A Critical Perspective on MNAR Data: Imputation, Generation, and the Path Toward a Unified Framework

Fatemeh Azad
fatemeh.azad@fri.uni-lj.si
University of Ljubljana
Ljubljana, Slovenia

Matjaž Kukar
matjaz.kukar@fri.uni-lj.si
University of Ljubljana
Ljubljana, Slovenia

Abstract

Missing Not at Random (MNAR) data remains one of the most difficult challenges in statistical analysis and machine learning. Despite the widespread availability of advanced imputation methods, most research continues to focus on Missing Completely at Random (MCAR) and partially on Missing at Random (MAR) scenarios. This paper provides a critical overview of existing approaches for MNAR imputation, methods for simulating MNAR data, and the limitations of current evaluation practices. We highlight the lack of standardized benchmarks, unrealistic missingness rates, and insufficient coverage of MNAR conditions in empirical studies. Finally, we propose a suitable framework for comprehensive testing of design principles, enabling robust and reproducible evaluation of imputation methods across mechanisms and missingness rates.

Keywords

Missing data, MNAR, data imputation, missingness mechanisms, data generation, machine learning, evaluation framework.

1 Introduction

Missing data is a pervasive challenge across various domains, from clinical diagnostics and bioinformatics to finance, sensor networks, and social sciences. Missing, damaged, or unrecorded data entries can negatively affect the accuracy of statistical analysis and machine learning models. They reduce predictive power, introduce bias, and often create incompatibilities with algorithms requiring complete inputs [8]. The impact is especially important in critical areas like healthcare decision support, where unreliable data or incorrect interpretation can lead to harmful conclusions.[14, 2].

A primary difficulty in handling missing data is understanding the underlying *missingness mechanism*. According to the taxonomy of Little and Rubin [10], We have three types of missingness: *Missing Completely at Random* (MCAR), *Missing at Random* (MAR), and *Missing Not at Random* (MNAR).

To formally describe the MCAR, MAR, and MNAR mechanisms, we first define the following notation, as per [9, 19]:

- X: the complete data matrix, which consists of two parts, with X_{obs} being the observed and X_{mis} the missing part of the data.
- R: an indicator matrix of the same dimensions as X, where $R_{ij} = 1$ if the value X_{ij} is missing, and $R_{ij} = 0$ if it is observed.
- ψ: a parameter or set of parameters that govern the missingness process.

- Data is *MCAR* if the probability of a value being missing is completely independent of both the observed and the unobserved data. The missingness is unrelated to the data itself — it is a purely random (Eq. 1) as the missingness pattern (R) depends neither on the observed data (X_{obs}) nor on the missing data (X_{mis}).

$$P(R|X_{obs}, X_{mis}, \psi) = P(R|\psi) \tag{1}$$

- Data is *MAR* if the probability of a value being missing depends only on the observed data, not on the missing data itself (Eq. 2). This means that the missingness could be predicted from available (non-missing) data. The probability of the missingness pattern (R) is conditionally independent of the actual missing values (X_{mis}) once the observed values (X_{obs}) are taken into account.

$$P(R|X_{obs}, X_{mis}, \psi) = P(R|X_{obs}, \psi) \tag{2}$$

- Data is *MNAR* if the probability of a value being missing depends on some unobserved (missing) value itself, even after accounting for all the observed data (Eq. 3). In this case (X_{mis}) can also include latent features, unobserved for all instances. This is the most complex scenario, as the missingness pattern itself is informative. The probability of the missingness pattern (R) is therefore dependent on the missing values (X_{mis}) in a way that cannot be explained by the observed values (X_{obs}).

$$P(R|X_{obs}, X_{mis}, \psi) \tag{3}$$

While MCAR and MAR have been extensively studied, MNAR remains the most difficult and least explored scenario, precisely because the missingness itself carries information about the data. For example, high-income individuals may systematically withhold reporting their wealth, or patients with severe conditions may drop out of longitudinal studies. In both cases, the very act of non-response encodes meaningful but hidden signals.

The prevalent imputation (replacing missing values) research has focused on MCAR and MAR settings, where assumptions about independence or conditional dependence simplify methodological development and evaluation [14, 23, 13]. In contrast, MNAR scenarios pose a dual challenge: not only is the missing information inherently dependent on unobserved values, but there are also very few benchmark datasets that explicitly model or annotate MNAR mechanisms. Consequently, evaluation standards remain incomplete. Reported missingness rates often underestimate or ignore MNAR effects, and even sophisticated models, such as generative adversarial networks [7, 24], graph neural approaches [25], or transformer-based imputers [3], rarely demonstrate systematic robustness in MNAR conditions. Recent works [11, 4] have shown the potential of ensemble or meta-imputation strategies, which combine diverse imputers

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.skui.0957

into robust pipelines. However, these frameworks are also mostly validated under MCAR or MAR assumptions.

