BMC Medical Research Methodology

RESEARCH

Open Access

Multi-metric comparison of machine learning imputation methods with application to breast cancer survival

Imad El Badisy[1,2,3*], Nathalie Graffeo[3], Mohamed Khalis[1,2] and Roch Giorgi[3,4]

Abstract

Handling missing data in clinical prognostic studies is an essential yet challenging task. This study aimed to provide a comprehensive assessment of the effectiveness and reliability of different machine learning (ML) imputation methods across various analytical perspectives. Specifically, it focused on three distinct classes of performance metrics used to evaluate ML imputation methods: post-imputation bias of regression estimates, post-imputation predictive accuracy, and substantive model-free metrics. As an illustration, we applied data from a real-world breast cancer survival study. This comprehensive approach aimed to provide a thorough assessment of the effectiveness and reliability of ML imputation methods across various analytical perspectives. A simulated dataset with 30% Missing At Random (MAR) values was used. A number of single imputation (SI) methods - specifically KNN, missMDA, CART, missForest, missRanger, missCforest - and multiple imputation (MI) methods - specifically miceCART and miceRF - were evaluated. The performance metrics used were Gower's distance, estimation bias, empirical standard error, coverage rate, length of confidence interval, predictive accuracy, proportion of falsely classified (PFC), normalized root mean squared error (NRMSE), AUC, and C-index scores. The analysis revealed that in terms of Gower's distance, CART and missForest were the most accurate, while missMDA and CART excelled for binary covariates; missForest and miceCART were superior for continuous covariates. When assessing bias and accuracy in regression estimates, miceCART and miceRF exhibited the least bias. Overall, the various imputation methods demonstrated greater efficiency than complete-case analysis (CCA), with MICE methods providing optimal confidence interval coverage. In terms of predictive accuracy for Cox models, missMDA and missForest had superior AUC and C-index scores. Despite offering better predictive accuracy, the study found that SI methods introduced more bias into the regression coefficients compared to MI methods. This study underlines the importance of selecting appropriate imputation methods based on study goals and data types in time-to-event research. The varying effectiveness of methods across the different performance metrics studied highlights the value of using advanced machine learning algorithms within a multiple imputation framework to enhance research integrity and the robustness of findings.

Keywords Machine learning, Imputation methods, Single and multiple imputation, Performance metrics, Breast cancer survival, Survival analysis

*Correspondence:
Imad El Badisy
ielbadisy@um6ss.ma

Full list of author information is available at the end of the article

Background

Models built using clinical, histological, biological, and radiological variables often face the challenge of missing data, as clinical databases may not have complete records for every patient. Understanding the nature of missing data and choosing the most appropriate imputation method is crucial to ensure reliable and accurate results. This is especially true in prognostic studies, where the presence of missing data poses significant challenges. Missing values in certain covariates for a subset of patients can lead to a substantial reduction in the sample size available for analysis. This data insufficiency can produce biases in parameter estimates, impacting the validity of the subsequent analytical conclusions.

Rubin's taxonomy provides a framework for understanding the nature of missing data, classifying them into three distinct mechanisms: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) [1, 2]. In the MCAR mechanism, the probability of missingness does not depend on either observed or unobserved data. In contrast, for MAR, the probability of missingness depends only on the observed data; it is independent of the unobserved data. Finally, in the MNAR mechanism, which is the most complex of the three, the probability of missingness relies on unobserved data, even when conditioning on observed data.

At its core, a missing data imputation method seeks to replace missing values using a specific model or algorithm to provide the most accurate representation of the original data. One frequent approach to handle missing data is complete-case analysis (CCA). This method limits the analysis to individuals who have complete data (i.e., all values for all the included variables). Consequently, the sample size may be dramatically reduced. CCA may not always be the optimal strategy to use due to the potential for biases and a reduction in statistical power [3].

Two alternative strategies have been suggested to overcome the limitations of CCA: single imputation (SI) and multiple imputation (MI). SI is a straightforward method that replaces a missing value with just one value, such as the mean or mode, resulting in a single dataset. Therefore, no post-imputation operations are necessary. In contrast, MI creates multiple copies of the dataset, with each missing value in each dataset copy replaced with independent random draws from the imputation model [4].The analysis model processes each of these datasets for the primary analysis, with results later pooled to account for imputation uncertainty. The subsequent step in MI is the application of the analysis model to each of the imputed datasets. After this step, Rubin's rules are employed to pool the results from all these analyses into a singular inference [5].

In order to eliminate potential biases in MI, it is essential to include all variables used in the analysis model as well as any interactions between variables in the imputation model [6]. Moreover, the imputation model can incorporate auxiliary variables, which are not included in the analysis model. These variables serve to make the MAR assumption more plausible and to provide additional information on missing values [7].

Given the increasing prevalence of machine learning (ML) imputation algorithms, a variety of strategies based on ML have emerged to tackle the challenge of missing data. However, the performance of these strategies varies greatly depending on the specific context of the problem at hand; this complicates the identification of a gold standard ML approach for handling missing data.

Numerous studies have sought to statistically compare the performance of these different ML-based imputation algorithms using a variety of performance metrics. These range from metrics related to the quantities estimated by the analysis model (post-imputation), to metrics assessing the predictive accuracy of the analysis model, and those assessing the imputation accuracy regarding the variables imputed. Depending on the purpose of the specific statistical analysis, one method may be preferable to another. However, it is often difficult to determine which method produces the most accurate results [8]. For example, the method with the highest predictive power does not necessarily produce the least biased estimates. Conversely, a method that minimizes bias does not necessarily guarantee the best predictive performance.

ML imputation algorithms are based on constructing a predictive model to predict missing values using the available data. Popular ML algorithms, such as RandomForest and KNN, have been extensively used as imputation methods in various prognostic studies in both SI and MI settings. Results from a wealth of comparative studies on these methods highlight the controversy surrounding their relative performance [9–12].

For instance, KNN, which imputes missing values by considering subjects with similar outcome patterns, exhibits the lowest imputation error when compared with other imputation approaches [13]. Another example is missForest, which is based on Random-Forest, an algorithm praised for consistently low imputation error and superiority when variables have high inter-correlations [14, 15]. While the performance of missForest tends to improve with increasing correlation levels, it does not always perform better than other methods when the missing data mechanism is MCAR [16].

Another interesting approach to use ML imputation algorithms is their integration into the MICE (Multiple Imputation by Chained Equations) framework. The mice-CART algorithm, a nonparametric approach for implementing MICE, uses sequential regression trees as the

conditional models. It is recognized for its adeptness in capturing complex relationships with minimal tuning by the data analyst, as conditional models do not need to be explicitly specified [10]. However, studies measuring its performance have observed that while its imputations for continuous values generated via recursive partitioning may preserve interactions, the miceCART algorithm may underestimate main effects [17].

Despite the numerous comparative studies on the performance of different ML imputation algorithms, their study designs tend to lack consistency; moreover, they often solely focus on either the precision of imputed data or the bias following imputation (the latter being tied to parameter estimates) [18]. Additionally, in health-related studies, discussions about how one imputation method might influence epidemiological interpretation, especially within the context of survival analysis, are rare.

The present study aimed to investigate whether the use of various metrics to assess the performance of machine learning-based algorithms can lead to significant differences in the interpretation of results. To this end, we compared eight ML algorithms adapted to missing data imputation. Specifically, we performed a simulation analysis to identify the most effective imputation method according to different performance metrics. Subsequently, we applied the chosen method to real-world breast cancer survival data. This comprehensive analysis provides insights into how different performance metrics can influence the understanding and effectiveness of machine learning algorithms in practical health-related applications.

