International Journal of Business and Technology Management
e-ISSN: 2682-7646 | Vol. 7, No. 9, 297-308, 2025
https://asianscholarsnetwork.com/asnet-journals
SPECIAL ISSUE: 13th International Conference on Engineering Business Management 2025

Handling Missing Data in Administrative Records Through Imputation Methods for Data Quality Improvement

Nur Azmina Osman[1*], Suraya Yaacob[1], Nurulhuda Firdaus Mohd Azmi[1]

[1] Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

*Corresponding Author: azmina98@graduate.utm.my

Received: 15 July 2025 | Accepted: 25 August 2025 | Published: 1 December 2025

DOI: https://doi.org/10.55057/ijbtm.2025.7.9.23

Abstract: *Administrative data is one of the valuable sources used to support research, policymaking, and decision-making processes. However, missing data remains a significant challenge that can affect the completeness, quality, and reliability of data. Missing data, whether due to errors in data collection or other factors, can lead to biased results if not properly addressed. While a common approach is to discard incomplete rows or columns, this leads to loss of valuable information. Imputation provides a better alternative by estimating missing data based on observed data. Thus, this study presents an experimental evaluation of four widely used imputation methods: mean imputation, median imputation, K-Nearest Neighbors (KNN), and Multiple Imputation using Chained Equations (MICE). The methods were applied to simulated administrative data consisting of continuous variables with missing data introduced completely at random at three levels of missingness (5%, 20%, and 50%). Performance was assessed using Normalized Root Mean Squared Error (NRMSE) to quantify imputation accuracy and distribution plots to visually compare the distributions of imputed versus original data. The results reveal that increasing dataset size reduces average NRMSE. Among the methods, MICE consistently achieves the highest accuracy and best preserves data distribution across all dataset sizes and missingness levels. KNN also performs well but is outperformed by mean imputation at 50% missingness in larger datasets. Overall, advanced methods like MICE and KNN maintain the integrity of the original distribution more effectively than simpler approaches such as mean and median imputation. These findings highlight the importance of selecting robust imputation techniques for improving data quality, especially in administrative datasets. Future research could explore alternative missing data mechanisms and extend evaluations to categorical and mixed-type data.*

Keywords: Data quality, Data completeness, Missing data, Imputation, Administrative records

---

1. Introduction

In today's digital world, data is an invaluable resource, serving as the backbone for decision-making, analysis, and innovation across various sectors. One of the key sources of data is administrative data, which is primarily collected by government agencies and organizations for administrative and operational purposes (Kline, 2022). These datasets are particularly valued for their extensive population coverage, cost efficiency, and ability to support long-term, cross-sectoral analyses in fields such as health, education, and labor (Milne et al., 2022).

Despite these advantages, administrative data is often not originally intended for research or analytical purposes. As a result, various data quality challenges arise, with missing data being among the most persistent and problematic (Soldatenkova et al., 2023). Missing values may result from incomplete reporting, data entry errors, or limitations within data collection systems. If not properly addressed, missing data can lead to biased analyses, reduced accuracy, and compromised validity of research outcomes, especially in surgery outcome studies, as emphasized by (Timofte et al., 2018).

To address this, a wide range of strategies has been developed to handle missing data, from basic deletion methods to advanced imputation methods. Simple approaches such as list-wise or pair-wise deletion are straightforward but can lead to substantial data loss and analytical distortion. Meanwhile, imputation techniques estimate and fill in missing values based on observed data patterns. These include statistical methods such as mean, median, and multiple imputation, as well as machine learning methods including K-Nearest Neighbors (KNN), regression, and clustering (Zhou et al., 2024).

With the growing reliance on administrative data for evidence-based decision-making, it is important to understand how different imputation methods improve data quality and preserve the integrity of the original data for future analyses. Therefore, this study aims to evaluate several widely used imputation methods to identify the most effective approaches for enhancing data quality, particularly in administrative records. The analysis focuses on simulated administrative data consisting of continuous variables and examines the performance of selected imputation methods in addressing missing data across different conditions.

2. Literature Review

2.1 Characteristics of Missing Data

Before applying any imputation technique, it is essential to understand the characteristics of missing data, as these directly influence the selection and effectiveness of the chosen method. Missing data is a common challenge in real-world datasets, especially in administrative records, and can significantly compromise the validity and reliability of statistical analyses. To address missingness effectively, it is important to understand its key characteristics: missing rate, pattern, and mechanism (Afkanpour et al., 2024; Zhou et al., 2024).

a. Missing Rate

The missing rate refers to the proportion of absent values within a dataset; higher rates reduce usable information and can distort model performance if left unaddressed.

b. Missing Patterns

Missing data patterns can be categorized into several types. *Univariate* missingness occurs in a single variable, while *multivariate* missingness affects multiple variables. *Monotone* patterns follow a systematic order, whereas *non-monotone* patterns are more randomly distributed.

c. Missing Mechanisms

Missing data mechanisms explain the underlying causes of missingness and are typically classified into three types. Missing Completely at Random (MCAR) occurs when the probability of a value being missing is unrelated to both observed and unobserved data. Missing at Random (MAR) implies that missingness is related only to observed variables. In contrast,

Missing Not at Random (MNAR) occurs when the likelihood of missingness is directly related to the unobserved data.

