MSR172

A COMPARATIVE ANALYSIS OF MISSING VALUE IMPUTATION TECHNIQUES: SPLINE VS. MARKOV CHAIN MONTE CARLO & MACHINE LEARNING ALGORITHMS

Paul Choudhury S[1], Dutta Majumdar A[1] , Sil A[1] , Dutta SK[1], Mahon R[2,3]

[1]PharmaQuant Insights Pvt. Ltd., Kolkata, WB, India, [2]PharmaQuant Insights Pvt. Ltd., Dublin, Ireland, [3]University of Galway, Galway, Ireland

INTRODUCTION

- Missing data on costs, health related outcomes and confounding variables may introduce bias in economic evaluation and real-world evidence (RWE) generation, which may misguide to take wrong policy decision.1,2,3]
- Machine learning-based imputation methods typically use modelling to extract valuable information from incomplete data, enabling reasonable inference of missing values. The general approach of the machine learning algorithms applied in this study involves using complete samples within the incomplete dataset as the training set to build a predictive model, which then estimates the missing values based on the trained model [4]. While machine learning (ML) algorithms offer solutions, they often require large, high-quality datasets, which may not always be available.
- Markov Chain Monte Carlo (MCMC) imputation also exists to handle missing data. [5,6] ML techniques predict missing values based on patterns in the observed data to achieve optimal accuracy, while MCMC imputation generates multiple plausible datasets by sampling from the posterior distributions, capturing uncertainty in the imputed values. Even though MCMC is a widely used and robust method, it can be computationally intensive with large datasets, and convergence issues may arise due to poor mixing and high autocorrelation between samples. MCMC methods may struggle with high-dimensional data due to the curse of dimensionality. As the number of parameters in the model increases, it becomes harder to explore the posterior distribution effectively.[7,8]
- We proposed a novel imputation approach using spline models, known for their flexibility in capturing complex, non-linear relationships, and compare their performance with established methods.

OBJECTIVES

- To evaluate the performance of standard MCMC and various machine learning algorithms in comparison to a proposed spline-based imputation method.

METHODS

Data preparation:

- Two publicly available datasets were selected for illustrative purposes.
- In the first example, data on the ages of 228 patients were extracted from the North Central Cancer Treatment Group (NCCTG) database (NLCD) [9].
- For the second illustration, data on the duration from diagnosis to randomization (DR), measured in months, was selected for 137 patients from the Veterans' Administration Lung Cancer Study (VACD) [10]. These datasets were chosen to demonstrate the application of imputation techniques across different patient characteristics.
- We created missing value datasets (MVD) by randomly ***removing 30% of the observations*** from the primary datasets. Little's (1988) test demonstrated that the data is missing completely at random (MCAR).

ML imputation procedure:

- The remaining 70% of the data was used to train various imputation models, including random forest (RF), decision tree (DT), support vector machine (SVM), gradient boosting model (GBM), and linear regression (LR).
- For the NLCD dataset, the independent variables included sex, performance score, Karnofsky performance score, patient-rated Karnofsky performance score (where 100 indicates a good score), calories consumed at meals, and weight loss over the last six months (in pounds). In the VACD dataset, the independent variables comprised the Karnofsky performance score (where 100 indicates a good score), prior therapy indicator, and age. These variables were selected to enhance the predictive power of the imputation models by leveraging relevant patient characteristics.

MCMC imputation procedure:

- The MCMC imputation method was used on MVD to create five different datasets by randomly selecting five different seeds.
- We replaced each missing value with the mean of one hundred samples drawn from the posterior distribution.

Spline imputation procedure:

Both natural spline models (NSM) and regression spline models (RSM) were methods used on the remaining 70% of the data to model relationships between independent and dependent variables with greater flexibility. Here's a breakdown of each and why NSM was chosen with hyperparameter tuning:

1. Regression Spline Models (RSM)
   - RSM divides data into segments and fits a polynomial function to each segment, connecting these segments at specific points called "knots." This approach helps capture non-linear relationships in the data.
   - However, RSM often struggles at the "ends" or "extremes" of the predictor values, where it tends to exhibit high variance (i.e., the model predictions fluctuate widely). This issue can result in wide confidence intervals and less reliable predictions, particularly if the sample size is small. [11,12]
2. Natural Spline Models (NSM)
   - NSM is a type of regression spline that adds boundary constraints, which forces the model to be linear near the edges of the predictor variable range. This adjustment reduces the high variance seen in RSM at the extremes, making NSM more stable and less prone to overfitting at the edges. [12]

Hyperparameter tuning was applied to NSM to adjust the shape and flexibility of the spline further. By optimizing certain parameters (e.g., the number and position of knots), the NSM can provide a better fit to the training data without unnecessary fluctuations, improving its ability to predict missing values accurately. [11]

Comparison between techniques:

Root-mean-square error (RMSE) was estimated to compare the accuracy of the techniques.

