Multiple imputation using multivariate adaptive regression splines

2025-01-10

**Author:**

Jerome Sepin

**Correspondence:**

Jerome Sepin, University of Lucerne, Alpenquai 4, Lucerne CH-6005, Switzerland, Email: jerome.sepin@unilu.ch

**Affiliations:**

University of Lucerne, Faculty of Health Sciences and Medicine

**ORCID:**

Jerome Sepin: <https://orcid.org/0000-0002-8893-8903>

**Targeted journal:**

Journal of Computational and Graphical Statistics

**Abstract:**

Handling incomplete data is a common challenge in quantitative research, and multiple imputation (MI) has become a widely adopted solution. However, traditional MI methods can struggle with complex data patterns, potentially leading to biased estimates. To address this, algorithms like random forests, classification and regression trees, and tree-based boosting techniques have been integrated into common imputation tools, improving the preservation of complex data patterns. Despite these advancements, tree-based methods typically require large sample sizes, limiting their efficiency in smaller datasets. Classical regression techniques, including generalized additive models for location, scale, and shape (GAMLSS), have also been applied to MI. While GAMLSS effectively model non-linearities, they require manual specification of interactions and become computationally intensive with many predictors. This paper explores the use of multivariate adaptive regression splines (MARS) within a fully conditional specification MI framework. MARS offers a balance between flexibility and computational speed, efficiently handling continuous variables without requiring the large datasets needed by tree-based models. A MARS-based MI algorithm was implemented and evaluated through a comprehensive simulation study, demonstrating that MARS provides promising results, with less bias and better coverage, particularly in scenarios involving complex data patterns.

**Keywords:**

multiple imputation; multivariate adaptive regression splines; simulation study; continuous variables; statistical efficiency

# 1. Introduction

Quantitative research relies on the availability of high-quality and complete data. Yet, in practice, datasets are often characterized by incomplete cases, leading to possibly biased estimates and reduced statistical power. Multiple imputation (MI) has become a widely used solution for addressing this issue. MI involves replacing an incomplete dataset with multiple imputed datasets and analyzing each separately, with the final estimates being pooled using Rubin’s rules (Rubin 1987). However, MI algorithms that fail to capture complex data patterns can also lead to biases (Seaman, Bartlett, and White 2012; Curnow et al. 2023). This issue is particularly pronounced for multivariate missingness and when using fully conditional specification (FCS) MI algorithms, such as those implemented in the popular mice (“Multivariate Imputation by Chained Equations”) R package (van Buuren and Groothuis-Oudshoorn 2011). FCS MI operates by using a series of univariate imputation models in a cyclic fashion, each of which requires careful specification of non-linear relationships and interactions, which can be impractical. Therefore, developing procedures that can automatically address these complexities is of significant interest.

With advancements in computational power, sophisticated algorithms have become essential for handling complex data structures with the aim to improve missing data imputation. Techniques like random forest (RF) (Breiman 2001) and classification and regression trees (CART) (Breiman et al. 2017) have been integrated into popular imputation packages like mice. These methods have demonstrated better preservation of complex data patterns compared to traditional approaches (Burgette and Reiter 2010; Shah et al. 2014; Doove, Van Buuren, and Dusseldorp 2014). More recently, Yongshi Deng and Lumley (2023) implemented an MI algorithm using the XGBoost technique (Chen and Guestrin 2016) within the mixgb R package, which outperformed RF and CART-based imputations. Briefly, XGBoost is a fast, tree-based boosting algorithm that builds a strong predictive model by combining multiple weak learners in a sequential manner.

Despite their benefits, models like CART and RF often require larger sample sizes to reliably learn the underlying data structure due to their flexibility in modeling interactions and categorizing continuous variables (Riley et al. 2020). While large datasets are sometimes available, such as in observational studies, randomized controlled trials (RCTs) or other quasi-experimental designs often face sample size limitations due to high costs of data collection. Nevertheless, addressing missing data remains crucial to maintaining the credibility of trial conclusions, as emphasized by Little et al. (2012). Studies such as those by Spekreijse et al. (2023) and Strandell et al. (2024) provide examples where MI with predictive mean matching (PMM) was applied to handle missing data in primary analyses. Classical regression-based methods, such as generalized additive models (GAM) (Wood 2017), offer a solution to capture potential non-linearities more effectively. GAM extensions, such as generalized additive models for location, scale, and shape (GAMLSS) (Stasinopoulos and Rigby 2007), have been explored for MI and are available via the mice add-on package imputerobust (Salfran and Spieß 2018). However, these methods can be computationally intensive when many predictors are involved.