Methods

Motivating example: breast cancer case study

To apply our most effective method to a real-world situation, we used data from a retrospective cancer survival study of women with breast cancer was conducted in Morocco, involving 711 incident cancers diagnosed in 2009 and followed until December 2014 [19]. That study aimed to identify prognostic variables, including epidemiological, clinical, pathological, biomarker expression, and treatment characteristics. Additionally, other risk factors such as oral contraceptive use, a family history of breast cancer, and obesity were evaluated. The outcome of interest was event-free survival, calculated from the date of surgery or initiation of chemotherapy to the earliest date for either locoregional recurrence or distant metastasis. Thirteen baseline covariates were included in the cancer study's dataset : age, body mass index (BMI), radiotherapy, mammographic size, Scarff-Bloom-Richardson (SBR) grade, nulliparity, lymph nodes (N0, N1, N2, N3), oral contraception, PgR, vascular invasion, trastuzumab, hormone therapy, HER2, and ER. Age refers to the patient's age in years. BMI is a measure of body fat based on height and weight. Radiotherapy indicates whether the patient received radiation treatment. Mammographic size measures the size of the tumor in centimeters. The Scarff-Bloom-Richardson (SBR) grade categorizes tumor aggressiveness into three grades: I, II, and III. Nulliparity refers to whether a woman has never given birth. Lymph nodes are classified into four stages (N0, N1, N2, N3) based on the extent of cancer spread, following the TNM classification system. Oral contraception indicates the use of birth control pills. PgR (Progesterone Receptor) indicates hormone therapy responsiveness. Vascular invasion refers to the presence of cancer cells in blood vessels. Trastuzumab is a targeted therapy drug. Hormone therapy involves treatments that block hormones to slow down or stop cancer growth. HER2 (Human Epidermal Growth Factor Receptor 2) identifies a protein that can promote cancer growth. ER (Estrogen Receptor) status helps determine the cancer's growth sensitivity to estrogen, guiding hormone therapy choices. Table 1 presents the distribution of the variables for the dataset along with the percentage of missing values for each variable. The proportion of missing values varied between covariates, ranging from 0.6 to 43%. Missing values were observed in 607 women, resulting in an almost 85% reduction in sample size in CCA, leaving a study sample of only 105 cases. This example highlights the need to appropriately handle missing data in order to optimally analyse the study data, while both minimizing the risks of loss of efficacy and reducing potential bias.

Machine learning imputation algorithms

Using various performance metrics, our analysis consisted in comparing the performances of eight different ML-based imputation methods using SI and MI (Table 2).

A comprehensive presentation follows for each of the eight algorithms investigated. We will explore SI strategies for KNN, missMDA, CART, missForest, missRanger, and missCforest, and MI strategies for miceCART and miceRandomForest.

K-nearest neighbours (KNN)

The KNN imputation method, which is similar to the hot-deck method, uses donor observations. The value imputed is an aggregation of the values of the k closest neighbors. The method of aggregation depends on the type of variable. For continuous variables, the default aggregation is the median, while for categorical variables it is the most frequent category among the k values.

This method computes a distance to determine the nearest neighbours using a version of Gower's distance that can handle different types of variables, specifically binary, categorical, ordered, continuous, and semi-continuous variables [20]. The distance between two observations is a weighted average of the contributions of each

Table 1 Summary of clinical characteristics for breast cancer patients

Variables	Overall (*N* = 711)
Time	
Mean (SD)	28.0 (22.0)
Median [Min, Max]	23.0 [0, 87.0]
Missing	0 (0%)
Status	
No event	603 (84.8%)
Event	108 (15.2%)
Missing	0 (0%)
Age	
Mean (SD)	48.9 (11.6)
Median [Min, Max]	48.0 [23.0, 89.0]
Missing	0 (0%)
BMI	
Mean (SD)	27.1 (5.18)
Median [Min, Max]	26.4 [16.3, 46.6]
Missing	310 (43.6%)
Radiotherapy	
No	338 (47.5%)
Ues	368 (51.8%)
Missing	5 (0.7%)
Mammographic size	
Mean (SD)	3.50 (2.44)
Median [Min, Max]	3.00 [0, 25.0]
Missing	290 (40.8%)
SBR grade	
SBR I	46 (6.5%)
SBR II	370 (52.0%)
SBR III	191 (26.9%)
Missing	104 (14.6%)
Nulliparity	
No	486 (68.4%)
Yes	157 (22.1%)
Missing	68 (9.6%)
Lymph nodes	
N0	216 (30.4%)
N1	173 (24.3%)
N2	89 (12.5%)
N3	61 (8.6%)
Missing	172 (24.2%)
Oral contraception	
No	280 (39.4%)
Yes	186 (26.2%)
Missing	245 (34.5%)
PgR	
Negative	168 (23.6%)
Positive	432 (60.8%)
Missing	111 (15.6%)
Vascular invasion	
No	366 (51.5%)
Yes	227 (31.9%)
Missing	118 (16.6%)
Trastuzumab	
No	648 (91.1%)

Table 1 (continued)

Variables	Overall (*N* = 711)
yes	58 (8.2%)
Missing	5 (0.7%)
Hormone therapy	
No	366 (51.5%)
Yes	341 (48.0%)
Missing	4 (0.6%)
HER2	
Negative	385 (54.1%)
Positive	116 (16.3%)
Missing	210 (29.5%)
ER	
Negative	191 (26.9%)
Positive	409 (57.5%)
Missing	111 (15.6%)

Table 2 Missing data imputation algorithms and their associated

Imputation Approach	Algorithm	Description	*R* package
Single imputation	KNN	k-Nearest Neighbour Imputation based on a variation of the Gower Distance for numerical, categorical, and ordered variables	VIM
	missMDA	Imputation with principal component analysis (PCA), multiple correspondence analysis (MCA) model or multiple factor analysis (MFA) model	missMDA
	CART	Imputation based on CART algorithm	simputation
	missForest	Nonparametric Imputation using Random Forest	missForest
	missRanger	Alternative implementation of missForest algorithm using predictive mean matching	missRanger
	missCforest	Imputation based on Ensemble Conditional Trees	missCforest
Multiple Imputation	MICE CART	Multiple Imputation based on CART algorithm	mice
	MICE RandomForest	Multiple Imputation based on RandomForest algorithm	mice

variable, with the weights reflecting the importance of each variable.

For continuous variables, KNN calculates the absolute distance between two observations and then divides it by the total range of that variable. The same approach is used for ordinal variables, after converting them to integer variables. For nominal and binary variables, a binary distance of 0/1 is used [21]. KNN imputation preserves the inherent structure and relationships within the data during the imputation process. As a non-parametric method, it refrains from making assumptions about data distribution, and consequently offers a non-parametric solution ideal for datasets with unknown or non-normal distributions. However, its effectiveness is closely tied to the critical choice of k. Additionally, its reliance on a distance metric that accommodates diverse variables introduces sensitivity to this choice, necessitating careful weighting.

Classification and regression trees (CART)

CART, often referred to as decision trees, is a versatile class of ML algorithms. The core distinction between classification and regression trees lies in the criteria used for data splitting and tree pruning; a more detailed discussion on these aspects can be found in [22, 23]. A key feature of the CART algorithm is its inherent ability to preserve interactions between the data in the same dataset.

CART identifies predictors and cut-off points within these predictors that can be used to divide the sample. These partitions split the sample into sub-samples of greater homogeneity. The splitting procedure is repeated on the sub-samples, producing a binary tree structure. The target variable in these models can either be discrete (classification trees) or continuous (regression trees).

When employing CART for imputation of missing data, the algorithm treats the variable with missing values

as the dependent variable (target), and the remaining variables as predictors. The procedure involves building a decision tree where the splits are based on the predictors, the aim being to predict the missing values of the target variable. For continuous variables, regression trees are employed, predicting a value that minimizes the variance within nodes. For categorical variables, classification trees are used, where the prediction corresponds to the most frequent class within a node.

Decision trees possess properties that make them particularly appealing for imputation tasks [24]. They are robust against outliers, can handle multicollinearity and skewed distributions, and are flexible enough to accommodate interactions and non-linear relationships. Moreover, many aspects of model fitting have been automated, meaning minimal tuning by the imputer [10].

Imputation with multivariate data analysis (missMDA)

missMDA is based on the Factor Analysis of Mixed Data (FAMD) method. It handles missing values in mixed data types, including both continuous and categorical variables. This method considers the similarities between individuals and the associations between variables. It performs a Principal Component Analysis (PCA) for continuous variables and a Multiple Correspondence Analysis (MCA) for categorical variables.

The imputation process begins with initial estimates for missing values, using the mean for continuous variables and modes for categorical variables. For mixed data, the Factor Analysis of Mixed Data (FAMD) method, which integrates the principles of PCA and MCA, is utilized to handle both data types cohesively. The algorithm iteratively updates these estimates as follows: in each cycle, FAMD predicts missing values based on observed data; these predictions replace previous estimates, and a new model incorporating these updates is then fitted. The process continues until a convergence criterion is met, typically when changes in imputed values between iterations fall below a set threshold. This ensures that the final imputed values are both statistically robust and contextually aligned with the dataset's structure [25].

