Research Square

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

Random draws from observed values for missing data: when do they support valid inference? A simulation study with a clinical illustration

Nyiko Muhluri Khoza[1]
University of South Africa
Bhekisipho Twala
twala.bhekisipho@mut.ac.za

Mangosuthu University of Technology

Research Article

Keywords: Missing data, multiple imputation, hot-deck imputation, approximate Bayesian bootstrap, Rubin's rules, simulation study, confidence-interval coverage

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11151581/v1

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

Additional Declarations: No competing interests reported.

Random draws from observed values for missing data: when do they support valid inference? A simulation study with a clinical illustration

Nyiko Muhluri Khoza[1], Department of Statistics, University of South Africa, Pretoria, South Africa
Bhekisipho Twala[2], Resources and Planning, Mangosuthu University of Technology, Durban, South Africa
[1]ngobenh@unisa.ac.za [2]twala.bhekisipho@mut.ac.za

Abstract

Imputation methods that draw replacement values from a variable's admissible domain are attractive because they are simple and never produce structurally invalid values. This study asks when such draws support valid statistical inference. The Random-Draw Observation Framework (RDOF) is examined in three forms: the original Random-Draw Observation Imputation (ROI), which draws uniformly between the observed minimum and maximum; ROI-E, which draws from the empirical distribution of the observed values via the approximate Bayesian bootstrap; and Random-Draw Class Observation Imputation (RCOI-E), which applies the same draws within cells defined by a class variable and the analysis variables. We combine completed datasets using Rubin's rules and also assess the common practice of retaining the imputation with the smallest standard error (SE). In a pre-specified simulation study of 72 scenarios (six marginal distributions; MCAR, MAR and MNAR mechanisms; 30% and 60% missingness; strong and weak class variables; n = 340; 500 replications per scenario), uniform-range ROI attenuated the regression slope by 55–83% on average, and its 95% confidence intervals covered the true slope in no more than 16% of replications in any scenario (6% when averaged over distributions). Selecting the minimum-SE imputation understated the mean SE by up to 68% without reducing bias. RCOI-E never produced an invalid value and achieved 93–96% slope coverage at 30% missingness under all three mechanisms. Still, coverage fell to 85% at 60% MAR missingness and to 58% when the class variable was weakly informative. Predictive mean matching (PMM), which also imputes observed values, was more robust under MAR, whereas RCOI-E was comparable or slightly better under MCAR and MNAR. A resampling study of 270 complete clinical records (Statlog Heart data) reproduced these patterns: at 60% missingness, uniform-range ROI attenuated the slope of age on maximum heart rate by 71–74% and covered the true slope in at most 1% of replications, whereas RCOI-E and PMM kept coverage between 91% and 95%. Value-format preservation is therefore necessary but not sufficient: random-draw imputation supports valid inference only when the draws condition on the variables that drive the analysis and the missingness, and when the imputations are pooled rather than selected.

Keywords

Missing data; multiple imputation; hot-deck imputation; approximate Bayesian bootstrap; Rubin's rules; simulation study; confidence-interval coverage

Highlights

- Uniform draws between the observed minimum and maximum bias slopes by 55–83% and give 95% CI coverage of at most 16%.
- Keeping the imputation with the smallest SE understates uncertainty by up to 68% and does not reduce bias.

- Class-conditional bootstrap draws (RCOI-E) pooled with Rubin's rules give near-nominal coverage at moderate missingness.
- RCOI-E degrades with high MAR missingness and weakly informative classes, where PMM is more robust.
- Preserving admissible values is necessary but not sufficient for valid inference.

1. Introduction

Missing data are a persistent problem in applied statistics. They arise from nonresponse, attrition, recording errors and incomplete data capture, and they can bias estimates, reduce efficiency and invalidate confidence intervals (Graham, 2009; Little and Rubin, 2019). The severity of these consequences depends on the missingness mechanism, the proportion and pattern of missing values, and the assumptions of the method used to handle them.

Many practitioners prefer imputation methods that respect a variable's value format: counts stay integers, bounded scores stay within their bounds, and categories stay valid. Mean imputation, regression imputation and several machine-learning imputers can violate these constraints by producing fractional counts or values outside the observed range. A natural response is to draw replacements at random from each variable's admissible domain. The Random-Draw Observation Framework (RDOF) formalises this idea through two methods: Random-Draw Observation Imputation (ROI), which draws from each variable's admissible domain, and Random-Draw Class Observation Imputation (RCOI), which draws within classes of a grouping variable.

Respecting the value format does not by itself guarantee valid inference. An imputation method must also reproduce the distribution of the incomplete variable and its relationships with the other variables, and the analysis must propagate the extra uncertainty that imputation introduces (Rubin, 1987; Schafer & Graham, 2002). Two design choices matter in particular. The first is the distribution from which values are drawn: a uniform draw over the observed range ignores the distribution's shape and the variable's association with other variables. The second is how to combine multiple imputations. Retaining the completed dataset with the smallest standard error (SE), a proposed convergence or selection rule, turns multiple imputation back into a data-dependent single imputation and discards the between-imputation variance (Khoza, Twala and Malela-Majika, unpublished manuscript).

Because precision and accuracy are distinct properties, a lower SE does not indicate a better method. The bias–variance decomposition makes this explicit:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \{\mathrm{Bias}(\hat{\theta})\}^2, \qquad \mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$$

An estimator can therefore have a small SE and a large bias, and a reported SE can itself understate the true sampling variability (Casella & Berger, 2002). Evaluating imputation methods requires assessing bias, root mean squared error (RMSE), the accuracy of the reported SE, and confidence-interval coverage together, against a known true value (Morris, White and Crowther 2019).

This paper makes five contributions:

- It defines three members of RDOF: the original uniform-range ROI; ROI-E, which draws from the empirical distribution of the observed values; and RCOI-E, which draws within cells defined by a class variable and the analysis variables. ROI-E and RCOI-E use the approximate Bayesian bootstrap so that the imputations are proper in Rubin's sense.
- It shows analytically and by simulation why selecting the minimum-SE imputation invalidates inference, and replaces that rule with Rubin's rules.
- It reports a pre-specified simulation study, following the ADEMP framework, that compares the RDOF methods with complete-case analysis, mean imputation, predictive mean matching, k-nearest-neighbour imputation and missForest across 72 scenarios.
- It illustrates the same comparison on real clinical data by repeatedly removing values from a complete dataset whose true values are known.
- It identifies the conditions under which value-preserving random draws support valid inference, and those under which they do not.