2.2 Imputation Methods and Application

Imputation is a widely adopted strategy to address missing data, offering an alternative to deletion by estimating missing values using observed patterns. These techniques are broadly categorized into traditional and advanced methods. In recent years, numerous empirical studies have assessed the performance of these approaches across diverse domains, datasets, and missing data characteristics, including mechanism (MCAR, MAR, MNAR), rate, and pattern.

Several studies have demonstrated the increasing relevance of imputation in large-scale administrative and government datasets. (Suresh A & B et al., 2019), using over eight million samples from the Business Longitudinal Analysis Data Environment (BLADE), evaluated twelve machine learning algorithms and found that ensemble methods such as Extra Trees, Bagging, and Random Forest consistently outperformed traditional regression-based techniques. Similarly, (Li et al., 2024) assessed multiple imputation methods including MICE, kNN, and Random Forest in a high-dimensional cohort study dataset to enhance cardiovascular disease prediction. Their results revealed that machine learning–based methods significantly improved accuracy under moderate missingness, specifically MAR at 20 percent. (Askinadze & Conrad, 2018) examined student-controlled school data with 25 to 75 percent MCAR and highlighted the predictive superiority of hot-deck (1-NN) over mean imputation, especially under high missingness levels. These findings support the practical feasibility and effectiveness of imputation techniques in administrative records where data privacy, missingness, and complexity are common challenges.

While advanced methods show promising performance in complex administrative datasets, traditional methods such as mean, median, and mode substitution remain widely used due to their computational simplicity and ease of implementation. These methods have been explored across various domains as baseline techniques for benchmarking. For example, (Jadhav et al., 2019) found that mean imputation performs reasonably well under low missingness levels, particularly when data is MCAR. Similarly, (Mohammed et al., 2021) using simulated continuous data with missingness rates of 15%, 30%, and 45%, observed that median imputation slightly outperformed mean in preserving data distribution, yet both methods degraded in accuracy under higher missingness. (Askinadze & Conrad, 2018) highlighted this limitation in administrative education data, where mean imputation was outperformed by hot-deck methods under 25–75% MCAR. These findings underscore that while traditional methods may offer a quick fix, they are less suitable for complex datasets where missingness is non-random or exceeds minimal thresholds.

However, the limitations of simple imputation, particularly its tendency to underestimate variance and ignore data structure, have led to widespread adoption of more advanced techniques. Among these, MICE and kNN have been widely applied across diverse domains due to their ability to maintain data integrity under varying missingness conditions. Several studies have highlighted the robust performance of kNN imputation. (Jadhav et al., 2019) demonstrated its superior accuracy across five UCI datasets, with consistent performance regardless of missingness rate. In financial time-series data, (Alwateer et al., 2024) found kNN to outperform six other techniques, including SVM and random forest, across five accuracy metrics. Similarly, (Gayathri & Kavitha, 2024) reported that kNN produced the lowest error in air quality datasets, outperforming traditional single imputation methods. These findings

converge on the reliability of kNN in preserving data quality, especially under moderate to high missingness. Nonetheless, its sensitivity to parameter tuning and computational complexity warrant consideration in large-scale or high-dimensional applications.

In parallel, MICE has emerged as a robust method for addressing multivariate missingness, particularly in datasets requiring preservation of inter-variable relationships. (Grigoroff et al., 2024), using multi-centre clinical pathology data, found that while MICErf (MICE with random forest regression) performed reasonably, it was notably less accurate than missForest under stratified data and higher missingness rates. This was attributed to MICE's sensitivity to batch effects and structural heterogeneity. Similarly, (Sun et al., 2023) compared MICE with deep learning methods (GAIN, VAE) and found that MICE maintained superior accuracy and stability in datasets with fewer than 30,000 samples, especially under MCAR and MAR conditions. In applied domains, (Saini & Nagpal, 2024) reported that MICE outperformed simpler statistical methods in crop yield and energy consumption data, reinforcing its practical utility across structured, tabular datasets. These studies collectively emphasize that while MICE may be less resilient to data stratification or complex non-linearities, it remains a dependable choice for moderate-sized datasets with structured missingness and the need for valid statistical inference.