An alternative approach proposed in this paper is multivariate adaptive regression splines (MARS), introduced by Friedman (1991). MARS balances computational efficiency and flexibility, modeling non-linearities and interactions without requiring large datasets as in tree-based models, which may suffer from a lack of smoothness (Hastie, Tibshirani, and Friedman 2017). Although the use of MARS for imputation tasks is not entirely new (Sánchez-Borrego, Mar Rueda, and Munoz 2011; Sánchez Lasheras et al. 2020), its application in a FCS MI framework remains unexplored.

This paper aims to implement an FCS MI algorithm based on MARS and evaluate its performance against existing models within the mice framework. [Section 2](#sec-mars) briefly introduces the MARS algorithm, and [Section 3](#sec-simulation) describes the simulation study, including the experimental settings, comparative models, and performance metrics. [Section 4](#sec-results) presents the results of the simulation, followed by a discussion of the findings in [Section 5](#sec-discussion). The paper concludes with a summary and future directions.

# 2. Brief introduction to MARS

To introduce the MARS algorithm, we draw primarily from the explanation provided in Hastie, Tibshirani, and Friedman (2017) (pp. 321-326), which offers a more extensive summary of the method and additional theoretical background.

MARS is a flexible, non-parametric regression technique introduced by Friedman (1991). It models the relationship between input variables and a response variable using linear combinations of basis functions. These basis functions, also known as hinge functions, are piecewise linear splines that adapt to the data by fitting different linear segments in regions of the input space where needed.

MARS uses hinge functions of the form:

where each observed value of serves as a knot , generating a set of candidate basis functions:

This set can become large, with up to basis functions when all values of are distinct. Therefore, MARS employs a stepwise model-building process in which serves as the pool of candidate predictors. The MARS model is then expressed as:

where represents a subset of , and the coefficients and are estimated through standard optimization methods, such as least squares estimation.

The fitting procedure begins with a **forward pass**, where basis functions are added iteratively until a stopping criterion is met (such as a maximum number of coefficients or an increase in ). Since this process can lead to overfitting, a **backward pass** is used afterward to prune the model by removing redundant terms.

MARS incorporates interaction terms hierarchically, meaning higher-order interactions are only introduced if the corresponding lower-order terms are already present in the model. The maximum degree of interaction is user-specified, which allows for control over the complexity of the interactions without having to manually specify which variables interact.

In essence, MARS combines elements of both CART and GAM. Compared to CART, MARS is more efficient at modeling continuous variables because hinge functions can adapt more smoothly than CART’s binary partitions. However, MARS is not as smooth as GAM models. Both CART and GAM handle interaction terms, with CART doing so recursively and GAM via tensor products that require manual specification. MARS simplifies this by only requiring the user to define the maximum interaction degree, not the specific variable interactions.

While GAM models offer superior smoothness and flexibility, they come with a higher computational cost due to the large number of parameters that must be optimized. MARS, on the other hand, strikes a practical balance between modeling flexibility and computational efficiency, making it a particularly suitable option for tasks like MI and for smaller datasets. [Figure 2.1](#fig-compMARS) illustrates the differences between CART, GAM, and MARS in a simple bivariate setting, simulating a continuous sine relationship with normal error.

|  |
| --- |
| Figure 2.1: Illustrative comparison of CART, MARS, and GAM predictions in a simple parametric setting. CART exhibits a stepwise prediction pattern, while the GAM provides a fully smooth prediction. MARS predictions offer a balance between the two. |

The R package mice allows for user-specified imputation functions in a straightforward manner, making it easy to implement and test alternative MI methods. We implemented the MARS method for MI in a function called mice.impute.mars (code available in the supplementary material) using the R package earth (“Enhanced Adaptive Regression Through Hinges”) (Milborrow 2024) to fit the MARS method.