In this paper, we take a critical perspective on the current state of missing data research, specifically focusing on MNAR. We argue that three gaps must be addressed: (i) the lack of effective imputation techniques designed specifically for MNAR, as current methods are limited in scope and seldom used in practice; (ii) the deficiency of datasets and generators that can faithfully represent MNAR patterns; and (iii) the insufficiency of reported missingness rates. To bridge these gaps, we outline the vision and design principles of a comprehensive *framework* for MNAR research that integrates data generation, imputation, and evaluation under standardized conditions. Such a framework would enable more robust comparisons of existing methods and guide the development of novel techniques tailored to the inherent challenges of MNAR.

The remainder of this paper is organized as follows. Section 2 reviews existing imputation approaches and discusses their applicability to MNAR. Section 3 examines methodologies for simulating and generating MNAR data, highlighting their limitations. Section 4 critiques how missingness is reported and motivates the need for standardized benchmarks. Finally, Section 5 presents our vision for a unified MNAR research framework and outlines open challenges for the community.

2 Imputation Methods for MNAR Data

A wide range of imputation techniques has been proposed in the literature, from simple statistical to advanced deep generative models. While these methods have demonstrated effectiveness under Missing Completely at Random (MCAR) or Missing at Random (MAR) assumptions, their suitability for Missing Not at Random (MNAR) scenarios remains highly questionable. This section reviews the main categories of imputation techniques and highlights their limitations when faced with MNAR data.

While it is often stated that there are almost no methods tailored for MNAR, several strands of work do exist … However, these remain underutilized and rarely integrated into mainstream imputation pipelines.

2.1 Statistical Imputation Methods

Statistical techniques such as mean, median, mode, or regression-based imputations are simple and computationally efficient but they mostly rely on strong assumptions about the independence or conditional dependence of missingness [8, 27]. These assumptions rarely hold under MNAR, where the missingness mechanism is informative itself. For example, imputing systematically underreported values (e.g., income, clinical severity) with central-tendency statistics introduces bias and distorts the true distribution. Maximum likelihood and Bayesian approaches attempt to capture uncertainty, but they typically assume that the missingness process can be ignored or is fully modeled by observed data [10], which is not the case for MNAR.

2.2 Machine Learning-Based Approaches

Machine learning methods, such as k-nearest neighbors (KNN) [14], matrix factorization [20], decision trees [21], and support vector machines (SVMs) [6], utilize feature dependencies to address missing data entries. While more flexible than statistical methods, they fail when the missingness depends on unobserved or latent variables. For instance, if severely ill patients systematically omit follow-up surveys, no observed features can explain this absence, and machine learning based imputers cannot recover the missing structure without explicitly modeling the mechanism.

2.3 Deep Learning Approaches

Deep generative models have significantly advanced imputation research. Variational Autoencoders (VAEs) [2] and Generative Adversarial Networks (GANs) [23, 7, 24] are capable of learning complex distributions and have shown robustness to high missingness rates. However, their performance in the context of MNAR conditions is not assured. While some frameworks, such as MisGAN, explicitly attempt to learn the missingness mask distribution alongside the data [7], they often rely on approximations that do not generalize across domains. Similarly, diffusion-based models [22, 26] and graph-based imputers [25] extend coverage to structured data but rarely test systematically against MNAR conditions. Transformers, such as ReMasker [3], provide context-aware imputations, but again, their evaluations are mostly limited to MCAR and MAR scenarios.

2.4 Ensemble Approaches

Recent efforts highlight the potential of combining multiple imputers in ensemble or meta-learning frameworks [11, 4]. Such methods leverage complementary strengths of diverse imputers and often achieve more stable performance across heterogeneous datasets. However, existing ensemble frameworks have been validated primarily under MCAR assumptions, and their ability to handle MNAR remains largely unexplored. Recent work has also explored meta-imputation strategies, such as the Meta-Imputation Balanced (MIB) framework, which combines multiple base imputers in a supervised setting [1].