While both kNN and MICE consistently outperform traditional approaches, their performance varies by context. kNN excels in capturing local data structures, making it ideal for nonlinear patterns, whereas MICE offers a model-based framework suited for complex, multivariate datasets with missingness assumed to be at random. Comparative findings by (Mohammed et al., 2021) reinforce MICE's superiority, demonstrating that it consistently yielded the lowest root mean squared error (RMSE) across four real-world datasets with varying levels of missingness. While kNN showed moderate reliability, MICE remained robust regardless of dataset type or missing data proportion. These results underscore MICE's strength as a dependable imputation strategy for continuous data where statistical accuracy and inter-variable coherence are critical.

Beyond kNN and MICE, other advanced techniques such as random forest-based methods and deep learning models have shown promise in handling complex missing data patterns. For instance, missForest, an ensemble-based approach, has demonstrated high accuracy and robustness across various missingness mechanisms and data types, particularly in heterogeneous or stratified datasets (Grigoroff et al., 2024). Similarly, deep learning techniques like Generative Adversarial Imputation Networks (GAIN) and Variational Autoencoders (VAE) offer potential for capturing nonlinear and high-dimensional structures. However, empirical evaluations by (Sun et al., 2023) reveal that these models often underperform compared to traditional methods like MICE and missForest in small to moderately sized tabular datasets ($n < 30,000$), primarily due to challenges such as mode collapse and sensitivity to the missingness mechanism. These findings highlight the need to align imputation strategy selection with dataset characteristics, computational resources, and the intended analytical goals.

In summary, existing studies converge on the conclusion that no single imputation method is universally optimal. Instead, the choice should be guided by data characteristics such as size, variable type, and missingness mechanism. Simple methods may suffice for preliminary analyses or fully random missingness, while kNN, MICE, and random forest-based methods are preferred for more complex scenarios. Deep learning models offer potential for future

applications, particularly in large-scale or unstructured datasets, but require further refinement before they can be considered practical alternatives in most applied research settings. Ultimately, selecting an appropriate imputation strategy requires careful consideration of dataset structure, missingness type, computational demands, and the intended use of the imputed data.

3. Methodology

In this study, four widely used imputation methods are employed: mean, median, k-nearest neighbors (kNN), and multiple imputation by chained equations (MICE). The analysis uses a synthetic dataset (*simulated_medical_records*) designed to mimic clinical administrative data, consisting of 20 continuous variables such as body mass index (BMI), blood pressure, glucose, and cholesterol. The dataset is generated in three different sample sizes (n = 100, 3,000, and 50,000) to represent small to large-scale administrative records. The experiment begins by introducing missing values using the mdatagen Python library (Mangussi et al., 2025), which simulates missingness at three levels (5%, 20%, and 50%) under a Missing Completely at Random (MCAR) mechanism. The selected imputation methods are then applied using appropriate Python packages to fill in the missing values. Finally, the imputed datasets are evaluated by comparing them with the original complete dataset using Normalized Root Mean Square Error (NRMSE) to assess accuracy, and distribution plots to evaluate how well each method preserves the original data structure. **Figure 1** illustrates the experimental workflow in this study.

