Afkanpour *et al.*
*BMC Medical Research Methodology* (2024) 24:188
https://doi.org/10.1186/s12874-024-02310-6

BMC Medical Research Methodology

RESEARCH

Open Access

# Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review

Marziyeh Afkanpour[1], Elham Hosseinzadeh[1] and Hamed Tabesh[1*]

## Abstract

**Background and objectives** Comprehending the research dataset is crucial for obtaining reliable and valid outcomes. Health analysts must have a deep comprehension of the data being analyzed. This comprehension allows them to suggest practical solutions for handling missing data, in a clinical data source. Accurate handling of missing values is critical for producing precise estimates and making informed decisions, especially in crucial areas like clinical research. With data's increasing diversity and complexity, numerous scholars have developed a range of imputation techniques. To address this, we conducted a systematic review to introduce various imputation techniques based on tabular dataset characteristics, including the mechanism, pattern, and ratio of missingness, to identify the most appropriate imputation methods in the healthcare field.

**Materials and methods** We searched four information databases namely PubMed, Web of Science, Scopus, and IEEE Xplore, for articles published up to September 20, 2023, that discussed imputation methods for addressing missing values in a clinically structured dataset. Our investigation of selected articles focused on four key aspects: the mechanism, pattern, ratio of missingness, and various imputation strategies. By synthesizing insights from these perspectives, we constructed an evidence map to recommend suitable imputation methods for handling missing values in a tabular dataset.

**Results** Out of 2955 articles, 58 were included in the analysis. The findings from the development of the evidence map, based on the structure of the missing values and the types of imputation methods used in the extracted items from these studies, revealed that 45% of the studies employed conventional statistical methods, 31% utilized machine learning and deep learning methods, and 24% applied hybrid imputation techniques for handling missing values.

**Conclusion** Considering the structure and characteristics of missing values in a clinical dataset is essential for choosing the most appropriate data imputation technique, especially within conventional statistical methods. Accurately estimating missing values to reflect reality enhances the likelihood of obtaining high-quality and reusable data, contributing significantly to precise medical decision-making processes. Performing this review study creates a guideline for choosing the most appropriate imputation methods in data preprocessing stages to perform analytical processes on structured clinical datasets.

*Correspondence:
Hamed Tabesh
tabesh79@gmail.com; TabeshH@mums.ac.ir
Full list of author information is available at the end of the article

## Highlights

- The evidence map emphasized the importance of considering missing data characteristics when choosing an imputation method, providing insights for researchers in method selection.
- The distinction between statistical and learning-based approaches highlighted the strengths and considerations of each method based on data structure.
- Understanding missing data structures can enhance the quality and reliability of data imputation techniques, and improve medical decision-making accuracy.
- Simulation studies are crucial for validating imputation techniques and enhancing their robustness in practical applications.
- Considering missing data mechanisms, patterns, and ratios can aid researchers in making informed decisions on selecting appropriate imputation methods, leading to high-quality, and reusable data for precise medical decision-making.

**Keywords** Imputation methods, Missing values, Mechanism of missingness, Pattern of missingness, Missing ratio, Clinical dataset, Simulation study

## Introduction

Missing data refers to values that not observed in one or more features in a dataset, but would have significance for analysis if they were. Essentially, a missing value conceals a valuable piece of information [1, 2]. The structure of missing values refers to the two concepts of mechanism and pattern of missing data in a dataset. The missingness mechanism(s) concentrates on the connection between missing data and the variables' values in the data set. While the pattern of missing data indicates which values are absent and which one of them is present in the data set [1], Rubin initially classified the mechanism of missingness into three main categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Additionally, the pattern of missing values includes univariate, multivariate, monotone, arbitrary or general, and file matching [1].

Missing data or Missing ness within a dataset considered a widespread problem in many real-world datasets for data analysis, especially in health domains such that it exists in almost all clinical and epidemiological research studies [3, 4]. In clinical settings, missing data can result from factors like lack of data observation, human and machine errors, attrition due to social or natural causes, user privacy concerns, missed clinic appointments, data transmission issues, incorrect measurements, and merging unrelated data [5–7].

