STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–19.
Published online in International Academic Press (www.IAPress.org)

A Sequential Hybrid Approach for Imputing Missing Values in Explanatory Variables of Multiple Linear Regression Models: A Monte Carlo Simulation Study and Application to Kidney Patient Data

Abbas Ahmed Hassoun, Qutaiba Nabil Nayef*

Department of Statistics, College of Administration and Economics, University of Baghdad, Baghdad, Iraq

Abstract Missing values in explanatory variables can affect multiple linear regression, particularly when nonlinear relationships or interactions exist. This study evaluates a two-stage hybrid framework combining statistical imputation (SRI, EM, or MICE) with machine-learning refinement (RF, XGBoost, or w-KNN). Comparing nine hybrid pathways with their statistical counterparts isolates the incremental contribution of refinement. A Monte Carlo simulation with 500 replications per condition examined missing at random (MAR), sample sizes of 50, 150, and 500, missing-data proportions of 5%, 10%, 20%, and 30% and error standard deviations of 0.75, 1.50, and 3.00 . Performance was evaluated using imputation RMSE and in-sample regression-fit error (RMSE Model). Based on simulation conditions resembling the empirical data , SRI+XGBoost was compared with SRI using 400 complete kidney-patient laboratory observations from Al-Karama Teaching Hospital, Baghdad. Missing values were artificially introduced under MAR to enable comparison with known values. Creatinine was the response; urea, glucose, phosphorus, and albumin were predictors. Hybrid pathways reduced imputation RMSE relative to corresponding statistical methods across the simulated conditions. At n = 50, however, traditional methods, particularly EM, generally performed better in terms of RMSE Model. In the empirical application, SRI+XGBoost achieved lower RMSE and RMSE Model than SRI at all missing-data proportions; gains in imputation accuracy exceeded those in regression fit. The best-performing pathway varied by sample size, missingness proportion, and evaluation criterion. The findings should be interpreted within the investigated MAR conditions, multiple linear regression setting, and fixed hyperparameter configurations.

Keywords Missing values, Multiple linear regression, Sequential hybrid imputation, Machine-learning, MAR, Creatinine

DOI: 10.19139/soic-2310-5070-4730

1. Introduction

In the estimation of multiple linear regression models, missing values in explanatory variables are an important problem because not only do they compromise the amount of information available, but they may also lead to inaccurate estimates of the regression model and to incorrect results from the model. This is more difficult to overcome when nonlinear relationships or interactions are present among variables. While some statistical techniques are straightforward and intuitive to interpret, their ability to capture these types of relationships can be constrained if these characteristics are not made explicit in the imputation model used. Machine-learning methods, in contrast, offer greater flexibility in capturing nonlinear patterns and relationships among variables, but the performance of the machine-learning methods may be affected by sample size, the proportion of missing data and the type of relationships between variables.
For this reason, the present study follows a sequential hybrid method, which entails statistical imputation followed by machine-learning methods. Three statistical procedures are used to calculate provisional estimates in the

*Correspondence to: Abbas Ahmed Hassoun (Email: Abbas.Ahmed2401m@coadec.ubaghdad.edu.iq). Department of Statistics, College of Administration and Economics, University of Baghdad, Baghdad, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)
Copyright © 202x International Academic Press

first stage: Stochastic Regression Imputation (SRI), stochastic imputation based on Expectation-Maximization algorithm (EM) estimates and Multiple Imputation by Chained Equations (MICE) using Predictive Mean Matching (PMM). In the second step, the initially imputed values at the originally missing locations are refined using machine-learning algorithms: Random Forest (RF), Extreme Gradient Boosting (XGBoost), or weighted k-Nearest Neighbors (w-KNN) while all originally observed values remain unchanged. The approach proposed therefore is a mixture of first statistical imputation and the flexibility of machine learning methods to deal with some degree of nonlinearity and interactions between variables.

This framework has been designed for machine-learning techniques to be used as a second step refinement process and not as a replacement for statistical imputation. This design enables computing the extra improvement that can be attributed to the refinement stage by comparing each hybrid pathway with its corresponding initial statistical method and allows for comparisons of the hybrid pathways performance, as a function of the sample size, proportion of missing data and dispersion level.

In this study, the following main contributions are made:

1. To propose a sequential hybrid comparative framework for filling in missing values in the explanatory variables: first, a statistical imputation method is used to fill in the missing values, and then, a machine-learning method is used to refine the imputation results in a separate second stage so that the incremental contribution of the refinement stage can be evaluated separately from the initial statistical imputation.
2. To directly evaluate the impact of hybridization by comparing each statistical method with its respective hybrid pathways using RMSE as a measure of imputation accuracy and the RMSE Model as a measure of in-sample regression-model fitting error. This comparison also examines whether improvements in imputation accuracy translate into corresponding improvements in regression-model fit. Paired comparisons and confidence intervals are used to assess the magnitude, direction, and uncertainty of the differences.
3. To perform a detailed comparison of the 12 methods to determine which methods achieve the lowest mean RMSE and RMSE Model, respectively under each simulation condition and to analyse the performance of the best methods as a function of sample size, proportion of missing data and the error standard deviation.
4. To conduct an empirical evaluation of the hybrid pathway selected based on its performance under simulation conditions resembling the empirical data (SRI+XGBoost) against its corresponding statistical method (SRI) using real laboratory data from kidney patients with artificially introduced missing values in an originally complete dataset, allowing imputation accuracy to be assessed against known values alongside in-sample regression-model fitting error.

Various methods of handling missing data have been discussed in the literature. The EM algorithm was developed to estimate parameters with incomplete data by Dempster, Laird and Rubin [1]. Van Buuren and Groothuis-Oudshoorn [2] proposed the MICE framework which goes through a series of conditional models for filling in missing values. Stekhoven and Bühlmann [3] introduced the nonparametric missForest algorithm, which uses Random Forest to impute missing values iteratively without requiring a specified parametric distribution.

Other studies have examined the use of machine learning within imputation methods. Shah et al. [4] compared parametric MICE with Random Forest-based MICE and found advantages for the latter in some settings involving nonlinear relationships. Tang and Ishwaran [5] also showed that the performance of these random forest approaches to imputation is influenced by the correlation structure and the percentage of missing data, yet they can still handle interactions and nonlinearities. Deng and Lumley [6] developed mixgb, an XGBoost-based multiple-imputation framework. Zhang et al. [7] combined initial prefilling with XGBoost-based imputation of medical laboratory data, illustrating the use of machine learning following an initial imputation step.

Saber et al. [8] examined the performance of several missing value imputation methods including K-Nearest Neighbors (KNN) and Random Forest (RF) across four datasets with varying proportions of missing data. Their findings indicate that imputation accuracy is influenced by both the extent of missingness and the nature of the data. Furthermore, Abdel-Fattah et al. [9] proposed two imputation frameworks based on ”stacking” (ensemble learning), which combine statistical techniques with machine learning to address missing values in the presence of outliers. Their results demonstrated an improvement in imputation accuracy in most of the cases studied.

The studies show that the methods used to deal with missing data have shifted from conventional statistical techniques to more flexible machine-learning techniques and attempts to integrate machine learning into the

imputation process itself. In the present study, statistical imputation is performed first, followed by the application of machine learning techniques to refine the imputed values. The study evaluates the added benefit of this second stage by comparing three statistical imputation methods along with their nine corresponding hybrid variants using the imputation error RMSE and the in-sample regression model fit error RMSE Model.

1.1. Theoretical Comparison of the Missing-Data Methods Used in the Study

The methods used to handle missing data differ in their underlying principles and in their ability to represent relationships among variables. Stochastic Regression Imputation (SRI) estimates a missing value using a fitted regression model and then adds a random component to the predicted value, thereby helping to reduce the shrinkage in the variance of the imputed values compared with deterministic imputation [10, 11]. The Expectation-Maximization (EM) algorithm, in contrast, relies on an iterative procedure that alternates between the expectation and maximization steps to obtain maximum-likelihood estimates in the presence of incomplete data [1, 10]. Multiple Imputation by Chained Equations (MICE) is based on constructing a sequence of conditional models for variables containing missing values. In the present study, Predictive Mean Matching (PMM) was used within the multiple-imputation procedure, thereby allowing uncertainty associated with the missing values to be incorporated into the imputation process. The Predictive Mean Matching (PMM) method was selected within the MICE algorithm due to its suitability for continuous explanatory variables. It relies on actual observed values with similar predicted values, which helps keep imputed values within the range of the observed data and reduces reliance on strict distributional assumptions [2, 11].

By contrast, machine-learning algorithms rely on more flexible structures for representing relationships among variables. Random Forest (RF) constructs a large number of decision trees using randomly selected samples and subsets of variables and then aggregates the results of these trees to obtain the final estimate, enabling it to represent nonlinear relationships and interactions [4, 12]. Extreme Gradient Boosting (XGBoost) builds regression trees sequentially, with each successive tree improving the predictions generated by the preceding trees, while model complexity is regularized and the learning process is controlled to reduce overfitting [13]. Weighted k-Nearest Neighbors (w-KNN), on the other hand, relies on the similarity among observations and assigns greater weights to nearer neighbors than to more distant ones when estimating the missing value [14, 15].

