

Impact of Missing Data Imputation on The Structure of Real-World Clinical Datasets: A Comparison of Median, K-Nearest Neighbours, and MICE Approaches

Karol Aguiar Quevedo

Fundación Instituto Valenciano de Oncología

Alilis Fontana Bellorín

Hospital Universitari i Politècnic La Fe

Carlos Jordá Jordá Aragón

Hospital Universitari i Politècnic La Fe

Cristóbal Aguilar-Gallardo

`cristobal_aguilar@iislafe.es`

Health Research Institute Hospital La Fe (IIS La Fe)

Research Article

Keywords: missing data, imputation, k-nearest neighbours, multiple imputation by chained equations, clinical datasets, distributional fidelity, multivariate structure, biomedical research

Posted Date: August 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10549611/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

In clinical research, median substitution remains widely used for handling missing data despite its known tendency to compress variance and distort distributions. We compared median imputation, k-nearest neighbours (KNN), and multiple imputation by chained equations with predictive mean matching (MICE/PMM) on two independent clinical cohorts: a lung transplantation registry ($n = 408$; 52 continuous variables; 0.2–77.9% missingness) and an acute myocardial infarction dataset ($n = 1700$; 18 continuous variables; 0.24–63.29% missingness). Distributional fidelity was assessed using the Kolmogorov–Smirnov D-statistic, and multivariate structural preservation was quantified through correlation matrix comparison. KNN achieved the lowest mean KS statistics in both cohorts (0.036 and 0.032) compared with median imputation (0.076 and 0.074) and MICE (0.111 and 0.123). The proportion of variables with meaningful distortion ($D > 0.10$) was 9.6–11.1% for KNN, 22.2–25.0% for median, and 34.6–38.9% for MICE. KNN also showed the shallowest increase in distortion with increasing missingness. Multivariate correlations were well preserved by KNN and median imputation but substantially altered by MICE. KNN preserved the statistical structure of clinical data more closely than median substitution, with performance comparable to MICE, while requiring considerably less methodological specification.

1. Introduction

Missing data are a persistent feature of clinical and biomedical research, arising from incomplete data entry, missed assessments, equipment failure, loss to follow-up, non-response, or fields that are routinely left blank in clinical registries and electronic health records. The amount and pattern of missingness can materially affect the validity, precision, and reproducibility of downstream statistical analyses, and inadequate handling of missing values is a well-documented source of bias and reduced statistical power in observational research [1].

Methodological frameworks for the principled handling of missing data have been available for decades, including Rubin's classification of missingness mechanisms missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) together with the corresponding analytical recommendations [2, 3]. Notwithstanding this body of methodological guidance, single imputation by the mean or median of the observed values remains one of the most frequently applied strategies in applied clinical research. Its persistence is largely attributable to its conceptual simplicity, minimal computational burden, and its availability as a default option in widely used statistical software [1, 3]. However, mean or median substitution is known to systematically reduce the variance of the imputed variable, distort its marginal distribution (typically by increasing the density of values around the central tendency), and attenuate the covariance and correlation structure between variables, because an identical value is assigned to every missing observation irrespective of each subject's individual covariate profile [2, 3]. Simulation studies have confirmed this effect empirically: under both MCAR and MAR scenarios, mean substitution produces downwardly biased standard errors because all substituted values are centred on the same global estimate, resulting in a narrower distribution than that of the true

underlying data, whereas multiple imputation recovers the original parameters and their associated uncertainty [4].

Multiple imputation by chained equations (MICE), implemented through fully conditional specification with univariate imputation models such as predictive mean matching (PMM), is widely regarded as a methodological reference standard, since it propagates the uncertainty associated with each imputed value and better preserves the joint distribution of the observed data [5]. From a theoretical standpoint, multiple imputation is preferable to both complete-case analysis and single imputation under any missingness mechanism, including MNAR, because it at least captures the dependencies of the missing data on observed variables; moreover, the predictive mean matching variant of fully conditional specification is particularly robust to violations of normality, as it imputes actually observed donor values rather than model-derived predictions [6]. Nevertheless, the practical implementation of MICE requires the specification of conditional imputation models for every variable, an adequate choice of the number of imputations and iterations, and a level of statistical expertise that may not always be readily available in applied clinical settings, particularly when working with small- to moderate-sized single-centre datasets [3, 5].

Between these two poles, k-nearest neighbours (KNN) imputation offers an intermediate, instance-based, non-parametric alternative. The algorithm was originally proposed for estimating missing values in DNA microarray data [7, 8], and most subsequent evaluations have been conducted in genomic or high-dimensional settings where missingness is primarily technical (failed probes, below-threshold signals) and the number of variables far exceeds the number of samples. Clinical datasets present a different challenge. They typically contain tens rather than thousands of variables, sample sizes are moderate, and the patterns of missingness reflect clinical workflows (unrequested laboratory tests, loss to follow-up, incomplete registries) that are often associated with disease severity or treatment decisions [9, 10]. Under these conditions, a MAR mechanism is more plausible than MCAR, which makes the choice of imputation strategy particularly consequential. KNN estimates each missing value from the weighted values of the most similar complete observations according to a chosen distance metric. Because each estimate is conditioned on the multivariate profile of comparable subjects rather than on a single global summary, KNN has the potential to preserve both marginal distributions and pairwise correlations more faithfully than median substitution, while requiring less specification than MICE. Comparative evaluations in clinical datasets have reported competitive performance of KNN relative to MICE, with considerably lower implementation complexity [11, 12].

Most of these comparisons, however, have measured the downstream impact of imputation on predictive model performance (discrimination, calibration of risk scores) rather than on the statistical structure of the dataset itself: the univariate distributions of individual variables and the multivariate correlations among them [11, 12]. This distinction matters for applied clinical research. Many analyses in observational studies, including descriptive epidemiology, exploratory correlation analysis, and dimensionality reduction, depend directly on the fidelity of these distributional properties. A clinician analysing a single-centre registry does not typically need a tuned predictive model; what is needed are

reliable descriptive statistics and undistorted relationships between variables. Whether KNN imputation preserves these properties in clinical datasets with the characteristic patterns of workflow-driven missingness has not been directly evaluated.

We designed this study to directly compare median imputation, KNN, and MICE/PMM on two independent clinical cohorts, focusing on their ability to preserve univariate distributions and the multivariate correlation structure of the data as originally collected. Our intent is not to declare KNN optimal in an absolute sense. The question we ask is more specific: does median imputation, still the default in many applied settings, distort the structure of clinical data to a greater extent than KNN? And if so, how closely does KNN approximate the results obtained with MICE? We analysed a single-centre lung transplantation registry and a publicly available acute myocardial infarction dataset (used for external validation), evaluating each method at both the univariate and multivariate levels.

2. Materials and methods

2.1. Study design and rationale

A methodological comparative study was conducted to evaluate the ability of different missing-data imputation strategies to preserve the statistical structure of real-world clinical datasets. This is an applied methodological study based on real-world biomedical data, not a study of algorithm development. Two independent observational datasets from different medical domains were analysed to assess the robustness and generalizability of the findings.