Nonparametric imputation using RandomForest (missForest)

The missForest algorithm, a non-parametric imputation technique, utilizes a Random Forest model to iteratively predict missing data values. Initially, it estimates the missing values, typically using the mean or mode, and then employs a Random Forest model to impute missing values for each variable based on observed values from the other variables. The process is iterative, with predictions being updated in each cycle. A key component of this process is the utilization of Out-Of-Bag (OOB) error estimates to assess the accuracy of imputations after each iteration. The algorithm replaces missing values with new predictions and refits the Random Forest models to these updated data, continuing until the OOB error suggests that further iterations would not result in more accurate imputations. This method is adept at handling mixed-type data and provides an estimate of the imputation error, thereby offering an indication of the reliability of the imputed values.

Random Forests differ from the CART methodology by generating multiple trees instead of just one. By averaging numerous trees, the variance of unstable trees is significantly reduced, leading to a more reliable and robust model [23]. The introduction of variability in individual trees yields a more resilient solution, which subsequently enhances the method's accuracy. This variability can be generated through different processes, including bootstrapping and random input selection [22].

Alternative implementation of missForest algorithm using predictive mean matching (missRanger)

An alternative to missForest is the implementation of missRanger, which incorporates in the imputation process the Predictive Mean Matching (PMM) option [26]. Missing values for a variable are imputed using predictions made by a Random Forest, which employs all remaining variables as covariates. The algorithm performs repeated iterations across all variables. This process continues until there is no further improvement observed in the average OOB prediction error of the models; this serves as the stopping criterion for the iterations. An option within this algorithm is the use of the PMM method [27]. For each missing value, PMM creates a donor pool, which consists of complete cases that have predicted values of the outcome closest to the predicted value of the missing entry. From this donor pool, PMM then randomly selects one donor case, and the actual observed value of this selected case is used to replace the missing value. This way, the PMM aims to restore the variance of resulting conditional distributions to a realistic level and preserves the original distribution of the data.

The missRanger imputation method provides a robust alternative to the widely-used missForest algorithm. In particular, the intergrated PMM option ensures the credibility of imputed values [27].

Imputation based on ensemble conditional trees (missCforest)

The missCforest can be utilized for the imputation of numerical, categorical, and mixed-type data [28]. Through ensemble prediction using Conditional Inference Trees (Ctree) as base learners, missing values are imputed [29]. Ctree is a non-parametric class of regression and classification trees that combines recursive

partitioning with conditional inference theory [30]. missCforest redefines the imputation problem as a prediction problem using a single imputation approach. The missing values are predicted iteratively based on the set of complete cases, which is updated at each iteration. There is no predefined stopping criterion, and the imputation procedure stops when all missing data have been imputed. Since the recursive partitioning of Conditional Trees is based on numerous test procedures, this algorithm is resistant to outliers and pays special attention to the statistical relationship between covariates (i.e., variables used for imputation) and the outcome (i.e., variable to be imputed).

Multiple imputation based on CART algorithm (miceCART)

The MICE (Multivariate Imputation by Chained Equations) algorithm is an iterative method for handling missing data. It operates in a variable-by-variable manner, using the observed data to generate initial random imputations for the missing data [31].

miceCART is a variation of the CART algorithm, designed to work within the framework of MI [24]. Initial missing values are estimated by drawing a random distribution from the observed values of each associated variable. A tree is then fitted to the first variable with at least one missing value, using the remaining variables as predictors.

Only individuals with observed values for the outcome are considered. This produces a tree with multiple leaves, each containing a subset of the data. An individual with a missing value for the outcome is placed inside one of these leaves. A random value from this leaf's subset is then selected and used for imputation. This procedure is carried out for each variable with missing data and is ultimately repeated multiple times, resulting in multiple imputed datasets.

Multiple imputation based on RandomForest algorithm (miceRF)

MICE RandomForest uses the Random-Forest algorithm to predict missing values based on observed data, updating imputations until convergence is reached. By combining Random-Forest with MI, a degree of uncertainty can be introduced into the imputation model, which makes it more suitable for generating parameter estimates with desirable characteristics [24].

The miceRFt approach can yield more accurate and reliable imputations when the data exhibits complex patterns that standard regression models (used in traditional MICE) might fail to capture [32].

However, as with any imputation method, the resulting imputations should be validated through methods such as sensitivity analyses or comparisons with CCA to ensure the robustness of the imputation procedure.

Simulation study

In this simulation study, we aimed to evaluate the performance of different ML methods for missing data imputation using different types of performance metrics.

Our simulation design was tailored to incorporate mixed-type covariates (i.e., both continuous and categorical), include non-linear interdependencies between variables, and address the key issue of missing data. The approach used for this comprehensive assessment was as follows:

(1) Generate survival data for all cases.
(2) Estimate the Cox Proportional Hazards (PH) model using complete data.
(3) Introduce missing values into covariates.
(4) Impute missing data using different methods.
(5) Re-estimate the Cox PH model using imputed data.
(6) Compare the performance of the imputation methods using several performance metrics.

Data generation

We generated a total of 500 all-case datasets, each containing 700 observations (no missing values). The datasets included three covariates X_1, X_2 and X_3- in addition to status and observed survival times - which were simulated as follows:

(i) X_1 a continuous covariate, drawn from a standard normal distribution $(mean = 0, sd = 1)$
(ii) X_2 a continuous covariate non-linearly dependent on X_1, with the following distribution: $X_2 = X_1^2 + X_1 + U$ where $U \sim \text{Uniform}\,(0, 1)$
(iii) X_3 a binary covariate, generated from a binomial distribution with $p = 0.5$

To generate individual survival times, we employed the cumulative hazard inversion method, as described by [33]. The method initiates with the specification of log-hazards ratios for covariates X_1, X_2, and X_3, which were set at 0.1, 0.3, and 0.6, respectively.

This technique is based on inverting the survival function, formulated as $S_i^{-1}(u) = H_0^{-1}(-log(u)\, exp(-X_i\beta))$, where $S_i^{-1}(u)$ denotes the inverted survival function for the i th individual, $H_0^{-1}(u)$ represents the inverted cumulative baseline hazard function, X_i is the vector of covariates for the individual, and β comprises the corresponding effect parameters on survival.

To generate individual simulated event times T_i, we applied this formula: $T_i = S_i^{-1}(U_i)$, where U_i is a random variable drawn from a uniform $U(0, 1)$ distribution.

For our analysis, we adopted a generalized Weibull distribution to model the baseline hazard, chosen for its clinically plausible risk shape which closely follows

the pattern observed for the risk of certain cancers. This parametric framework is particularly suitable for representing the aggressive nature of cancer risks in the early observation stages, with a notable decrease in risk following treatment. The generalized Weibull distribution is therefore an ideal fit for scenarios that require a realistic depiction of baseline hazards, as it effectively captures the dynamic risk pattern over time, a trend that aligns with observations highlighted in the present work [34].

The individual censoring times C_i were drawn from an exponential distribution. We calibrated the parameter of exponential distribution to obtain a censoring rate of approximately 30%. The individual observed survival times were then determined as the minimum of the uncensored (T_i^*) and censored survival times $T_i = min\left(T_i^*, C_i\right)$, with the event status adjusted accordingly. We applied administrative censoring at seven years.

Missing values

All complete datasets were subjected to the generation of missingness in variables X_2 and X_3. Specifically, we introduced missing values under the MAR mechanism, which resulted in 30% of the observations in the datasets exhibiting missing values. The probability of missingness in variables X_2 and X_3 was determined by the linear predictor of a logistic model, which comprised the fully observed X_1, the event status indicator, and the observed time [9].

$$l_{p_i} = 0.1 \times \ X_{1i} + 0.1 \times \ \text{Event}_i + 0.1 \times \ T_i$$

Imputation models

All imputation models included the fully observed covariate X_1, event status, survival times, and marginal Nelson-Aalen cumulative hazard, in accordance with the approach detailed in [35].

Estimation models

Irrespective of the method used for imputation, the substantive model, which is the model used in the estimation step (post-imputation), was a multivariate Cox Proportional Hazards [36]. In the case of SI methods, a single model per dataset was estimated; for MI methods, we created ten imputed datasets $(m = 10)$, and combined the log-hazard ratios from the Cox model using Rubin's rules [1].

Performance metrics

To compare the performance of the imputation methods, we considered three distinct types of performance metrics, each based on different criteria. We describe the different metrics of all three in the following sections.

Substantive model-free performance metrics

This set of metrics assesses the performance of imputation methods independently of any estimation model. They provide a direct measure of the quality of the imputed data.