In the case of EM, the conditional mean obtained after convergence was not adopted as a fixed final imputed value because doing so may lead to shrinkage in the variance of the imputed values and may inadequately represent the uncertainty associated with missingness. Therefore, Missing values were randomly drawn from the estimated conditional distribution following the convergence of the Expectation-Maximization (EM) algorithm [10]. Moreover, relying solely on the best point prediction, even when it reduces RMSE, does not necessarily imply that the imputation adequately reflects the uncertainty associated with missing values [11]. This implementation also makes the comparison more consistent with SRI and MICE-PMM , both of which include a stochastic component in the imputation process. This distinction also provides a rationale for the sequential hybrid approach adopted in this study, whereby statistical methods first generate initial imputations and machine-learning algorithms are subsequently used to refine the imputed values while leaving the observed values unchanged. However, subsequent refinement using machine learning produces point estimates and does not, in itself, account for the uncertainty associated with missing values. These methods were selected to compare various approaches to statistical imputation and to evaluate the improvement in their results—using techniques based on trees, boosting, and distance metrics within a unified framework.

1.2. Research Problem

The research problem arises from the fact that missing values in explanatory variables need to be addressed before fitting a multiple linear regression model, and that the accuracy of the imputed values may vary according to the imputation method used, which may in turn affect the fit of the model after the data have been completed. Furthermore, the performance of imputation methods may vary according to sample size, the proportion of missing values, and the standard deviation of the random error. Accordingly, the problem centers on determining the magnitude of the improvement added by the machine-learning refinement stage to the initial statistical imputation,

whether improvements in imputation accuracy are reflected to the same extent in regression-model fit, and which methods achieve the lowest error under the different design conditions.

1.3. Research Objective

This study aims to evaluate a two-stage hybrid framework for imputing missing values in the explanatory variables of a multiple linear regression model, beginning with initial statistical imputation and followed by refinement of the imputed values using machine-learning techniques. It also aims to measure the effect of the refinement stage on imputation accuracy and regression-model fit through paired comparisons, and subsequently to compare the statistical methods and hybrid pathways. This is done to identify the methods that yield the lowest level of error under various study conditions, based on the Root Mean Square Error RMSE and the in-sample regression model fit error RMSE Model.

1.4. Significance of the Study

The significance of this study lies in combining the statistical properties of traditional imputation methods with the flexibility of machine learning techniques within a single sequential framework, thereby allowing the effect of refining the imputed values to be evaluated independently of the initial imputation stage. This framework also makes it possible to examine how the effectiveness of hybridization changes across different data conditions. The application of this approach was demonstrated using real laboratory data from kidney patients data that were originally complete, but into which missing values were subsequently introduced artificially.

1.5. Research Hypotheses

The study is based on two interrelated hypotheses. The first assumes that adding a machine-learning refinement stage may reduce the error in imputing missing values, as measured by RMSE, compared with the initial statistical imputation, depending on the study conditions. This hypothesis is evaluated using paired differences and 95% confidence intervals. The second hypothesis assumes that improvements in imputation accuracy are not necessarily reflected to the same extent in regression-model fit, as measured by RMSE Model; rather, the effect of the refinement stage varies according to sample size, the proportion of missing values, and the standard deviation of the random error.

1.6. Research Gap

The literature has demonstrated considerable diversity in methods for handling missing data, ranging from statistical approaches to machine-learning-based techniques. Nevertheless, the question remains as to the extent of the additional benefit provided by machine-learning-based refinement, compared with the statistical imputation method on which each hybrid approach is based. The present study differs from these approaches by adopting a sequential two-stage hybrid framework that allows the additional effect of refinement to be evaluated through paired comparisons between each statistical method and its corresponding hybrid pathways, followed by a comprehensive comparison of the statistical methods and hybrid pathways according to RMSE and RMSE Model.

2. Methodology

Within the evaluation framework the regression model is estimated using Ordinary Least Squares (OLS) for SRI, EM and their corresponding hybrid pathways. In these single-imputation approaches, the imputed values are treated as fixed in the subsequent (OLS) analysis; consequently, the estimated variances of the regression coefficients may not fully reflect the uncertainty associated with the imputation process. For MICE and its hybrid pathways the regression model is estimated within each imputed dataset after which the coefficient estimates and their variances are combined according to Rubin's rules. Performance is evaluated at two levels: imputation accuracy using RMSE and in-sample regression model fit using RMSE Model. The following sections present the statistical imputation methods and machine learning techniques used within the proposed framework.

2.1. Statistical Imputation Methods

Three statistical methods were used in the study to obtain initial estimates of missing values in explanatory variables: Stochastic Regression Imputation (SRI) random imputation based on the estimates of the Expectation-Maximization (EM) algorithm and Multiple Imputation by Chained Equations (MICE) using Predictive Mean Matching (PMM). Each method was applied independently before the refinement stage using machine learning techniques.

2.1.1. Stochastic Regression Imputation : Stochastic Regression Imputation is based on estimating the missing value from a regression model and then adding a random component to the predicted value rather than using the prediction alone as the final imputation [16]. In this study when the explanatory variable $\mathbf{X}_i$ contains missing values it is treated as the target variable while the available information from the fully observed response variable $\mathbf{Y}$ and the available explanatory variables is used in the imputation model according to the adopted imputation order.

When stochastic imputation is used for missing covariates, the imputation model should include the dependent variable (Y) to obtain unbiased regression coefficient estimates [17]. The regression model used to impute $\mathbf{X}_i$ can be expressed as follows:

$$X_{ij} = \beta_{0j} + Z_{ij}^T \beta_j + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma_j^2) \tag{1}$$

where $\mathbf{X}_{ij}$ represents the value of the explanatory variable $\mathbf{X}_i$ for observation j, β_{0j} represents the intercept of the regression model, ϵ_{ij} represents the random error, Z_{ij} represents the vector of predictors available when imputing these predictors and β_j represents the regression coefficients corresponding to these predictors.

The model coefficients are estimated using the cases in which the value of $\mathbf{X}_i$ is observed in the original incomplete dataset. The expected value for each missing location is then calculated from the fitted model [2]:

$$\hat{X}_{ij} = \hat{\beta}_{0j} + Z_{ij}^T \hat{\beta}_j$$

To perform stochastic imputation a random component with a mean of zero and a variance equal to the estimated residual variance is generated:

$$\epsilon_{ij}^* \sim N(0, \hat{\sigma}_j^2) \tag{2}$$

Accordingly, the imputed value is given by:

$$X_{ij}^{(imp)} = \hat{X}_{i,j} + \epsilon_{ij}^* \tag{3}$$

Thus the resulting imputed value does not depend solely on the prediction obtained from the regression model; rather a random component is added to it allowing a certain degree of variability to be preserved in the imputed values [10]. When more than one explanatory variable has missing values SRI was implemented sequentially according to the order $X_3 \to X_1 \to X_4 \to X_2$ without mean initialization. At each step the fully observed Y together with the variables that had become available in the preceding steps was used. Imputation was then restricted to the missing locations while the observed values were kept unchanged.

2.1.2. EM-Based Stochastic Imputation : The Expectation-Maximization (EM) algorithm is used to obtain maximum likelihood estimates in the presence of incomplete data. It is an iterative algorithm consisting of two successive steps: the Expectation step (E-step) and the Maximization step (M-step) [1].

In the E-step the conditional expectation of the log-likelihood function for the complete data is calculated based on the observed data and the current estimates of the parameters:

$$Q(\theta|\theta(t)) = E_{\theta^{(t)}} \hat{\ell} \tag{4}$$

where ℓ denotes the log-likelihood function for the complete data and $\theta^{(t)}$ denotes the parameter estimates at iteration t.

In the M-step the parameter estimates are updated by selecting the value that maximizes the Q function.

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}) \tag{5}$$

The two steps are repeated until convergence is achieved. In the implementation adopted in this study the maximum number of iterations was set to 5000 iterations and convergence was considered to have been achieved when the relative change in the observed-data log-likelihood became less than 10^{-6}. The expected values obtained during the EM iterations were not treated as the final imputed values [1, 10].

The EM algorithm was applied within a joint multivariate normal model for the vector $(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4)$ with Y remaining fully observed and being included among the information used in the conditional distribution. After convergence was achieved the EM estimates were used to determine the conditional distribution of the missing values given the observed values. The resulting conditional mean was not adopted as a fixed final imputation; instead a random component from the estimated conditional distribution was added according to the following formula:

$$X_{\text{mis}}^{(imp)} = E\left[X_{\text{mis}}|X_{\text{obs}}\right] + \eta \tag{6}$$

where:

$$\eta \sim N(0, \hat{\Sigma}_{\text{mis/obs}}) \tag{7}$$

Here $\text{E}(X_{\text{mis}}|X_{\text{obs}})$ represents the estimated conditional mean of the missing values given the observed values while $\hat{\Sigma}_{\text{mis/obs}}$ represents the estimated conditional covariance matrix. The term η represents the random component added to the conditional mean whose degree of dispersion depends on the estimated conditional variance of the missing values [10].