The objective was to compare the extent to which three missing-data imputation strategies, median substitution, k-nearest neighbours (KNN) imputation, and multiple imputation by chained equations (MICE) with predictive mean matching (PMM) preserved (i) the original univariate distributions and (ii) the multivariate correlation structure of two independent, real-world incomplete clinical datasets. Median imputation was treated as the primary comparator, representing a simple strategy that remains widely used in applied clinical research; KNN imputation was the principal method under evaluation; and MICE/PMM served as a secondary methodological reference, used to assess whether KNN approximated the results obtained with a method considered a standard of reference in the literature [5].

2.2. Data sources

2.2.1. Primary dataset: lung transplantation cohort

The primary dataset comprised a retrospective, single-centre cohort of 408 lung transplant recipients treated at a tertiary referral centre. The original database included 193 variables, comprising demographic characteristics, pre-transplant clinical data, donor information, intraoperative variables, postoperative outcomes, and follow-up data. After preprocessing, the analytical dataset contained 408 observations and 191 variables, including 66 numeric continuous variables, of which 52 contained at

least one missing observation and were therefore included in the comparative evaluation of imputation methods. Missingness among these 52 variables ranged from 0.2% to 77.9%.

2.2.2. Secondary dataset: acute myocardial infarction cohort (external validation)

The secondary dataset was a publicly available cohort of patients with acute myocardial infarction complications [13], comprising 1700 observations. After preprocessing and restoration of variable coding according to the original data dictionary, the analytical dataset contained 94 variables, comprising 18 numeric continuous variables and 76 categorical variables. All 18 numeric variables contained at least one missing observation and were included in the imputation analyses; missingness ranged from 0.24% to 63.29%. This dataset was used exclusively as an independent external validation of the findings obtained in the primary cohort and was not used for any clinical inference regarding lung transplantation.

2.3. Data preprocessing

Preprocessing procedures were standardized across both datasets. Variables used exclusively as patient identifiers were removed before any analysis. Numeric variables with more than 80% missing observations were excluded, because reliable imputation cannot reasonably be expected under such extreme levels of incompleteness; this criterion led to the exclusion of one variable in each dataset (postoperative ECMO duration in the lung transplantation cohort, and one numeric variable in the myocardial infarction cohort).

Categorical variables were retained in both datasets but were not subjected to imputation themselves. Instead, missing values in categorical variables were recoded as an additional category labelled “Unknown”, preserving the original sample size and allowing these variables to contribute as auxiliary predictors in the multivariate imputation procedures, without introducing artificial assumptions regarding their underlying distribution.

2.4. Variable selection for the comparative evaluation

The comparative evaluation of imputation methods was restricted to numeric continuous variables containing at least one missing observation in each dataset (52 of 66 numeric variables in the lung transplantation cohort; all 18 numeric variables in the myocardial infarction cohort). This restriction follows directly from the study’s central objective, namely to assess the preservation of continuous univariate distributions and of the correlation structure among continuous variables, rather than the imputation of categorical fields.

2.5. Assessment of the missing-data mechanism

The pattern of missingness was initially explored by quantifying the amount and proportion of missing observations for each numeric variable; these proportions were presented in tabular form and visualized using missingness heatmaps for both datasets.

To assess whether the observed pattern of missingness was consistent with a completely random mechanism, Little's test for Missing Completely at Random [14, 15] was applied to the full set of continuous variables in each dataset, implemented through the `nanian` package in R. A statistically significant result ($p < 0.05$) was interpreted as evidence against the MCAR assumption, suggesting that the missingness was at least partly dependent on observed covariates and thus consistent with a Missing at Random (MAR) mechanism.

Given the observational nature of the datasets, subsequent analyses were conducted under the assumption of a predominantly MAR mechanism. No formal distinction between MAR and MNAR was attempted, as such a distinction cannot be verified from observed data alone [2, 3, 6].

2.6. Imputation methods

Three imputation strategies, representing different levels of methodological complexity, were evaluated and compared.

2.6.1. Median imputation

For each continuous variable, missing values were replaced by the median of the observed values for that variable. This deterministic, single-value approach was treated as the primary comparator, representing a simple imputation strategy that remains widely used in applied clinical research due to its simplicity and ease of implementation [1, 3].

2.6.2. K-nearest neighbours (KNN) imputation

KNN imputation was performed using the `kNN()` function from the `VIM` package in R [16], applied to the numeric continuous variables identified in Section 2.4. Distances between observations were computed using Gower's distance, the default metric of the `VIM` implementation, which internally rescales each variable by its range; no prior standardization was therefore applied. Missing values were estimated from the five most similar cases ($k = 5$), identified according to the multivariate structure of the dataset. The value $k = 5$ was selected as the default of the `VIM` implementation and is consistent with the values most commonly used in clinical datasets of moderate dimensionality; [7, 11, 12] no optimization of k was performed, as the objective of this study was to evaluate the method under standard conditions rather than to tune its performance. The parameter `imp_var = FALSE` was used to avoid creating auxiliary indicator variables during the imputation process, and a fixed random seed (`set.seed(123)`) was applied to ensure reproducibility. KNN imputation constituted the principal method evaluated in this study.

2.6.3. Multiple imputation by chained equations (MICE) with PMM

As a secondary methodological reference, missing values were additionally imputed using MICE under a fully conditional specification, implemented with the `mice` package in R (version 3.16.0) [5]. Only numeric variables containing missing values were imputed, while all other variables were left unchanged.

Predictive mean matching (PMM) was selected as the univariate imputation model for continuous variables because it preserves plausible observed values and avoids unrealistic extrapolation [5, 6]. Five multiply imputed datasets ($m = 5$) were generated using 10 iterative cycles ($\text{maxit} = 10$), with a fixed random seed ($\text{seed} = 123$).

Unlike the median and KNN approaches, which each produced a single completed dataset, MICE generated five independently completed datasets. Each evaluation metric described in Section 2.7 was calculated separately on each of the five imputed datasets, treating each as a standalone completed dataset, exactly as would be obtained by a researcher selecting any single imputation for downstream analysis and the resulting values were averaged across imputations to obtain a single summary estimate per variable and per metric. This approach evaluates the expected distributional fidelity of a typical MICE-completed dataset, rather than the pooled inferential properties for which MICE was originally designed [5, 17].

MICE/PMM was not the primary objective of this study; rather, its purpose was to determine whether KNN produced results sufficiently close to those obtained with a method generally considered a reference standard in the missing-data literature.

2.7. Evaluation of the effect of imputation

The performance of the three imputation methods was assessed by comparing each completed dataset with the corresponding original dataset using several complementary approaches. For median and KNN imputation, each metric was calculated directly on the single completed dataset produced by that method; for MICE/PMM, each metric was calculated separately on each of the five imputed datasets and then averaged across imputations, as described in Section 2.6.3.

2.7.1. Preservation of univariate distributions (primary outcome)

For each continuous variable and each imputation method, four descriptive statistics: mean, median, standard deviation, and interquartile range (IQR), were calculated before and after imputation, and the relative error was computed as: $\text{Relative error (\%)} = 100 \times |S_{\text{imputed}} - S_{\text{original}}| / |S_{\text{original}}|$, where S denotes the corresponding summary statistic.