Gower's distance

Gower's Distance serves as a measure to quantify the dissimilarity between datasets, as outlined by [20]. Specifically, it is adept at handling datasets composed of mixed variable types (i.e., including both categorical and continuous variables). In our study, Gower's Distance was employed to evaluate the discrepancy between two distinct datasets: one with complete cases and the other with imputed data. The calculation of Gower's Distance, denoted as D, between any two data points i and j is given as:

$$D\left(i, j\right) = \frac{1}{p}\sum\nolimits_{k=1}^{p} \delta_{\ ijk} \cdot \ d_{ijk}$$

Where p is the total number of variables within the dataset; $\delta_{\ ijk}$ is a binary variable that is assigned a value of 1 when both data points i and j possess a non-missing value for variable k, and 0 when either data point has a missing value for that variable; finally, d_{ijk} represents the normalized distance between data points i and j for variable k.

For continuous variables, d_{ijk} is often calculated as the absolute difference between the values for i and j, normalized by the range of variable across all data points. For categorical variables, d_{ijk} is typically 0 if i and j have the same category for variable k, and 1 otherwise.

To obtain the overall Gower's distance for the dataset, we calculate $D\left(i, j\right)$ for all pairs of data points. The overall Gower's distance is then computed by averaging the pairwise dissimilarities across all pairs of data points, which can be expressed as:

$$\text{Overall Gowers distance} = \frac{2}{n\left(n-1\right)}\sum\nolimits_{i=1}^{n-1}\sum\nolimits_{j=i+1}^{n} D\left(i, j\right)$$

where n is the total number of data points in the dataset. The factor of $\frac{2}{n(n-1)}$ ensures that the average is computed correctly by considering all unique pairs without redundancy. This provides a comprehensive measure of the dissimilarity between the complete (original) and imputed datasets.

Normalized root mean squared error (NRMSE)

For continuous variables, the NRMSE is utilized to quantify the divergence of imputed values from their actual counterparts [37]. This metric serves as an indicator of

imputation accuracy, with a lower NRMSE signifying greater precision in imputation.

$$NRMSE = \sqrt{\frac{\text{mean}\left(\left(X_{\text{true}} - X_{\text{imp}}\right)^2\right)}{\text{var}\left(X_{\text{true}}\right)}}$$

Here, X_{true} represents the actual values of the covariate, and X_{imp} denotes the corresponding imputed values. This normalized measure offers a standardized assessment of the deviation between original and the imputed values.

Proportion of false classified (PFC)

For categorical covariates, the PFC is a metric designed to evaluate the accuracy of imputation. It metrics the proportion of instances in which the imputed binary values deviate from their original counterparts.

$$\text{PFC} = \frac{1}{n}\sum_{i=1}^{n} I\left(X_{\text{true},i} \neq X_{\text{imp},i}\right)$$

Where n represents the total number observations; $X_{\text{true},i}$ and $X_{\text{imp},i}$ denote the actual and imputed values for the i-th observation, respectively; $I(\cdot)$ is an indicator function, equaling 1 when actual and imputed values differ, and 0 otherwise.

The PFC metric therefore provides a straightforward and intuitive measure of imputation accuracy specifically tailored to categorical data, quantifying the frequency of misclassification introduced through the imputation process.

Post-imputation bias, accuracy, and reliability of regression estimates

This set of metrics evaluates the impact of imputation on the bias, accuracy, and precision of regression model estimates.

Post-imputation bias

The post-imputation Bias is a metric used to evaluate the extent to which the imputation process influences the estimation of regression coefficients. It is computed by comparing the estimated coefficients derived from imputed data with those obtained from all cases data.

$$\text{Post-imputation bias} = \beta_{\text{true}} - \hat{\beta}_{\text{imp}}$$

Where β_{true} is the coefficient estimated from the all cases data; and $\hat{\beta}_{\text{imp}}$ is the average estimated coefficient for a given covariate obtained from the imputed data.

This measure quantifies the deviation in the regression coefficients due to the imputation of missing data, thereby assessing the impact of the imputation technique on the regression analysis. A smaller value of post-imputation bias indicates that the imputation process has minimal distortion on the regression estimates, suggesting a more accurate and reliable imputation methodology.

Empirical standard error (empirical SE)

The empirical SE is a statistical metric that quantifies the variability of estimated regression coefficients derived from imputed data. It serves as an indicator of the precision of these estimates, with a smaller SE suggesting greater precision.

$$\text{Empirical SE} = \sqrt{\frac{1}{m-1}\sum_{k=1}^{m}\left(\beta_{\text{imp},k} - \hat{\beta}_{\text{imp}}\right)^2}$$

Where m represents the number of imputed datasets; $\beta_{\text{imp},k}$ is the estimated coefficient from the k-th imputed dataset; and $\hat{\beta}_{\text{imp}}$ is the average of the estimated coefficients across all m imputed datasets.

This measure essentially calculates the standard deviation of the regression coefficients across multiple imputed datasets, providing insight into the spread or dispersion of the coefficient estimates. A lower Empirical Standard Error implies that the coefficient estimates across different imputed datasets are more consistent, indicating a higher level of reliability and stability in the imputation process.

Empirical coverage rate (ECR)

The ECR is a metric used to evaluate the accuracy of imputed data in terms of statistical inference. It measures how often the 95% CI, calculated from the imputed data, encompasses the true regression coefficients.

$$\text{ECR} = \frac{1}{p}\sum_{i=1}^{p} I\left(\beta_{\text{true},k} \in \text{CI}_{95\%,k}\right)$$

Where p is the total number of coefficients being estimated; $\beta_{\text{true},k}$ represents the true value of the k-th coefficient; $\text{CI}_{95\%,k}$ is the 95% confidence interval for the k-th coefficient, as calculated from the imputed data; finally, $I(\cdot)$ is an indicator function that equals 1 if the true coefficient $\beta_{\text{true},k}$ falls within the corresponding 95% confidence interval $\text{CI}_{95\%,k}$, and 0 otherwise.

The ECR quantifies the proportion of times these confidence intervals accurately capture the true parameter values. A value close to 95% indicates that the confidence intervals derived from the imputed data are reliable and effectively represent the uncertainty surrounding the parameter estimates. This metric is crucial for assessing the validity of statistical inferences made from imputed datasets.

Post-imputation predictive accuracy

This set of metrics assesses how well a predictive model performs when trained on imputed data, compared to its performance on all cases.

Time-dependent area under the ROC curve (AUC)

The AUC is a metric used to assess the discriminatory power of a predictive model in a time-dependent context, which is particularly relevant in cancer survival analysis. This measure evaluates the ability of the model to correctly distinguish between different outcome classes (e.g., event vs. no event) at various time points [38]. The time-dependent AUC is calculated as follows.

$$\mathrm{AUC}(t) = \int_0^t \mathrm{Sensitivity}(u) \times \mathrm{Specificity}(u)\, du$$

Where $\mathrm{AUC}\,(t)$ is the area under the ROC curve up to time t; $\mathrm{Sensitivity}\,(u)$ is the true positive rate (sensitivity) at time u; and $\mathrm{Specificity}(u)$ is the derivative of the true negative rate (specificity) with respect to time at u.

In practical terms, this metric integrates the sensitivity and the rate of change of specificity over time, providing a comprehensive measure of the model's performance in accurately classifying individuals at risk over the entire follow-up period. A higher Time-Dependent AUC indicates better discriminative ability of the model at different time points, which is crucial for the accurate prediction of cancer survival outcomes in clinical research.

C-index

The C-index, or Concordance index, is a widely used metric in cancer survival analysis to evaluate the concordance between predicted and observed outcomes. It provides a measure of the predictive accuracy of a model, particularly in the context of censored survival data [39]. The C-index is computed as follows.

$$\text{C-index} = \frac{\sum_{i<j} I(T_i < T_j) \cdot I(H_i > H_j) + 0.5 \cdot I(H_i = H_j)}{\sum_{i<j} I(T_i < T_j)}$$

Where T_i and T_j are the observed survival times for pairs of individuals i and j, respectively; H_i and H_j are the predicted instant hazards (or risks) for individuals i and j, respectively; and $I\,(\cdot\,)$ is an indicator function that equals 1 if the condition is true and 0 otherwise.