In the implementation adopted in this study these steps were applied within the framework of a multivariate normal model whereby the distribution parameters were first estimated using the EM algorithm. Subsequently the missing values were randomly sampled from the estimated conditional distribution after convergence had been achieved while the observed values were left unchanged.

2.1.3. Multiple Imputation by Chained Equations Using PMM : Multiple Imputation by Chained Equations (MICE) is based on the Full Conditional Specification (FCS) framework in which a conditional imputation model is constructed for each variable containing missing values. The imputed values are then updated sequentially based on the observed data and the current values of the other variables.

The conditional imputation process for the variable X_j can be represented as follows:

$$X_{j,\text{mis}}^{(t,r)} \sim p\left(X_{j,\text{mis}} \mid X_{j,\text{obs}}, X_{-j}^{(t,r)}, \phi_j^{(r)}\right) \tag{8}$$

Here $X_{j,\text{mis}}$ represents the missing values in the variable X_j whereas $X_{j,\text{obs}}$ represents its observed values. The term $X_{-j}^{(t,r)}$ denotes the current values of the other variables used in the imputation model during cycle t and in imputed dataset r while $\phi_j^{(r)}$ represents the parameters of the imputation model for the variable X_j. The symbol t denotes the cycle number whereas $(r = 1, \ldots, m)$ denotes the number of the imputed dataset.

The updating process is repeated across the variables containing missing values to obtain several complete datasets. In the implementation adopted in this study $(m = 20)$ imputed datasets and maxit=10 updating cycles were used with PMM following a preliminary assessment of the convergence and stability of the MICE chains. The fully observed variable (Y) was also used as a predictor when imputing the explanatory variables while the observed values were left unchanged [2].

After obtaining the complete datasets the regression model was estimated independently for each dataset. The resulting estimates of the regression coefficients and their variances were then combined using Rubin's rules which integrate the information obtained from the different imputed datasets while accounting for both the within-imputation variance and the between-imputation variance [10]. The datasets themselves were not pooled before estimating the regression model. In the MICE method and its hybrid pathways, the RMSE and RMSE Model metrics were calculated independently for each of the twenty completed datasets, and their averages were then computed across these datasets. Rubin's rules were applied to combine the regression coefficient estimates and their variances, but not to combine the error metrics.

Predictive Mean Matching (PMM) was used within MICE for continuous variables. Its principle is based on calculating the predicted value for a case containing a missing value then identifying a set of observed cases whose

predicted values are closest to the predicted value of that case. Subsequently one of these donor cases is selected at random and its observed value is used as the imputed value.

One of the advantages of PMM is that the final imputed value is not merely a predicted value generated by the regression model but rather an observed value that actually exists in the data. This helps keep the imputed values within the range of the observed values and reduces the generation of implausible values [18].

Compared to imputation based on parametric linear regression, the PMM method relies less on distributional assumptions; meanwhile, logistic and multinomial logistic regression imputation methods are used for categorical variables, rather than for the continuous explanatory variables addressed in this study.

2.2. Machine Learning Techniques

Three machine learning techniques were used in the second stage of the hybrid approach: Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Weighted k-Nearest Neighbors (w-KNN). These techniques provide flexibility in representing relationships among variables particularly nonlinear relationships and interactions. For Random Forest and XGBoost, the models were trained using only the originally observed values for each target variable, while the w-KNN method selected its neighbors from cases where the target variable's value was observed. Other variables—including Y—were used as predictors based on their current values, which may have included values resulting from initial imputation or values updated during previous refinement steps. The target variable itself was excluded from the predictor variables at each step of the refinement process.

The refinement process was conducted sequentially, with updated values being used as the procedure moved on to subsequent variables. Prior to conducting the final simulation, preliminary sensitivity analyses were performed to examine various settings for algorithm parameters, the number of refinement cycles (1,3,5, and 10) and the sequences for updating variables. For XGBoost, the initial tests included tree depths ranging from 1 to 4, learning rates from 0.01 to 0.20, and a number of boosting rounds ranging from 50 to 500. The selected settings for the corresponding hybrid pathways were then fixed, and the refinement process was terminated after a specified number of cycles, rather than relying on an automated convergence criterion.

2.2.1. Random Forest : Random Forest is based on constructing an ensemble of regression trees. Each tree is constructed using a Bootstrap sample drawn from the training data and at each node a random subset of candidate variables is selected to determine the split. The predictions of the individual trees are then aggregated to obtain the final prediction of the forest [12]. In the regression setting the final prediction is calculated as the average of the predictions of the individual trees:

$$\hat{f}_K(x) = \frac{1}{K}\sum_{k=1}^{K} h_k(x, \Theta_k) \tag{9}$$

Here K represents the total number of trees, $h_k(x, \Theta_k)$ represents the prediction produced by tree k for the vector x while Θ_k represents the random components associated with the construction of tree k. This structure enables Random Forest to represent nonlinear relationships and interactions without specifying their form in advance [3, 12].

In this study a regression forest was trained for each explanatory variable containing missing values using the cases in which the target variable was observed [19]. The fully observed variable Y together with the available explanatory variables was used as predictors. The forest predictions were then used to refine the missing entries only while the observed values were left unchanged.

In the final implementation mtry=1 and three refinement cycles were used according to the order $(X_3 \to X_1 \to X_4 \to X_2)$ while the number of trees was set to 500 trees.

2.2.2. Extreme Gradient Boosting (XGBoost) : The Extreme Gradient Boosting XGBoost algorithm is based on constructing regression trees sequentially such that each new tree improves the predictions produced by the preceding trees. It also relies on an objective function that combines prediction error with a regularization term for model complexity which helps control the learning process and reduce overfitting [13].

The prediction in the XGBoost model is calculated by summing the contributions of the added trees as follows:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \tag{10}$$

Here k represents the number of trees in the model, f_k represents the function of tree k while F represents the space of regression trees to which the functions f_k belong.

The algorithm learns these trees by minimizing a regularized objective function that combines the loss function with a penalty for tree complexity:

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{11}$$

Here $l(y_i, \hat{y}_i)$ represents the loss function that measures the difference between the actual values y_i and the predicted values $\hat{y}_i$ while $\Omega(f_k)$ represents the regularization term for tree f_k which penalizes model complexity and contributes to reducing overfitting [13].

In this study an XGBoost model was trained for each explanatory variable containing missing values using its observed cases with Y and the remaining available variables used as predictors. Its predictions were then used to refine the missing entries only while the observed values were left unchanged. A maximum depth of 2, a learning rate of 0.10, and 50 boosting rounds were used along with three refinement cycles in the order ($X_3 \rightarrow X_1 \rightarrow X_4 \rightarrow X_2$). The subsample and colsample_bytree ratios were both fixed at 0.80, with the regularization parameters set to $\alpha = 1$ and $\alpha = 0$. These are the fixed settings used in the final code.

2.2.3. Weighted k-Nearest Neighbors (w-KNN) : Weighted k-Nearest Neighbors (w-KNN) estimates a value based on the observations that are closest to the case being estimated assigning greater weights to closer neighbors than to more distant [15]. In the regression setting, the estimated value is calculated using a weighted mean [21]:

$$haty\ (q) = \frac{\sum_{r=1}^{k} w_r\, y_r}{\sum_{r=1}^{k} w_r} \tag{12}$$

In the implementation adopted in this study the weight of each neighbor was calculated using inverse-distance weighting:

$$w_r = \frac{1}{(d_{r(q)} + \varepsilon)^{\alpha}} \tag{13}$$

Here y_r represents the observed value of neighbor r, w_r represents the weight assigned to that neighbor and $d_{r(q)}$ represents the distance between the target case q and neighbor r. The parameter α represents the magnitude of the distance effect while ε is added as a small value to avoid division by zero.

Before calculating the distances, the predictor variables were standardized and the Euclidean distance was then used to identify the nearest neighbors. The selection of donors was restricted to cases in which the target variable was originally observed with Y and the remaining explanatory variables excluding the target variable used in calculating the distance. The values k=10, α=2, and ε=10^{-8} were used with one refinement cycle according to the order (X_3ßX_1ßX_4ßX_2) while the observed values were left unchanged.