Distributional similarity was additionally assessed using the two-sample Kolmogorov–Smirnov (KS) statistic (D), which quantifies the maximum difference between the empirical cumulative distribution functions of the original and imputed data and ranges from 0 (identical distributions) to 1 (no overlap); lower values indicate greater distributional similarity. For each method, the mean, median, and maximum KS statistic across variables were summarized, together with the proportion of variables exceeding a predefined threshold ($D > 0.10$), which was adopted as the primary outcome of the study. The D -statistic was preferred over the KS p -value as the primary metric because, in large samples, even trivially small distributional shifts can reach statistical significance, whereas the D -statistic provides a scale-

independent measure of effect size [2]. The variables showing the largest KS distances were identified for each method, and representative density plots were generated for variables with low, moderate, and high proportions of missing data.

2.7.2. Effect of increasing missingness

To evaluate the influence of missingness severity on imputation performance, the KS statistic for each variable was related to its percentage of missing observations. Spearman correlation coefficients were calculated to assess the monotonic association between missingness and distributional distortion. Simple linear regression models were additionally fitted separately for each imputation method, using the KS statistic as the dependent variable and the proportion of missing values as the independent variable; the regression slope was interpreted as the expected increase in distributional distortion for each additional 1% increase in missingness.

2.7.3. Preservation of the multivariate correlation structure (secondary outcome)

For each dataset, pairwise Spearman correlation matrices among continuous variables were computed using (i) the originally observed data (pairwise complete observations) and (ii) the completed datasets obtained after median imputation, KNN imputation, and MICE/PMM. Matrices were visualized as heatmaps, and matrix similarity was quantified by computing the Spearman correlation between the vectorized upper-triangular portions of the original and imputed matrices. The mean absolute error (MAE), root mean square error (RMSE), and maximum absolute difference between corresponding correlation coefficients were also calculated, with lower values indicating better preservation of the original correlation structure.

2.7.4. Secondary comparison with MICE

As a secondary analysis, the results obtained with KNN imputation (univariate KS statistics and multivariate correlation-preservation metrics) were compared descriptively with those obtained using MICE/PMM, in order to assess whether KNN approximated the reference standard despite its substantially simpler implementation. This comparison was not intended to test the superiority of either method, but to evaluate the practical equivalence of KNN as an accessible alternative to MICE.

2.7.5. Comparison across datasets

The findings obtained from the two cohorts were compared narratively to assess the consistency and generalizability of the results. This comparison consisted of contrasting the relative ranking and overall pattern of performance of the three imputation methods between the lung transplantation and myocardial infarction cohorts, rather than applying a separate statistical test.

2.8. Statistical software

All analyses were performed using R version 4.3.2 and RStudio version 2026.05.0 + 218 (R Foundation for Statistical Computing, Vienna, Austria). Data manipulation and preprocessing were conducted using

the tidyverse ecosystem (version 2.0.0), and janitor (version 2.2.0) was used for variable name standardization. Missing-data visualization and MCAR testing were performed with naniar (version 1.0.0). Median imputation was implemented using base R functions, KNN imputation was performed with the VIM package (version 6.2.2) [16], and multiple imputation by chained equations was carried out with the mice package (version 3.16.0) [5]. Correlation analyses, descriptive summaries, and graphical representations were generated using standard R functions and the ggplot2 package (version 3.4.4); tables were created using the flextable (version 0.9.4) and officer (version 0.6.3) packages. Two-sided statistical tests were used throughout, with statistical significance set at $p < 0.05$.

3. Results

3.1. Dataset characteristics and missing-data patterns

The characteristics of both analytical datasets are summarized in Table 1. The lung transplantation cohort comprised 408 patients and 191 variables after preprocessing, including 66 numeric continuous variables, of which 52 (78.8%) contained at least one missing observation. Missingness ranged from 0.2% (baseline serum creatinine, donor PaO₂/FiO₂ ratio) to 77.9% (pulmonary artery systolic pressure by echocardiography). Ten variables had more than 30% missing values, and five exceeded 50%. One variable (postoperative ECMO duration) surpassed the predefined 80% exclusion threshold and was removed before imputation.

Table 1
Summary of dataset characteristics.

Characteristic	Lung transplantation	Myocardial infarction
Patients (n)	408	1700
Total variables after preprocessing	191	94
Numeric continuous variables	66	18
Variables with ≥ 1 missing value	52 (78.8%)	18 (100%)
Range of missingness (%)	0.2–77.9	0.24–63.29
Variables with $> 30\%$ missing	10	2
Little’s MCAR test	$p < 0.001$	$p < 0.001$

The myocardial infarction cohort comprised 1700 patients and 94 variables after preprocessing, including 18 numeric continuous variables, all of which contained at least one missing observation. Missingness ranged from 0.24% (previous myocardial infarction) to 63.29% (systolic and diastolic blood pressure at emergency medical service arrival). The distribution of missing values across numeric variables for both cohorts is shown in Fig. 1; variable-level missing data profiles are provided in Supplementary Tables S1 and S2.

Little’s MCAR test rejected the assumption of completely random missingness in both datasets (lung transplantation: $\chi^2 = 46,697.18$, $df = 17,433$, $p < 0.001$; myocardial infarction: $\chi^2 = 2784.5$, $df = 2159$, $p < 0.001$). All subsequent analyses were conducted under a predominantly MAR framework, acknowledging that MNAR mechanisms cannot be entirely excluded.

3.2. Preservation of univariate distributions

3.2.1. Relative error analysis

Mean relative errors for four descriptive statistics (mean, median, standard deviation, and interquartile range) are presented in Table 2 for both cohorts. Across both datasets, KNN consistently showed the lowest relative errors for measures of variability (standard deviation and interquartile range), while maintaining low errors for the mean. Median imputation preserved the median exactly by construction but introduced substantially larger distortions in measures of dispersion, with mean relative errors for the interquartile range of 26.2% and 20.8% in the lung transplantation and myocardial infarction cohorts, respectively. MICE exhibited the largest relative errors overall, particularly for the standard deviation (31.0% and 24.3%) and interquartile range (38.0% and 42.4%).

Table 2
Mean relative error (%) of descriptive statistics across numeric variables after imputation.

Statistic	Median	KNN	MICE	Median	KNN	MICE
Mean	1.19	1.15	8.19	1.45	1.20	6.38
Median	0.00	1.17	3.66	0.00	6.45	7.67
Standard deviation	8.89	6.39	31.04	8.77	6.43	24.29
Interquartile range	26.19	10.74	38.01	20.79	11.20	42.39
<i>Note: Columns 2–4 correspond to the lung transplantation cohort; columns 5–7 to the myocardial infarction cohort. Relative error was calculated as $100 \times S_{\text{imputed}} - S_{\text{original}} / S_{\text{original}}$. KNN, k-nearest neighbours; MICE, multiple imputation by chained equations with predictive mean matching.</i>						

3.2.2. Distribution comparison

Representative density plots comparing the original and imputed distributions for variables with low, moderate, and high proportions of missing data are shown in Fig. 2. In both cohorts, all imputation methods produced distributions nearly indistinguishable from the original for variables with minimal missingness (baseline serum creatinine, 0.2% missing; age, 0.5% missing). As the proportion of missing data increased, differences became more pronounced: median imputation generated a visible concentration of observations around central values, reducing variability; KNN more closely preserved the shape of the original distribution; and MICE produced broader distributions with noticeable shifts in location for some variables.