To synthesize the discussion above, Table 1 summarizes the main categories of imputation approaches, their representative methods, applicability to missingness mechanisms, and key references.

3 Generation of MNAR Data

A persistent challenge in missing data research is the lack of reliable and reproducible benchmarks for handling MNAR scenarios. While MCAR and MAR can be easily simulated by random masking or conditioning on observed features, MNAR requires masking rules that depend on unobserved or latent variables, which makes the generation process more challenging. Consequently, most experimental studies rely on oversimplified masking strategies that do not capture the complexity of real-world MNAR mechanisms [18, 5].

3.1 The Role of Data Amputation

Deliberately injecting missing values into fully complete datasets, referred to as *data amputation*, plays a crucial role in evaluating imputation techniques. However, until recently, implementations of amputation were highly heterogeneous and often insufficiently documented, preventing fair comparisons across studies [18]. This problem is particularly acute for MNAR, where even slight differences in implementation can lead to very different conclusions.

Table 1: Comparison of Imputation Approaches from Literature

Approach	Representative Methods	Missingness Types Addressed	Representative References
Traditional Statistical	Mean, Median, Mode, Regression-based, Maximum Likelihood, Bayesian Approaches	MCAR only (rarely MAR)	Schafer & Graham [27], Little & Rubin [10], Lin & Tsai [8]
Machine Learning	KNN, Matrix Factorization, Decision Trees, SVM	MCAR, partially MAR	Murti et al. [14], Lee et al. [20], Song & Lu [21], Feng et al. [6]
Deep Learning	VAEs, GANs, Diffusion Models, Graph-based Models, Transformers	MCAR, MAR (limited MNAR)	Collier et al. [2], Yoon et al. [23, 24], Li et al. [7], Du et al. [3], Tashiro et al. [22], Zheng & Charoenphakdee [26], You et al. [25]
Meta-Learning / Ensembles	Meta Learning, Meta-Regressio, MIB Frameworkn	MCAR, partially MAR; potential for MNAR	Liu et al. [11], Ellington et al. [4], Azad et al. [1]

3.2 Artificial MNAR Generation Strategies

The most common way to simulate MNAR is by masking values as a function of their own magnitude or distribution. For instance, removing a feature's highest or lowest values mimics non-disclosure of extreme outcomes (e.g., very high glucose levels) [18]. Stochastic variants extend this idea by assigning missingness probabilities proportional to the unobserved value itself, enabling flexible control over the intensity of missingness [16]. While intuitive, such strategies remain oversimplified, often restricted to univariate rules that fail to capture the multidimensional dependencies of real domains [5].

Recent work has proposed standardized libraries for data amputation to address reproducibility concerns. The `mdatagen` package provides a broad set of implementations for MCAR, MAR, and MNAR, supporting univariate and multivariate scenarios [12]. In particular, it incorporates advanced MNAR mechanisms such as Missingness Based on Own Values (MBOV), Missingness Based on Own and Unobserved Values (MBOUV), and Missingness Based on Intra-Relations (MBIR) [15]. These implementations move beyond ad hoc thresholding by systematically encoding missingness processes and offering reproducible pipelines. In addition, `mdatagen` includes visualization and evaluation modules, allowing researchers to inspect missingness patterns and assess their impact on imputation performance.

Together, these synthetic and standardized approaches form the current toolkit for MNAR data generation. However, despite their usefulness, they remain abstractions of real-world processes and should ideally be complemented by domain-informed simulations.

3.3 Domain-Inspired Simulation

Beyond standardized libraries, domain knowledge remains critical for realistic MNAR generation. In healthcare, dropout is often linked to disease severity, side effects, or socioeconomic constraints. In socioeconomic surveys, non-response may be strongly correlated with privacy-sensitive attributes such as income or debt. Encoding these mechanisms requires integrating causal assumptions with probabilistic masking rules [17]. However, such domain-specific approaches are difficult to generalize, limiting their utility as benchmarks.

4 Toward a Unified Framework for MNAR Research

Two key insights emerge from the previous sections: (i) current imputation methods are not explicitly designed for MNAR, and (ii) the lack of realistic MNAR generators inhibits effective evaluation. To address these gaps, we anticipate a unified framework integrating generation, imputation, and evaluation of MNAR data under standardized and reproducible conditions.