The presence of missing values in a dataset complicates data preprocessing and analysis, leading to a reduction in statistical power, potential bias in treatment effect estimates, decreased sample size, and compromised precision of confidence intervals, ultimately resulting in an underestimation of variability [2, 8, 9]. This affects data analysis and predictive accuracy, and introduces bias in decision-making processes [8, 10]. Missing values have an impact on data quality, affecting analysis outcomes and presenting challenges for predictive tasks and statistical studies. Addressing missing values is crucial for obtaining clinically collected data with research reusability and ensuring results with high generalizability, despite the biases and uncertainties they may introduce in analytical findings [11, 12].

Therefore, identifying and carefully managing missing values in a dataset is crucial [13]. Data imputation, a method for handling missing data [1], involves predicting missing values by estimating them based on the data context [7, 12, 14]. This process aims to replace missing attributes with estimated values to establish meaningful relationships among all dataset values [15], preserving completeness and data quality for analytics [16]. Hence, reliable imputation methods are essential for addressing missing data issues, with three main approaches: single imputation, multiple imputation, and predictive imputation [17]. A single imputation adds a plausible value to each missing value [10], while multiple imputations assign multiple plausible values to each missing value for more unbiased estimates [2]. Predictive imputation utilizes machine learning deep learning techniques to create prediction models for estimating missing values [17, 18]. Various techniques like regression [19, 20], hot-deck imputation [21, 22], expectation maximization [23], support vector machine [24], decision tree [25], ensemble learning [26, 27], and neural networks [28, 29] can be employed within these approaches to impute missing values effectively.

It is crucial to consider the nature and structure of missing values in a dataset to determine the most appropriate data imputation method, including the mechanism, pattern, and ratio of missingness. This systematic review aims to identify the most imputation methods

based on the hypothesized mechanism, pattern, and ratio of missingness in clinical tabular datasets. According to this objective, the research questions come from as following:

> RQ1- What is the utilization pattern of different data imputation techniques in a clinical tabular dataset during the timeframe (2000–2023)?
> RQ2- What are common techniques used for data imputation, considering the mechanism of missing values, the pattern of missing values, and the missing ratios in a clinical tabular dataset.
> RQ3- What is the most suitable data imputation method, considering the nature and characteristics of the missing values in a tabular (structured) clinical dataset?

In general, answering this question can lead to the development of a clear guideline for data analysts to employ appropriate imputation techniques based on the conditions and characteristics of the missing values in a clinically structured dataset.

## Materials and methods

This systematic review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology guideline [30].

### Search strategy

A systematic literature search performed to identify relevant research articles concentrating on keywords related to missing values, imputation methods, and healthcare. The automated search encompassed four research databases: PubMed, Scopus, Web of Science, and IEEE Xplore. Search queries were crafted using keywords and logical operators to enhance search results. The search strategies employed detailed below. Specifics of the search strategies used for each database found in Additional File 1.

("Data gaps*" OR "Incomplete data*" OR "Unobserved values*" OR "Not recorded values*" OR "Data omissions*" OR "Nonexistent values*" OR "missing value*" OR "missing data*" OR "data missingness*").

AND

("Data imputation*" OR "Data interpolation*" OR "Data completion*" OR "Data augmentation*" OR "Data repair*" OR "imputation*" OR "imputing*").

AND

("Medical*" OR "Healthcare*" OR "Patient care*" OR "Clinical practice*" OR "clinical*" OR "Medical treatment*" OR "Patient management*" OR "Health services*" OR "medicine*" OR "health*").

### Inclusion criteria

Our inclusion criteria emphasized selecting primary studies conducted in medical or clinical fields, addressing missing values in clinical datasets, evaluating the proposed imputation methods in a clinical tabular dataset, and writing in English.

### Selection procedure and data extraction

The relevant articles retrieved on September 20, 2023, and managed using EndNote 21 software. Duplicate references eliminated through EndNote's function, followed by a manual check to ensure the thorough removal of duplicates. The selection process consisted of two phases: screening titles/abstracts and a full-text review by three reviewers (M.A., E.H., and H.T.). Two reviewers (M.A. and E.H.) independently screened the titles and abstracts according to the eligibility criteria. Any discrepancies were resolved through discussions with a third reviewer (H.T.). Key findings on missing data, imputation methods, and data imputation performance extracted. Initially, the reviewers independently screened the titles of the identified studies based on the exclusion criteria outlined in Table 1. Studies that passed the title-screening phase advanced to the abstract screening stage, where they evaluated against the specified exclusion criteria.