3.2.3. Kolmogorov–Smirnov analysis

The KS statistic was used to quantify the similarity between the original and imputed distributions for each numeric variable. Results are summarized in Table 3 for both cohorts; the variables with the largest KS statistics for each method are reported in Supplementary Tables S3 and S4.

Table 3
Summary of Kolmogorov–Smirnov statistics by imputation method and dataset.

Method	Mean D	Median D	Max D	% D > 0.10	Mean D	Median D	Max D	% D > 0.10
KNN	0.036	0.016	0.249	9.6	0.032	0.024	0.135	11.1
Median	0.076	0.037	0.390	25.0	0.074	0.062	0.303	22.2
MICE	0.111	0.056	0.521	34.6	0.123	0.078	0.617	38.9
<i>Note: Columns 2–5 correspond to the lung transplantation cohort; columns 6–9 to the myocardial infarction cohort. Lower KS statistics indicate better preservation of the original distributions. KNN, k-nearest neighbours; MICE, multiple imputation by chained equations.</i>								

KNN achieved the lowest mean KS statistic in both datasets (0.036 and 0.032, respectively), indicating the closest agreement with the original distributions. Median imputation showed intermediate performance (0.076 and 0.074), whereas MICE produced the largest deviations (0.111 and 0.123). The proportion of variables with KS values exceeding 0.10 was 9.6% and 11.1% for KNN, 25.0% and 22.2% for median imputation, and 34.6% and 38.9% for MICE, in the lung transplantation and myocardial infarction cohorts, respectively. The maximum KS statistic was also consistently lowest for KNN (0.249 and 0.135) compared with median (0.390 and 0.303) and MICE (0.521 and 0.617).

3.3. Effect of increasing missingness

A strong positive association was observed between the proportion of missing values and the magnitude of distributional distortion for all three imputation methods in both cohorts (Fig. 3). Spearman correlation coefficients ranged from 0.964 to 0.999 (all $p < 0.001$), indicating that variables with higher proportions of missing data consistently experienced greater distributional distortion after imputation, regardless of the method applied.

Linear regression models confirmed that the rate of increase in distributional distortion differed across methods (Table 4). KNN showed the lowest regression slope in both datasets (0.0023 and 0.0021, respectively), indicating the slowest increase in KS distortion per additional 1% of missing data. Median imputation showed intermediate slopes (0.0048 and 0.0047), and MICE exhibited the steepest increase (0.0062 and 0.0093). All models showed excellent goodness of fit (R^2 from 0.871 to 0.998).

Table 4
Association between missingness and distributional distortion.

Method	ρ	Slope	R^2	ρ	Slope	R^2
KNN	0.985	0.0023	0.903	0.973	0.0021	0.990
Median	0.999	0.0048	0.998	0.993	0.0047	0.997
MICE	0.964	0.0062	0.871	0.969	0.0093	0.979
<i>Note: Columns 2–4 correspond to the lung transplantation cohort; columns 5–7 to the myocardial infarction cohort. The slope represents the expected increase in the KS statistic for each additional 1% of missing values. ρ, Spearman's correlation coefficient; KNN, k-nearest neighbours; MICE, multiple imputation by chained equations.</i>						

3.4. Preservation of the multivariate correlation structure

Pairwise Spearman correlation matrices were computed for the original and imputed datasets in both cohorts. Table 5 summarizes the similarity metrics for each imputation method.

Table 5
Preservation of the multivariate correlation structure after imputation.

Method	Matrix r	MAE	RMSE	Max diff	Matrix r	MAE	RMSE	Max diff
KNN	0.971	0.022	0.040	0.357	0.986	0.017	0.034	0.188
Median	0.980	0.017	0.035	0.397	0.979	0.020	0.045	0.251
MICE	0.815	0.056	0.098	0.915	0.790	0.056	0.112	0.721
<i>Note: Columns 2–5 correspond to the lung transplantation cohort; columns 6–9 to the myocardial infarction cohort. Matrix r, Spearman correlation between the vectorized upper triangles of the original and imputed correlation matrices; MAE, mean absolute error; RMSE, root mean square error; Max diff, maximum absolute difference. KNN, k-nearest neighbours; MICE, multiple imputation by chained equations.</i>								

In the lung transplantation cohort, median imputation achieved the highest matrix correlation with the original structure (0.980), followed closely by KNN (0.971); both methods showed low MAE and RMSE values. In the myocardial infarction cohort, KNN outperformed all methods, achieving the highest matrix correlation (0.986), the lowest MAE (0.017), RMSE (0.034), and maximum absolute difference (0.188). In both datasets, MICE produced substantially lower matrix correlations (0.815 and 0.790) and larger errors across all metrics, indicating a greater alteration of the multivariate structure.

3.5. Cross-dataset consistency

The relative ranking of the three imputation methods was remarkably consistent across the two independent cohorts. KNN achieved the lowest mean KS statistics and the shallowest regression slopes in both datasets. Median imputation provided intermediate performance, and MICE produced the largest alterations in both univariate distributions and multivariate correlation structures. This pattern held

despite substantial differences in sample size (408 vs. 1700 patients), number of imputed variables (52 vs. 18), clinical domain, and missing-data profiles, suggesting that the observed performance differences reflect intrinsic properties of the methods rather than idiosyncratic features of a single dataset.

4. Discussion

4.1. Principal findings

This study compared median substitution, k-nearest neighbours (KNN), and multiple imputation by chained equations with predictive mean matching (MICE/PMM) on their ability to preserve the statistical structure of two independent real-world clinical datasets: a single-centre lung transplantation registry ($n = 408$; 52 continuous variables, 0.2–77.9% missingness) and a publicly available acute myocardial infarction cohort ($n = 1700$; 18 continuous variables, 0.24–63.29% missingness). Across both datasets and all evaluation criteria, KNN produced the closest approximation to the original observed distributions and maintained the multivariate correlation structure with the lowest or near-lowest error. Median substitution showed intermediate performance, distorting measures of dispersion while preserving central tendency. MICE/PMM produced the largest distributional deviations in both cohorts, a finding that is counterintuitive given its status as a reference method in the missing-data literature and which requires careful contextual interpretation.