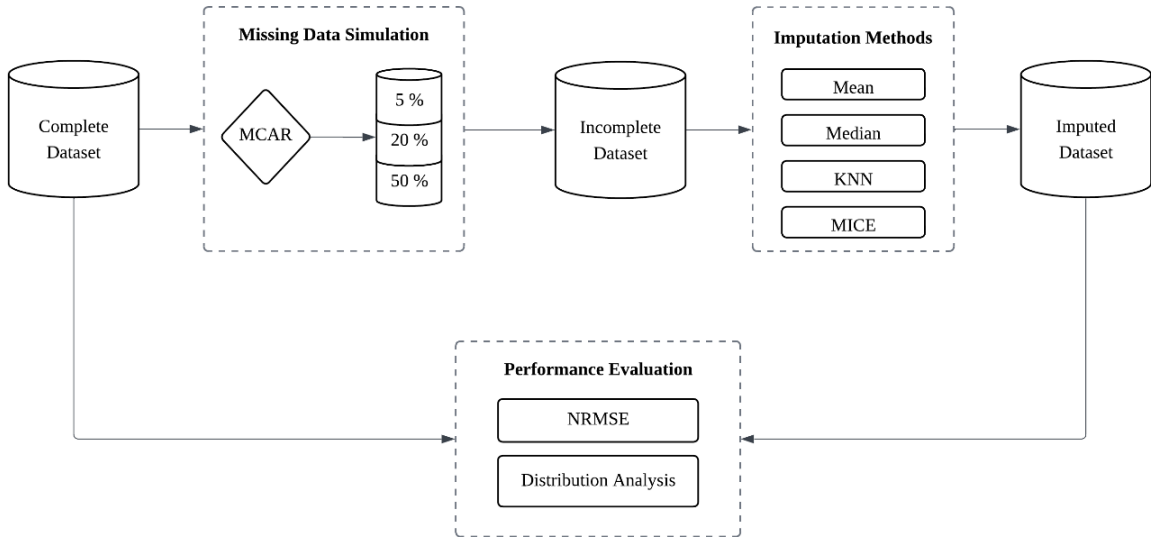

Figure 1: Overview of the experimental setup workflow

3.1 Imputation Methods

To systematically assess how different imputation techniques perform under controlled missingness conditions, this study selected a combination of simple and advanced methods. The selection was based on their prevalence in academic literature, computational feasibility, and suitability for continuous variables in administrative records. Each method is described below, detailing its mathematical foundation, assumptions, and the corresponding Python libraries used for implementation to ensure reproducibility and practical application.

a. Mean Imputation

Mean imputation replaces missing values with the mean of observed values for the variable. This approach assumes that the data are missing completely at random (MCAR) (Zhou et al., 2024). While it is simple and computationally efficient, it can reduce variability and introduce bias, especially when the proportion of missing data is large, potentially distorting the distribution (Alwateer et al., 2024; Jadhav et al., 2019). This method was implemented using SimpleImputer(strategy='mean') from the scikit-learn library and corresponds to formula (3.1).

$$x_{imputed} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.1}$$

where $x_{imputed}$ is the imputed value, x_i are the observed values for the variable, n is the number of observed (non-missing) entries.

b. Median Imputation

Median imputation substitutes missing values with the median of observed values in each variable. This method is robust to outliers and skewed distributions, making it more suitable than mean imputation for non-normal or right-skewed data (Alwateer et al., 2024). Similar to mean imputation, it treats variables independently and assumes MCAR, potentially limiting its ability to preserve multivariate structure (Zhou et al., 2024). This method was implemented with SimpleImputer(strategy='median') from scikit-learn Python library. The median is calculated by sorting the observed data $x_1 \leq x_2 \leq \ldots \leq x_n$, and defined as:

$$x_{imputed} = \begin{cases} x\left(\frac{n+1}{2}\right) & \text{if } n \text{ is odd} \\ \frac{x\left(\frac{n}{2}\right) + x\left(\frac{n}{2}+1\right)}{2} & \text{if } n \textit{ is even} \end{cases} \tag{3.2}$$

where $x_{imputed}$ is the imputed value, x_i is the i-th smallest value in the sorted list of observed values, n is the number of observed (non-missing) entries

c. k-Nearest Neighbor (kNN) Imputation

kNN imputation is a multivariate method that estimates missing values based on the average of the k nearest samples, identified through a distance metric over observed features. This method captures relationships between variables better than univariate imputations but requires careful selection of k and distance metrics (Alwateer et al., 2024; Zhou et al., 2024). It was implemented using the KNNImputer class from scikit-learn, with default parameters: n_neighbors=5, weights='uniform', and metric='nan_euclidean', which computes distances over shared non-missing features.

For each incomplete instance, distances are computed using the nan-aware Euclidean formula:

$$d_{(i,j)} = \sqrt{\sum_{f \in F} (x_{if} - x_{if})^2} \tag{3.3}$$

where F is the set of non-missing features shared between records i and j

Missing values are then imputed using uniform weights (default), where the mean is calculated over the *k* nearest neighbors:

$$x_{imputed} = \bar{x}\,\frac{1}{k}\sum_{j=1}^{k} x_j \tag{3.4}$$

This procedure is repeated for all missing entries until the entire dataset is imputed.

d. Multiple Imputations Using Chained Equation (MICE)

MICE is a widely used technique for addressing missing data, especially when the data are assumed to be Missing at Random (MAR). It employs an iterative process where each variable with missing values is modeled based on the others in a sequential, conditional manner. This multivariate approach helps maintain the inherent relationships between variables (Alwateer et al., 2024; Zhou et al., 2024). In this study, the implementation was carried out using the IterativeImputer from Python's scikit-learn package, leveraging Bayesian Ridge Regression for estimation. Despite being computationally, MICE is recognized for its superior accuracy and lower bias compared to traditional methods (Mohammed et al., 2021)

3.2 Evaluation Metrics

NRMSE is used to quantify the accuracy of imputed values for each variable. The Root Mean Squared Error (RMSE) is first calculated per variable and then normalized by that variable's original range. This normalization is essential for enabling meaningful comparisons across variables with different units and scales (e.g., blood pressure vs. age) within the dataset (Gayathri & Kavitha, 2024; Jadhav et al., 2019). This process was implemented in Python using the mean_squared_error function from sklearn.metrics. The formula used for NRMSE is:

$$NRMSE_j = \frac{RMSE}{Range} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}}{x_{max} - x_{min}} \tag{3.5}$$

where x_i and $\hat{x}_i$ are the true and imputed values, n is the number of observations, and x_{max} and x_{min} are the maximum and minimum values of the original variable.