The rest of this paper is organised as follows. The remainder of the Background covers the missingness mechanisms, multiple imputation with Rubin's combining rules, and the donor-based and machine-learning imputation methods that serve as comparators. The Methods section defines the admissible value domain on which the Random-Draw Observation Framework (RDOF) is built. It then describes the original uniform-range Random-Draw Observation Imputation (ROI), its empirical approximate Bayesian bootstrap version (ROI-E) and the class-conditional version (RCOI-E). It explains how to pool imputations under Rubin's rules and why choosing the imputation with the smallest standard error invalidates inference. It then sets out the simulation study following the ADEMP framework and the empirical illustration on the Statlog Heart data. The Results section reports inference for the regression slope, the mean and distributional estimands, and the effects of distribution shape and of class-variable strength. It also covers the consequences of minimum-SE selection, the imputation diagnostics and the Heart data illustration. The Discussion interprets these findings in light of existing imputation practice. It sets out the study's strengths and limitations, and the Conclusions provide practical recommendations on when value-preserving random draws support valid inference.

2. Background

2.1. Missingness mechanisms

Rubin (1976) distinguished three missingness mechanisms. Data are missing completely at random (MCAR) when missingness is unrelated to any data, missing at random (MAR) when it depends only on observed data, and missing not at random (MNAR) when it depends on unobserved values. Under MCAR, complete-case analysis is unbiased but inefficient. Under MAR, valid inference generally requires a method that conditions on the variables associated with missingness. Under MNAR, no method based on the observed data alone is guaranteed to be unbiased, and sensitivity analysis is needed (Little & Rubin, 2019).

The distinction between MAR and MNAR cannot be tested from the observed data alone, because it concerns values that were never seen. In practice, analysts make MAR more plausible by including in the imputation model the variables that predict both the incomplete variable and its missingness, an inclusive strategy recommended by Schafer and Graham (2002). For RDOF, this has a direct consequence: a random draw that ignores the other

variables is valid only under MCAR, and the class and cell variables used by RCOI-E should be chosen because they predict the incomplete variable or its missingness, not only because they form convenient groups.

2.2. Multiple imputation and Rubin's rules

Multiple imputation replaces each missing value with M > 1 plausible values drawn from a predictive distribution, analyses each completed dataset, and combines the results (Rubin, 1987; van Buuren, 2018). The combined estimate is the average of the M estimates. Its variance adds the average within-imputation variance to an inflated between-imputation variance, which reflects uncertainty about the missing values. Imputations are *proper* when the draws also reflect uncertainty about the parameters of the imputation model; improper imputations tend to understate the total variance.

Rubin's rules are valid only when every completed dataset is kept and analysed. The between-imputation variance is the part of the total variance that reflects uncertainty about the missing values, so any rule that discards completed datasets, or that chooses one of them by looking at its results, removes that uncertainty from the analysis. The number of imputations also matters: small values of M were once considered adequate for point estimates, but more imputations are needed for stable standard errors and confidence intervals when the fraction of missing information is large, and a common rule of thumb sets M at least equal to the percentage of incomplete records (White, Royston and Wood 2011).

2.3. Donor-based imputation

Hot-deck methods impute a missing value with an observed value from a similar record, the donor (Andridge & Little, 2010). Because donors are real observations, imputed values are always admissible. The approximate Bayesian bootstrap (ABB) makes random hot-deck imputation proper. For each imputation, the donor pool is first resampled with replacement and imputations are then drawn from the resampled pool (Rubin & Schenker, 1986). Predictive mean matching (PMM) selects donors whose predicted values from a regression model are close to those of the incomplete record (Little 1988; Morris, White and Royston 2014). PMM is the default method for continuous variables in several multiple-imputation packages. RCOI is closely related to random hot-deck imputation within adjustment cells, and PMM is therefore its most relevant benchmark.

The main design choice in donor-based imputation is how donors match recipients. Fine adjustment cells or close predictive matches reduce bias, because donors resemble the records they replace. However, they leave fewer donors per recipient and can understate uncertainty when the same donor is used repeatedly (Andridge & Little, 2010). For PMM, the number of candidate donors has a similar effect, and too small a donor pool can make intervals too narrow (Morris, White and Royston 2014). RCOI-E faces the same trade-off through its cell definition and its rule for collapsing sparse cells.

2.4. Machine-learning imputation

k-nearest-neighbour (KNN) imputation replaces a missing value with an average of the values of the k most similar complete records (Troyanskaya et al., 2001). missForest iteratively fits random forests to predict each

incomplete variable from the others (Stekhoven & Bühlmann, 2012), and generative adversarial approaches such as GAIN learn the joint distribution with neural networks (Yoon, Jordon and van der Schaar, 2018). These methods often achieve good predictive accuracy (Jadhav, Pramod and Ramanathan, 2019; Emmanuel et al., 2021; Luo, 2022), but they are usually applied as single imputation. Treating a single completed dataset as fully observed understates uncertainty, which leads to confidence intervals that are too narrow (Zhang, 2016).

These methods are usually evaluated by how closely the imputed values match the true values, for example, by the root mean squared error of the imputations. That criterion rewards predictions close to the conditional mean, which is exactly what makes a single completed dataset look less variable than real data. Inference about population parameters requires a different standard: estimates must be unbiased, and confidence intervals must reach their nominal coverage (Morris, White and Crowther 2019). This study therefore evaluates all methods, including the machine-learning comparators, by bias, estimator RMSE, SE accuracy, and coverage.

3. The Random-Draw Observation Framework

3.1. Admissible value domain

Let $X = \{X_1, \ldots, X_p\}$ be a dataset of n records, and write $X_k = X_{ok} \cup X_{mk}$ for the observed and missing values of variable X_k. The admissible domain V_k is the set of values X_k can take. For a continuous variable estimated from the data, V_k is the interval between the observed extremes:

$$V_k = [a_k, b_k], \qquad a_k = \min(X_{ok}), \quad b_k = \max(X_{ok})$$

For a count or other discrete variable, V_k is the set of integers in $[a_k, b_k]$; for a categorical variable, it is the set of observed categories. When measurement definitions give tighter or wider bounds, use those instead. An imputation method *preserves the value format* if every imputed value lies in V_k. This property does not imply that the imputed values follow the distribution F_k of X_k, nor that they preserve the joint distribution of X_k and the other variables.