4.2. Median substitution: expected limitations confirmed

The limitations of median substitution observed here are consistent with well-established methodological principles. Single-value substitution methods are known to be unbiased for central tendency but systematically reduce variance and attenuate the covariance structure, because an identical value is assigned to all missing observations regardless of each subject's covariate profile [2, 17]. This theoretical prediction was confirmed empirically: median imputation yielded mean relative errors for the interquartile range of 26.2% and 20.8% in the lung transplantation and myocardial infarction cohorts, respectively, compared with 10.7% and 11.2% for KNN (Table 2). The KS statistic confirmed moderate distributional distortion (mean $D = 0.076$ and 0.074), with the degree of distortion increasing linearly with the proportion of missing data at a rate approximately twice that of KNN (slopes of 0.0048 and 0.0047 vs. 0.0023 and 0.0021 per 1% additional missingness; Table 4). These findings quantify why median substitution, despite its computational convenience and widespread use in applied clinical research [18], produces systematically biased representations of continuous variables when missingness exceeds trivial proportions.

4.3. Interpretation of MICE performance in the present context

The observation that MICE/PMM produced the largest distributional deviations across both datasets, despite being widely considered the methodological reference standard for missing-data imputation [2,

17], is the most notable finding of this study.

Two distinct considerations are relevant to its interpretation. The first is methodological. MICE is designed to produce valid inferences about population parameters by propagating the uncertainty inherent in the imputation process across multiple completed datasets and pooling estimates following Rubin's combining rules [17]. Its primary objective is not to produce a single completed dataset that closely resembles the observed data in terms of marginal distributions, but to yield unbiased parameter estimates and valid standard errors for inferential analyses. In the present study, each of the five MICE-completed datasets was evaluated individually as a standalone imputed dataset, exactly as would be obtained by a researcher selecting any single imputation for downstream analysis, and the metrics were then averaged. The reported results therefore reflect the expected distributional fidelity of a typical MICE output, not an artefact of cross-imputation averaging. Nevertheless, the stochastic between-imputation variability that is a deliberate feature of MICE in inferential contexts manifests as greater apparent distributional distortion when each completed dataset is compared individually against the original observed data, a scenario that is common in applied research where a single imputed dataset is used for analyses that do not involve formal Rubin pooling [2].

The second consideration is empirical. The conditions prevailing in the datasets analysed here are known to challenge MICE performance independently of the evaluation criterion. MICE operates by fitting a series of univariate conditional regression models iteratively until convergence. When the number of variables requiring simultaneous imputation is large relative to the available sample, as in the lung transplantation cohort where 52 variables were imputed in 408 patients, specifying compatible conditional models and achieving stable convergence become increasingly difficult [19, 20]. Furthermore, the robustness of MICE under high proportions of missing data is limited. A systematic evaluation in longitudinal health indicator data demonstrated high robustness up to approximately 50% missingness, with progressively increasing distortion between 50–70% and compromised reliability beyond 70% [21]. Independent simulation evidence from cancer registry data has shown severe bias and poor coverage in scenarios combining small sample sizes with substantial missingness proportions [22]. In the present study, ten variables in the lung transplantation cohort exceeded 30% missing values and five exceeded 50%, conditions that fall within the range where MICE performance has been documented to degrade.

Notably, the underperformance of MICE relative to KNN persisted in the myocardial infarction dataset ($n = 1700$), where sample size constraints were less stringent and the number of imputed variables was substantially smaller. The persistence of the same ranking in a larger dataset with lower overall missingness suggests that the performance difference is not attributable solely to the extreme conditions of the lung transplantation cohort but may also reflect the sensitivity of MICE to highly skewed clinical variables and to the evaluation criterion employed.

These findings are consistent with previous reports. Li et al [12]. compared eight imputation methods in a cardiovascular cohort of 10,164 subjects and found that KNN achieved the lowest MAE and RMSE,

outperforming MICE, which ranked behind both KNN and random forest imputation. Joel et al [23]. reported that KNN ranked second only to missForest across multiple healthcare datasets and missingness levels, consistently outperforming MICE. Hegde et al [24]. found MICE to perform poorly in a clinical dataset with approximately 50% missingness, attributing this explicitly to the departure from the 2.5–30% range where MICE is known to perform best.

4.4. Mechanistic basis for the distributional advantage of KNN

The consistent advantage of KNN has a straightforward mechanistic explanation. Because each missing value is estimated from the weighted profile of the k most similar complete observations ($k = 5$ in this study), KNN inherently captures local multivariate structure and avoids the global smoothing that characterizes single-value imputation. Unlike MICE, it does not require the specification of univariate conditional regression models, does not rely on iterative convergence, and is less affected by a high variable-to-sample ratio. Its non-parametric nature avoids distributional assumptions that may be violated in clinical data, particularly for variables with skewed or bimodal distributions, precisely those showing the largest KS statistics in the present study (TAPSE, pulmonary artery systolic pressure, ischaemia durations).

The relationship between missingness and distributional distortion was notably shallower for KNN in both datasets. Regression slopes of 0.0023 and 0.0021 per 1% additional missingness were approximately half those of median imputation and less than a quarter of those for MICE (Table 4). This resistance to progressive distortion is clinically relevant: in single-centre registries, variables such as echocardiographic or haemodynamic measures collected only in selected patients routinely exceed 40–60% missingness, and the choice of imputation method under these circumstances has a non-trivial impact on the resulting data structure.

At the level of multivariate correlation structure, the results were more nuanced. In the lung transplantation cohort, median imputation marginally outperformed KNN in terms of matrix correlation (0.980 vs. 0.971), whereas in the myocardial infarction cohort KNN clearly outperformed all methods (0.986 vs. 0.979 for median). The near-equivalence of KNN and median imputation for correlation preservation in the first cohort reflects a known property of single-value imputation: by concentrating values around the central tendency, median substitution may inadvertently preserve rank-based global correlations even as it distorts marginal distributions [2]. This dissociation between univariate and multivariate metrics reinforces the importance of evaluating imputation methods across complementary criteria rather than relying on a single summary measure.

4.5. Consistency across datasets

The relative ranking of the three methods was remarkably stable across two cohorts from different clinical domains, different sample sizes (408 vs. 1700), different numbers of imputed variables (52 vs. 18), and different missingness profiles. The Spearman correlation between missingness and KS

distortion was consistently strong for all methods in both datasets (p from 0.964 to 0.999; all $p < 0.001$), confirming a predictable linear relationship between the degree of missingness and the magnitude of distributional distortion irrespective of method or dataset. This stability strengthens confidence that the observed performance differences reflect intrinsic properties of the methods rather than idiosyncratic features of a single cohort, and supports the generalizability of the conclusions to other real-world clinical registries with analogous characteristics.

4.6. Limitations

This study has some limitations that should be considered when interpreting the results. The evaluation criterion adopted, preservation of marginal distributions and pairwise correlations relative to the observed data, directly quantifies structural fidelity, which is the property of interest when a single completed dataset is used for subsequent analysis. It does not, however, assess the recovery of the true data-generating process, and should not be interpreted as a general performance ranking for inferential use cases where MICE retains its theoretical advantages [2]. The comparative evaluation was restricted to continuous numeric variables; the imputation of categorical variables was not formally assessed. KNN was evaluated at a fixed parameterization ($k = 5$, Gower's distance), and the number of MICE imputations was set at $m = 5$, which is generally considered sufficient for most purposes [5, 17]. Finally, prediction-oriented outcomes such as discrimination or calibration were not evaluated, as they represent a distinct use case for imputation [12].