4.1 Design Principles

A comprehensive MNAR framework should have the following principles:

- **Synthetic realism:** Data generators should simulate MNAR scenarios that mirror real-world domains (e.g., systematic dropout in healthcare, self-censoring in socio-economic data), either by extending existing functionality (e.g., mdatagen [12]) or by incorporating custom plug-in modules. To balance interpretability with scalability, both threshold-based rules and learned mechanisms should be supported.
- **Comprehensive evaluation:** Benchmarks must test across all three missingness mechanisms (MCAR, MAR, MNAR) and a full spectrum of missingness rates.
- **Cross-domain applicability:** The framework should support diverse data types (tabular, sequential, multimodal) and allow integration of domain knowledge for context-specific MNAR simulation.

4.2 Proposed Framework Components

We propose that a unified MNAR framework should consist of three interdependent modules:

(1) **MNAR Data Generators:** Domain-informed and probabilistic tools for simulating missingness patterns that depend on latent or unobserved values, using existing libraries ([12] or incorporating custom plug-in functions.
(2) **Imputation Engines:** A modular interface with plug-in adapters for existing methods that support statistical, machine learning, deep learning, and ensemble methods [14, 23, 1]. By isolating imputers within a common framework, researchers can test their robustness under controlled MNAR scenarios.
(3) **Evaluation Suite:** Standardized protocols that combine direct metrics (e.g., Mean Absolute Error (MAE), Root Mean Squared Error (RMSE)) with indirect metrics (downstream predictive performance, such as accurracy, RMSE/MAE, or domain relevant metrics such as interpretability, reliability, fairness, ...) [1].

4.3 Benefits and Impact

Developing such a framework would enable several advances:

- *Reproducibility:* Common benchmarks and generators ensure that different imputation methods can be fairly compared.
- *Realism:* Domain-specific MNAR mechanisms bring evaluations closer to real-world conditions, reducing the gap between research and practice.
- *Innovation:* By exposing the weaknesses of existing methods under MNAR, the framework incentivizes the development of mechanism-aware imputers.
- *Generalization:* Unified treatment of MCAR, MAR, and MNAR encourages methods that adapt to unknown or mixed missingness mechanisms without prior assumptions.

5 Conclusion

Missing data remains one of the most persistent challenges in machine learning and statistical analysis. While decades of research have produced numerous imputation techniques, ranging from simple statistical estimators to deep generative models, most methods have been designed and evaluated under the more tractable MCAR and MAR mechanisms. In contrast, the most realistic and challenging setting, MNAR, remains critically underexplored.

Our review highlights three major gaps in the current state of the field. First, existing imputation methods rarely model the dependence of missingness on unobserved values, making them unsuitable for MNAR scenarios. Second, generating realistic MNAR data is crucial because most benchmarks use ad hoc or overly simplistic masking strategies, which fail to capture the complexity of real-world missingness. Third, evaluation standards remain incomplete, with reported missingness rates often conflating MCAR/MAR assumptions and failing MNAR realities. Together, these shortcomings hinder fair comparisons and limit methodological innovation.

To address these challenges, we propose the vision and design principles of a unified MNAR framework that integrates three components: (i) data generators that are aware of mechanisms and can create realistic MNAR patterns, (ii) modular imputation engines that enable thorough testing of various methods, and (iii) extensive evaluation suites that include direct metrics and indirect metrics. Such a framework would provide reproducibility, realism, and a strong foundation for developing next-generation imputation techniques.

Future research should move toward principled, mechanism-aware imputers and adopt standardized benchmarks for MNAR generation and evaluation. To advance MNAR research, we need more powerful algorithms and standardized tools and protocols that enhance rigor and comparability in the field.

Acknowledgements

The research and development presented in this paper were funded by the Research Agency of the Republic of Slovenia (ARIS) through the ARIS Young Researcher Programme (research core funding No. P2-209). While preparing this work, the authors used Grammarly to check the correctness of grammar and improve the fluency of the writing, aiming to enhance the clarity and impact of the publication. The authors reviewed and edited the content produced with this tool/service and accept full responsibility for the final published content.

References

[1] Fatemeh Azad, Zoran Bosnić, and Matjaž Kukar. 2025. Meta-imputation balanced (mib): an ensemble approach for handling missing data in biomedical machine learning. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBE)*. submitted.

[2] Mark Collier, Bayan Mustafa, and Mihaela van der Schaar. 2020. VAEs in the presence of missing data. *arXiv:2006.05301*.

[3] Meng Du, Gábor Melis, and Zhaozhi Wang. 2023. Remasker: imputing tabular data with masked autoencoding. In *The Eleventh International Conference on Learning Representations*.