3.2. Original ROI: uniform draws on the observed range

The original ROI draws each missing value independently from a uniform distribution on the admissible domain, rounding to the nearest integer for discrete variables:

$$\tilde{x}_{ik}^{(m)} \sim \mathrm{U}(a_k, b_k), \qquad m = 1, \ldots, M$$

Uniform draws are centred on the mid-range $(a_k + b_k)/2$ rather than the mean, so they bias the mean whenever the distribution is asymmetric. Because each draw ignores the rest of the record, the imputed values are unrelated to the other variables, and regression coefficients involving X_k are attenuated towards zero in proportion to the fraction of missing values.

3.3. ROI-E: empirical approximate Bayesian bootstrap

ROI-E replaces the uniform draw with an ABB draw from the observed values. For imputation m, a bootstrap sample of size n_{ok} is drawn with replacement from X_{ok}, and each missing value is replaced by a value drawn at random from that bootstrap sample. The imputed values are observed values, so the value format is preserved by construction, and their distribution matches the observed marginal distribution. ROI-E still ignores the other variables, so it is valid for the marginal distribution only under MCAR.

The bootstrap step is what makes ROI-E proper. Drawing directly from the observed values, without first resampling them, treats the observed empirical distribution as known exactly; the between-imputation variance is then too small and the pooled intervals too narrow (Rubin & Schenker, 1986). Resampling the donor pool before each set of draws adds the uncertainty about that distribution. ROI-E can also be seen as a special case of RCOI-E with a single cell, making it a useful reference for measuring how much conditioning in RCOI-E contributes.

3.4. RCOI-E: class-conditional approximate Bayesian bootstrap

RCOI-E applies the ABB draw within imputation cells. The cells are formed by crossing the class variable C with coarse categories of the other analysis variables, including the analysis model outcome. In this study, the cells were C × tertile(Y) × half(Z), where Y is the outcome, and Z is an auxiliary variable. If a cell contains fewer than five donors, it is merged into a coarser cell (C × tertile(Y), then C, then the whole sample). Including the analysis variables in the cell definition is essential: an imputation model that omits the outcome attenuates associations with it (van Buuren, 2018).

We fixed the cell structure before running the simulation. Crossing the class variable with tertiles of the outcome and halves of the auxiliary variable gives 18 cells, which is enough to capture the main associations at n = 340 while usually leaving at least five donors per cell at 30% missingness. The collapsing rule matters mainly at 60% missingness, where sparse cells merge more often, and the conditioning becomes coarser. Data-adaptive cells, such as cells formed from a predicted mean, would move RCOI-E towards PMM; the fixed design keeps the method simple and transparent, at a cost that the results quantify.

3.5. Combining imputations

We analyse each of the M completed datasets, yielding estimates $\hat{Q}_m$ with variance estimates U_m. The combined estimate and its total variance follow Rubin's rules:

$$\bar{Q} = \frac{1}{M}\sum_{m=1}^{M} \hat{Q}_m\,, \qquad \bar{U} = \frac{1}{M}\sum_{m=1}^{M} U_m\,, \qquad B = \frac{1}{M-1}\sum_{m=1}^{M} \left(\hat{Q}_m - \bar{Q}\right)^2$$

$$T = \bar{U} + \left(1 + \frac{1}{M}\right) B$$

Confidence intervals use a t reference distribution with the Barnard–Rubin degrees of freedom (Barnard & Rubin, 1999). Following White, Royston and Wood (2011), we set M equal to the percentage of incomplete records (M = 30 or 60).

3.6. Why selecting the minimum-SE imputation fails

An alternative rule, used in earlier versions of RDOF, generates M completed datasets and reports the one whose estimate has the smallest SE:

$$m^* = \underset{m}{\operatorname{argmin}}\ \widehat{\mathrm{SE}}\big(\hat{\theta}^{(m)}\big), \qquad \hat{\theta}_{\mathrm{sel}} = \hat{\theta}^{(m^*)}$$

This rule has two defects. First, the reported SE is the minimum of M noisy estimates, so it is biased downward, and the bias grows with M. Second, it discards the between-imputation variance B, the component of T that reflects uncertainty about the missing values. Selection cannot remove bias in the draws, because the same mechanism generates every candidate dataset. The result is an estimate as biased as the imputation mechanism, but with overstated reported precision.

4. Simulation study

We designed and reported the simulation following the ADEMP framework (Morris, White and Crowther 2019). All code, the full results and the scripts that produce every table and figure are available as supplementary material.

The study compares nine methods over 72 scenarios formed by crossing six marginal distributions, three missingness mechanisms, two proportions of missing data and two strengths of the class variable. Each method received its own random-number stream, so adding or removing a comparator could not change any other method's draws. The analysis was carried out in Python with NumPy, SciPy, pandas and scikit-learn, and every scenario can be regenerated from the supplied scripts and seeds.

4.1. Aims

The aims were (i) to assess whether ROI-E and RCOI-E, pooled with Rubin's rules, give unbiased estimates and calibrated confidence intervals under MCAR and MAR; (ii) to identify the conditions under which they fail, including MNAR, non-normal distributions and weakly informative classes; and (iii) to quantify the effect of uniform-range draws and minimum-SE selection.

We stated three expectations before running the study. First, uniform-range ROI would attenuate associations with the incomplete variable and bias its mean whenever the distribution is asymmetric. Second, selecting the minimum-SE imputation would understate the SE without reducing bias. Third, RCOI-E would approach nominal coverage under MAR when the class variable is informative, and lose accuracy when it is not. The results sections report how far each expectation was met.

4.2. Data-generating mechanisms

Each dataset contained an always-observed auxiliary variable Z ~ N(0, 1), an incomplete variable X, a three-level class variable C and an outcome Y. A Gaussian copula gave every marginal distribution of X the same dependence structure:

$$L = 0.5\,Z + \sqrt{0.75}\,U, \qquad X = F^{-1}\{\Phi(L)\}, \qquad U \sim N(0,1)$$
$$Y = 1 + \beta X + 0.3\,Z + \varepsilon, \qquad \beta = 0.5/\sigma_X, \qquad \varepsilon \sim N(0,1)$$

We formed C from the tertiles of ρ_c L + (1 − ρ_c^2)^½ V, with V ~ N(0, 1), and ρ_c = 0.9 (strong class) or 0.2 (weak class). Table 1 lists the factors varied: six marginal distributions of X, three missingness mechanisms, two proportions of missing data and two strengths of the class variable, which cross to give 72 scenarios. The sample size was n = 340, which matches the Life Expectancy subsample used in earlier work on these methods.