4.7. Practical implications

The central practical contribution of this study is the empirical demonstration that KNN imputation preserves the statistical structure of clinical datasets, both univariate distributions and the multivariate correlation matrix, substantially better than median substitution, the single imputation strategy most widely used in applied clinical research [18]. Preserving the underlying data structure matters for any subsequent analysis that relies on the completed dataset: regression modelling, dimensionality reduction, clustering, variable selection, or any procedure sensitive to the distributional properties and correlation structure of the input data. Distortions introduced at the imputation stage propagate into all downstream analyses, and the present results demonstrate that KNN minimizes these distortions more effectively than median substitution across all missingness levels studied.

From an implementation standpoint, KNN offers a compelling balance between methodological rigour and practical accessibility. It requires no specification of conditional imputation models, no iterative convergence assessment, and produces a single deterministic completed dataset at substantially lower computational cost than MICE. In settings where the number of variables with missing data is large relative to the available sample, a common situation in single-centre registries, KNN avoids the model specification and convergence challenges that limit MICE performance [19, 20]. These considerations support KNN as a methodologically sound and computationally simple alternative to median substitution, and as a practical option when the conditions for reliable MICE implementation, adequate sample size, moderate missingness, and feasible model specification for all imputed variables are not

met. When the analytical objective requires formal statistical inference with propagation of imputation uncertainty, MICE with Rubin's combining rules remains the appropriate choice, provided the dataset characteristics fall within the method's documented operational range.

5. Conclusions

Across two independent real-world clinical cohorts from different medical domains, sample sizes, and missing-data profiles, KNN imputation consistently preserved univariate distributions and the multivariate correlation structure of the original observed data more closely than either median substitution or MICE/PMM. The proportion of variables showing meaningful distributional distortion ($KS D > 0.10$) was two- to threefold lower for KNN than for median imputation, and three- to fourfold lower than for MICE. The rate of increase in distributional distortion per additional percentage of missing data was approximately half that of median imputation and less than a quarter that of MICE. Median substitution confirmed its known tendency to compress variance and distort dispersion, whereas MICE showed greater sensitivity to the high missingness proportions and variable-to-sample ratios characteristic of clinical registries.

These findings support KNN as a methodologically sound, computationally accessible alternative to median substitution for producing a single completed clinical dataset that faithfully preserves the statistical structure of the originally observed data. When the analytical objective requires formal inferential pooling with propagation of imputation uncertainty, MICE with Rubin's combining rules remains the appropriate framework, provided dataset conditions fall within its documented operational range.

Declarations

Ethical considerations

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Comité de Ética de la Investigación con Medicamentos (CEIm) of the Hospital Universitari i Politècnic La Fe de Valencia (protocol number 2020-492-1; date of approval: 16 December 2020).

Consent to participate

Patient consent was waived by the CEIm due to the retrospective design of the study and the use of fully anonymized clinical data. The myocardial infarction dataset is publicly available and does not require individual consent.

Consent for publication

Not applicable. No individual person's data in any form is included in this manuscript.

Declaration of conflicting interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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

Data availability statement

The lung transplantation dataset is available from the corresponding author on reasonable request, subject to institutional and data-protection restrictions. The acute myocardial infarction dataset is publicly available from the UCI Machine Learning Repository (<https://doi.org/10.24432/C53P5M>).

Author contributions

Conceptualization: C.A.G. and K.D.A.Q.; Methodology: K.D.A.Q. and C.A.G.; Software: K.D.A.Q.; Formal analysis: K.D.A.Q. and C.A.G.; Investigation: K.D.A.Q., A.F.B. and C.J.A.; Resources: A.F.B. and C.J.A.; Data curation: K.D.A.Q.; Writing original draft: K.D.A.Q. and C.A.G.; Writing, review and editing: all authors; Visualization: K.D.A.Q.; Supervision: C.A.G. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

Not applicable.

References

1. Newgard, C.D., Lewis, R.J.: Missing data: how to best account for what is not known. *JAMA* 314, 940–941 (2015). <https://doi.org/10.1001/jama.2015.10516>
2. Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*, 3rd edn. Wiley, Hoboken (2019). <https://doi.org/10.1002/9781119482260>
3. Jakobsen, J.C., Gluud, C., Wetterslev, J., Winkel, P.: When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. *BMC Med Res Methodol* 17, 162 (2017). <https://doi.org/10.1186/s12874-017-0442-1>
4. de Goeij, M.C.M., van Diepen, M., Jager, K.J., Tripepi, G., Zoccali, C., Dekker, F.W.: Multiple imputation: dealing with missing data. *Nephrol Dial Transplant* 28, 2415–2420 (2013). <https://doi.org/10.1093/ndt/gft221>
5. van Buuren, S., Groothuis-Oudshoorn, K.: mice: multivariate imputation by chained equations in R. *J Stat Softw* 45, 1–67 (2011). <https://doi.org/10.18637/jss.v045.i03>
6. van Ginkel, J.R., Linting, M., Rippe, R.C.A., van der Voort, A.: Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *J Pers Assess* 102, 297–308 (2020).

- <https://doi.org/10.1080/00223891.2018.1530680>
7. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 520–525 (2001). <https://doi.org/10.1093/bioinformatics/17.6.520>
 8. Jadhav, A., Pramod, D., Ramanathan, K.: Comparison of performance of data imputation methods for numeric dataset. *Appl Artif Intell* 33, 913–933 (2019). <https://doi.org/10.1080/08839514.2019.1637138>
 9. Sterne, J.A.C., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., Carpenter, J.R.: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 338, b2393 (2009). <https://doi.org/10.1136/bmj.b2393>
 10. Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M.: Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 59, 1087–1091 (2006). <https://doi.org/10.1016/j.jclinepi.2006.01.014>
 11. Luo, Y.: Evaluating the state of the art in missing data imputation for clinical data. *Brief Bioinform* 23, bbab489 (2022). <https://doi.org/10.1093/bib/bbab489>
 12. Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., Ding, Y., Li, Y., Jian, L., Cheng, J., Guo, H.: Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Med Res Methodol* 24, 41 (2024). <https://doi.org/10.1186/s12874-024-02173-x>
 13. Golovenkin, S.E., Shulman, V.A., Rossiev, D.A., Shesternya, P.A., Nikulina, S.Y., Orlova, Y.V., Voino-Yasenetsky, V.F.: Myocardial infarction complications [Dataset]. UCI Machine Learning Repository (2020). <https://doi.org/10.24432/C53P5M>
 14. Little, R.J.A.: A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc* 83, 1198–1202 (1988). <https://doi.org/10.1080/01621459.1988.10478722>
 15. Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*, 2nd edn. Wiley, Hoboken (2002). <https://doi.org/10.1002/9781119013563>
 16. Kowarik, A., Templ, M.: Imputation with the R package VIM. *J Stat Softw* 74, 1–16 (2016). <https://doi.org/10.18637/jss.v074.i07>
 17. Rubin, D.B.: *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York (1987). <https://doi.org/10.1002/9780470316696>
 18. Heymans, M.W., Twisk, J.W.R.: Handling missing data in clinical research. *J Clin Epidemiol* 151, 185–188 (2022). <https://doi.org/10.1016/j.jclinepi.2022.08.016>
 19. Deng, Y., Chang, C., Ido, M.S., Long, Q.: Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Sci Rep* 6, 21689 (2016). <https://doi.org/10.1038/srep21689>
 20. Costantini, E., Lang, K.M., Sijtsma, K., Reeskens, T.: Solving the many-variables problem in MICE with principal component regression. *Behav Res Methods* 56, 1715–1737 (2024). <https://doi.org/10.3758/s13428-023-02117-1>