**Table 1** Exclusion criteria at the title and abstract level for screening primary studies in the first phase

| | |
|---|---|
| **exclusion criteria at the title level** | (i) Review (Systematic review/Meta-Analysis, Scoping Review, Narrative Review)<br>(ii) Protocol of Study, Erratum, Book/Section of Book<br>(iii) Proceeding and Conference Review<br>(iv) non-Relevant context<br>(v) Out of date (2000–2023) studies |
| **exclusion criteria at the abstract level** | (i) non-tabular static data (image data, signal data, voice data, microarray or genomics data, longitudinal or time series data)<br>(ii) Develop a prediction model (as part of the preprocessing of data analysis in observational or interventional studies)<br>(iii) Review (Systematic review/Meta-Analysis, Scoping Review, Narrative Review)<br>(iv) Comparison of imputation methods<br>(v) non-Relevant context<br>(vi) Not Abstract |

The included studies then progressed to the full-text review phase, where all reviewers independently evaluated them. Microsoft Excel used to review the full text, with reviewers recording their decisions and justifications for exclusion. During this phase, four key aspects of information extracted: the article's missingness mechanism, missingness pattern, missingness ratio, and the imputation methods employed.

### Quality assessment of studies and data analysis

According to data analysis, it is crucial to interpret the study results based on current literature evidence to derive reliable findings regarding the most effective imputation methods for handling missing data. Tabular datasets organized with observations as rows and features as columns. Understanding the structure and characteristics of missing values requires considering the missing data mechanism, pattern, and percentage of missing data in a dataset. This review study created an evidence map [31] to illustrate suitable imputation methods across various structures and characteristics of missing values in tabular datasets. Seven cases represent combinations of missingness mechanisms, patterns, and ratios. Based on the employed imputation methods, articles categorized into conventional statistical imputation methods, machine learning-based methods, deep learning-based methods [18], and hybrid methods. Imputation approaches further classified into single imputation, multiple imputation, a combination of both, and predictive imputation. The evidence map presented as a 7×4 matrix, with rows representing different states of missing data structures and columns containing the main categories of imputation methods. Data analysis performed using R (version 4.0.2).

## Results

### Study selection

Our search across four databases yielded 2,955 studies, of which 58 selected for analysis. The selection process outlined in the PRISMA flowchart in Fig. 1.

According to the flowchart, out of the 58 studies considered for full-text review and data extraction to address the research questions, 90% conducted between 2016 and 2023, while only 10% conducted from 2003 to 2015. This highlights the significance of missing values imputation in modeling and data analysis, particularly in the medical field. Figure 2 illustrates the trend of studies conducted between 2016 and 2023.

### Study characteristics

The study characteristics encompass data sources, clinical context, software for implementing data imputation