Table 1. Simulation factors

Factor	Levels
Marginal distribution of X	Normal (mean 50, SD 10); lognormal (σ = 0.8, right-skewed); normal mixture 0.6 N(0, 1) + 0.4 N(4, 1) (bimodal); t with 3 df (heavy-tailed); Poisson with mean 3 (count); 100 × Beta(2, 5) (bounded, skewed)
Missingness mechanism	MCAR; MAR on standardised Y and Z (logistic, coefficients 1.0 and 0.5); MNAR on standardised X (logistic, coefficient 1.0)
Proportion missing	30%; 60% (logistic intercept solved for the target proportion)
Class strength	Strong (ρ_c = 0.9); weak (ρ_c = 0.2)
Sample size	n = 340
Replications	500 per scenario; 72 scenarios

4.3. Estimands

Four estimands were studied: the mean of X; the standard deviation of X; the proportion of X above the 90th percentile of its true distribution, which is sensitive to the shape of the imputed distribution; and the coefficient β of X in the linear regression of Y on X and Z. True values were computed from 2,000,000 draws of the data-generating mechanism.

The four estimands probe different properties of an imputation method. The mean of X was the target of earlier work on RDOF and is sensitive to where imputed values are centred. The standard deviation measures whether the imputations preserve spread; because we did not compute a model-based SE, we report only its bias and RMSE. The tail proportion is sensitive to the shape of the imputed distribution, and the regression slope is sensitive to whether the imputations preserve the association between X and the other variables. The slope model was correctly specified, so differences between methods reflect imputation alone.

4.4. Methods compared

Table 2 lists the nine methods compared. We analysed the ROI in two ways from the same draws to separate the effect of the uniform draw from the effect of minimum-SE selection.

The methods fall into three groups. Complete-case analysis serves as the reference that uses no imputation. Mean imputation, KNN and missForest are single-imputation methods analysed as if the completed data were fully observed, which is how they are commonly used in practice. ROI, ROI-E, RCOI-E and PMM generate multiple completed datasets; for ROI, the same draws were analysed twice, once by selecting the minimum-SE dataset and

once by Rubin's rules, so that the effect of the uniform draw and the effect of selection can be separated. GAIN was not included because its training cost across 36,000 simulated datasets was prohibitive, and its imputations are also single predictions.

Table 2. Methods compared

Method	Imputation	Inference
Complete case (CC)	None; incomplete records deleted	Standard analysis
Mean imputation (MOI)	Observed mean	Single dataset, naive SE
ROI (min-SE)	Uniform on [min, max]; keep dataset with smallest SE of the mean	Single dataset, naive SE
ROI (Rubin)	Same uniform draws	Rubin's rules
ROI-E	ABB from all observed values	Rubin's rules
RCOI-E	ABB within C × tertile(Y) × half(Z) cells, collapsing if < 5 donors	Rubin's rules
PMM	Proper PMM on Y, Z and C, 5 donors	Rubin's rules
KNN	Mean of 5 nearest neighbours on (Y, Z, C)	Single dataset, naive SE
missForest	Random forest (50 trees) of X on (Y, Z, C)	Single dataset, naive SE

ABB, approximate Bayesian bootstrap; PMM, predictive mean matching; KNN, k-nearest neighbours. With a single incomplete variable, missForest reduces to one random-forest regression. Each method used an independent random-number stream so that no comparator could change another method's draws.

4.5. Performance measures

For each method and estimand, the study reports bias; relative bias (bias divided by the true value, or by the true SD of X for the mean, because the true mean is near zero for two distributions); empirical SE; the relative error of the average model-based SE against the empirical SE; RMSE; and the coverage of nominal 95% confidence intervals. Two diagnostics describe the imputed values themselves: the proportion of imputed values outside the admissible domain (non-integer values for the Poisson variable), and the Kolmogorov–Smirnov distance between imputed and true missing values. With 500 replications, the Monte Carlo SE of a coverage near 95% is about 1.0 percentage point, and never exceeded 2.2 points.

Coverage is the proportion of replications in which the 95% confidence interval contains the true value; a valid method should give coverage close to 95%, and values well below it signal either bias or an understated SE. The relative error of the model SE separates these two causes: a large negative value means the reported uncertainty is too small, whatever the bias. Results are presented for individual scenarios and, where stated, as averages over the six distributions with a strong class variable; averaging is used only to summarise patterns that hold across distributions, as Table 4 shows.

5. Results

5.1. Regression slope

Table 3 and Figure 1 summarise inference for the regression slope, averaged over the six distributions with a strong class variable. Averaged over distributions, uniform-range ROI attenuated the slope by 55–83% in every setting, and its intervals covered the true slope in no more than 6% of replications (16% in the best single scenario).

Pooling the uniform draws with Rubin's rules inflated the SE but did not help, because the bias was far larger than any interval. ROI-E performed almost as badly, because its draws also ignore Y.

Table 3. Inference for the slope β: relative bias, 95% CI coverage and relative error of the model SE (strong class, averaged over six distributions)

Method	Rel. bias MAR 30%	Rel. bias MAR 60%	Coverage MCAR 30%	Coverage MCAR 60%	Coverage MAR 30%	Coverage MAR 60%	Coverage MNAR 30%	SE error MAR 60%
CC	−9.7	−11.2	94.8	94.6	89.7	91.6	94.7	−2
MOI	−30.7	−39.3	92.3	88.5	55.9	59.3	89.8	+7
ROI (min-SE)	−56.3	−73.6	1.1	0.0	4.2	1.0	5.8	−12
ROI (Rubin)	−58.1	−74.9	0.4	0.0	3.7	0.9	4.4	+79
ROI-E	−53.0	−76.2	26.0	0.4	5.9	0.5	31.2	+112
RCOI-E	−8.0	−14.7	95.2	91.3	93.0	85.0	95.5	−3
PMM	−1.5	−1.6	93.7	92.4	94.7	92.2	91.9	−7
KNN	+11.1	+23.4	83.7	55.3	82.7	61.0	72.8	−37
missForest	+9.5	+19.3	84.9	65.4	83.4	64.6	72.5	−39

All values are in %. Relative bias = 100 × bias / β. SE error = 100 × (mean model SE / empirical SE − 1); negative values mean the reported SE is too small. Monte Carlo SE of coverage ≤ 2.2 percentage points.