21. Junaid, K.P., Kiran, T., Gupta, M., Kishore, K., Siwatch, S.: How much missing data is too much to impute for longitudinal health indicators? A preliminary guideline for the choice of the extent of missing proportion to impute with multiple imputation by chained equations. *Popul Health Metr* 23, 2 (2025). <https://doi.org/10.1186/s12963-025-00364-2>
22. Kästner, A., Hoffmann, W., Hüsing, J., Stang, A., Hüsing, A.: Multiple imputation for missing values in ordinal variables from cancer registry data when performing Cox proportional hazards regression. *BMC Med Res Methodol* 26, 47 (2026). <https://doi.org/10.1186/s12874-026-02790-8>
23. Joel, L.O., Doorsamy, W., Paul, B.S.: A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. *Int J Data Sci Anal* 20, 6357–6373 (2025). <https://doi.org/10.1007/s41060-025-00825-9>
24. Hegde, H., Shimpi, N., Panny, A., Glurich, I., Christie, P., Acharya, A.: MICE vs PPCA: missing data imputation in healthcare. *Inform Med Unlocked* 17, 100275 (2019). <https://doi.org/10.1016/j.imu.2019.100275>

Figures

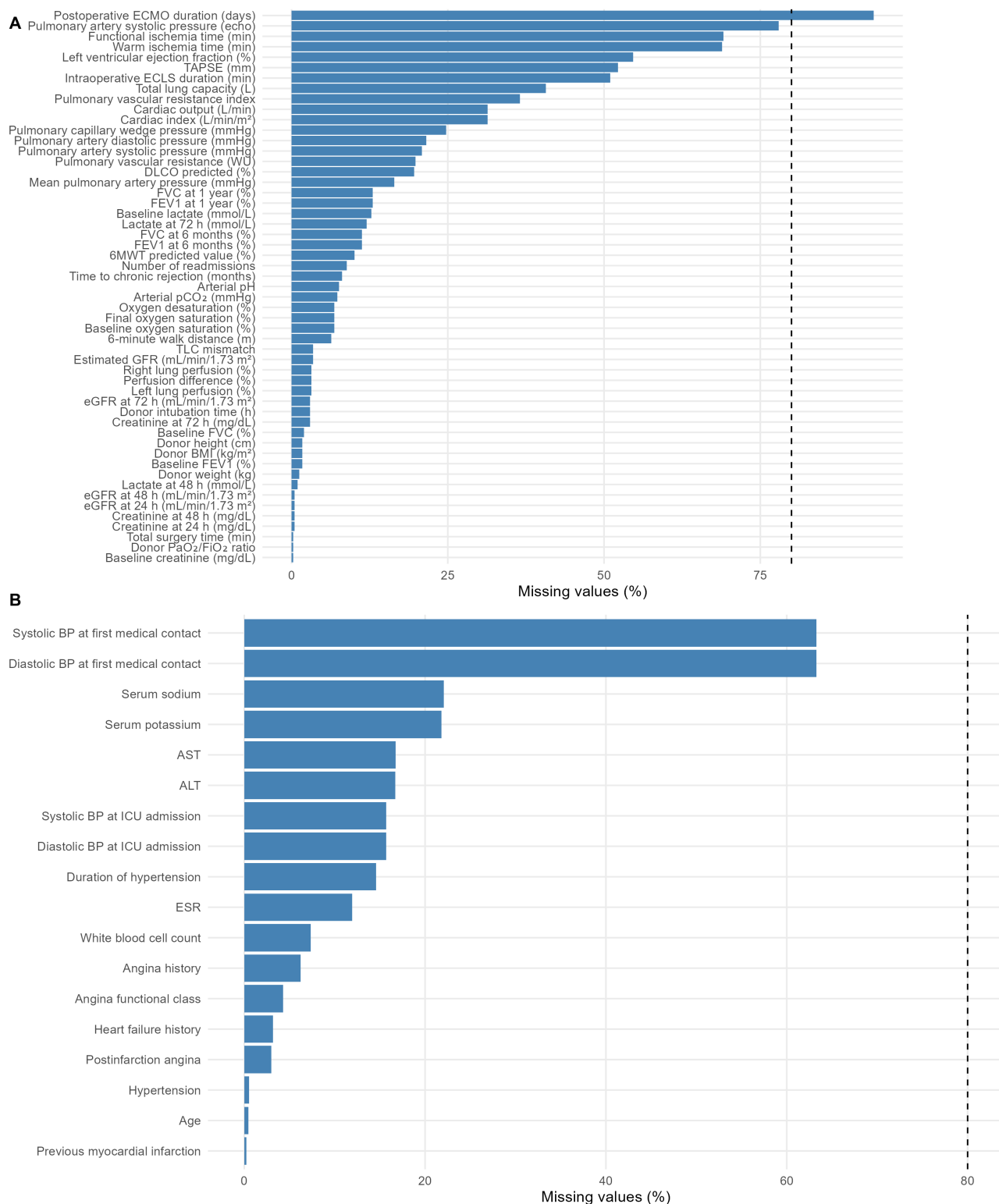


Figure 1

Percentage of missing values across numeric variables in the lung transplantation cohort (A) and the myocardial infarction cohort (B). Dashed vertical lines indicate the predefined exclusion threshold (80% missing values).