[4] E. Ellington, Guillaume Bastille-Rousseau, Cayla Austin, Kristen Landolt, Bruce Pond, Erin Rees, Nicholas Robar, and Dennis Murray. 2014. Using multiple imputation to estimate missing data in meta-regression. *Methods in Ecology and Evolution*, 6, (Dec. 2014). DOI: 10.1111/2041-210X.12322.

[5] Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. 2021. A survey on missing data in machine learning. (May 2021). DOI: 10.21203/rs.3.rs-535520/v1.

[6] Hao Feng, Lihui Chen, and Ke Wang. 2005. A svm regression based approach to filling in missing values. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 581–587.

[7] Shun-Chuan Li, Bingsheng Jiang, and Benjamin M Marlin. 2019. Misgan: learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*.

[8] Wei-Chao Lin and Chih-Fong Tsai. 2020. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53, 1487–1509.

[9] Roderick J. A. Little and Donald B. Rubin. 1986. *Statistical Analysis with Missing Data*. John Wiley & Sons. ISBN: 978-0471802545.

[10] Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

[11] Qian Liu and Manfred Hauswirth. 2020. A provenance meta learning framework for missing data handling methods selection. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. DOI: 10.1109/UEMCON51285.2020.9298089.

[12] Arthur Mangussi, Miriam Santos, Filipe Loyola Lopes, Ricardo Cardoso Pereira, Ana Lorena, and Pedro Henriques Abreu. 2025. Mdatagen: a python library for the artificial generation of missing data. *Neurocomputing*, 625, (Apr. 2025), 129478. DOI: 10.1016/j.neucom.2025.129478.

[13] Pierre-Alexandre Mattei and Jes Frellsen. 2019. Miwae: deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*. PMLR, 4413–4423.

[14] Dinar M P Murti, I N A Ramatryana, and A P Wibawa. 2019. K-nearest neighbor (k-nn) based missing data imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)*. IEEE, 83–88.

[15] 2023. *Siamese autoencoder-based approach for missing data imputation*. (June 2023), 33–46. ISBN: 978-3-031-35994-1. DOI: 10.1007/978-3-031-35995-8_3.

[16] 2023. *Automatic delta-adjustment method applied to missing not at random imputation*. (June 2023), 481–493. ISBN: 978-3-031-35994-1. DOI: 10.1007/978-3-031-35995-8_34.

[17] Ricardo Cardoso Pereira, Joana Cristo Santos, José Amorim, Pedro Rodrigues, and Pedro Henriques Abreu. 2020. Missing image data imputation using variational autoencoders with weighted loss. In (Apr. 2020).

[18] Miriam Seoane Santos, Ricardo Cardoso Pereira, Adriana Fonseca Costa, Jastin Pompeu Soares, João Santos, and Pedro Henriques Abreu. 2019. Generating synthetic missing data: a review by missing mechanism. *IEEE Access*, 7, 11651–11667. DOI: 10.1109/ACCESS.2019.2891360.

[19] Joseph L. Schafer and John W. Graham. 2002. Missing data: our view of the state of the art. *Psychological methods*, 7 2, 147–77. https://api.semanticscholar.org/CorpusID:7745507.

[20] Nandana Sengupta, Madeleine Udell, Nathan Srebro, and James Evans. 2022. Sparse data reconstruction, missing value and multiple imputation through matrix factorization. *Sociological Methodology*.

[21] Ying-Ying Song and Ying Lu. 2015. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27, 2, 130.

[22] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csdi: conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*. Vol. 34, 24804–24816.

[23] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GAIN: missing data imputation using generative adversarial nets. In *International conference on machine learning*. PMLR, 5689–5698.

[24] Sanghoon Yoon and Sanghoon Sull. 2020. Gamin: generative adversarial multiple imputation network for highly missing data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8456–8464.

[25] Jiaxuan You, Xiaobai Ma, Daisy Ding, Mykel Kochenderfer, and Jure Leskovec. 2020. Handling missing data with graph representation learning. In *Advances in Neural Information Processing Systems*. Vol. 33, 18357–18368.

[26] Shuhan Zheng and Nontawat Charoenphakdee. 2022. Diffusion models for missing value imputation in tabular data. *arXiv preprint arXiv:2210.17128*.

[27] Yuyang Zhou, Sarjyt Aryal, and Mohamed Reda Bouadjenek. 2024. Review for handling missing data with special missing mechanism. *arXiv preprint arXiv:2404.04905*.