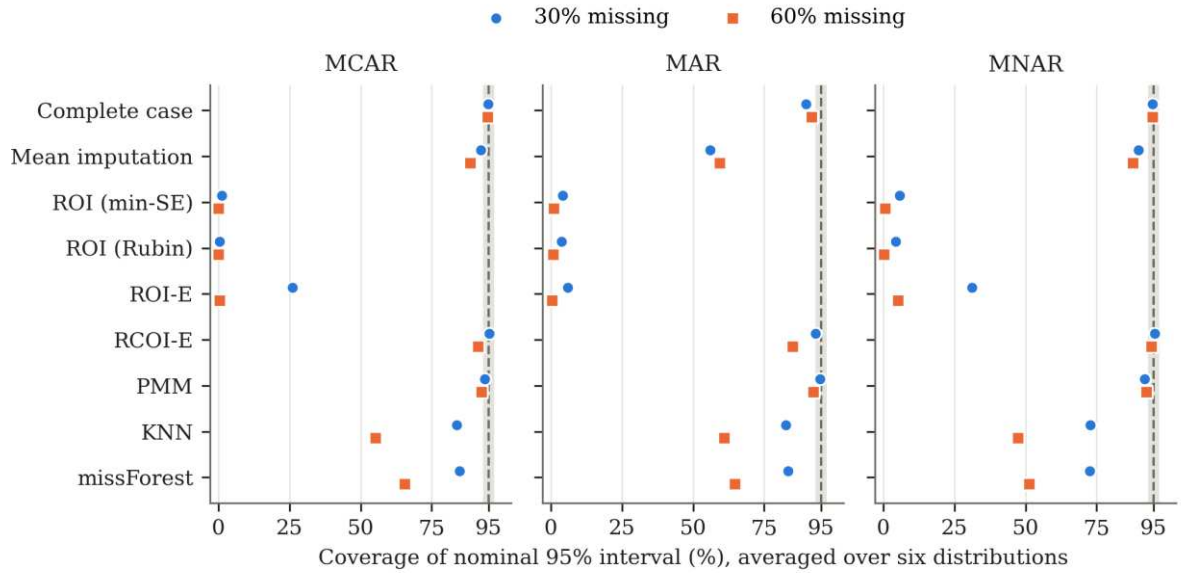


Figure 1. Coverage of nominal 95% confidence intervals for the slope β, by method, mechanism and proportion missing (strong class, averaged over six distributions). The dashed line marks 95%, and the grey band marks the Monte Carlo error around it.

In Figure 1, the points for ROI and ROI-E sit near zero coverage in every panel, while complete-case analysis, RCOI-E, and PMM cluster near the 95% line, and KNN and missForest drift further below it as missingness rises. RCOI-E and PMM both gave near-nominal coverage at 30% missingness under every mechanism (RCOI-E 93.0–95.5%, PMM 91.9–94.7%). At 60% MAR missingness, RCOI-E undercovered (85.0%) because the coarse cells leave residual confounding by Y and Z, and the slope was attenuated by 14.7%. PMM, whose regression model uses Y and Z as continuous predictors, stayed closer to nominal (92.2%). Under MNAR, where missingness depended only on X, the conditional slope remains identifiable, and both methods were well covered.

KNN and missForest overestimated the slope by 9–38% and reported SEs 11–39% too small, so their coverage fell to 47–65% at 60% missingness. These methods predict missing values rather than drawing them, which understates residual variation in X and inflates the slope (Zhang, 2016).

5.2. Mean and distributional estimands

Figure 2 shows coverage for the mean of X. Under MCAR, ROI-E, RCOI-E and PMM were all close to nominal (94.1–95.1%). Uniform-range ROI was biased towards the mid-range of the observed values by +18% to +22% of an SD at 30% MCAR missingness and +34% to +38% at 60%, and its coverage was 26–44%. Under MAR, only methods that condition on Y and Z reduced the bias of complete-case analysis; RCOI-E covered 87.5% and 75.8% at 30% and 60% missingness, compared with 90.5% and 85.4% for PMM. Under MNAR, no method recovered the mean, as expected: the best coverage was 47%.

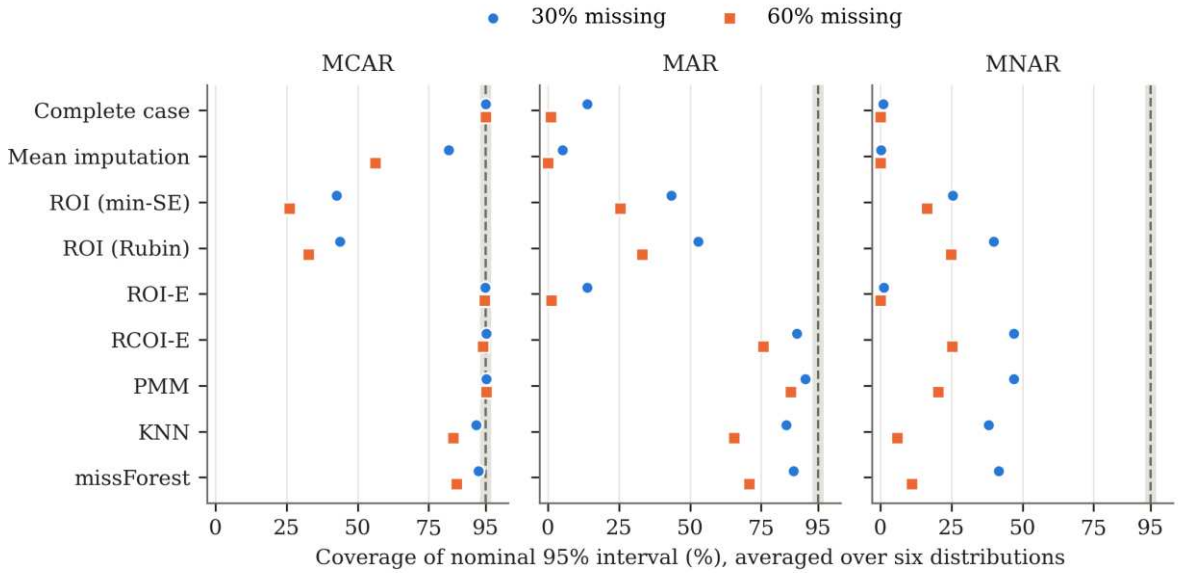


Figure 2. Coverage of nominal 95% confidence intervals for the mean of X, by method, mechanism and proportion missing (strong class, averaged over six distributions).

For the tail proportion $P(X > q_{90})$, RCOI-E gave the best coverage of any method under MAR (88.2% at 30% and 80.0% at 60% missingness, against 86.8% and 71.3% for PMM). Drawing observed values within cells reproduces the local distribution shape, including skewness and bimodality, without relying on a linear model.