The C-index quantifies the proportion of all usable patient pairs in which the predictions and outcomes are concordant. A pair is considered usable if one of the pair's members experiences the event of interest and the other is either censored or experiences the event at a later time. A C-index of 0.5 suggests no predictive discrimination (random chance), while a C-index of 1.0 indicates perfect discrimination. In practice, a high C-index indicates that the model has good predictive accuracy, successfully ranking individuals in the order of their observed times to the event, which is particularly important in the analysis of imputed datasets in survival studies.

Importantly, in our evaluation approach, each of these metrics offers a unique perspective on the performance of imputation methods, covering aspects from the precision of imputed data to the impact on subsequent statistical analyses and predictive modeling. This comprehensive evaluation ensures a robust understanding of the strengths and limitations of the different imputation methods studied here.

Results

Simulation study: performance metrics

Substantive model-free performance metrics

Gower's distance The evaluation of imputed versus fully observed datasets revealed that CART and missCforest yielded the most accurate results. In other words, they exhibited the smallest Gower's distance values, at 0.0128 and 0.013, respectively, indicating high fidelity of their imputations to the fully observed datasets (Fig. 1).

NRMSE In the context of the continuous covariate (X_2), missForest and miceCART outperformed the other six methods, with the smallest NRMSE values at 0.2964 and 0.3065, respectively, demonstrating their effectiveness at imputing continuous data (Fig. 1).

PFC For the binary covariate (X_3), missMDA and CART were the most effective methods, achieving the lowest PFC values at 0.4231 and 0.4328, respectively, suggesting superior performance in accurately classifying binary data (Fig. 1).

Post-imputation Bias and Accuracy of Regression estimates

Bias For the continuous covariate (X_2), CART and miceRF showed minimal bias towards the null, with bias metrics of -0.0012 and −0.0042, respectively, and relative biases of -0.3931% and −1.4110%, respectively. Conversely, missMDA exhibited the highest bias at 0.0455 with a relative bias of 15.1613% (Table 3). For the binary covariate (X_3), miceRF and miceCART were the least biased methods, with bias metrics of 0.005 and −0.0077, respectively, and relative biases of 0.8383% and −1.29%, respectively, while missMDA showed the highest bias at 0.1780 with a relative bias of 29.6638% (Table 4).

Efficacy The empirical SE indicated that all imputation methods offered more efficient estimates than the CCA, the latter yielding the lowest average efficiency estimates (Tables 3 and 4).

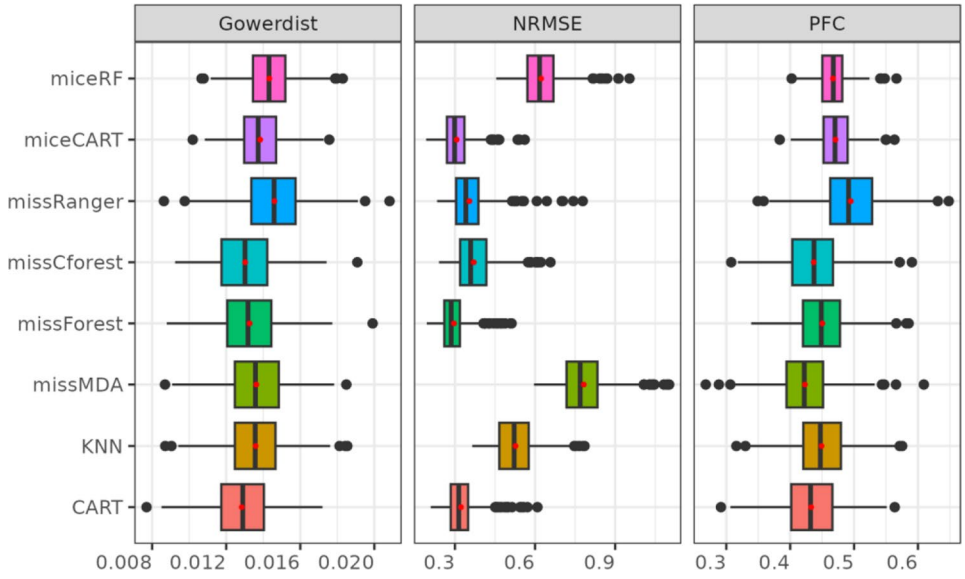

Fig. 1 Box plot of imputation methods across multiple datasets, evaluating Gower distance for all variable types, NRMSE for continuous variables, and PFC for binary variables. Each box represents the interquartile range of the corresponding performance metric, with the red point indicating the mean performance for each method

Table 3 Comparative analysis of imputation methods on regression estimates for variable X_2 (continuous) using bias, empirical SE, relative bias, ECR

Method	Bias	Empirical SE	Relative bias	ECR
miceRF	-0.0042	0.0353	-1.4110	0.970
miceCART	-0.0055	0.0334	-1.8466	0.956
missRanger	0.0056	0.0315	1.8605	0.930
missCforest	0.0173	0.0321	5.7782	0.882
missForest	0.0181	0.0320	6.0396	0.890
missMDA	0.0455	0.0324	15.1613	0.686
KNN	0.0212	0.0315	7.0813	0.882
CART	-0.0012	0.0313	-0.3931	0.918
complete cases	-0.0046	0.0376	-1.5364	0.932
all cases	0.0024	0.0313	0.7850	0.962

Table 4 Comparative analysis of imputation methods on regression estimates for variable X_3 (binary) using bias, empirical SE, relative bias, ECR

Method	Bias	Empirical SE	Relative bias	ECR
miceRF	0.0050	0.1034	0.8383	0.948
miceCART	-0.0077	0.1018	-1.2900	0.960
missRanger	-0.1118	0.0936	-18.6250	0.718
missCforest	0.1263	0.0949	21.0544	0.690
missForest	0.0744	0.0946	12.4073	0.816
missMDA	0.1780	0.0965	29.6638	0.532
KNN	0.0715	0.0948	11.9212	0.818
CART	0.1365	0.0951	22.7439	0.662
complete cases	-0.0224	0.1130	-3.7395	0.954
all cases	-0.0106	0.0940	-1.7727	0.962

ECR Except for miceCART and miceRF, all six other imputation methods provided suboptimal ECR for both covariate types. miceCART achieved 95.6% and 96% coverage for continuous (X_3) and binary covariates (X_3), respectively, while miceRF ensured 97% and 94.8% coverage, respectively (Tables 3 and 4).

Post-imputation Predictive Accuracy of Cox Models

AUC and C-index In predictive accuracy, missMDA and missCforest provided the best results. The missMDA attained an AUC of 0.734 and a C-index of 0.675, showcasing high predictive power, while the missCforest had an AUC of 0.726 and a C-index of 0.667. In contrast, missRanger scored lowest for both the AUC (0.704) and C-index (0.652) metrics (Fig. 2).

Overview of the three type of performances metrics

For the continuous covariate, the plot (Fig. 3) visualizes the trade-offs between bias and the NRMSE for each imputation method, with the size of the points proportional to the C-index. The SI methods CART and missForest managed to maintain a delicate equilibrium, highlighting low bias and NRMSE, while achieving a moderate to high C-index. Conversely, although missMDA presented a higher bias and NMRSE, it achieved a high C-index, as indicated by its larger point size. Additionally, of the eight MI methods, miceCART and miceRF stood out by providing the least biased estimates. Nevertheless, miceRF exhibited a relatively high NRMSE.