2.3. Performance Evaluation and Statistical Comparison

2.3.1. Imputation Accuracy : The RMSE criterion was used to measure imputation accuracy by comparing the imputed values with the true values only at the locations where missingness had been artificially introduced. A smaller value indicates higher imputation accuracy [22]:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{Nm}} \sum_{(\mathrm{i,j}) \in \mathrm{M}} \left(\widehat{\mathrm{X}}_{\mathrm{ij}} - \mathrm{X}_{\mathrm{ij}}\right)^2} \tag{14}$$

Here $\widehat{\mathrm{X}}_{\mathrm{ij}}$ represents the imputed values X_{ij} represents the true values preserved before the missingness was introduced M denotes the set of locations that were originally missing and Nm represents the number of missing values evaluated.

2.3.2. Regression Model Fit After Imputation : RMSE Model was also used to measure the in-sample regression model fit error after completing the data by comparing the actual values of the response variable with the values predicted by the model itself [23, 26]:

$$\text{RMSE Model} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{Y}_{\mathrm{i}} - \widehat{\mathrm{Y}}_{\mathrm{i}}\right)^{2}} \tag{15}$$

Here Y_{i} represents the actual values of the response variable while $\widehat{\mathrm{Y}}_{\mathrm{i}}$ represents the values estimated by the regression model and n represents the sample size. Thus the first criterion directly measures the accuracy of imputing the missing values whereas the second criterion reflects the effect of the imputation process on the regression model fit after completing the data.
For MICE and its hybrid pathways both RMSE and RMSE Model were calculated independently for each imputed dataset and then averaged across the 20 datasets for comparison purposes.

2.3.3. Replication Summarization and Paired Comparison : Since each simulation condition was replicated 500 times the performance of each method was summarized using the mean standard deviation Monte Carlo Standard Error (MCSE) [24] and the 95% confidence interval for the mean. This allows the stability of the results across replications to be assessed rather than relying solely on the mean .The MCSE was calculated as follows:

$$\mathrm{MCSE}(\overline{\mathrm{M}}) = \frac{SD(M_r)}{\sqrt{R}} \tag{16}$$

Here $SD(M_r)$ represents the standard deviation of the performance criterion values across the replications while (R=500) represents the number of Monte Carlo replications.
To compare each statistical method with its corresponding hybrid counterpart a paired analysis was performed within the same replication. Thus both methods were applied to the same incomplete dataset and the difference between their respective criterion values was then calculated .

$$\mathrm{D}_r = \mathrm{Q}_{\mathrm{H},r} - \mathrm{Q}_{\mathrm{S},r} \tag{17}$$

Here $\mathrm{Q}_{\mathrm{H},r}$ represents the value of the criterion for the hybrid method while $\mathrm{Q}_{\mathrm{S},r}$ represents its value for the statistical method in replication (r). According to the order of subtraction a negative value indicates a reduction in error for the hybrid method whereas a positive value indicates a reduction in error for the original statistical method. A 95% confidence interval for the paired difference was also used. If the interval lay entirely below zero the direction of the difference favored the hybrid method; if it lay entirely above zero it favored the statistical method; and if it included zero no clear direction of the difference between the two methods was observed [20, 25].

3. Experimental and Empirical Analysis

3.1. Simulation Study

We did a Monte Carlo experiment to compare the performance of the methods of dealing with missing data under various data conditions. The sample size levels tested ($n = 50, 150, 500$), the levels tested for the standard deviation of the random error ($\sigma = 0.75, 1.50, 3.00$) and the proportion of missing data ($5\%, 10\%, 20\%, 30\%$) were chosen as part of the simulation design. The replication for each set of design conditions was made 500 times in order to minimize the impact of random variation that could arise from the generation of data and to provide more stable results.

The four explanatory variables (X_1, X_2, X_3, X_4) were produced from a multivariate normal distribution with a mean of zero and a variance of one for each variable, and a correlation coefficient of 0.40 for the explanatory variables. This level was chosen to be a moderate degree of correlation, such that there was a sufficient degree of association between the variables for the imputation methods to take advantage of, but not so high that multicollinearity would become the major factor in the comparison.

Some degree of nonlinearity and weak interaction between variables were included in the simulation. The regression model was used to estimate the coefficients of the linear component, which were then fixed to some finite values suitable to the simulation design. The response variable was created based on the following model:

$$\begin{aligned} Y_i = 2 &+ 1.00X_{1i} - 0.30X_{2i} + 0.60X_{3i} + 0.20X_{4i} \\ &+ 0.15(X_{1i}^2 - 1) + 0.2(X_{2i}^2 - 1) + 0.20(X_{1i}X_{3i} - 0.40) + \epsilon_i \end{aligned}$$

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad \sigma = 0.75, 1.50, 3.00$$

Two quadratic terms, (X_{1i}, X_{3i}) and (X_{1i}, X_{2i}) were added with comparatively low coefficients to indicate a small nonlinearity, and a small cross term of $(X_{1i})(X_{3i})$ was added to represent a weak interaction between the two quadratics. The centered terms $(X_{1i}^2 - 1)$, $(X_{2i}^2 - 1)$, and $(X_{1i}X_{3i} - 0.40)$ were added to the model so that the influence of nonlinearity and interaction would not affect the theoretical mean of the response variable.

Once the full data was available, the missingness was added to the four explanatory variables, and the response variable (Y) was fully observed. A calibrated logistic-Bernoulli model for the fully-observed response variable (Y) was used to generate missingness. Specifically, the probability of missingness was modeled as a function of (Y) with the effect coefficient ($\gamma = 0.70$) fixed; the intercept was calibrated to obtain the desired proportion of missing data in each design condition. Hence, missingness was only related to fully observed information, which corresponds to the Missing at Random (MAR) principle. The Missing at Random (MAR) mechanism was selected because the occurrence of missing data may depend on observed patient characteristics, such as serum creatinine levels. Furthermore, this assumption provides a controlled setting for comparing imputation methods, without necessarily implying that MAR represents the actual mechanism of missingness in the medical records. The target missingness rates were distributed unevenly across the four explanatory variables, such that their mean equaled the overall target missingness rate specified for each scenario. Specifically, the target missingness rates for the explanatory variables (X_1), (X_2), (X_3), and (X_4) were (8%, 3%, 5%, 4%) at an overall missingness level of 5%; (16%, 6%, 10%, 8%) at a level of 10%; (30%, 12%, 20%, 18%) at a level of 20%; and (48%, 18%, 30%, 24%) at a level of 30%. The locations of missing values for each variable were generated using a Logistic-Bernoulli model conditioned on the observed dependent variable (Y), allowing for missingness to occur in more than one variable within a single observation. Furthermore, the same random numbers were used across different missingness levels within each iteration to ensure that the locations of missing values were nested across these levels.

The statistical methods SRI, EM, and MICE were evaluated separately and in combination with the nine hybrid pathways created by combining each of these methods with RF, XGBoost, and w-KNN for evaluation. Thus, under each of the simulation design conditions, twelve methods were compared.

3.1.1. Simulation Results : The simulation results have been arranged into two levels of comparisons. The first level was on the impact of incorporating an imputation refinement step based on a machine learning approach to the statistical imputation methods. In order to compare the direction of advantage for each statistical method and its related hybrid pathway, this assessment was carried out by making paired comparisons between each statistical method and the related hybrid pathway to determine the direction of the advantage, based on the paired difference and its 95% confidence interval, as well as to identify the conditions under which the refinement stage added a clear advantage and those under which no clear advantage was observed. For economy of presentation, to avoid the necessity of a proliferation of tabular statements, the results of these comparisons were expressed in a compact analytical form.

The second level focused on the numerical superiority of the twelve methods based on RMSE and RMSE Model. The data for the complete ranking of all methods under each combination was not reported due to the number of design combinations. Rather, only the method that yielded the lowest mean of all methods for each criterion is

reported here in Tables (1) and (2) along with measures of variability and uncertainty to emphasize the variation of the method's mean value by sample size, dispersion level and missing-data proportion.

The impact of the machine-learning refinement stage was evaluated by comparing each statistical imputation method with its corresponding hybrid pathways :

First: Effect of Refinement on Imputation Accuracy (RMSE) The paired comparisons revealed that the incorporation of the machine learning refinement stage gave consistently better imputation accuracy. The 95% confidence intervals for all paired RMSE differences were found to be entirely negative when SRI, EM, or MICE were compared with their respective hybrid pathways, using RF, XGBoost, and w-KNN for all sample sizes, dispersion levels, and missing data proportions used in the study. Consequently, none of the comparisons favored the original statistical methods; instead, all comparisons favored the hybrid approaches in terms of the root-mean-square error RMSE criterion.

The second effect is the impact of the refinement on the regression-model fit (RMSE Model) : RMSE Model was affected differently with a change in sample size. For each sample size, 108 paired comparisons were conducted, encompassing nine hybrid approaches, four proportions of missing data, and three levels of standard deviation for the random error.

Of the 108 comparisons, 77 favoured traditional statistical approaches, with 15 showing a preference for the hybrid pathways and 16 ($n = 50$) showing no clear preference. This means that the impact of the refinement on the fit of the regression model was modest with small sample sizes. The pattern shifted at ($n = 150$) with 91 of the 108 comparisons favoring the hybrid pathways and 7 of the 108 comparisons favoring the original statistical methods; there were 10 comparisons that showed no clear advantage. For all 108 comparisons at ($n = 500$), the hybrid pathways were favored, and the 95% Confidence Intervals for the paired differences were completely negative.