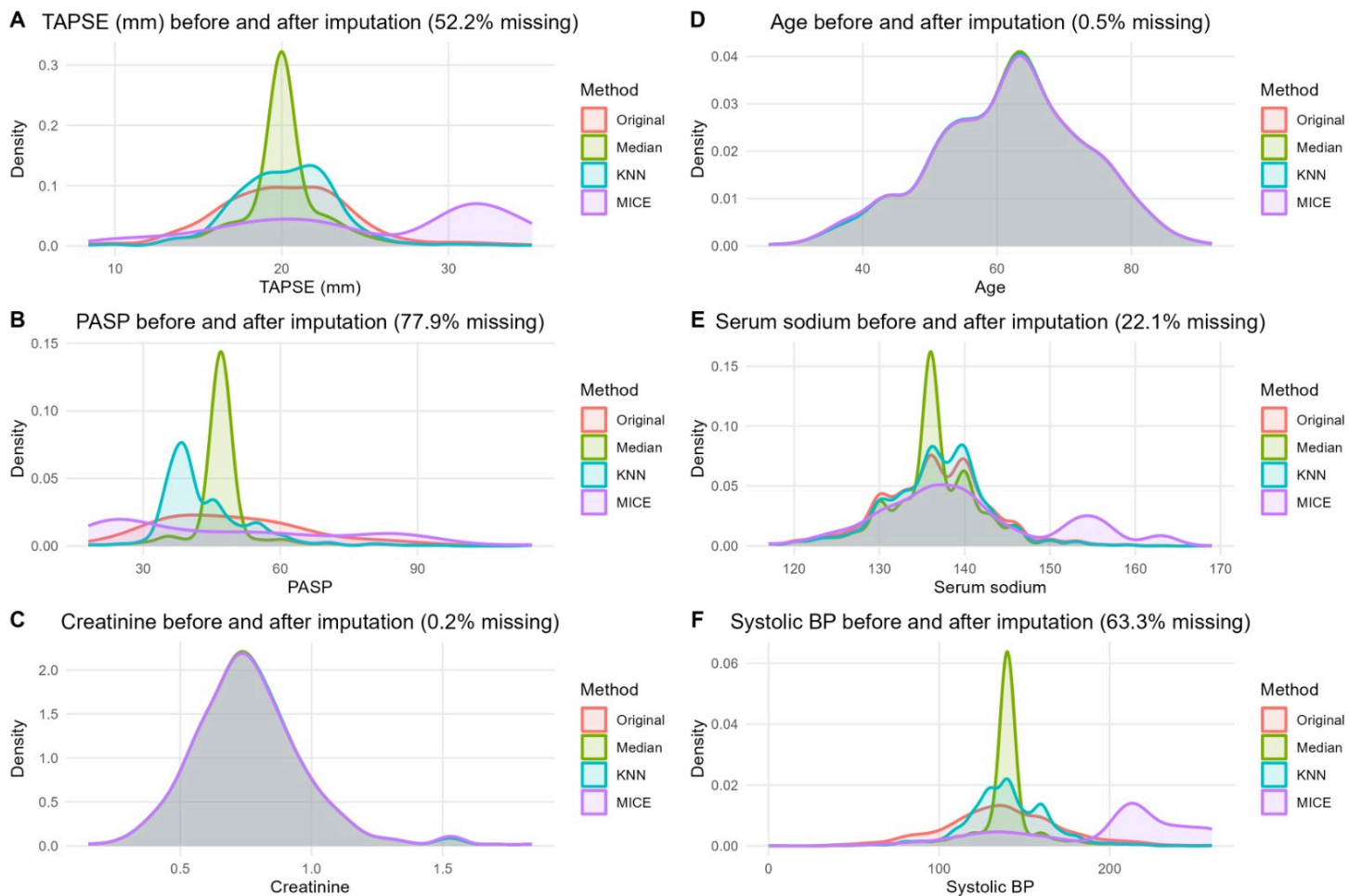


Figure 2

Distribution of representative numeric variables before and after imputation. Panels show variables with low, moderate, and high proportions of missing values in the lung transplantation cohort (A–C: baseline serum creatinine, 0.2% missing; TAPSE, 52.2% missing; PASP, 77.9% missing) and the myocardial infarction cohort (D–F: age, 0.5% missing; serum sodium, 22.1% missing; systolic blood pressure at presentation, 63.3% missing).

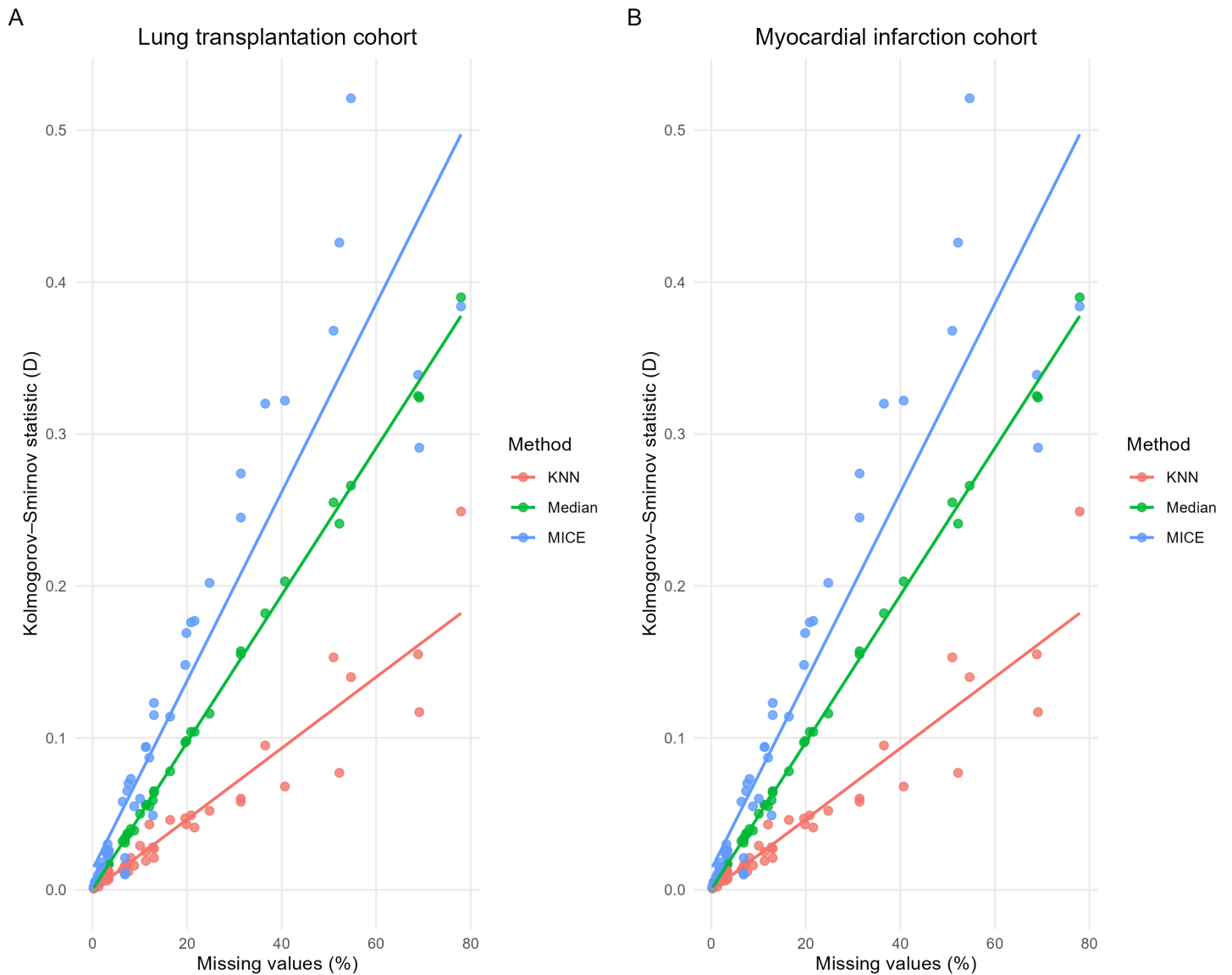


Figure 3

Relationship between the proportion of missing values and distributional distortion in the lung transplantation cohort (A) and the myocardial infarction cohort (B). Each point represents a numeric variable; solid lines represent linear regression fits for each imputation method.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTable.docx](#)