5.3. Distribution shape

Table 4 breaks down coverage at 30% MAR missingness by distribution. RCOI-E and PMM gave slope coverage of 92–96% for every distribution, including the lognormal, bimodal and heavy-tailed cases. Coverage of the mean was lowest for the lognormal distribution (RCOI-E 77.0%, PMM 81.8%), where a few large, poorly represented values dominate the mean. Uniform-range ROI failed for all six distributions.

Table 4 also shows that the failure of uniform-range ROI depends predictably on the shape of the distribution. Its coverage of the mean ranged from 17.4% for the t distribution and 23–24% for the normal and lognormal distributions to 74.0% for the Poisson count and 82.0% for the bounded Beta variable, reflecting how far the

midpoint of the observed range lies from the true mean. By contrast, its slope coverage was 13.8% or lower for every distribution because slope attenuation does not depend on where the draws are centred. KNN performed worst on the lognormal distribution, with 66.0% coverage for the mean and 70.2% for the slope, where averaging neighbours pulls imputations away from the long right tail.

Table 4. Coverage (%) of 95% confidence intervals by distribution, MAR, 30% missing, strong class

Estimand	Method	Normal	Lognormal	Bimodal	t (3 df)	Poisson	Beta
Slope β	ROI (min-SE)	0.0	13.8	3.2	0.0	3.6	4.6
	RCOI-E	93.8	93.4	92.4	92.0	93.2	93.4
	PMM	95.6	95.4	95.0	92.2	95.8	94.2
	KNN	86.6	70.2	88.6	87.0	82.2	81.4
Mean of X	ROI (min-SE)	24.2	23.2	39.6	17.4	74.0	82.0
	RCOI-E	90.6	77.0	91.6	88.8	89.8	87.4
	PMM	93.4	81.8	93.4	94.4	90.4	89.6
	KNN	89.0	66.0	89.4	87.0	86.4	85.4

5.4. Strength of the class variable

Figure 3 shows how RCOI-E depends on the class variable. With a strong class, RCOI-E and PMM behaved similarly. With a weak class, the cells captured less of the variation in X, and at 60% MAR missingness, RCOI-E coverage fell to 57.7% for the slope and 40.4% for the mean, compared with 85.7% and 78.4% for PMM. Across all 72 scenarios, the RMSE of RCOI-E relative to PMM ranged from 0.87 (MNAR, strong class) to 1.83 (mean of X, MAR, weak class, 60% missing). RCOI-E had the lower slope RMSE in 46% of scenarios, mostly under MCAR and MNAR.

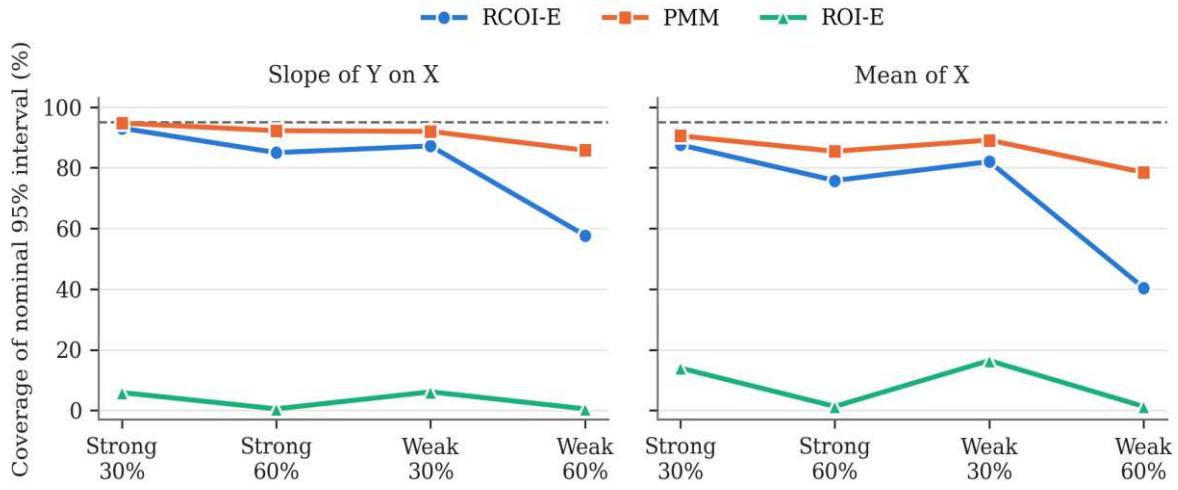


Figure 3. Coverage of nominal 95% intervals for RCOI-E, PMM and ROI-E under MAR, by class strength and proportion missing (averaged over six distributions).

The class variable matters much less for PMM, because PMM also uses Y and Z as continuous predictors. Its slope coverage under MAR fell from 94.7% to 92.0% at 30% missingness and from 92.2% to 85.7% at 60% when

the class variable was weak. ROI-E, which uses no conditioning at all, covered the slope in about 6% of replications at 30% missingness and under 1% at 60%, regardless of class strength. RCOI-E is therefore best seen as a method whose validity rests on the quality of its cells: with an informative class variable it matches PMM, and without one it moves towards the unconditional ROI-E.

5.5. Minimum-SE selection

Selecting the minimum-SE imputation behaved as Section 3.6 predicts. For the mean of X under MCAR, the reported SE was 48% too small at 30% missingness and 68% too small at 60%, compared with 34% and 56% for the same uniform draws pooled with Rubin's rules. Selection did not reduce bias in any scenario. The lower SEs reported for ROI in earlier work therefore come from the selection step, not from better estimation.

For the slope, the two ways of analysing the same uniform draws failed in opposite directions. Selection understated the SE by 12% at 60% MAR missingness, whereas Rubin's rules overstated it by 79%, because the between-imputation variance of uniform draws is large. Neither helped: coverage was no more than 6% in either case, because the bias of the uniform draws was far larger than any interval. Pooling is necessary for valid inference, but it cannot correct draws that come from the wrong distribution; for the mean under MCAR, pooling raised coverage only from 43% to 44% at 30% missingness and from 26% to 33% at 60%.

5.6. Value format and computation

Table 5 summarises the diagnostics. All draw-based methods (ROI, ROI-E, RCOI-E and PMM) preserved the value format by construction. For the Poisson count variable, mean imputation produced non-integer values in 99% of imputations, KNN in 80% and missForest in 98%. RCOI-E and PMM imputed values closest to the true missing values in distribution (mean Kolmogorov–Smirnov distances of 0.18 and 0.16). Computation was not a constraint: with M = 30 or 60, RCOI-E took on average 39 ms per dataset and PMM 33 ms.