The advantage of hybrid approaches in terms of RMSE Model became more consistent as the sample size increased. Given that the study involved a large number of paired comparisons (108 comparisons for each sample size), the 95% confidence intervals were interpreted as exploratory rather than confirmatory, as no correction for multiple comparisons was applied. Consequently, the results should be interpreted with caution, particularly regarding the potential for false-positive findings. The ranking of the methods was based on mean error values and is considered descriptive, rather than evidence of statistically significant superiority. At a sample size of 50, statistical methods were favored in most comparisons. In contrast, hybrid approaches were favored in most comparisons at a sample size of 150, and in all comparisons at a sample size of 500.

3.1.2. The methods were compared using numerical Superiority Based on RMSE and RMSE Model : The effect of the refinement stage was assessed by paired comparisons and the analysis was then continued to find the superior numerical method of the 12 methods in each combination of the design conditions. The numerical superiority was obtained by the smallest mean value of RMSE and RMSE Model obtained from 500 replications. Table (1) shows the method with the lowest mean RMSE while Table (2) shows the method with the lowest mean RMSE Model, along with the standard deviation (SD), Monte Carlo standard error (MCSE), and the 95% confidence interval for the mean performance measure.

The pattern of numerical leadership for the methods in terms of imputation accuracy is shown in Table (1) which displays the method which had the lowest mean RMSE under each design condition. As far as numbers of leaders were concerned this situation differed depending on the sample size:

At a small sample size (50), RF-based approaches ranked first in 10 out of 12 cases, while XGBoost-based approaches ranked first in the remaining 2 cases. For medium sample size ($n = 150$), RF-based pathways ranked first in 9 conditions, while XGBoost-based pathways ranked first in the remaining 3 conditions. When the big sample size was considered ($n = 500$), the numerical leadership over the pathways switched entirely to the XGBoost-based pathways, in which they became the first ranked pathway in all 12 design conditions. The results are summarized in a heatmap (Figure (1)), which displays the number of design conditions for which each method had the highest ranking.

Table 1

n	σ	Missingness (%)	Best-Performing Method	Mean RMSE	SD	MCSE	95% CI
50	0.75	5	SRI+XGBoost	0.8452	0.2124	0.0095	[0.8266, 0.8638]
		10	EM+XGBoost	0.869	0.1617	0.0072	[0.8549, 0.8831]
		20	EM+RF	0.8805	0.1078	0.0048	[0.8711, 0.8899]
		30	EM+RF	0.896	0.0929	0.0042	[0.8878, 0.9042]
	1.5	5	EM+RF	0.8802	0.2085	0.0093	[0.8620, 0.8984]
		10	EM+RF	0.9038	0.1595	0.0071	[0.8899, 0.9177]
		20	MICE+RF	0.9173	0.1143	0.0051	[0.9073, 0.9273]
		30	SRI+RF	0.9331	0.1021	0.0046	[0.9241, 0.9421]
	3	5	SRI+RF	0.9307	0.2173	0.0097	[0.9117, 0.9497]
		10	SRI+RF	0.9398	0.1482	0.0066	[0.9269, 0.9527]
		20	EM+RF	0.9519	0.1149	0.0051	[0.9419, 0.9619]
		30	SRI+RF	0.9711	0.1078	0.0048	[0.9617, 0.9805]
150	0.75	5	EM+XGBoost	0.8064	0.1175	0.0053	[0.7960, 0.8168]
		10	SRI+XGBoost	0.8093	0.0855	0.0038	[0.8019, 0.8167]
		20	SRI+XGBoost	0.8355	0.0663	0.003	[0.8296, 0.8414]
		30	SRI+RF	0.8517	0.0539	0.0024	[0.8470, 0.8564]
	1.5	5	EM+RF	0.8626	0.1168	0.0052	[0.8524, 0.8728]
		10	MICE+RF	0.8699	0.0801	0.0036	[0.8628, 0.8770]
		20	MICE+RF	0.8865	0.0613	0.0027	[0.8812, 0.8918]
		30	MICE+RF	0.9046	0.0522	0.0023	[0.9001, 0.9091]
	3	5	MICE+RF	0.902	0.1169	0.0052	[0.8918, 0.9122]
		10	EM+RF	0.9057	0.0872	0.0039	[0.8981, 0.9133]
		20	MICE+RF	0.9214	0.0641	0.0029	[0.9157, 0.9271]
		30	EM+RF	0.9438	0.0583	0.0026	[0.9387, 0.9489]
500	0.75	5	EM+XGBoost	0.7621	0.0629	0.0028	[0.7566, 0.7676]
		10	MICE+XGBoost	0.7679	0.043	0.0019	[0.7642, 0.7716]
		20	SRI+XGBoost	0.7868	0.0315	0.0014	[0.7841, 0.7895]
		30	EM+XGBoost	0.8003	0.0295	0.0013	[0.7978, 0.8028]
	1.5	5	SRI+XGBoost	0.8312	0.0583	0.0026	[0.8261, 0.8363]
		10	SRI+XGBoost	0.8386	0.0421	0.0019	[0.8349, 0.8423]
		20	MICE+XGBoost	0.8562	0.032	0.0014	[0.8535, 0.8589]
		30	EM+XGBoost	0.8745	0.0293	0.0013	[0.8720, 0.8770]
	3	5	EM+XGBoost	0.8777	0.0631	0.0028	[0.8722, 0.8832]
		10	MICE+XGBoost	0.8861	0.0445	0.002	[0.8822, 0.8900]
		20	MICE+XGBoost	0.9067	0.0339	0.0015	[0.9038, 0.9096]
		30	EM+XGBoost	0.9318	0.0313	0.0014	[0.9291, 0.9345]

The top ranking does not necessarily mean a significant difference in the top pathways. At $n = 500$, $\sigma = 0.75$, and 30% missing data, the mean RMSE values are 0.8003, 0.8006, and 0.8010, respectively, for the EM+XGBoost, SRI+XGBoost, and MICE+XGBoost methods. To identify the method with the numerically lowest RMSE, then, the ranking of the mean values was used, while the paired comparison and the confidence interval for the paired difference were used to evaluate the effect of adding the machine-learning refinement stage to the original statistical method.

In the context of the stability measures: The standard deviation (SD) values reflect the variability of the RMSE across the 500 iterations, while the Monte Carlo standard error (MCSE) values indicate the precision of the estimated mean RMSE. In general, MCSE values were smaller and confidence intervals were narrower with larger sample sizes. These metrics provide additional information regarding the variability and precision of the results, alongside the mean RMSE values.

Table (2) shows the method that obtained the lowest mean RMSE Model for each design condition and thus represents the numerical dominance of the methods in terms of regression-model fit. In regard of number of leaders, the pattern did not remain consistent:

- For the small sample size n=50, the standalone statistical methods continued to dominate the 12 design conditions, ranking first in 11 out of 12 design conditions. EM performed better than any other for 10