To obtain a single performance score for each dataset, the NRMSEs for all variables are then averaged:

$$Average\ NRMSE = \frac{1}{n}\sum_{i=1}^{n} NRMSE \tag{3.6}$$

where *n* is the number of variables in the dataset.

4. Results and Discussion

Based on the results presented in **Table 2** and visualized in **Figure 2**, Multiple Imputation by Chained Equations (MICE) consistently achieved the highest imputation accuracy across all sample sizes (n = 100, 3,000, 50,000) and missingness levels (5%, 20%, 50%). K-Nearest Neighbors (kNN) followed closely, showing strong performance particularly in larger datasets. In contrast, traditional methods such as Mean and Median imputation yielded higher

Normalized Root Mean Square Error (NRMSE) values, with error magnitudes increasing alongside the proportion of missing data. These findings are consistent with previous studies by (Mohammed et al., 2021) and (Sun et al., 2023), who emphasized the robustness of MICE in handling mixed-type data under Missing at Random (MAR) conditions. Similarly, kNN has proven effective in domains involving nonlinear patterns and heterogeneous data (Alrawajfi et al., 2024; Jadhav et al., 2019; Li et al., 2024), whereas simpler methods like Mean or Median imputation are only suitable when missingness is minimal (Jadhav et al., 2019).

Table 1: Summary of average NRMSE.

Missing (%)	Method	n = 100	n = 3,000	n = 50,000
5	KNN	0.0373	0.0325	0.0300
	MICE	0.0302	0.0274	0.0257
	Mean	0.0432	0.0386	0.0356
	Median	0.0446	0.0398	0.0368
20	KNN	0.0863	0.0687	0.0634
	MICE	0.0849	0.0574	0.0533
	Mean	0.0934	0.0769	0.0713
	Median	0.0979	0.0794	0.0736
50	KNN	0.1462	0.1248	0.1163
	MICE	0.1329	0.1030	0.0955
	Mean	0.1477	0.1218	0.1126
	Median	0.1569	0.1256	0.1164

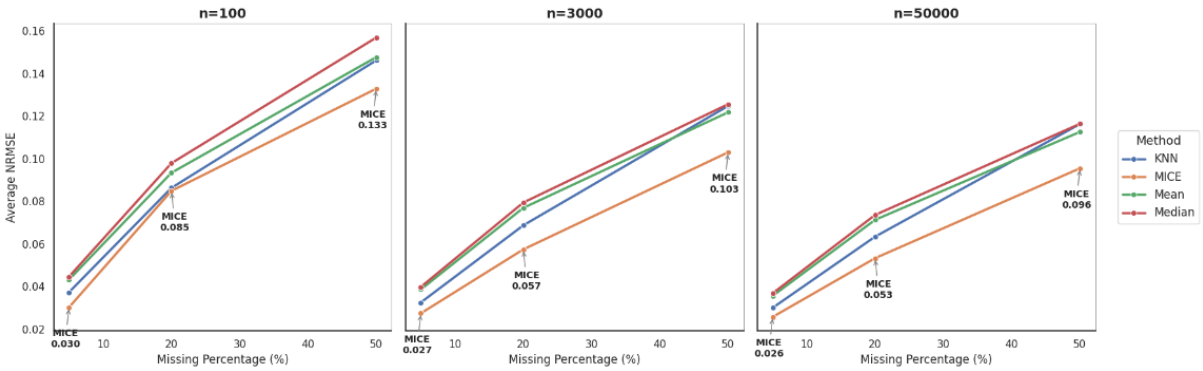

Figure 2: Average NRMSE by missing percentage, method & dataset size.

In terms of distributional accuracy, **Figure 3** illustrates how each imputation technique impacts the shape of the BMI variable. MICE consistently preserved the original distribution across all experimental scenarios, even under 50% missingness. KNN also maintained a close approximation, particularly as sample size increased. In contrast, Mean and Median imputations introduced visible distortions—most notably sharp central peaks—indicating a tendency to oversmooth and bias results toward the mean. These patterns were especially pronounced in smaller datasets (n = 100), confirming the sensitivity of basic methods to data sparsity. This observation aligns with the findings of (Jadhav et al., 2019), who caution against using central tendency-based imputations for large-scale or complex datasets.