Table 5. Imputation diagnostics, averaged over all scenarios

Method	Non-integer imputations, Poisson (%)	KS distance to true values	SE error, mean, MCAR 60% (%)	Time per dataset (ms)
MOI	99	0.68	−61	0.2
ROI (min-SE)	0	0.29	−68	6.3
ROI (Rubin)	0	0.31	−56	6.3
ROI-E	0	0.27	−2	8.0
RCOI-E	0	0.18	−8	38.8
PMM	0	0.16	−1	33.2
KNN	80	0.23	−27	3.4
missForest	98	0.21	−26	51.0

KS, Kolmogorov–Smirnov. Times are for all M imputations and analyses of one dataset on a single CPU core; ROI times cover both the min-SE and Rubin variants, which share the same draws.

For the five continuous distributions, no method produced values outside the observed range: mean imputation uses the observed mean, and KNN and random-forest predictions average observed values. The value-format

advantage of draw-based methods is therefore specific to discrete and bounded variables in this design, where non-integer or out-of-range values would be invalid. The Kolmogorov–Smirnov distances in Table 5 show a more general advantage of drawing observed values within cells: RCOI-E and PMM reproduced the missing-value distribution more closely than uniform ROI (0.29–0.31) or mean imputation (0.68), which places every imputed value at a single point.

6. Empirical illustration: Statlog Heart data

We checked the simulation results on real data by removing values from a complete clinical dataset so that the true values remain known. The Statlog Heart data contain 270 patients with 13 clinical measurements and a binary heart-disease outcome, and no missing values (UCI Machine Learning Repository, Statlog (Heart)).

The dataset suits this purpose for three reasons. It is complete and publicly available, so the true values are known and the analysis can be reproduced. Maximum heart rate is left-skewed (median 153.5, mean 149.7 beats per minute, range 71–202), which tests the midpoint problem of uniform draws on real data. Moreover, it relates to other variables in a clinically meaningful way: it declines with age (correlation −0.40). It is lower in the 120 patients with heart disease than in the 150 without (138.9 against 158.3 beats per minute), so a MAR mechanism based on age and disease status is plausible and gives RCOI-E an informative class variable.

6.1. Design

We treated the 270 records as the population. The incomplete variable was maximum heart rate achieved during exercise (population mean 149.7 beats per minute). We studied two estimands: its mean and the slope of age on maximum heart rate (population value −0.158 years per beat per minute). Each of 1,000 replications per scenario drew a bootstrap sample of 270 records and removed 30% or 60% of heart-rate values, either completely at random (MCAR) or at random given age and disease status (MAR; logistic model on standardised age plus 0.8 for disease present).

The methods matched the simulation study. ROI drew uniform values between the observed minimum and maximum, rounded to whole beats, and was analysed using both minimum-SE selection and Rubin's rules. RCOI-E used cells formed by disease status and age tertile, collapsing to disease status when a cell had fewer than five donors. PMM used age and disease status as predictors. All multiple-imputation methods used M = 20.

6.2. Results

Table 6 summarises the results at 60% missingness; the real-data results reproduced the simulation. Under MCAR, complete-case analysis, ROI-E, RCOI-E and PMM were unbiased for the mean and covered close to 95%. Uniform-range ROI underestimated the mean by 4.4–5.1 beats per minute, because maximum heart rate is left-skewed and the midpoint of its observed range lies below its mean; its coverage was 39–47%. Mean imputation was unbiased but covered only 55% because its SE was too small.

Under MAR, patients with missing heart rates were older and more often had heart disease, so complete-case analysis, mean imputation and ROI-E overestimated the mean by about 6.6 beats per minute and covered in 16% of replications or fewer. RCOI-E and PMM, which condition on age and disease status, reduced bias to 1.0 and 0.5 beats per minute and covered 89% and 88% of replications, respectively.

Table 6. Statlog Heart data, 60% missing: mean of maximum heart rate and slope of age on maximum heart rate (1,000 replications)

Method	Mean: bias, MCAR (bpm)	Mean: coverage, MCAR (%)	Mean: coverage, MAR (%)	Slope: rel. bias, MCAR (%)	Slope: coverage, MCAR (%)	Slope: coverage, MAR (%)
Complete case	0.00	94.2	13.8	+1.5	95.8	94.7
MOI	0.00	54.9	1.2	+1.5	96.8	97.3
ROI (min-SE)	−4.39	39.1	56.1	−70.8	0.0	1.0
ROI (Rubin)	−5.12	46.9	75.9	−73.5	0.0	0.2
ROI-E	+0.02	94.5	16.0	−58.7	5.6	5.9
RCOI-E	+0.05	95.1	88.9	−8.7	93.7	91.4
PMM	−0.15	94.3	87.8	−0.8	94.9	94.4

bpm, beats per minute. Relative bias = 100 × bias / β; negative values mean attenuation towards zero. Monte Carlo SE of a coverage near 95% is about 0.7 percentage points. Results at 30% missingness follow the same pattern and are in the supplementary material.

For the slope, ROI attenuated the association by 71–74% at 60% missingness and never covered the true value under MCAR. ROI-E attenuated it by 59%, because its draws ignore age. RCOI-E attenuated it by 9% and covered 94%, and PMM was unbiased with 95% coverage. As in the simulation, RCOI-E's coarse cells left residual attenuation that PMM avoided.

7. Discussion

Preserving the value format was necessary to avoid invalid imputations, but it was far from sufficient for valid inference. The original ROI never produced an invalid value. However, it gave the worst inference of any method studied: uniform draws on the observed range bias the mean towards the mid-range and break the relationship between the imputed variable and every other variable. Selecting the imputation with the smallest SE made matters worse, because it concealed the problem behind an understated SE. Earlier conclusions that ROI improves precision, which were based on smaller reported SEs, should therefore not be relied on.

Two changes turn random-draw imputation into a sound method. First, drawing from the observed values with the approximate Bayesian bootstrap, rather than from a uniform distribution, reproduces the marginal distribution and makes the imputations proper. Second, drawing within cells that include the outcome and the predictors of missingness preserves the key associations and addresses MAR missingness. RCOI-E, which combines both changes, gave near-nominal coverage for the slope at moderate missingness under all three mechanisms, and the best coverage of any method for a tail proportion under MAR.