Table 2

n	σ	Missingness (%)	Best-Performing Method	Mean RMSE Model	SD	MCSE	95% CI
50	0.75	5	EM	0.7015	0.0843	0.0038	[0.6941, 0.7089]
		10	EM	0.7106	0.0949	0.0042	[0.7023, 0.7190]
		20	EM	0.7194	0.115	0.0051	[0.7092, 0.7295]
		30	EM	0.7069	0.1521	0.0068	[0.6935, 0.7202]
	1.5	5	EM	1.375	0.161	0.0072	[1.3608, 1.3891]
		10	EM	1.3725	0.1739	0.0078	[1.3572, 1.3877]
		20	EM	1.3684	0.1917	0.0086	[1.3516, 1.3853]
		30	EM	1.3464	0.2295	0.0103	[1.3263, 1.3666]
	3	5	SRI	2.7127	0.2986	0.0134	[2.6864, 2.7389]
		10	EM	2.7046	0.3099	0.0139	[2.6773, 2.7318]
		20	EM+XGBoost	2.6752	0.3274	0.0146	[2.6464, 2.7040]
		30	EM	2.6078	0.3675	0.0164	[2.5755, 2.6401]
150	0.75	5	SRI+XGBoost	0.7522	0.0547	0.0024	[0.7474, 0.7570]
		10	SRI+XGBoost	0.7724	0.0679	0.003	[0.7665, 0.7784]
		20	EM+XGBoost	0.7915	0.0805	0.0036	[0.7845, 0.7986]
		30	EM	0.8077	0.0845	0.0038	[0.8003, 0.8151]
	1.5	5	SRI+XGBoost	1.4645	0.094	0.0042	[1.4562, 1.4728]
		10	MICE+XGBoost	1.4656	0.1008	0.0045	[1.4568, 1.4745]
		20	SRI+XGBoost	1.4578	0.1125	0.005	[1.4479, 1.4677]
		30	SRI+XGBoost	1.4471	0.1332	0.006	[1.4354, 1.4588]
	3	5	SRI+XGBoost	2.9147	0.1681	0.0075	[2.9000, 2.9295]
		10	MICE+XGBoost	2.9093	0.1721	0.0077	[2.8942, 2.9244]
		20	MICE+XGBoost	2.8857	0.1811	0.0081	[2.8698, 2.9016]
		30	EM+XGBoost	2.8466	0.2015	0.009	[2.8289, 2.8643]
500	0.75	5	SRI+XGBoost	0.76	0.0312	0.0014	[0.7572, 0.7627]
		10	EM+XGBoost	0.7659	0.0342	0.0015	[0.7628, 0.7689]
		20	SRI+XGBoost	0.7684	0.0434	0.0019	[0.7646, 0.7722]
		30	SRI+XGBoost	0.7613	0.0531	0.0024	[0.7566, 0.7659]
	1.5	5	SRI+XGBoost	1.4864	0.0494	0.0022	[1.4820, 1.4907]
		10	EM+XGBoost	1.4792	0.0539	0.0024	[1.4744, 1.4839]
		20	SRI+XGBoost	1.461	0.0612	0.0027	[1.4557, 1.4664]
		30	SRI+XGBoost	1.4309	0.0758	0.0034	[1.4242, 1.4376]
	3	5	MICE+XGBoost	2.9716	0.098	0.0044	[2.9630, 2.9802]
		10	MICE+XGBoost	2.964	0.1008	0.0045	[2.9551, 2.9729]
		20	SRI+XGBoost	2.9422	0.1065	0.0048	[2.9328, 2.9515]
		30	SRI+XGBoost	2.8934	0.1267	0.0057	[2.8823, 2.9045]

conditions, and SRI performed best for 1 condition. The EM+XGBoost pathway only performed best in one condition, i.e. for ($\sigma = 3.00$) and a missing-data proportion of 20%. At this sample size the paired comparisons suggest that the RMSE Model did not always decrease when the machine-learning stage was added. The traditional statistical method stayed superior under some conditions and it was found that, under other conditions, the confidence interval for the paired difference contained the number zero, suggesting that there was no clear advantage. This result demonstrates that substantial gains in imputation accuracy do not necessarily result in the same amount of improvement in imputation model fit.

- The pattern became very different as the sample size grew to n=150. XGBoost-based pathways outperformed EM in 11 out of the 12 conditions and EM only outperformed XGBoost-based pathways in one condition. The paired comparisons also showed a stronger preference for the hybrid pathways as the percentage of missing data increased, with some cases performing close to each other that did not show a clear preference to one pathway or another based on confidence intervals. Therefore, the effect of the refinement stage became more uniform with the increase of samples.
- Pathways based on XGBoost were the top-performing for each of the 12 design conditions at the large sample size n=500. Among these approaches, the SRI+XGBoost configuration ranked first in 8 design scenarios, while EM+XGBoost and MICE+XGBoost each ranked first in 2 scenarios. Specifically, MICE+XGBoost ranked first at missingness rates of 5% and 10% when the error standard deviation was 3.00, whereas

SRI+XGBoost ranked first at missingness rates of 20% and 30% across all three standard deviation levels. These rankings are based on the lowest mean RMSE Model; they do not necessarily imply the existence of statistically significant differences between the pathways. These results are summarized in Figure (1) which presents a heatmap of the proportion of time each method was chosen as the best method by sample size.

As far as the stability measures are concerned:

- The confidence intervals were narrower at n=500 than at n=50, and the MCSE values at n=500 ranged from 0.0014 to 0.0057. Furthermore, the higher the dispersion level σ, the greater the increase in SD. For instance, when the leading method is used, the MCSE for n=50 was 0.0164 at a 30% missing data rate, whereas for n=500 it was 0.0057 at the same missing data rate. As the sample size increased, the estimates of the mean RMSE Model generally became more accurate, even at higher levels of error dispersion.
- The benefit of using hybridization in terms of RMSE Model was not as consistent as was seen for the imputation accuracy. The statistical methods were not far behind at the small sample size, especially EM. As the sample size increased, however, the numerical advantage of the XGBoost-based pathways gradually increased and at n=500 the pathways with XGBoost were ranked as the most accurate one under all design conditions. This finding helps to validate the importance of considering the impact of the data imputation procedure on the statistical model to be applied after the data has been imputed. The limited appearance of w-KNN among the top-performing methods under the conditions of this study does not necessarily imply poor performance in general, as its effectiveness may depend on the nature of the data and the degree of non-linearity.

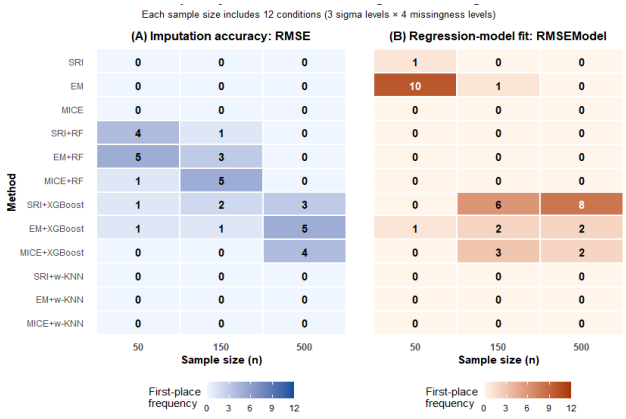


Figure 1. Frequency with Which the Methods Ranked First According to RMSE and RMSE Model by Sample Size.

3.1.3. A supplementary comparison with Direct XGBoost Imputation To assess the utility of preliminary statistical imputation, an additional comparison was conducted between a "direct" XGBoost approach—which relies on the algorithm's built-in capacity to handle missing values without prior imputation—and three hybrid approaches: SRI+XGBoost, EM+XGBoost, and MICE+XGBoost. For each variable containing missing values, an XGBoost model was trained using its observed values, with Y and the remaining explanatory variables serving as predictors. XGBoost handles missing values within the predictors directly, without the need for prior statistical imputation. The comparison covered 36 design conditions, with 500 Monte Carlo replications per condition, using the same simulation settings. In each experimental case, the hybrid approach that yielded the lowest average error was compared against a standard XGBoost model—separately for both RMSE and RMSE Model. This resulted in 36 comparisons for each evaluation metric.

Table 3. Missingness Levels at Which Hybrid Pathways Outperformed Direct XGBoost

n	σ	RMSE	RMSE Model
		Missingness Levels (%)	Missingness Levels (%)
50	0.75	5, 10, 20, 30	5, 10, 20, 30
50	1.50	5, 10, 20, 30	5, 10, 20, 30
50	3.00	5, 10, 20	5, 10, 20, 30
150	0.75	10, 20, 30	5, 10, 20
150	1.50	5, 10	5, 10, 20, 30
150	3.00	5, 20, 30	5, 10
500	0.75	5, 10, 20, 30	—
500	1.50	5, 10	—
500	3.00	—	—

Note: The figures represent the missingness levels at which a given hybrid approach achieved a lower average error than the standalone XGBoost model; they do not represent the magnitude of improvement. The best-performing approach for each criterion was selected independently. The symbol (—) indicates that no hybrid approach outperformed the standalone model.

Table (3) shows that the hybrid approaches achieved a lower RMSE than the standalone XGBoost model in 27 out of 36 cases, and also outperformed it in 22 cases according to the RMSE Model criterion. The advantage of the hybrid approach in terms of imputation accuracy was most pronounced with a small sample size ($n = 50$), whereas the standalone XGBoost model became more competitive as the sample size increased. At a sample size of 500, the direct XGBoost approach achieved a lower average error in 5 out of 12 cases according to RMSE, and in 9 out of 12 cases according to RMSE Model.

These results suggest that preliminary statistical imputation can enhance the performance of XGBoost; however, its utility varies depending on the data conditions and the evaluation criterion. Consequently, the hybrid approach cannot be considered superior in all scenarios. Furthermore, these findings reflect a numerical advantage in terms of average error, but they do not, in themselves, establish the statistical significance of the observed differences.

3.1.4. Supplementary Evaluation Using Ten Explanatory Variables To examine the performance of the hybrid method when the number of explanatory variables is increased, a supplementary experiment was conducted using ten variables, comparing the SRI method with SRI+XGBoost. The experiment covered two sample sizes (150 and 500) and two missingness proportions (10% and 20%), with an error standard deviation of 1.50, with 500 Monte Carlo replications per design condition. The same datasets and missing data locations were used for both methods to ensure a fair comparison. The ten explanatory variables were generated from a multivariate normal distribution with a correlation coefficient of 0.40. Six variables were added to the original model—which already contained four variables—with each additional variable having a non-zero regression coefficient.