More formally, focusing on univariate missingness, we defined as the partially missing outcome variable, with corresponding predictors that were fully observed. The missingness indicator vector was denoted by , where indicated that was missing and indicated that was observed. Thus, and represent the observed subset of and (i.e., where ), while and represent the missing subset (i.e., where ).

Data can be missing completely at random (MCAR) if the probability of missingness is independent of both observed and unobserved data, missing at random (MAR) if it depends only on observed data, or missing not at random (MNAR) if it depends on unobserved information such as the partially missing variable itself. Our implementation of the MARS method was designed to handle missing data under the MAR assumption, implying that it could also operate under the MCAR assumption.

To ensure the imputations were statistically proper as defined by Rubin (1987) (p. 118), we accounted for the uncertainty arising from missing data. We achieved this by fitting the MARS model on a bootstrapped sample , with the same number of observations () as the observed data . The predictions from the fitted MARS model for the missing observations, , are then augmented by an error term, which is drawn from a standard normal distribution and scaled by the mean squared residuals from the MARS fit:

Hence, the imputed values are given by:

where is a vector of size containing random draws from the standard normal distribution.

By default, the MARS method in this implementation is specified to include interactions up to a maximum degree of 10, as allowed by the current implementation of the earth function (i.e., degree=10). However, users can adjust this setting to restrict the model to lower-degree interactions or entirely additive terms.

# 3. Simulation study

## 3.1 Data and missingness generation

The data generation process of the simulation study was closely based on the second simulation study conducted by Little and An (2004), with the exception that we included situations with a large number of candidate predictors (high compared to ). We focused on a single partially missing variable and examined the simple estimator of the mean of , hence the estimand of interest was . For each simulation run, the data generation process was newly executed.

To make the imputation task more challenging, we introduced noise variables generated from a multivariate standard normal distribution that had no impact on the data generation process. Specifically, independent variables were generated by drawing 100 observations from a multivariate standard normal distribution where was one of the experimental factors in our simulation study taking on the values 10 or 100.

To align more closely with Little and An (2004), the first two variables and , were transformed into realizations from a uniform distribution on the range using the transformation for where is the cumulative distribution function of the standard normal distribution.

The partially missing variable was then generated from the two fully observed variables and under the following conditions:

1. constant:
2. linear:
3. additive:
4. non-additive:

which means that the expected value for all mean structures was 10 (). The missingness in was generated under the following probability functions , which included both MAR and MCAR mechanisms:

1. constant: (MCAR)
2. linear: with
3. additive:
4. non-additive:

where realizations indicated as missing, and indicated as non-missing. Hence, the expected missingness probability was 0.5 across all conditions.

In summary, the simulation study design included 2 different amounts of noise variables, 4 data generation strategies, and 4 missingness generation strategies, resulting in a full factorial design with 32 experimental conditions.

## 3.2 Multiple imputation methods

The dataset provided to the MI algorithms included the fully observed variables and the partially missing variable . To minimize unnecessary variation, the dataset was kept identical across all imputation methods within a single simulation run.

The MI framework used throughout this study was based on the mice package (van Buuren and Groothuis-Oudshoorn 2011). Since we dealt with univariate missingness, the number of iterations (maxit) in mice was set to 1. The number of imputed datasets (argument m in mice) was set to cart, cc, full, lasso.select.norm, mars, mixgb, pmm, rf. Additionally, setting the eps argument in mice to 0, we prevented preliminary filtering of collinear variables.

The MI methods used within mice include:

* **MARS** (Multivariate Adaptive Regression Splines)
* **CART** (Classification and Regression Trees)
* **RF** (Random Forest)

In addition to these methods, we also applied more classical methods:

* **Linear Regression with LASSO-selected variables (lasso.select.norm)**, as described in Zhao and Long (2016) and Yi Deng et al. (2016) (indirect use of regularized regression).
* **Predictive Mean Matching (PMM)**.

Furthermore, we applied the XGBoost algorithm for MI with subsampling as implemented in the mixgb package (Yongshi Deng and Lumley 2023).