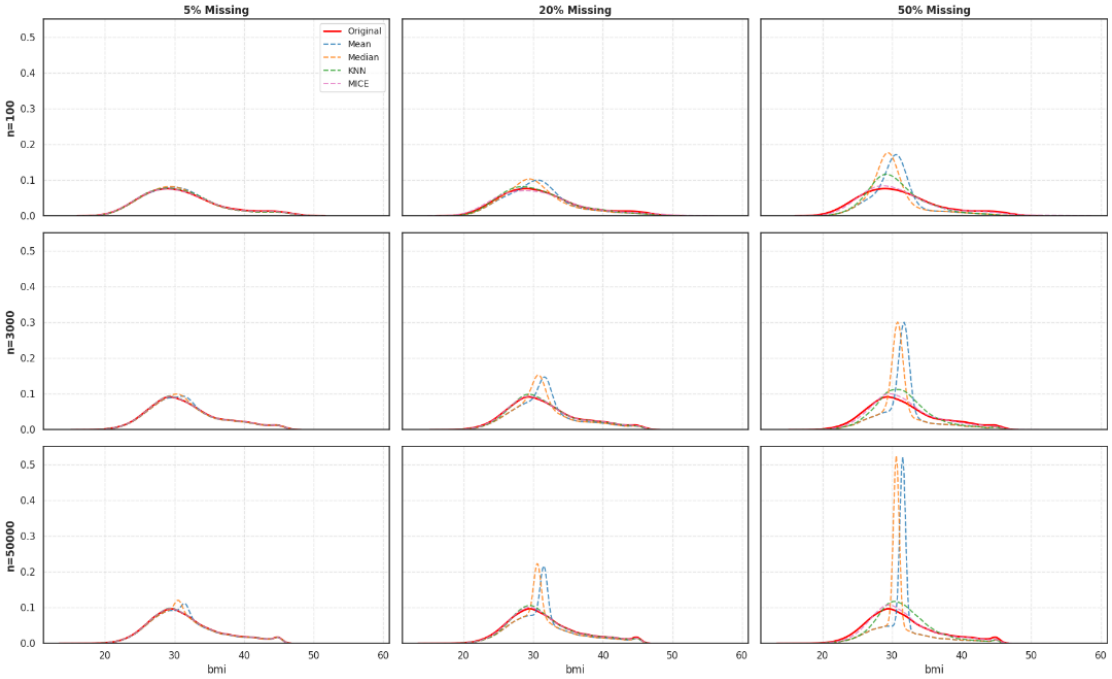

Figure 3: Distribution comparison of all imputation methods across dataset sizes and missingness levels.

These differences can be seen more clearly in **Figure 4** and **Figure 5**, which isolate the distributional effects of simple and advanced imputation methods, respectively. **Figure 4** highlights how both Mean and Median imputation distort the underlying distribution, particularly under high missingness. In contrast, **Figure 5** demonstrates that KNN and MICE better preserve distributional characteristics across dataset sizes, with MICE showing greater stability under severe missingness.

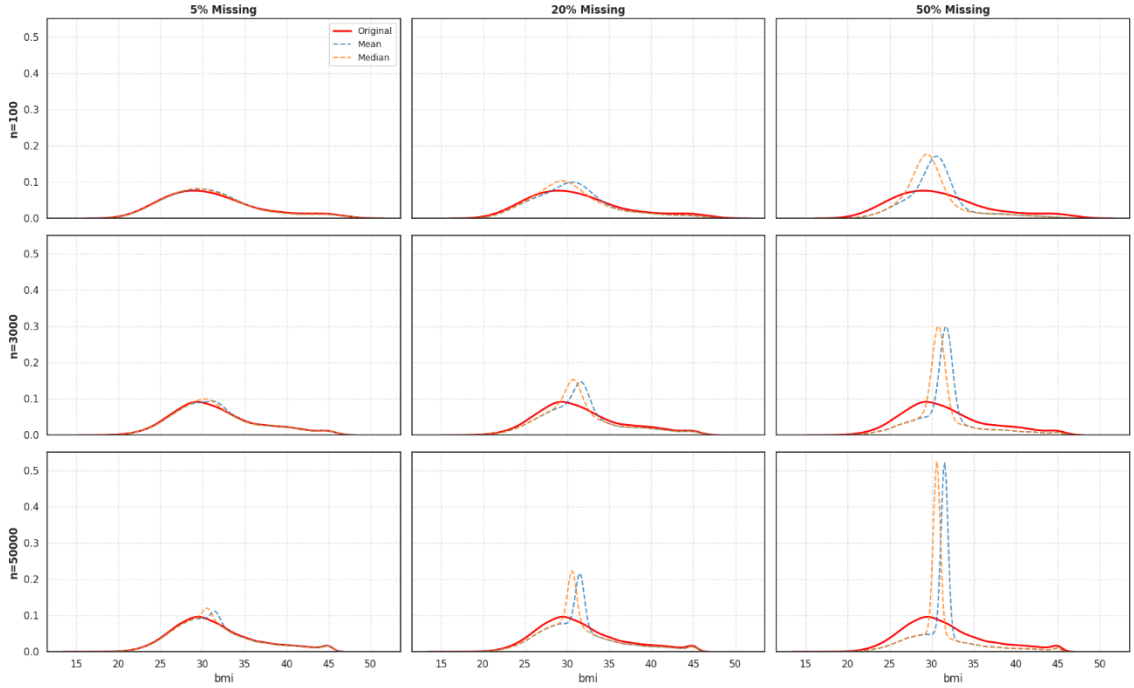

Figure 4: Distribution comparison of simple imputation methods (mean vs median) across dataset sizes and missingness levels.