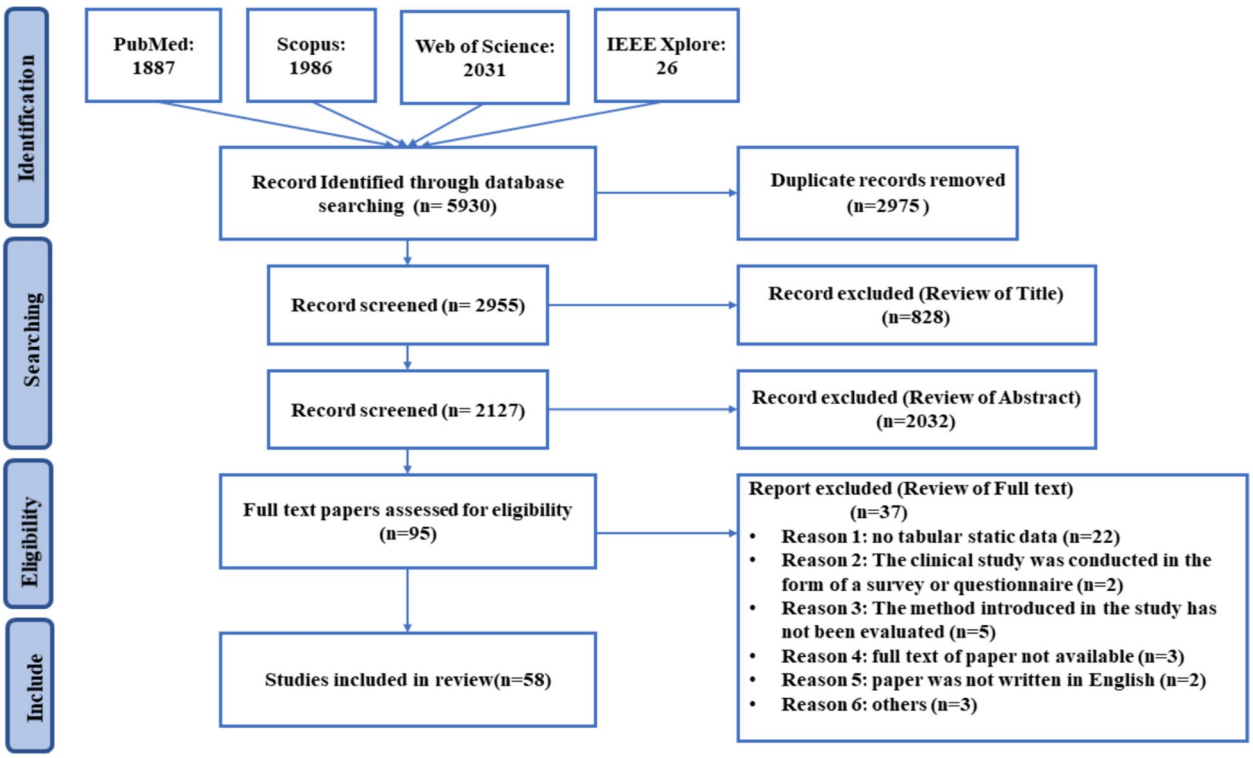


**Fig. 1** The flow diagram of Preferred Reporting Items for Systematic Review (PRISMA)

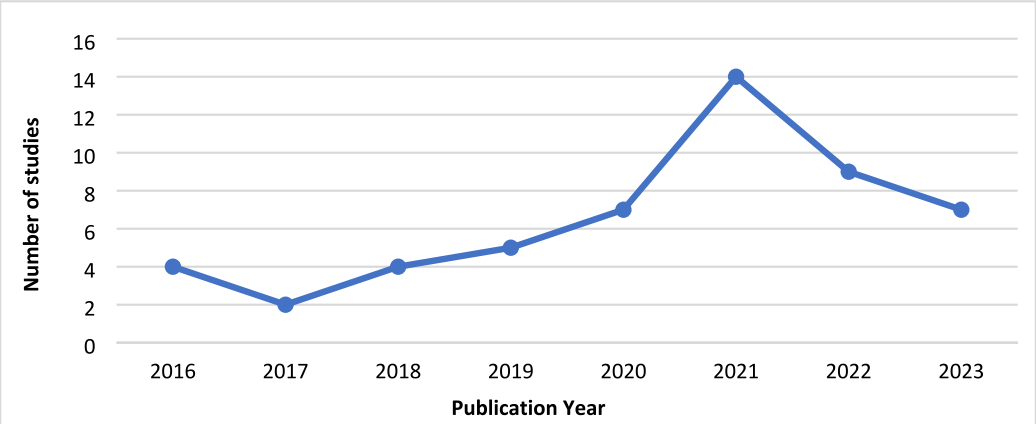


**Fig. 2** Trend of primary studies over the last 8 years

techniques, types of missing variables, patterns of missing values, and mechanisms of missing values extracted from each primary study, as depicted in Fig. 3. Specialized fields such as cardiology, diabetes, and cancer have notably addressed missing values in data pre-processing. Among the 58 studies, 27 did not provide information on missing value patterns (i.e., univariate, multivariate, and arbitrary patterns). Seven potential cases examined for the missing value mechanism, including a combination of MCAR, MAR, and MNAR mechanisms. Sixteen studies did not specify the mechanism of missing values.

### Primary studies analysis

In this review study, based on the classification in study [35] illustrated in Fig. 4, data imputation methods were grouped into four main categories Conventional statistical methods, machine learning-based methods, newly developed deep learning methods, and hybrid methods. A summary of the included studies presented in Additional File 2.