## 3.3 Performance evaluation

Let denote the pooled estimate for after multiple imputations for any simulation repetition , where was set to 500. To evaluate the performance of each MI method, we used the following summary statistics:

* **Mean and Standard Deviation of Bias**: The bias was calculated as , where represents the full-data estimate (i.e., before the introduction of missingness). We reported the mean and standard deviation of the biases for each method and experimental condition.
* **Mean and Standard Deviation of the Standard Error (SE)**: The standard error of , denoted by , was computed. We reported the mean and standard deviation of the standard errors for each method and experimental condition.
* **Coverage of the 95% Confidence Interval**: Coverage was assessed by checking whether the full-data estimate, (treated as fixed), fell within the 95% confidence interval of . This was evaluated using the criterion: , where is the 0.975 quantile of the standard normal distribution. The overall coverage for each method and experimental condition was quantified as the percentage of simulations where this criterion was satisfied. A 95% Wilson confidence interval was computed to quantify the uncertainty in the observed coverage rate.

In addition, we recorded the computation time required for each imputation method.

## 3.4 Secondary simulation: Increasing sample sizes

Preliminary results indicated that the most challenging experimental condition involved the combination of a non-additive mean structure, a non-additive missingness mechanism, and 100 noise explanatory variables. To explore whether increasing the sample size could mitigate these difficulties, we expanded the initial sample size of 100 to include larger sizes of 200, 400 and 800 for this experimental condition. The MI methods and performance evaluations remained consistent with those used in the primary simulation study.

# 4. Results

## 4.1 Performance of MI methods in the primary design

We conducted a simulation study following closely the methodology outlined by Little and An (2004) to compare the performance of the MARS method for MI against other widely-used MI methods. These methods were chosen for their flexibility and ability to handle complex data patterns. [Table 4.1](#tbl-qualitative_assessment) presents a preliminary qualitative assessment and summary of the performance of the different MI methods compared here.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4.1: **Qualitative overall assessment across MI methods:** (++) very good performance, (+) good performance, (-) satisfactory performance, (–) least satisfactory performance. Overall, MARS achieved the lowest bias and best coverage, though it had relatively large standard errors and moderate computation time. The XGBoost-based MI method (MixGB) and CART showed moderate bias, but CART was considerably faster. RF and PMM displayed the least satisfactory performance overall. The linear regression-based MI method with LASSO-selected variables (lasso.select.norm) performed well in terms of bias and coverage but had relatively long computation times.   | **Method** | **Bias** | **Standard error** | **Coverage** | **Time** | | --- | --- | --- | --- | --- | | mars | ++ | - | ++ | -- | | mixgb | - | + | + | -- | | cart | - | + | + | ++ | | rf | -- | - | - | ++ | | pmm | -- | - | + | ++ | | lasso.select.norm | + | - | + | - | |

To assess the bias of each imputation method, we compared the coefficient estimates from different MI implementations against the full-data estimate, which was expected to be 10. As a baseline, we also computed estimates from the complete cases (cc), where missing observations were simply omitted. All imputation methods produced complete imputed datasets, with no persisting missing values.

[Figure 4.1](#fig-bias) provides a visual comparison of these biases. As expected, the complete case estimates generally exhibited the largest biases across most experimental conditions when compared to those obtained from MI methods. The MI methods, including MARS, consistently produced estimates closer to the true parameter value, indicating their effectiveness in reducing bias that arise from missing data.

|  |
| --- |
| Figure 4.1: Bias in coefficient estimates across MI methods: This figure shows the bias of coefficient estimates across different MI methods over 500 simulation runs under various experimental conditions. Bias was calculated as the difference between the pooled estimated coefficient and the full-data estimate. The figure also includes comparisons to biases from complete cases (cc; where missing data were omitted). As expected, complete case estimates generally display the largest biases, indicating a significant deviation from the true value except in the MCAR scenarios, while most MI methods, including MARS, demonstrated lower biases. |

[Figure 4.2](#fig-se) illustrates the standard errors associated with each imputation method. As expected, most MI methods showed larger standard errors compared to the full-data standard error. This increase is typical, as proper imputation incorporates both within- and between-imputation variance, leading to higher uncertainty.