Table (4) shows that SRI+XGBoost continued to achieve lower imputation errors than SRI when the number of explanatory variables was extended to ten; it achieved a lower mean RMSE than the SRI method across all four studied scenarios, with 95% confidence intervals for the paired differences confirming this advantage. Furthermore, the mean RMSE Model decreased in all cases, although the confidence interval included zero at a sample size of 150 with a 20% missingness rate, indicating no clear advantage in that specific instance. These results suggest that the refinement stage using XGBoost remains effective—particularly regarding imputation accuracy—even as the number of explanatory variables increases, within the conditions covered by the experiment.

3.2. Empirical Application

3.2.1. Data The empirical application was performed with actual laboratory data obtained from the records of the Laboratory of the nephrology ward at Al-Karama teaching hospital in Baghdad, with regard to patients with renal

Table 4. Comparison of SRI and SRI+XGBoost performance using ten explanatory variables

n	Missingness (%)	Criterion	Mean SRI	Mean SRI+XGBoost	95% CI for paired difference
150	10	RMSE	1.2094	0.8749	[-0.3409, -0.3280]
150	10	RMSE Model	1.4491	1.4442	[-0.0096, -0.0002]
150	20	RMSE	1.2298	0.8897	[-0.3447, -0.3354]
150	20	RMSE Model	1.4513	1.4483	[-0.0089, 0.0028]
500	10	RMSE	1.1958	0.8319	[-0.3671, -0.3609]
500	10	RMSE Model	1.5127	1.4831	[-0.0322, -0.0271]
500	20	RMSE	1.2137	0.8445	[-0.3717, -0.3667]
500	20	RMSE Model	1.5279	1.4763	[-0.0551, -0.0482]

disorders, from January 2024 until June 2026. The data comprised 5 laboratory variables and 400 observations, and was complete prior to the introduction of artificially missing data for evaluation. Creatinine level in the serum was the dependent variable while urea, glucose, phosphorus and albumin were independent variables. All variables were recorded in their respective units in the laboratory, such as: mg/dL for creatinine, urea, glucose and phosphorus, and g/dL for albumin. Only laboratory records containing complete measurements for all five variables were included, while records lacking measurements for any of these variables were excluded, in order to provide a complete reference dataset for evaluating the accuracy of the imputation. The required ethical approvals were obtained prior to the issuance of administrative authorization to access hospital records for research purposes from the Baghdad/Al-Karkh Health Directorate, under the Iraqi Ministry of Health.

There were no worrisome multicollinearity problems as suggested by the variance inflation factor (VIF) values that ranged from 1.006 to 1.318. In addition, the Breusch-Pagan and Shapiro-Wilk tests applied to the residuals showed no clear evidence of violations of the homoscedasticity and normality assumptions, respectively ($p = 0.477$ and $p = 0.742$ respectively). Conversely, the Ramsey RESET test was statistically significant ($p < 0.001$), suggesting that there was some departure from linearity. The improvement in model fit was also statistically significant in a model with quadratic terms in comparison to the linear model [F(4,391)=15.993, $p < 0.001$]. The coefficient of determination rose from 0.447 to 0.525 and the effect size was around 0.164. Thus, the proposed method was deemed appropriate for the evaluation of data under limited-to-moderate degree of nonlinearity.

3.2.2. Application of the Selected Method The data consisted of ($n = 400$) observations and the standard deviation of the residuals from the model that included the quadratic terms was around 1.653. The simulation results, which were closest to these data characteristics, were used to choose the SRI+XGBoost pathway for empirical application along with SRI. This enabled the extra effect of the refinement stage to be examined, whilst keeping the first imputation method unchanged.

The complete dataset was stored as a reference, and missingness was simulated on the four explanatory variables at levels of 5%, 10%, 20% and 30% without affecting the creatinine measurement. The same MAR mechanism used in the simulation study was assumed, in which the probability of missing was a function of the observed creatinine values. This mechanism is not intended to be a complete list of all causes of missingness that might exist in real laboratory records.

The first application of SRI was followed by the use of XGBoost to refine imputed values at the sites where they were originally missing, while keeping the observed values as is. The imputation errors on each of the explanatory variables were scaled to create a uniform scale and the overall RMSE was computed this way to ensure that variables with large measurement scales were not overwhelming the overall measurement scale. To assess the sensitivity of the results to the random placement of missing values, 500 replications were performed at each missing-data proportion using a different random seed for each replication, while keeping the original 400 observations unchanged. Within each replication, SRI and SRI+XGBoost were evaluated using the same incomplete dataset. Performance was then summarized using the mean and standard deviation. The accuracy of

Table 5. Mean Performance of SRI and SRI+XGBoost, with Standard Deviations Across 500 Replications in the Empirical Data Application

Missingness (%)	method	RMSE			RMSE Model		
		Mean	SD	Improvement (%)	Mean	SD	Improvement (%)
	Original OLS	-	-	-	1.762487	–	-
5%	SRI	1.242275	0.115527		1.749198	0.025193	
	SRI + XGBoost	0.869524	0.105042	30.01	1.74481	0.020561	0.25
10%	SRI	1.246331	0.070131		1.730944	0.032872	
	SRI + XGBoost	0.875376	0.072538	29.76	1.725232	0.030182	0.33
20%	SRI	1.250305	0.048653		1.702197	0.044972	
	SRI + XGBoost	0.887545	0.047458	29.01	1.682004	0.044744	1.19
30%	SRI	1.24787	0.040041		1.66551	0.06125	
	SRI + XGBoost	0.895886	0.03755	28.21	1.625455	0.073185	2.4

the imputation was evaluated using the root mean square error (RMSE), while RMSE Model was used to assess the in-sample fit error of the regression model following imputation.

Table (5) shows that SRI+XGBoost achieved a lower imputation error than SRI at all four missing-data proportions. The percentage improvement in RMSE ranged from 28.2% to 30.0%. The rates of improvement were similar across the four missing-data proportions, although they decreased slightly as the proportion of missing data increased.

However, the improvement of RMSE Model was lesser in magnitude between 0.25% to 2.40%. The improvement was minimal at low missing data rates but increased as the proportion of missing data increased to 20% and 30%. This discrepancy may be explained by the distribution of imputation errors across the explanatory variables, the potential for bias in the regression coefficient estimates, and the constraints associated with the model specification. Consequently, improving the accuracy of the imputation does not necessarily lead to a corresponding improvement in the fit of the regression model.

Also, the values of the standard deviation (SD) indicated that variability between replications varied depending on the percentage of data missing and the performance criterion. The mean performance values can then be viewed with the understanding of their stability over the 500 replications. Across the 500 iterations for each missingness rate, the SRI+XGBoost approach yielded a lower value in 308, 293, 358, and 370 iterations at missingness rates of 5%, 10%, 20%, and 30% respectively. For RMSE Model, the 95% confidence intervals for the paired differences between the SRI+XGBoost and SRI methods were [-0.0060, -0.0028], [-0.0081, -0.0033], [-0.0235, -0.0169], and [-0.0453, -0.0348], corresponding to missingness rates of 5%, 10%, 20%, and 30% respectively. These differences were calculated by subtracting the error value of the SRI method from the error value of the hybrid SRI+XGBoost method.

In general, the empirical results imply that the main advantage of the hybrid pathway was the increased accuracy of missing-value imputation, while the same was true to a lesser extent for the regression model. This helps to assess the effectiveness of the imputation methods in terms of both the imputation accuracy and the impact of the imputations on the regression model following imputation.

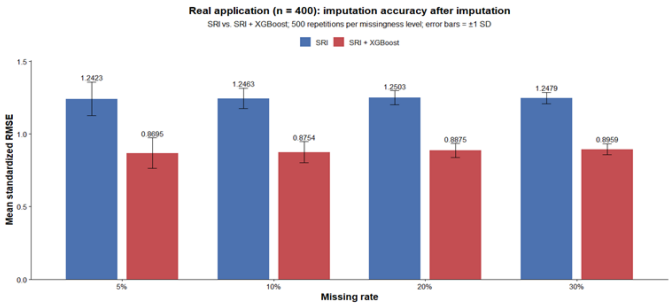


Figure 2. Comparison of SRI and SRI+XGBoost Performance According to RMSE Across Different Missing-Data Proportions in the Empirical Application.

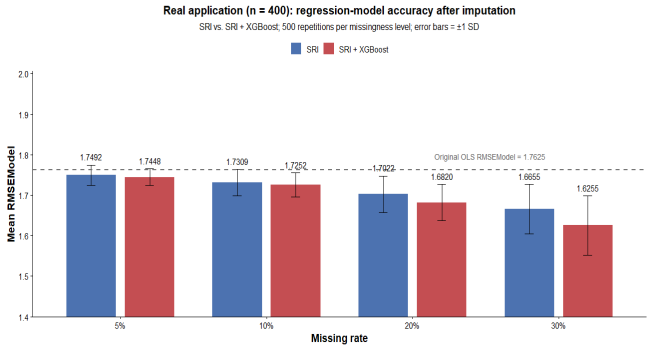


Figure 3. Comparison of SRI and SRI+XGBoost Performance According to RMSEModel Across Different Missing-Data Proportions in the Empirical Application.