Figure 5 presents the evidence map illustrating the relationship between the primary categories of imputation methods and the various types of missing value structures. Among the 58 studies reviewed, 23 focused on imputation methods for the mechanisms, patterns, and ratios of missingness hypotheses. None addressed both the mechanism and pattern assumptions. Sixteen studies concentrated on the mechanism and ratio of missingness assumptions, five examined the pattern and ratio of missingness, one explored the mechanism hypothesis, another focused on the pattern assumption, and twelve investigated the ratio of missingness hypothesis.

The majority of studies (48%, 28/58) employed the "predictive" approach, while (36%, 21/58) utilized a multiple approach for data imputation methods. In terms of classifying data imputation methods, (41%, 24) studies relied on conventional statistical methods, (24%, 14) studies on machine learning methods, (24%, 14) studies on hybrid imputation methods, and (11%, 6) studies on deep learning methods for assigning data values. Further details on the data imputation methods used in the studies can be found in Fig. 4 and Additional File 3.

The following subsections include algorithms used in each main category of data imputation methods for seven distinct cases of structural characteristics of missing values.

#### *Mechanism, pattern, ratio of missingness*

Among the 58 primary studies, 23 studies [3, 36, 38, 42–44, 46, 50, 51, 55–59, 61, 70, 71, 74, 75, 81, 82, 85, 87] examined the mechanisms, patterns, and ratios of the missingness hypothesis. Most of these studies (60%, 14/23) employed conventional statistical methods (i.e., regression [36, 70], multiple imputations by chained equations (MICE) [42, 3], miss ranger [46], logistic regression (logit) [50, 51], joint multiple imputations (JMI), conditional multiple imputations (CMI) [71, 74], matrix completion methods [81], bias-corrected multiple imputations (BCMI) [82], probabilistic principal component analysis (PPCA) [36], and multilevel and stratified imputation methods [55, 87]) for imputing missing values. Thirteen percent (i.e., 3/23) of the studies utilized machine learning-based imputation methods, including support vector machine (SVM) [38], ensemble learning (EL) [58], Bayesian classification and regression trees (CART), and Bayesian additive regression trees (BART) [85]. Another thirteen percent (i.e., 3/23) of studies employed hybrid imputation methods; for instance, study [44] used fuzzy principal component analysis (FPCA), SVM, and fuzzy c-means (FCM) clustering methods, while study [59] applied utility-based regression (UBR) with synthetic minority over-sampling technique for