The MARS-based MI algorithm exhibited relatively high standard errors in certain experimental conditions, suggesting a bias-variance trade-off in the selection of MI methods. This observation calls for further investigation of the conditions under which MARS excels or underperforms relative to other MI methods.

|  |
| --- |
| Figure 4.2: Standard error across MI method: This figure shows the standard errors of coefficient estimates across different MI methods over 500 simulation runs under various experimental conditions. Compared to the full-data standard error, most MI methods exhibited larger standard errors due to the incorporation of both within- and between-imputation variance. The MARS-based MI algorithm, in particular, demonstrated relatively high standard errors in certain experimental conditions, reflecting a higher degree of variability in its estimates. |

We evaluated the coverage of the 95% confidence intervals for each imputation method, as shown in [Figure 4.3](#fig-coverage). Ideally, the coverage should be close to 95%. However, in many cases, the coverage for some imputation methods fell below this threshold, indicating potential issues with underestimation of uncertainty of imputing missing data. In contrast, the MARS-based MI algorithm did not underestimate uncertainty and consistently achieved coverage rates closer to the desired 95% target.

|  |
| --- |
| Figure 4.3: Coverage of 95% confidence intervals across MI method: This figure displays the coverage rates of 95% confidence intervals for coefficient estimates across different MI methods over 500 simulation runs under various experimental conditions. The plot illustrates how well the coverage of these intervals meets the 95% target. In many experimental conditions, some imputation methods fell short of this target, indicating potential underestimation of uncertainty. The MARS-based MI algorithm, however, generally achieved coverage rates closer to the 95% target. |

## 4.2 Results of secondary simulation: Increasing sample size

In the secondary simulation study, we investigated whether increasing sample size could improve MI performance in the most challenging scenario identified in preliminary simulations. This scenario involved a non-additive mean structure, a non-additive missingness mechanism, and 100 noise explanatory variables.

As illustrated in [Figure 4.4](#fig-p_secondary_combined), increasing the sample size did not considerably improve the performance of more traditional approaches (lasso.select.norm and pmm). However, methods with the potential to adapt to complex data patterns (rf, cart, mixgb, and mars) showed marked improvements, including reduced bias, lower standard errors, and higher coverage. The degree of improvement varied across methods, with the MARS-based MI algorithm demonstrating the fastest convergence to nearly unbiased estimates, while consistently maintaining coverage around 95%.

|  |
| --- |
| Figure 4.4: Impact of the sample size: This figure displays the impact of increasing sample size on the bias, standard error, and coverage rates of 95% confidence intervals for coefficient estimates across different MI methods over 500 simulation runs under the most challenging experimental conditions. The figure illustrates that increasing sample size was only beneficial for MI methods with sufficient flexibility to adapt to complex data patterns (such as mars, rf, cart, and mixgb), resulting in reduced bias, lower standard error and improved coverage. Notably, the MARS-based MI algorithm achieved the fastest convergance to satisfactory performance. |

## 4.3 Results of imputation time

Finally, we recorded the time required for each imputation method to complete, as shown in [Figure 4.5](#fig-time). The imputations were performed on a desktop computer equipped with an Intel Core i7-14700 CPU (max. 5.4 GHz), 32 GB RAM, and running Ubuntu 24.04.1 LTS (64-bit). The analyses were conducted in R version 4.4.2, with parallel processing across 28 cores for all computations. The entire simulation study took approximately three hours to complete. Overall, the time taken for imputations increased with the number of explanatory variables. While most of the MI methods performed similarly in terms of imputation time, the MARS- and XGBoost-based MI methods had relatively long imputation times, which also increased with sample size. The LASSO-based MI method was relatively slow on average, but its performance remained stable with respect to sample size.

|  |
| --- |
| Figure 4.5: Time required for imputation methods: This boxplot displays the distribution of individual imputation times for each method across 500 simulation runs, showing the time taken to construct cart, cc, full, lasso.select.norm, mars, mixgb, pmm, rf imputed datasets for univariate missingness. |

# 5. Discussion

In this study, we developed and evaluated a FCS MI algorithm based on the MARS method and compared its performance against other established MI methods through a simulation study. Our findings suggest that the MARS-based MI algorithm performed well, particularly in terms of bias, and is the only method that consistently achieves the targeted coverage level. This is a noteworthy result, as coverage accuracy is a critical criterion for evaluating the validity of MI methods, ensuring that confidence intervals are appropriately calibrated to reflect true parameter uncertainty (Rubin 1987).