RCOI-E is not a replacement for PMM. With a weak class variable or 60% MAR missingness, RCOI-E's coarse cells left residual bias, while PMM, which uses the continuous predictors directly, was more robust. The two methods share the donor principle, so the choice between them depends on how donors are matched. Cells are

transparent and make no linearity assumption, which helps for distributional estimands; a regression-based match uses more of the available information. For most applied analyses, we recommend PMM with Rubin's rules as the default. RCOI-E is a reasonable alternative when a strong, substantively meaningful class variable exists, missingness is moderate, and the analysis targets distributional features such as tail proportions.

Use single-imputation machine-learning methods with caution for inference. KNN and missForest produced plausible-looking values but overestimated the slope and understated the SE, because a predicted value carries less variation than a real one. Their intervals were therefore too narrow.

The results also show why method comparisons need a known true value. SE and statistics derived from it cannot rank imputation methods: mean imputation and minimum-SE ROI both had some of the smallest reported SEs in this study and some of the worst coverage. Bias, RMSE, the accuracy of the reported SE and coverage, estimated by simulation against a known truth, are the appropriate criteria (Morris, White and Crowther 2019).

8. Limitations

The simulation considered a single incomplete variable, one sample size and one analysis model. Multivariate missingness, smaller and larger samples, categorical incomplete variables and interaction terms may change the relative performance of the methods. RCOI-E used a fixed cell definition; finer or data-adaptive cells might reduce bias under MAR, at the cost of fewer donors per cell. GAIN and other deep generative imputers were not included. Under MNAR, all methods were biased for the mean, and the results should not be read as support for any of them in that setting. Finally, the empirical illustration used one dataset, one incomplete variable and a sample of 270; applications with multivariate missingness and larger samples would strengthen the evidence.

The simulation used 500 replications per scenario, which gives a Monte Carlo SE for coverage of about 1 percentage point and never more than 2.2 points. This is precise enough to separate valid methods from the failures reported here. However, differences of one or two points between RCOI-E and PMM in individual scenarios should not be over-interpreted. The empirical illustration used a single dataset of 270 records and one incomplete variable, and the cell and PMM specifications were fixed rather than tuned; applications with multivariate missingness, larger samples and other clinical outcomes would test whether the patterns generalise.

9. Conclusion

Random-draw imputation that respects each variable's admissible values is simple and avoids invalid imputations, but drawing uniformly on the observed range and keeping the imputation with the smallest SE produces biased estimates with overstated precision. Drawing observed values with the approximate Bayesian bootstrap, within cells that include the analysis variables, and pooling the imputations with Rubin's rules gives valid inference at moderate missingness when the class variable is informative. In other settings, predictive mean matching is the more robust choice.

Three practical recommendations follow. First, judge imputation methods by bias and confidence-interval coverage against a known truth, not by the size of the reported SE. Second, never select completed datasets by

their results; pool them all. Third, where a simple, transparent, value-preserving method is wanted, RCOI-E with an informative class variable is a defensible choice at moderate missingness, and PMM is the safer default otherwise. Future work should examine data-adaptive cells for RCOI-E, multivariate missingness and sensitivity analyses for MNAR.

CRediT authorship contribution statement

Nyiko Muhluri Khoza: Conceptualisation, Methodology, Writing – original draft, Writing – review and editing. **Bhekisipho Twala:** Conceptualisation, Methodology, Software, Formal analysis, Visualisation, Supervision, Writing – review and editing.

Acknowledgements

We thank the University of South Africa, the New Generation of Academics Programme of the Department of Higher Education and Training, and Mangosuthu University of Technology for their support.

Declaration of competing interests

The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Declaration of generative AI use

During preparation of this manuscript, the authors used OpenAI tools for language refinement and Anthropic's Claude to assist with the simulation design, writing and reviewing simulation code, producing tables and figures, and drafting and editing text. The authors specified the research questions, checked the code and results, and took full responsibility for the manuscript's content.

Data and code availability

The supplied code can regenerate the simulated data. The Statlog Heart data are publicly available from the UCI Machine Learning Repository. The simulation and resampling code, the full results for every scenario, method and estimand, and the scripts that produce each table and figure are provided as supplementary material and will be deposited in a public repository before publication.

Funding

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

References

1. Andridge, R.R. and Little, R.J.A., 2010. A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1), pp.40–64.
2. Barnard, J. and Rubin, D.B., 1999. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4), pp.948–955.
3. Casella, G. and Berger, R.L., 2002. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury.
4. Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. and Tabona, O., 2021. A survey on missing data in machine learning—Journal *of Big Data*, 8(1), p.140.
5. Graham, J.W., 2009. Missing data analysis: making it work in the real world. *Annual Review of Psychology*, 60, pp.549–576.
6. Jadhav, A., Pramod, D. and Ramanathan, K., 2019. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10), pp.913–933.
7. Khoza, N.M., Twala, B. and Malela-Majika, J.-C., unpublished manuscript. The smallest standard error is the wrong one: selection among multiple imputations invalidates inference.
8. Little, R.J.A., 1988. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3), pp.287–296.
9. Little, R.J.A. and Rubin, D.B., 2019. *Statistical Analysis with Missing Data*. 3rd ed. Hoboken, NJ: Wiley.
10. Luo, Y., 2022. Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, 23(1), p.bbab489.
11. Morris, T.P., White, I.R. and Crowther, M.J., 2019. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), pp.2074–2102.
12. Morris, T.P., White, I.R. and Royston, P., 2014. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14, p.75.
13. Rubin, D.B., 1976. Inference and missing data. *Biometrika*, 63(3), pp.581–592.
14. Rubin, D.B., 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
15. Rubin, D.B. and Schenker, N., 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81(394), pp.366–374.
16. Schafer, J.L. and Graham, J.W., 2002. Missing data: our view of the state of the art. *Psychological Methods*, 7(2), pp.147–177.
17. Stekhoven, D.J. and Bühlmann, P., 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), pp.112–118.
18. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), pp.520–525.
19. UCI Machine Learning Repository, n.d. Statlog (Heart) [dataset]. Irvine, CA: University of California, Irvine. Available at: https://archive.ics.uci.edu/ml/datasets/statlog+(heart).
20. van Buuren, S., 2018. *Flexible Imputation of Missing Data*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.
21. White, I.R., Royston, P. and Wood, A.M., 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4), pp.377–399.

22. Yoon, J., Jordon, J. and van der Schaar, M., 2018. GAIN: missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80, pp.5689–5698.
23. Zhang, Z., 2016. Missing data imputation: focusing on single imputation. *Annals of Translational Medicine*, 4(1), p.9.