For the binary covariate, the plot (Fig. 4) visualizes the trade-offs between bias and the PFC for each imputation

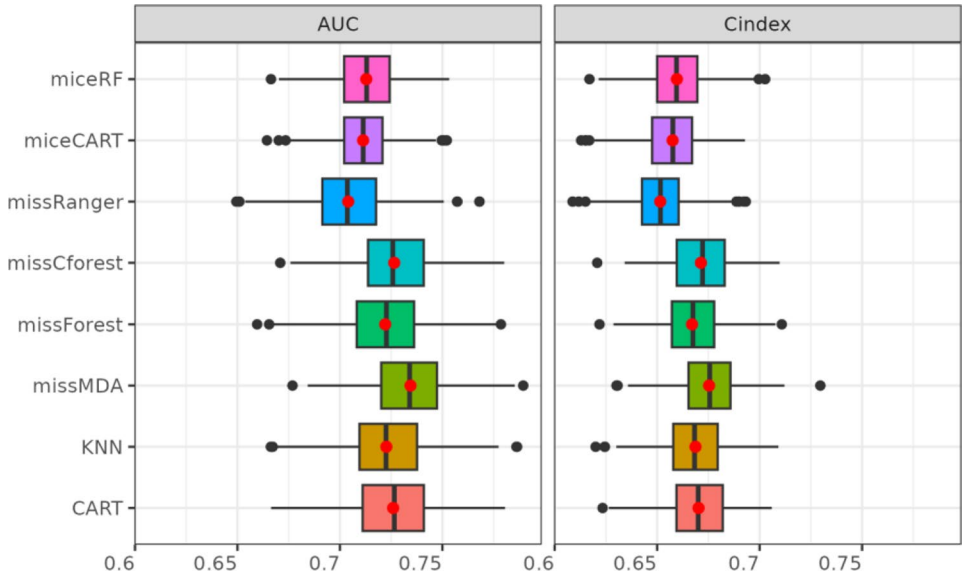

Fig. 2 Box plots comparing post-imputation predictive accuracy of various imputation methods using time-dependent AUC and Concordance Index (Cindex) metrics. Each box represents the interquartile range of the corresponding performance metric, with the red point indicating the mean performance for each method

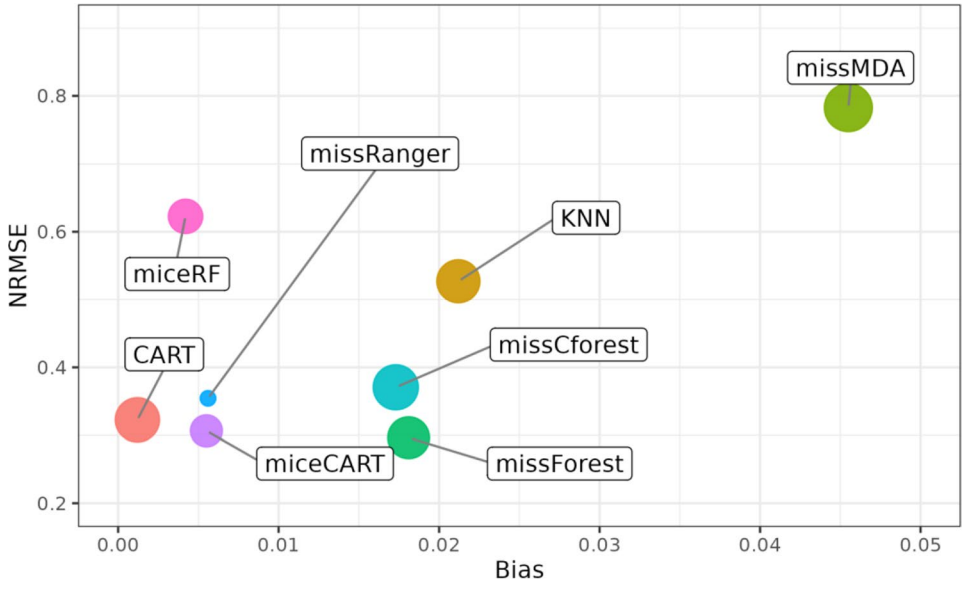

Fig. 3 : Overview comparison of ML imputation methods: Bias vs. NRMSE trade-off weighted by C-index metric for continuous covariates

method, with the size of the points reflecting the C-index value. Notably, the SI methods missForest and KNN exhibited a balanced performance with relatively low bias and moderate PFC, while maintaining a moderate to high C-index. At the other end of the spectrum, missMDA demonstrated stronger bias but compensated with a low PFC, and a high C-index (reflected by its larger point size). Moreover, the MI methods miceCART and miceRF provided the less biased estimates but relatively high PFC.

Application to the motivating example: Cox PH model with imputed dataset

In this section we describe the application of our work to real-world breast cancer study. We used the Cox PH model to estimate the effect of one or several variables on the time to event occurrence without requiring the specification of the underlying hazard function. This model is defined as:

$$h\left(t\right) = h_0\left(t\right) exp\left(\beta^T X\right)$$

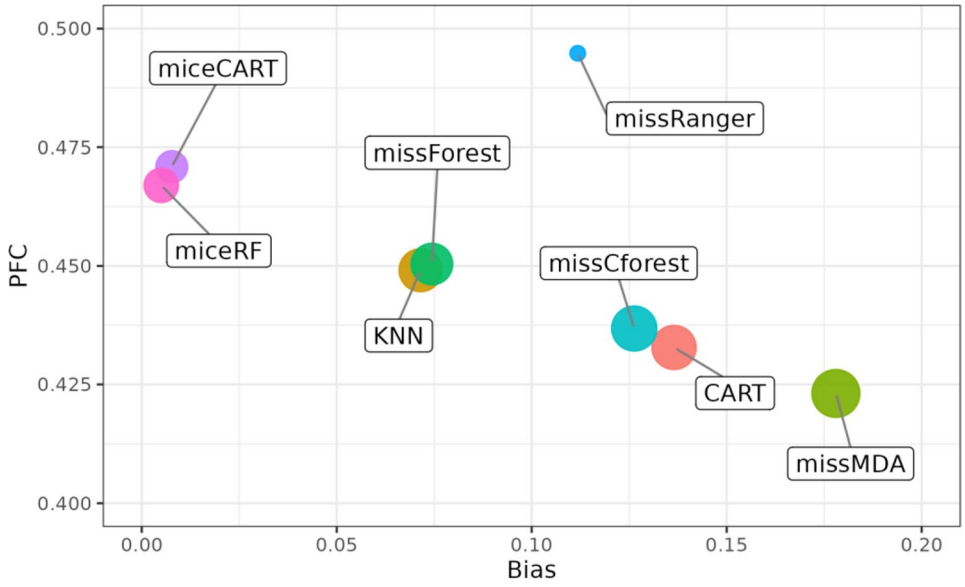

Fig. 4 Overview comparison of ML imputation methods: Bias vs. PFC trade-off weighted by C-index metric for the binary covariate

where $h(t)$ is the instantaneous hazard function at time t, $h_0(t)$ the baseline hazard at time t, β is a column vector of coefficients $(\beta_1, \beta_2, ., \beta_p)^T$, and X is a column vector of covariates $(X_1, X_2, ., X_p)^T$. These coefficients, estimated from the beta coefficients, describe how the risk of the event changes per one-unit increase in covariates. A positive β reflects a higher risk, while a negative reflects a lower risk. The model assumes proportional hazards and log linearity, indicating constant covariate effects over time and linear relationships on the log-hazard scale. In our application, we used a simple Cox model without interaction terms and assumed a proportional hazard similar to that reported in the original article [19].

Our objective in this application was to highlight the variability in outcomes produced by CCA (i.e., with no imputation) and miceRF, as any variability could lead to contradictory epidemiological and clinical interpretations. We focused on five prognostic factors widely recognized in existing literature as follows: Radiotherapy, Lymph nodes, PgR, Hormone therapy, and ER status (see above) [40, 41].

To handle missing data in our motivating example, we used the miceRF algorithm for imputation. This method, when investigated by [9], demonstrated that it produced unbiased estimates with better ECR. Our simulation results comfirmed this. Moreover, they indicated that thanks to the algorithm's tendency to yield less biased estimation coefficients with the most credible coverage rates, it is very effective, particularly in contexts where the impact on estimation coefficients implies significant consequences for the validity and reliability of research findings.

With no imputation (i.e., CCA), a positive ER status was associated with a favorable prognosis (HR=0.09, 95% CI: 0.02–0.38, p=0.001), indicating a potential protective effect. Similarly, hormone therapy was also linked to a positive prognosis (HR=0.11, 95% CI: 0.02–0.64, p=0.015), acting as a protective factor. Furthermore, a negative PgR had a significant protective effect (HR=0.01, 95% CI: 0.00-0.21, p=0.003). In contrast, radiotherapy did not exhibit a significant impact on event-free survival (HR=2.91, 95% CI: 0.48–17.50, p=0.244) (Table 5).

Interestingly, pre-imputation results (i.e., CCA) diverged significantly from post-imputation findings after imputing with miceRF. For example, after imputation, radiotherapy was significantly associated with better event-free survival (HR=0.49, 95% CI: 0.31–0.78, p=0.003), suggesting a protective role. Moreover, unlike the CCA results, ER after imputation no longer had a significant prognostic effect (HR=0.82, 95% CI: 0.47–1.47, p=0.493). This was also true for PgR status (HR=0.79, 95% CI: 0.40–1.59, p=0.509), indicating no clear impact on event-free survival. Similarly, hormone therapy no longer showed a significant prognostic effect after imputation (HR=0.88, 95%CI: 0.51–1.51, p=0.635). However, a significant effect on event-free survival was observed in the case of N3 lymph node status (HR=2.64, 95% CI: 1.27–5.5, p=0.011) (Table 5).