One aspect that likely contributed to the strong coverage performance of the MARS-based MI algorithm is its relatively higher standard errors compared to other MI methods. Achieving a balance between accuracy and efficiency remains a key challenge in the development of imputation algorithms, and future research should investigate the bias-variance trade-off in more detail to explore ways to improve the efficiency of the MARS-based MI algorithm without sacrificing its coverage properties.

Despite the promising results for the MARS-based MI algorithm, our study also has limitations. First, we focus exclusively on univariate missingness in a continuous outcome, which may not fully reflect the complexity of real-world data where missingness can occur across multiple variables, including both continuous and categorical types. Second, our simulation study is restricted to continuous explanatory variables. While continuous variables are prominent in many fields, categorical variables also play an important role and may pose distinct challenges for imputation, where other MI methods could perform better. Future research should evaluate the robustness of the MARS-based MI algorithm in more varied settings, including multivariate missingness and datasets with mixed variable types.

Furthermore, our study does not address situations where the missingness mechanism is not MAR, such as when missingness depends on the outcome variable itself. This presents a significant area for further exploration, as MI methods often assume MAR, and deviations from this assumption can lead to biased results. Investigating the application and performance of the MARS-based MI method under MNAR conditions could provide valuable insights into its broader applicability.

In conclusion, our study demonstrates that the MARS-based MI algorithm is a promising addition to the existing toolkit of MI methods, particularly in scenarios where achieving accurate coverage is paramount. However, further research is needed to enhance its efficiency and evaluate its performance in more complex and varied data settings.

## Acknowledgments

I would like to thank Stefan Boes for valuable discussions.

## Funding

This research received no specific funding.

## Disclosure statement

The author declares no conflicts of interest.

## Data availability statement

All material to reproduce the simulation study is available on Github: <https://github.com/jsepin/mi-using-mars>

# References

Breiman, Leo. 2001. *Machine Learning* 45 (1): 5–32. <https://doi.org/10.1023/a:1010933404324>.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 2017. *Classification and Regression Trees*. Routledge. <https://doi.org/10.1201/9781315139470>.

Burgette, L. F., and J. P. Reiter. 2010. “Multiple Imputation for Missing Data via Sequential Regression Trees.” *American Journal of Epidemiology* 172 (9): 1070–76. <https://doi.org/10.1093/aje/kwq260>.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. ACM. <https://doi.org/10.1145/2939672.2939785>.

Curnow, Elinor, James R. Carpenter, Jon E. Heron, Rosie P. Cornish, Stefan Rach, Vanessa Didelez, Malte Langeheine, and Kate Tilling. 2023. “Multiple Imputation of Missing Data Under Missing at Random: Compatible Imputation Models Are Not Sufficient to Avoid Bias If They Are Mis-Specified.” *Journal of Clinical Epidemiology* 160 (August): 100–109. <https://doi.org/10.1016/j.jclinepi.2023.06.011>.

Deng, Yi, Changgee Chang, Moges Seyoum Ido, and Qi Long. 2016. “Multiple Imputation for General Missing Data Patterns in the Presence of High-Dimensional Data.” *Scientific Reports* 6 (1). <https://doi.org/10.1038/srep21689>.

Deng, Yongshi, and Thomas Lumley. 2023. “Multiple Imputation Through XGBoost.” *Journal of Computational and Graphical Statistics* 33 (2): 352–63. <https://doi.org/10.1080/10618600.2023.2252501>.

Doove, L. L., S. Van Buuren, and E. Dusseldorp. 2014. “Recursive Partitioning for Missing Data Imputation in the Presence of Interaction Effects.” *Computational Statistics &Amp; Data Analysis* 72 (April): 92–104. <https://doi.org/10.1016/j.csda.2013.10.025>.

Friedman, Jerome H. 1991. “Multivariate Adaptive Regression Splines.” *The Annals of Statistics* 19 (1): 1–67.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2017. “The Elements of Statistical Learning: Data Mining, Inference, and Prediction.” Springer.

Little, Roderick J., and Hyonggin An. 2004. “Robust Likelihood-Based Analysis of Multivariate Data with Missing Values.” *Statistica Sinica*, 949–68.