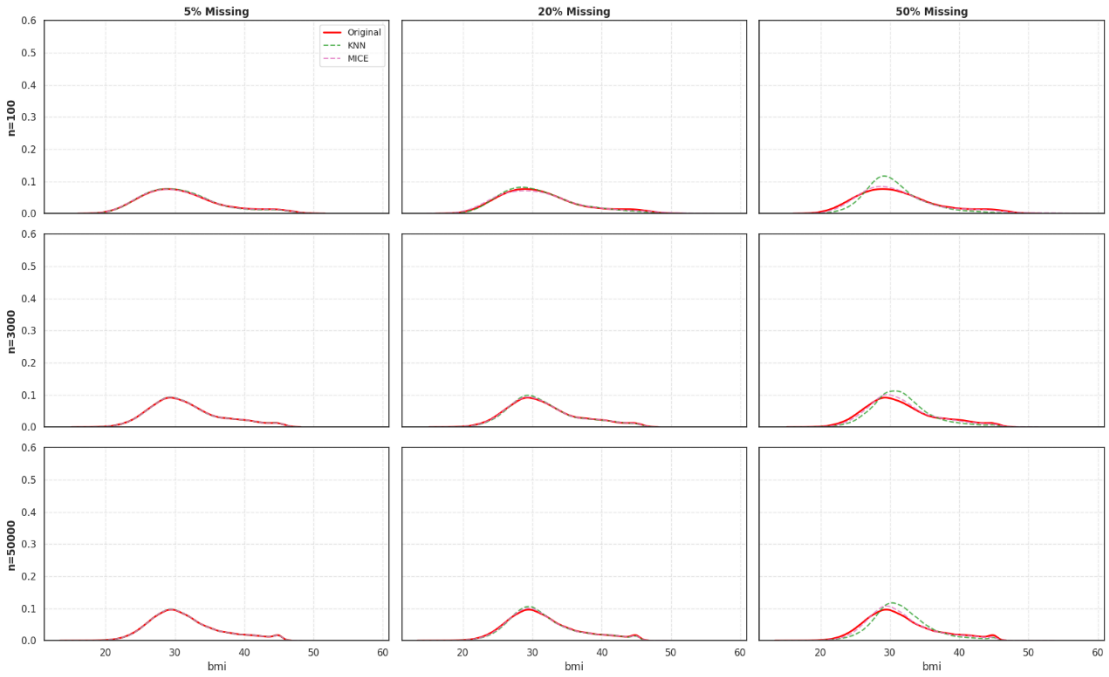

Figure 5: Distribution comparison of advanced imputation methods (kNN vs MICE) across dataset sizes and missingness levels.

Overall, the evidence reinforces that advanced imputation techniques such as MICE and KNN offer better performance not only in minimizing error but also in preserving the underlying distributional characteristics of the data, which is essential for accurate downstream analysis and informed decision-making—particularly in administrative and policy-relevant datasets.

4. Conclusion

This study highlights that both dataset size and missingness level significantly influence the performance of imputation methods. Larger datasets generally produced lower NRMSE values, indicating improved imputation accuracy. Among the four methods evaluated, MICE consistently achieved the highest accuracy and best preserved the original data distribution across all scenarios, followed by KNN. However, under high levels of missingness, KNN was occasionally outperformed by mean imputation, particularly in moderate to large datasets. Simple methods such as mean and median tended to distort the data distribution as missingness increased. Despite these insights, the study has certain limitations. It relied on simulated datasets that mimic administrative records, focused on continuous numerical variables, and assumed a Missing Completely at Random (MCAR) mechanism. These constraints may not fully capture the complexities of real-world data. Future work should explore other missingness mechanisms, such as MAR and MNAR, and evaluate imputation performance on categorical and mixed-type datasets. Such efforts would offer a more comprehensive understanding of imputation performance in diverse and complex data environments.

Acknowledgement

This research was supported by the Universiti Teknologi Malaysia under the UTM Fundamental Research Grant (UTMFR) with project number [Q.K130000.3856.23H47]. The authors would like to express their gratitude for the financial and administrative support that made this study possible.