Discussion

Missing data imputation methods that are easy to implement but which are potentially biased, such as CCA, are still commonly used, even if they rely on strong assumptions and despite the availability of potentially more appropriate techniques [42]. In order to investigate this

Table 5 Multivariate analysis results for breast cancer: comparison between complete cases and miceRF imputed data

Label	Levels	All	HR (Complete Cases)	HR (mice RF)
Age	Mean (SD)	48.9 (11.6)	0.94 (0.88-1.00, $p=0.051$)	0.99 (0.97–1.01, $p=0.289$)
BMI	Mean (SD)	27.1 (5.2)	1.01 (0.87–1.17, $p=0.944$)	0.99 (0.94–1.05, $p=0.775$)
Radiotherapy	no	338 (47.9)	-	-
	yes	368 (52.1)	2.91 (0.48–17.50, $p=0.244$)	0.49 (0.31–0.78, $p=0.003$)
Mammographic size	Mean (SD)	3.5 (2.4)	1.07 (0.85–1.35, $p=0.550$)	1.06 (0.98–1.15, $p=0.154$)
SBR grade	SBR I	46 (7.6)	-	-
	SBR II	370 (61.0)	0.62 (0.09–4.22, $p=0.628$)	0.95 (0.40–2.28, $p=0.909$)
	SBR III	191 (31.5)	1.35 (0.21–8.87, $p=0.755$)	1.30 (0.52–3.28, $p=0.574$)
Nulliparity	no	486 (75.6)	-	-
	yes	157 (24.4)	1.24 (0.22–7.01, $p=0.808$)	1.03 (0.60–1.75, $p=0.922$)
Lymph nodes	N0	216 (40.1)	-	-
	N1	173 (32.1)	0.69 (0.16–3.05, $p=0.624$)	0.98 (0.53–1.83, $p=0.951$)
	N2	89 (16.5)	0.70 (0.09–5.31, $p=0.731$)	1.68 (0.78–3.59, $p=0.175$)
	N3	61 (11.3)	1.68 (0.28–9.95, $p=0.568$)	2.64 (1.27–5.50, $p=0.011$)
Oral contraception	no	280 (60.1)	-	-
	yes	186 (39.9)	0.74 (0.19–2.87, $p=0.659$)	0.88 (0.52–1.49, $p=0.630$)
PgR	positive	432 (72.0)	-	-
	negative	168 (28.0)	0.01 (0.00-0.21, $p=0.003$)	0.79 (0.40–1.59, $p=0.509$)
Vascular invasion	no	366 (61.7)	-	-
	yes	227 (38.3)	1.98 (0.51–7.66, $p=0.324$)	1.10 (0.65–1.87, $p=0.703$)
Trastuzumab	no	648 (91.8)	-	-
	yes	58 (8.2)	0.21 (0.02–2.02, $p=0.177$)	0.66 (0.26–1.63, $p=0.358$)
Hormone therapy	no	366 (51.8)	-	-
	yes	341 (48.2)	0.11 (0.02–0.64, $p=0.015$)	0.88 (0.51–1.51, $p=0.635$)
HER2	negative	385 (76.8)	-	-
	positive	116 (23.2)	2.85 (0.46–17.69, $p=0.260$)	1.63 (0.90–2.95, $p=0.103$)
ER	negative	191 (31.8)	-	-
	positive	409 (68.2)	0.09 (0.02–0.38, $p=0.001$)	0.82 (0.47–1.45, $p=0.493$)

issue, in the present study, we evaluated the performance of eight ML imputation methods for missing data through simulations using several performance metrics.

An imputation method is typically deemed superior when it demonstrates certain key characteristics: minimal bias and relative bias in regression coefficient estimates, coverage rates closely aligning with the nominal coverage probability, and lower values of PFC and NRMSE. However, an important distinction arises in the performance of SI versus MI methods. The former often result in more biased regression coefficients compared to the latter. On the other hand, they tend to outperform the latter in terms of predictive accuracy. Therefore, it is essential to carefully consider the specific performance metrics being used to evaluate imputation methods. These metrics can range from the precision of regression estimates and the predictive accuracy of the model to the accuracy of variable imputation itself [8]. The choice of one metric over another can significantly influence the perceived performance of an imputation method.

Our study revealed that for binary covariate imputation, the algorithms missMDA and CART were the most precise, whereas missForest and miceCART excelled in imputing continuous covariates. Concerning the regression estimates, while the lowest bias in imputation of continuous covariates was observed with CART and miceRF; for binary covariates, the lowest bias was found with miceRF and miceCART. Additionally, miceCART and miceRF were superior in achieving optimal ECR. With regard to predictive accuracy, missMDA and missForest stood out for their high accuracy.

Overall, our findings underline that Random Forest within an MI framework is superior to SI techniques and CCA. This finding justifies the choice of applying this method to breast cancer in a context where the prognostic effect of factors is of great importance.

The discrepancy found in the prognostic effects of the four covariates examined before and after imputation is striking. Indeed, this discrepancy could potentially lead to misleading clinical conclusions, if not properly accounted for. Accordingly, we strongly recommend choosing robust imputation methods when handling missing data in clinical prognostic studies. Of course this choice depends on whether the objective of the study itself is predictive or inferential in nature. Moreover, this discrepancy demonstrates the importance of explicitly indicating whether the analysis is based on complete

cases or whether imputation methods were used, and if so, which one.

Although the impact of missing data on a study's final results can be quite significant, many studies do not sufficiently address it. A review of time-to-event studies in oncology highlighted this issue [43]. Specifically, out of 148 studies reviewed, 79 (53%) reported using complete cases, compared to 33 (22%) for MI. Importantly, of all the studies, 128 (86%) did not specify the assumptions their chosen analysis method made regarding missing data. These findings would suggest that while missing data are a common issue in many studies, there seems to be a lack of transparency as to how they were addressed or what assumptions were made.

Surprisingly, empirical studies investigating the performance of different imputation methods use inherently different metrics. For instance, missForest was reported as a effective tool for handling missing values, providing highly accurate imputations, and outperforming other common imputation methods. Yet these evaluations of missForest were based on its predictive accuracy [14, 44]. This approach may be insufficient, especially in epidemiological studies which aim to identify factors significantly associated with the outcome. An imputation method that merely minimizes the prediction error can be problematic, as it does not attempt to preserve the joint distribution of the data. Such an approach could lead to biased parameter estimates [8, 9], as found in our simulation study. Therefore, studies comparing imputation methods should systematically report post-imputation bias and its coverage rate, and not just post-imputation predictive accuracy.

Our comparative study provides a succinct summary of the performance metrics associated with the eight evaluated imputation methods. It focuses on minimizing bias and imputation errors while simultaneously aiming to maximize predictive accuracy. Indeed, the choice of metrics to use when an imputation method is selected greatly depends on the analytical goal, be it estimation or prediction. Accordingly, the appropriateness of using a metric in an imputation method is intrinsically tied to the specific objectives of the analysis.

SI methods are typically discouraged for handling cases where more than 10% of data is missing, as they tend to underestimate the variability in the data, which leads to inaccuracies in coverage rates [13]. However, when incorporated into a MI framework - such as CART combined with MICE (Multiple Imputation by Chained Equations) - the application of SI methods becomes more feasible and straightforward. MICE works by iteratively fitting a predictive model to each variable with missing values, while using other variables as predictors. In this context, any predictive model can be employed to estimate missing values. This flexibility is evident in the use of various MICE models, including miceCART and miceRF.This demostrates that even methods which are not traditionally recommended in situations where there is a high level of missingness can be effectively adapted in MI.

Moreover, in the face of model specification difficulties and/or the presence of complex interactions among variables, non-parametric algorithms like tree-based algorithms (CART, Cforest, Random Forest) present an excellent alternative. These types of algorithms do not require the same assumptions as conventional parametric models; this makes them more flexible and more able to capture complex relationships in the data [8, 9, 17].

Furthermore, a promising approach to the challenge of misspecification in an imputation model is the heterogeneous ensemble imputation strategy. One such approach is Super Learning, a method grounded in heterogeneous ensemble learning that integrates multiple algorithms within a single meta-algorithm. This strategy might effectively tackle the issue of imputation model miss-specification. Imputers such as MISL (Multiple Imputation by Super Learning) and SuperMICE exemplify the application of the Super Learner algorithm, a meta-algorithm that aggregates predictions from an array of base algorithms, thereby offering compelling properties [45, 46]. A more in-depth exploration is required to evaluate its potential advantages over the methodologies currently in use.

