.
17. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J Stat Softw. 2017;77(1):1-17.
18. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. Stat Methods Med Res. 2015;24(4):462-487.
19. Leacy FP, Floyd S, Yates TA, White IR. Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. Am J Epidemiol. 2017;185(4):304-315.
20. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
21. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, Marrero J, Zhu J, Higgins PDR. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open. 2013;3(8):e002847.