4. Conclusion

1. According to the simulation results, the performance of the missing-data imputation methods is influenced by the proportion of missing data, dispersion level and sample size. It was also determined that the method that imputes the most accurately does not necessarily have the lowest regression-model error after data imputation. This highlights the need to assess the imputations techniques with respect to the accuracy of the imputations and the impact of imputing on the following regression model.
2. In terms of imputation accuracy, the hybrid pathways generally outperformed the traditional statistical methods across the simulation conditions considered. However, the pattern of numerical leadership varied with sample size. RF-based pathways were more frequently ranked first at the small and medium sample sizes, whereas XGBoost-based pathways became more prominent at the large sample size.
3. With regard to regression-model-fit (RMSE Model), at the small sample size, the statistical methods were competitive, while pathways based on XGBoost became more prominent with larger sample sizes.
4. The SRI+XGBoost pathway resulted in a clear improvement of the RMSE for the medical data application at all missing data proportions. In contrast, the improvement in the RMSE model was much smaller, particularly at lower missing-data proportions. This suggests that it is not always possible to improve the accuracy of missing value imputation and obtain a comparable improvement in the fit of the regression model.

In general, the sequential hybrid approach contributed to improving imputation accuracy under the conditions studied; however, the extent of this improvement depended on the data characteristics and the evaluation criterion.

Thus, the outcomes are not sufficient to justify a single technique as the best one for all situations. Rather, the choice of imputation method should take into account the characteristics of the data and the purpose of the analysis. Both the simulation study and the practical application were limited to the Missing At Random (MAR) mechanism, with the practical application relying on a single clinical dataset into which missing values were artificially introduced. Furthermore, the final comparisons utilized machine-learning hyperparameters that were selected through preliminary sensitivity analyses and subsequently held constant; therefore, the extent to which these results generalize to other datasets, missingness mechanisms, and hyperparameter settings remains uncertain. The simulation also assumed that the variables follow a multivariate normal distribution; therefore, the performance of hybrid approaches under skewed or heavy-tailed distributions remains to be studied. Furthermore, the simulation design was limited to a fixed correlation coefficient of 0.40 and a restricted degree of non-linearity; therefore, caution should be exercised when generalizing the results to different correlation relationships or non-linear patterns. The evaluation focused on the accuracy of imputing missing values (RMSE) and the goodness-of-fit of the regression model within the sample (RMSE Model), in alignment with the study's primary objectives. Although the bias of regression coefficients, the coverage of confidence intervals, and out-of-sample prediction are also relevant criteria, a comprehensive assessment of these factors falls outside the scope of the comparisons presented here and is left for future research. The current study was limited to continuous explanatory variables. Extending the hybrid framework to incorporate data containing both continuous and categorical variables requires the use of imputation and refinement methods tailored to the nature of each variable. Such an extension represents a direction for future research. Future studies could examine whether additional refinement cycles or adaptive stopping criteria lead to improved performance under various data conditions.

5. Recommendations

Based on the results, we recommend that the method used for the missing values be determined by the data type, the missingness, and the purpose of the analysis rather than an all-or-nothing approach. Also, the method should be assessed based on how well it imputes the values as well as the impact it has on the regression model; if the values are improved, this does not necessarily mean that the model is improved. When the characteristics of the data sets to be compared are similar to the ones analyzed in the empirical application, the SRI+XGBoost pathway could be a suitable alternative to SRI, provided that it is tested in the specific characteristics of the available data sets.

6. Future Research

The hybrid approach could be further tested with more complex data structures that differ in the number of variables in the data and the degree of correlation between them, and with other missing data mechanisms like Missing Completely at Random (MCAR) and Missing Not at Random (MNAR) data. The simulation framework could also be extended to encompass higher levels of nonlinearity and interaction, as well as the application of the proposed approach to other datasets and statistical models, such as generalized linear models and mixed-effects, models to determine how robust the approach is in other settings than those utilized in the current study.

REFERENCES

1. A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society: Series B, vol. 39, no. 1, pp. 1–22, 1977.
2. S. Van Buuren and K. Groothuis-Oudshoorn, *mice: Multivariate imputation by chained equations in R*, Journal of Statistical Software, vol. 45, no. 3, pp. 1–67, 2011.
3. D. J. Stekhoven and P. Bühlmann, *MissForest—non-parametric missing value imputation for mixed-type data*, Bioinformatics, vol. 28, no. 1, pp. 112–118, 2012.
4. A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway, *Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study*, American Journal of Epidemiology, vol. 179, no. 6, pp. 764–774, 2014.

5. F. Tang and H. Ishwaran, *Random forest missing data algorithms*, Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 10, no. 6, pp. 363–377, 2017.
6. Y. Deng and T. Lumley, *Multiple imputation through XGBoost*, Journal of Computational and Graphical Statistics, vol. 33, no. 2, pp. 352–363, 2024.
7. X. Zhang, C. Yan, C. Gao, B. A. Malin, and Y. Chen, *Predicting missing values in medical data via XGBoost regression*, Journal of Healthcare Informatics Research, vol. 4, no. 4, pp. 383–394, 2020.
8. M. M. Saber et al., *The effect of the missing rate and its mechanism on the performance of the imputation methods on different real data sets*, Statistics, Optimization & Information Computing, vol. 14, no. 5, pp. 2131–2141, 2025.
9. M. Abdel-Fattah, M. Mohsen, and A. Mousa, *On a new stacked ensemble framework for imputing missing data in the presence of outliers*, Statistics, Optimization & Information Computing, vol. 14, no. 6, pp. 3526–3545, 2025.
10. R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed., John Wiley & Sons, 2019.
11. S. Van Buuren, *Flexible Imputation of Missing Data*, Chapman and Hall/CRC, 2018.
12. L. Breiman, *Random forests*, Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
13. T. Chen and C. Guestrin, *Xgboost: A scalable tree boosting system*, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
14. P. Cunningham and S. J. Delany, *K-nearest neighbour classifiers: (with Python examples)*, arXiv Preprint arXiv2004.04523, 2020.
15. R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, *Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications*, Journal of Big Data, vol. 11, no. 1, p. 113, 2024.
16. B. D. Leiby and D. K. Ahner, *Multicollinearity applied stepwise stochastic imputation: a large dataset imputation through correlation-based regression*, Journal of Big Data, vol. 10, no. 1, p. 23, 2023.
17. L. D'Agostino McGowan, S. C. Lotspeich, and S. A. Hepler, *The 'Why' behind including 'Y' in your imputation model*, Statistical Methods in Medical Research, vol. 33, no. 6, pp. 996–1020, 2024.
18. P. C. Austin and S. van Buuren, *Imputation of incomplete ordinal and nominal data by predictive mean matching*, Statistical Methods in Medical Research, vol. 34, no. 11, pp. 2163–2182, 2025.
19. E. Albu, S. Gao, L. Wynants, and B. Van Calster, *missForestPredict—Missing data imputation for prediction settings*, PLoS One, vol. 20, no. 11, p. e0334125, 2025.
20. A. C. Schütz and K. R. Gegenfurtner, *Within-subject confidence intervals for pairwise differences in scatter plots*, Psychonomic Bulletin & Review, vol. 32, no. 6, pp. 3238–3251, 2025.
21. S. Kim, H. Kim, and Y. Namkoong, *Ordinal classification of imbalanced data with application in emergency and disaster information services*, IEEE Intelligent Systems, vol. 31, no. 5, pp. 50–56, 2016.
22. T. Shadbahr, M. Roberts, J. Stanczuk, J. Gilbey, P. Teare, S. Dittmer, et al., *The impact of imputation quality on machine learning classifiers for datasets with missing values*, Communications Medicine, vol. 3, no. 1, p. 139, 2023.
23. J. Sinsomboonthong and S. Sinsomboonthong, *New adjusted missing value imputation in multiple regression with simple random sampling and rank set sampling methods*, PLoS One, vol. 20, no. 3, p. e0316641, 2025.
24. A. G. Adetunji and E. Obeng-Gyasi, *Benchmarking statistical methods for environmental chemical mixtures: Prediction, interaction detection, and an applied analysis of metals, essential elements and diabetes*, Stats, vol. 9, no. 5, p. 96, 2026.
25. A. N. Alkhateeb, A. M. Algboory, Z. A. Hadied, O. E. Al-Saqal, and Z. Y. Algamal, *Enhancing surface water potability assessment through a v-SVM hybrid statistical learning model*, Statistics, Optimization & Information Computing, vol. 15, no. 6, pp. 5481–5494, 2026.
26. Al Bakal, S. M., *Adaptive Liu estimator modified for estimating regression coefficients in the presence of Multicollinearity*, Statistics, Optimization & Information Computing, vol. 15, no. 6, pp. 4699-4717, 2026.