Little, Roderick J., Ralph D’Agostino, Michael L. Cohen, Kay Dickersin, Scott S. Emerson, John T. Farrar, Constantine Frangakis, et al. 2012. “The Prevention and Treatment of Missing Data in Clinical Trials.” *New England Journal of Medicine* 367 (14): 1355–60. <https://doi.org/10.1056/nejmsr1203730>.

Milborrow, Stephen. 2024. *Earth: Multivariate Adaptive Regression Splines*. <https://CRAN.R-project.org/package=earth>.

Riley, Richard D, Joie Ensor, Kym I E Snell, Frank E Harrell, Glen P Martin, Johannes B Reitsma, Karel G M Moons, Gary Collins, and Maarten van Smeden. 2020. “Calculating the Sample Size Required for Developing a Clinical Prediction Model.” *BMJ*, March, m441. <https://doi.org/10.1136/bmj.m441>.

Rubin, Donald B. 1987. “Multiple Imputation for Nonresponse in Surveys.” *Wiley Series in Probability and Statistics*, June. <https://doi.org/10.1002/9780470316696>.

Salfran, Daniel, and Martin Spieß. 2018. *ImputeRobust: Robust Multiple Imputation with Generalized Additive Models for Location Scale and Shape*. <https://CRAN.R-project.org/package=ImputeRobust>.

Sánchez Lasheras, Fernando, Paulino José García Nieto, Esperanza García-Gonzalo, Francisco Argüeso Gómez, Francisco Javier Rodríguez Iglesias, Ana Suárez Sánchez, Jesús Daniel Santos Rodríguez, et al. 2020. “Missing Data Imputation for Continuous Variables Based on Multivariate Adaptive Regression Splines.” In, 73–85. Springer International Publishing. <https://doi.org/10.1007/978-3-030-61705-9_7>.

Sánchez-Borrego, Ismael, Marı́a del Mar Rueda, and Juan F Munoz. 2011. “Imputation and Inference with Multivariate Adaptive Regression Splines.” In *Modern Mathematical Tools and Techniques in Capturing Complexity*, 341–53. Springer.

Seaman, Shaun R, Jonathan W Bartlett, and Ian R White. 2012. “Multiple Imputation of Missing Covariates with Non-Linear Effects and Interactions: An Evaluation of Statistical Methods.” *BMC Medical Research Methodology* 12 (1). <https://doi.org/10.1186/1471-2288-12-46>.

Shah, Anoop D., Jonathan W. Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. 2014. “Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study.” *American Journal of Epidemiology* 179 (6): 764–74. <https://doi.org/10.1093/aje/kwt312>.

Spekreijse, Lindsay, Rob Simons, Bjorn Winkens, Frank van den Biggelaar, Carmen Dirksen, Marjolijn Bartels, Ronald de Crom, et al. 2023. “Safety, Effectiveness, and Cost-Effectiveness of Immediate Versus Delayed Sequential Bilateral Cataract Surgery in the Netherlands (BICAT-NL Study): A Multicentre, Non-Inferiority, Randomised Controlled Trial.” *The Lancet* 401 (10392): 1951–62. <https://doi.org/10.1016/s0140-6736(23)00525-1>.

Stasinopoulos, D. Mikis, and Robert A. Rigby. 2007. “Generalized Additive Models for Location Scale and Shape (GAMLSS) in*R*.” *Journal of Statistical Software* 23 (7). <https://doi.org/10.18637/jss.v023.i07>.

Strandell, Annika, Leonidas Magarakis, Karin Sundfeldt, Mathias Pålsson, Per Liv, and Annika Idahl. 2024. “Salpingectomy Versus Tubal Occlusion in Laparoscopic Sterilisation (SALSTER): A National Register-Based Randomised Non-Inferiority Trial.” *The Lancet Regional Health - Europe*, August, 101026. <https://doi.org/10.1016/j.lanepe.2024.101026>.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. “mice: Multivariate Imputation by Chained Equations in r.” *Journal of Statistical Software* 45 (3): 1–67. <https://doi.org/10.18637/jss.v045.i03>.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with r*. Chapman; Hall/CRC. <https://doi.org/10.1201/9781315370279>.

Zhao, Yize, and Qi Long. 2016. “Multiple Imputation in the Presence of High-Dimensional Data.” *Statistical Methods in Medical Research* 25 (5): 2021–35. <https://doi.org/10.1177/0962280213511027>.