Conflict of Interest Statement

The authors declare that there is no conflict of interest regarding the publication of this study.

References

Afkanpour, M., Hosseinzadeh, E., & Tabesh, H. (2024). Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02310-6

Alrawajfi, A., Ismail, M. T., Al Wadi, S., Atiewi, S., & Awajan, A. (2024). Multiple imputation methods: a case study of daily gold price. *PeerJ Computer Science*, *10*. https://doi.org/10.7717/PEERJ-CS.2337

Alwateer, M., Atlam, E.-S., El-Raouf, M. M. A., Ghoneim, O. A., & Gad, I. (2024). Missing Data Imputation: A Comprehensive Review. *Journal of Computer and Communications*, *12*(11), 53–75. https://doi.org/10.4236/jcc.2024.1211004

Askinadze, A., & Conrad, S. (2018). Respecting data privacy in educational data mining: An approach to the transparent handling of student data and dealing with the resulting missing value problem. *Proceedings - 2018 IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE 2018*, 165–170. https://doi.org/10.1109/WETICE.2018.00037

Gayathri, M., & Kavitha, V. (2024). A Comparative Analysis of the Imputing of Missing Data on Air Pollution. *Proceedings of the 2nd IEEE International Conference on Networking and Communications 2024, ICNWC 2024*. https://doi.org/10.1109/ICNWC60771.2024.10537524

Grigoroff, L., Masuda, R., Lindon, J., Kadyrov, J., Nicholson, J. K., Holmes, E., & Wist, J. (2024). Evaluation of imputation strategies for multi-centre studies: Application to a large clinical pathology dataset. *Iranian Journal of Public Health*, *50*(7), 1372–1380. https://doi.org/10.21203/rs.3.rs-5308928/v1

Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, *33*(10), 913–933. https://doi.org/10.1080/08839514.2019.1637138

Kline, N. (2022). *Using Administrative Data in Social Policy Research*. https://acf.gov/sites/default/files/documents/opre/administrative_data_brief_feb2023.pdf

Li, J. H., Guo, S. X., Ma, R. L., He, J., Zhang, X. H., Rui, D. S., Ding, Y. S., Li, Y., Jian, L. Y., Cheng, J., & Guo, H. (2024). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02173-x

Mangussi, A. D., Santos, M. S., Lopes, F. L., Pereira, R. C., Lorena, A. C., & Abreu, P. H. (2025). mdatagen: A python library for the artificial generation of missing data. *Neurocomputing*, *625*. https://doi.org/10.1016/j.neucom.2025.129478

Milne, B. J., Souza, S. D., Andersen, S. H., & Richmond-Rakerd, L. S. (2022). Use of Population-Level Administrative Data in Developmental Science. *Annual Review of Developmental Psychology*, *4*, 447–468. https://doi.org/10.1146/annurev-devpsych-120920

Mohammed, M. B., Zulkafli, H. S., Adam, M. B., Ali, N., & Baba, I. A. (2021). Comparison of five imputation methods in handling missing data in a continuous frequency table. *AIP Conference Proceedings*, *2355*. https://doi.org/10.1063/5.0053286

Saini, P., & Nagpal, B. (2024). Analysis of missing data and comparing the accuracy of imputation methods using wheat crop data. *Multimedia Tools and Applications*, *83*(14), 40393–40414. https://doi.org/10.1007/s11042-023-17178-9

Soldatenkova, A., Calabrese, A., Ghiron, N. L., & Tiburzi, L. (2023). Emergency department performance assessment using administrative data: A managerial framework. *PLoS ONE*, *18*(11). https://doi.org/10.1371/journal.pone.0293401

Sun, Y., Li, J., Xu, Y., Zhang, T., & Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, *227*. https://doi.org/10.1016/j.eswa.2023.120201

Suresh A & B, M., Taib, R., Zhao, Y., & Jin, W. (2019). Sharpening the BLADE: Missing Data Imputation using Supervised Machine Learning. *AI 2019: Advances in Artificial Intelligence*, 215–227. https://doi.org/https://doi.org/10.1007/978-3-030-35288-2_18

Timofte, D., Stoian, A. P., Hainarosie, R., Diaconu, C., Iliescu, D. B., G. Balan, G., Ciuntu, B., & Neagoe, R. M. (2018). A Review on the Advantages and Disadvantages of Using Administrative Data in Surgery Outcome Studies. *Journal of Surgery (Jurnalul de Chirurgie)*, *14*(3), 105–107. https://doi.org/10.7438/1584-9341-14-3-3

Zhou, Y., Aryal, S., & Bouadjenek, M. R. (2024). Review for Handling Missing Data with Special Missing Mechanism. *ArXiv*, *abs/2404.04905*. http://arxiv.org/abs/2404.04905