RESULT

NLCD Dataset Observations:

- **NSM (RMSE: 8.1)**: Achieves the lowest RMSE, signifying strong predictive capability.
- **RSM (RMSE: 8.6)**: Slightly higher than NSM but still performs relatively well.

Figure 1: *RMSE values for NLCD data*

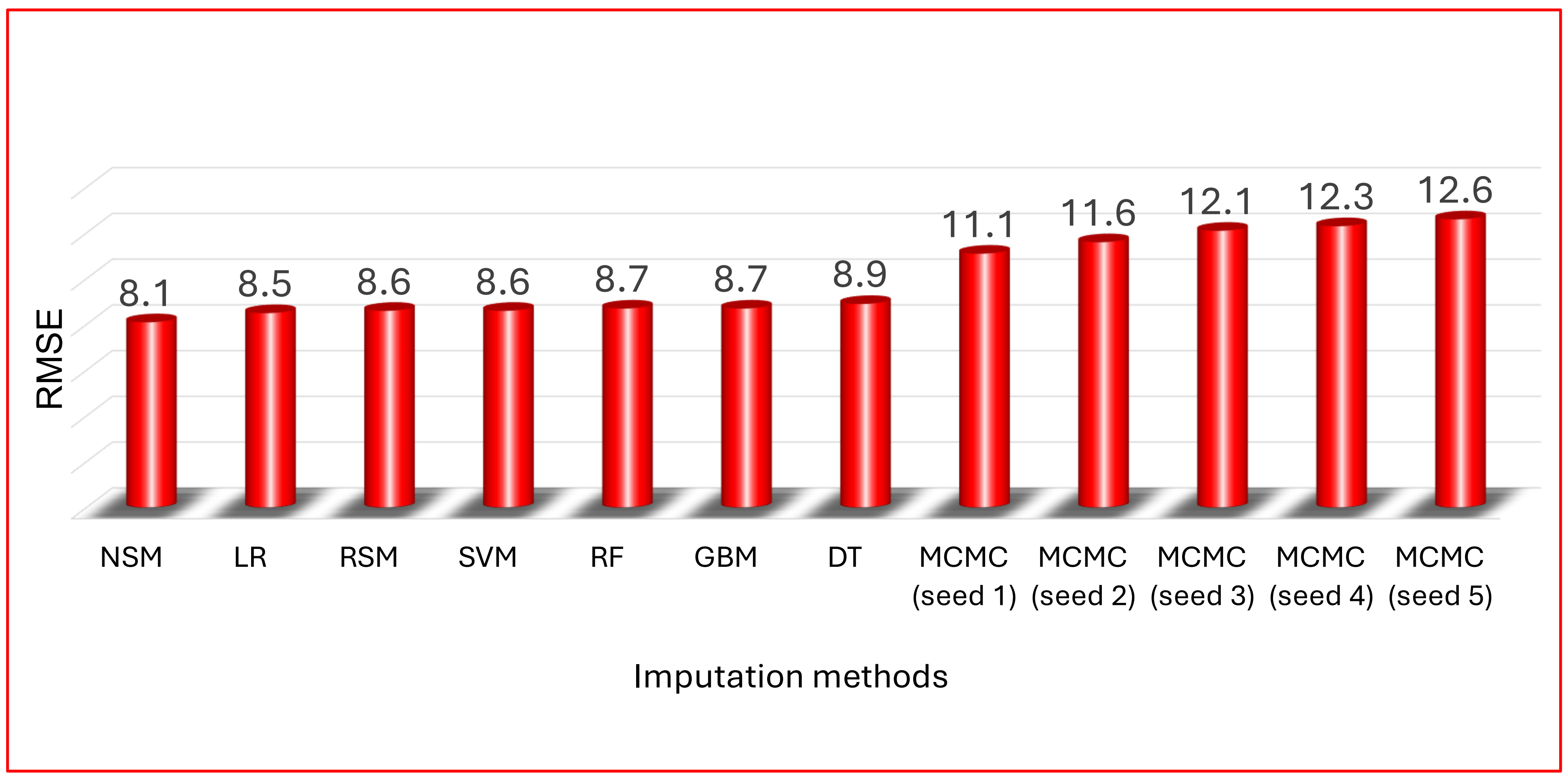
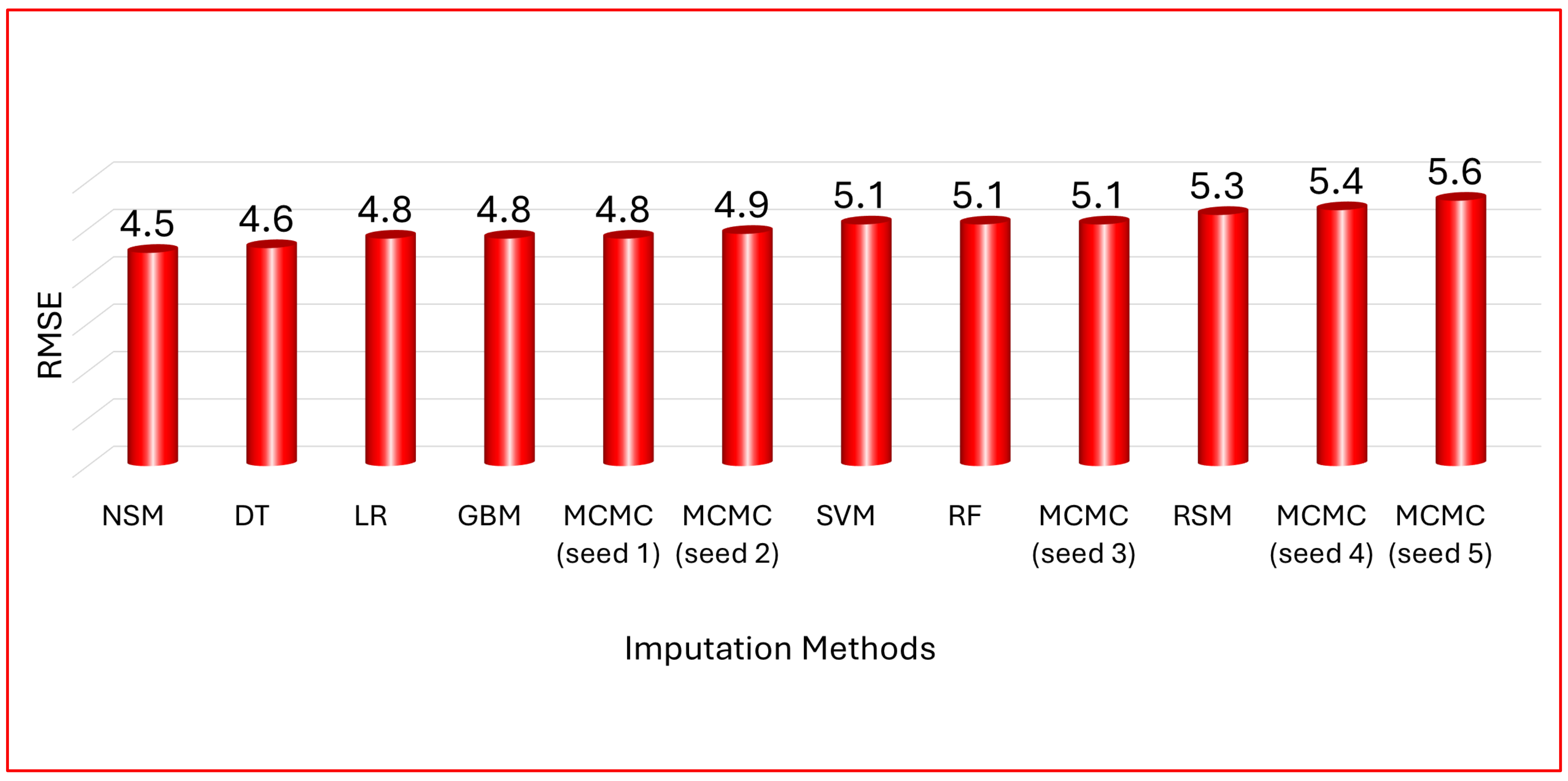