It is worth noting that our analyses and conclusions are based on the assumption that the data are MAR. If the missing data mechanism deviates from MAR, especially in the case of MNAR, the effectiveness of the imputation methods and the applicability of our conclusions may differ. While our results would not significantly change under an MCAR assumption, the generalizability of our findings could be limited under different missingness conditions, particularly MNAR, which is difficult to check in practice.

Future extensions of this research could focus on exploring the imputation performance of MLO methods, especially in contexts where covariates exhibit time-varying and non-linear effects. These complex scenarios, which pose considerable challenges for traditional statistical methods, might significantly benefit from the adaptability and sophistication offered by ML algorithms.

Conclusion

Addressing missing data in observational time-to-event studies is crucial for preserving the integrity and accuracy of findings. Different performance metrics can evaluate imputation methods from various perspectives; this highlights the importance of method selection based on the study's final objective. Whether the goal is inference, predictive performance, or minimizing original data distortion independent of any analysis model, the choice

of a particular imputation method should be carefully considered.

When a significant predictor has a high proportion of missing values, the results should be interpreted as a foundation for hypothesis generation rather than as definitive conclusions. This nuanced approach highlights the necessity of a robust strategy for handling missing data in cancer survival studies.

Although CCA remains a common strategy, we encourage researchers to explore modern techniques, including ML algorithms like Random Forest within an MI framework. These advanced methods offer advantages over SI techniques and CCA by adeptly managing complex missing data scenarios. Adopting a consistent methodology for managing missing data can significantly reduce bias in parameter estimates, thereby enhancing the credibility, reliability, and robustness of research findings.

Acknowledgements
We extend our gratitude to Meriem Slaoui for generously providing the Moroccan Breast Cancer dataset, which was originally published as open data (https://datadryad.org/stash/dataset/doi:10.5061/dryad.2pc43). This dataset served as the motivating example in our study. Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its High Performance Computing resource.

Author contributions
IE: Conceptualization, Formal Analysis, Writing, Review, and Editing. NG: Review and Editing. MK: Supervision. RG: Review and Editing, Supervision, Validation.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability
All data used in the simulation part were generated following our specific simulation design. The necessary code to reproduce our simulation is available via https://github.com/ielbadisy/ML-imputers-BC-simulation. The data used in the working example were kindly provided by its original author.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Mohammed VI Center For Research and Innovation, Rabat, Morocco
[2]International School of Public Health, Mohammed VI University of Sciences and Health, Casablanca, Morocco
[3]Aix Marseille Univ, INSERM, IRD, ISSPAM, SESSTIM, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Marseille, France
[4]Aix Marseille Univ, APHM, INSERM, IRD, SESSTIM, Hop Timone, Biostatistique et Technologies de l'Information et de la Communication, Sciences Economiques & Sociales de la Santé & Traitement de l'Information Médicale, ISSPAM, Hop Timone, BioSTIC, Biostatistique et Technologies de l'Information et de la Communication, Marseille, France

Received: 13 May 2024 / Accepted: 8 August 2024
Published online: 30 August 2024

References

1. Rubin DB. Inference and missing data. Biometrika. 1976;63:581–92.
2. Graham JW. Missing data analysis: making it work in the real world. Ann Rev Psychol. 2009;60:549–76.
3. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med. 2010;29:2920–31.
4. Little RJA, Rubin DB. Single imputation methods. In: Statistical analysis with missing data. 2002. pp. 59–74.
5. Rubin DB. Multiple imputation for nonresponse in surveys. John Wiley; Sons; 2004.
6. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338.
7. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30:377–99.
8. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Methodol. 2020;20:1–12.
9. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am J Epidemiol. 2014;179:764–74.
10. Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. Am J Epidemiol. 2010;172 9:1070–6.
11. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010;50:105–15.
12. Lakshminarayan K, Harp SA, Goldman RP, Samad T. Imputation of missing data using machine learning techniques. In: KDD. 1996.
13. Schwender H. Imputing missing genotypes with weighted k nearest neighbors. J Toxicol Environ Health Part A. 2012;75:438–46.
14. Waljee AK, Mukherjee A, Singal AG, Zhang Y, Warren J, Balis U, et al. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open. 2013;3:002847.
15. Tang F, Ishwaran H. Sci J. 2017;10:363–77. Random forest missing data algorithms. Statistical Analysis and Data Mining: The ASA Data.
16. Solaro N, Barbiero A, Manzi G, Ferrari PA. A simulation comparison of imputation methods for quantitative data in the presence of multiple data patterns. J Stat Comput Simul. 2018;88:3588–619.
17. Doove LL, Van Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. Comput Stat Data Anal. 2014;72:92–104.
18. Oberman HI, Vink G. Toward a standardized evaluation of imputation methodology. Biom J. 2024;66:2200107.
19. Slaoui M, Mouh FZ, Ghanname I, Razine R, Mzibri ME, Amrani M. Outcome of breast cancer in Moroccan young women correlated to clinic-pathological features, risk factors and treatment: a comparative study of 716 cases in a single institution. PLoS ONE. 2016;11:0164841.
20. Gower JC. A general coefficient of similarity and some of its properties. Biometrics. 1971;27:857.
21. Kowarik A, Templ M. Imputation with the r package VIM. J Stat Softw. 2016;74:1–16.
22. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC; 1984.
23. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: Data IntroductionMining. Inference Prediction. 2009;2.
24. Doove LL, Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. Comput Stat Data Anal. 2014;72:92–104.
25. Josse J, Husson F, missMDA. A package for handling missing values in multivariate data analysis. J Stat Softw. 2016;70:1–31.
26. Mayer M, Mayer MM, Package. missRanger. R Package; 2019.
27. Wright MN, Ziegler A, Ranger. A fast implementation of random forests for high dimensional data in c++and r. 2015.
28. EL Badisy I, missCforest. Ensemble conditional trees for missing data imputation. 2023.

29. Hothorn T, Hornik K, Zeileis A, Ctree. Conditional inference trees. Compr R Archive Netw. 2015;8.
30. Strasser H, Weber C. On the asymptotic theory of permutation statistics. 1999.
31. Buuren S, Groothuis-Oudshoorn K, Mice. Multivariate imputation by chained equations in r. J Stat Softw. 2011;45:1–67.
32. Slade E, Naylor MG. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. Stat Med. 2020;39:1156–66.
33. Bender R, Augustin T, Blettner M. Generating survival times to simulate cox proportional hazards models. Stat Med. 2005;24:1713–23.
34. Giorgi R, Belot A, Gaudart J, Launoy G. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. Stat Med. 2008;27:6310–31.
35. White IR, Royston P. Imputing missing covariate values for the cox model. Stat Med. 2009;28:1982–98.
36. Cox DR. Regression models and life-tables. J Royal Stat Soc Ser B. 1972;34:187–202.
37. Oba S, Sato M, Takemasa I, Monden M, Matsubara K, Ishii S. A bayesian missing value estimation method for gene expression profile data. Bioinformatics. 2003;19:2088–96.
38. Chambless L, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. Stat Med. 2006;25.
39. Gerds TA, Kattan MW, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Stat Med. 2013;32:2173–84.
40. Group EBCTC, et al. Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. Lancet. 2005;366:2087–106.
41. Karihtala P, Jääskeläinen A, Roininen N, Jukkola A. Prognostic factors in metastatic breast cancer: a prospective single-centre cohort study in a Finnish university hospital. BMJ open. 2020;10:e038798.
42. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a cox proportional hazards model: a resampling study. BMC Med Res Methodol. 2010;10:1–10.
43. Carroll OU, Morris TP, Keogh RH. How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review. BMC Med Res Methodol. 2020;20:1–15.
44. Ramosaj B, Pauly M. Predicting missing values: a comparative study on non-parametric approaches for imputation. Comput Stat. 2019;34:1741–64.
45. Carpenito T, Manjourides J. MISL: multiple imputation by super learning. Stat Methods Med Res. 2022;31:1904–15.
46. Laqueur HS, Shev AB, Kagawa RMC. SuperMICE: an ensemble machine learning approach to multiple imputation by chained equations. Am J Epidemiol. 2022;191:516–25.

Publisher's Note