Figure 2: *RMSE values for VACD data*



- **MCMC (RMSE: 11.1 to 12.6)**: Exhibits significantly higher RMSE values, suggesting it is less effective in this context. The wide range indicates variability in results depending on the random seed used.
- **Other Models (RF, DT, LR, SVM, GBM)**: All fall within a narrow range (8.5 to 8.9), showing similar predictive performance but still inferior to NSM. [**Figure 1**]

VACD Dataset Observations:

- **NSM (RMSE: 4.5)**: Demonstrates superior accuracy with the lowest RMSE.
- **RSM (RMSE: 5.3)**: Higher than NSM, indicating less precision in predictions.
- **MCMC (RMSE: 4.8 to 5.6)**: Performs reasonably well but not as effectively as NSM, with results also varying based on seed selection. . [**Figure 2**]
- **Other Models (RF, DT, LR, SVM, GBM)**: Show RMSE values ranging from 4.6 to 5.1, suggesting they are more comparable in this dataset than in NLCD, yet all are less effective than NSM.

NSM's consistent lowest RMSE across both datasets underscores its effectiveness as a missing value imputation approach. The MCMC model exhibits notable variability in its RMSE values, indicating its sensitivity to random seed choices. This variability could affect the reliability of its predictions as well as affecting the quality of missing value imputation.

CONCLUSIONS

In conclusion, the results indicate that spline models outperform other methods in terms of RMSE, highlighting their potential as a viable alternative for data modelling. Their ability to effectively capture complex, nonlinear relationships without overfitting makes them a strong choice for various applications. This analysis supports the adoption of spline models for improved predictive accuracy in practical settings.

REFERENCES

1. Li J, Yan XS, Chaudhary D, Avula V, Mudiganti S, Husby H, et al. Imputation of missing values for electronic health record laboratory data.npj Digital Medicine. 2021;4(1):147
2. 2. Liu M, Li S, Yuan H, Ong MEH, Ning Y, Xie F, et al. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. Artificial Intelligence in Medicine. 2023;142:102587..
3. Mukherjee K, Gunsoy NB, Kristy RM, Cappelleri JC, Roydhouse J, Stephenson JJ, Vanness DJ, Ramachandran S, Onwudiwe NC, Pentakota SR, Karcher H, Di Tanna GL. Handling Missing Data in Health Economics and Outcomes Research (HEOR): A Systematic Review and Practical Recommendations. Pharmacoeconomics. 2023 Dec;41(12):1589-1601. doi: 10.1007/s40273-023-01297-0. Epub 2023 Jul 25. PMID: 37490207; PMCID: PMC10635950.
4. Wang, H., Tang, J., Wu, M. et al. Application of machine learning missing data imputation techniques in clinical decision making taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an example. BMC Med Inform Decis Mak 22, 13 (2022). https://doi.org/10.1186/s12911-022-01752-6
5. Schunk D. A Markov chain Monte Carlo algorithm for multiple imputation in large surveys. AStA Advances in Statistical Analysis. 2008;92(1):101-14
6. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. Statistics in medicine. 2007 Jul 20;26(16):3057-77.
7. Brooks, S., & Gelman, A. (1998). "General methods for monitoring convergence of iterative simulations." Journal of Computational and Graphical Statistics, 7(4), 434-455.
8. Robert CP. The Bayesian choice: from decision-theoretic foundations to computational implementation. New York: Springer; 2007 Jun.
9. Loprinzi CL. Laurie JA. Wieand HS. Krook JE. Novotny PJ. Kugler JW. Bartel J. Law M. Bateman M. Klatt NE. et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology. 12(3):601-7, 1994.
10. D Kalbfleisch and RL Prentice (1980), *The Statistical Analysis of Failure Time Data*. Wiley, New York.
11. Shelevytsky I, Shelevytska V, Semenova K, Bykov I. Regression Spline-Model in Machine Learning for Signal Prediction and Parameterization. InLecture Notes in Computational Intelligence and Decision Making: Proceedings of the XV International Scientific Conference "Intellectual Systems of Decision Making and Problems of Computational Intelligence"(ISDMCI'2019), Ukraine, May 21–25, 2019 15 2020 (pp. 158-174). Springer International Publishing.
12. Perperoglou, A., Sauerbrei, W., Abrahamowicz, M. *et al.* A review of spline function procedures in R. *BMC Med Res Methodol* **19**, 46 (2019). https://doi.org/10.1186/s12874-019-0666-3