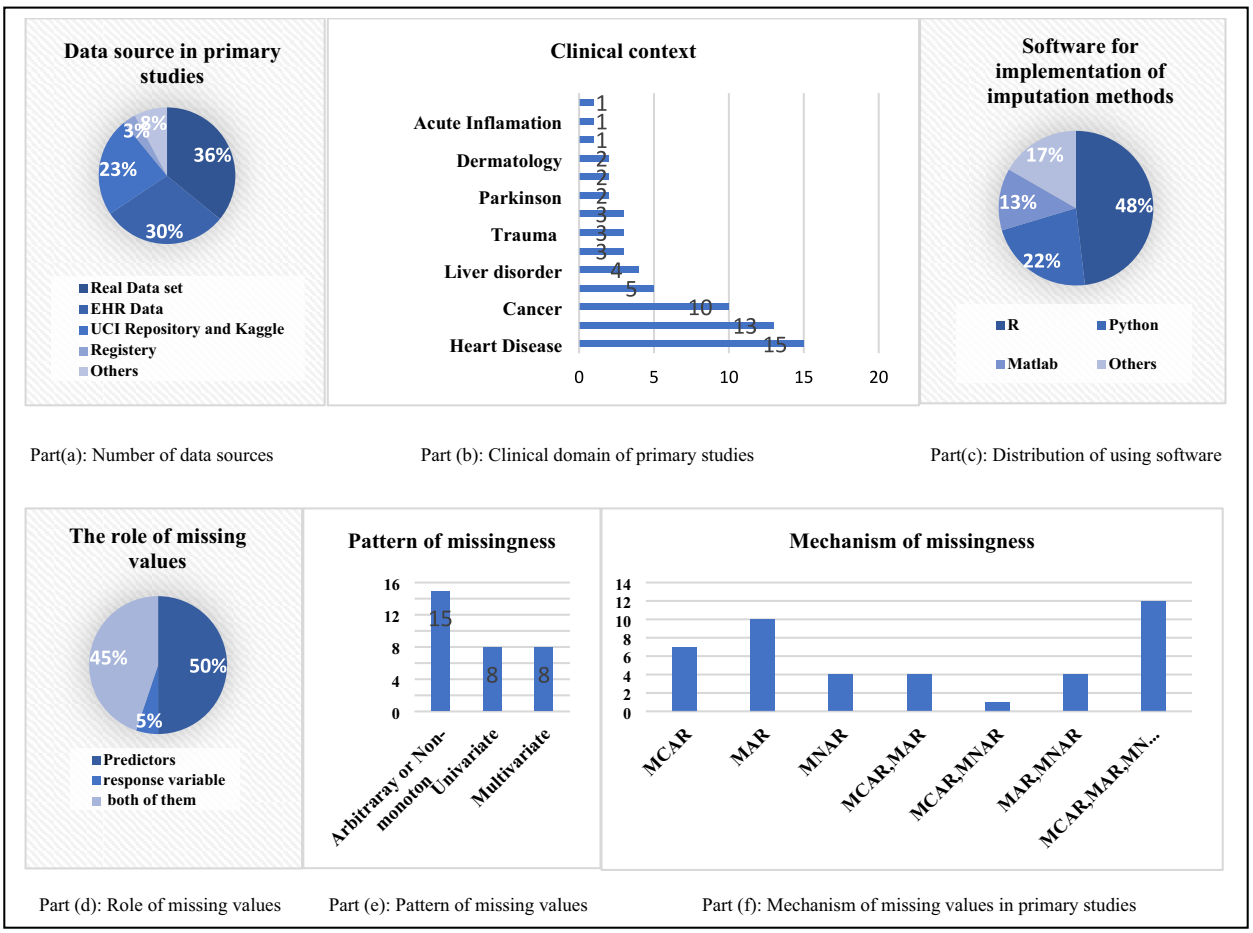


**Fig. 3** Part (**a**) shows the percentage of primary studies based on their data sources. Part (**b**) displays the distribution of clinical contexts, emphasizing the importance of managing missing data in medical decision-making processes. Part (**c**) illustrates the software used for implementing data imputation techniques. Part (**d**) presents the types of missing variables extracted from each study. In part (**e**) distribution of missing value patterns is shown. Part (**f**) depicts the frequency of missing value mechanisms mentioned in the primary studies

regression (SMOTER). Finally, thirteen percent (i.e., 3/23) of studies implemented deep learning-based imputation methods, including clinical condition generative adversarial network (CCGAN) [43, 56] and partial multiple imputation with variational auto-encoders (PMI-VAE) [75]. In this category, 65% of the studies utilized simulation approaches to implement algorithms for data imputation.

### *Mechanism and pattern hypothesis*

Among the 58 included studies, none had been conducted considering both the mechanism and the pattern conditions in the missing data.

### *Mechanism and ratio of missingness hypothesis*

Among the 58 included studies, 16 studies [32, 35, 37, 39, 41, 45, 48, 49, 53, 54, 62, 64, 65, 73, 78, 86] addressed the missing value hypothesis in the mechanisms and ratios of messiness. This category of missing values, comprising 50% of studies [32, 37, 41, 45, 64, 65, 78, 86], includes the use of conventional statistical methods for data imputation, such as predictive mean matching (PMM) [32], Bayesian ridge regression (BRR) model [37, 65, 86], MICE [41, 45], missing Gaussian processes (GP) [64], and joint modeling multiple imputation, as well as full conditional specification multiple imputation [78]. Four studies (i.e., 25%) [35, 48, 49, 53] focused on imputation methods based on machine learning methods, including extremely randomized trees (Extra Trees) [35], sequential random forest method [39], distance threshold nearest neighbor imputation method [48], correlation-weighted nearest neighbor imputation, and correlation-weighted regression imputation [53]. Additionally, four studies (i.e., 25%) [49, 54, 62, 73] examined hybrid imputation methods, such as k-means clustering with purity-based k nearest neighbor (KNN) imputation and distance

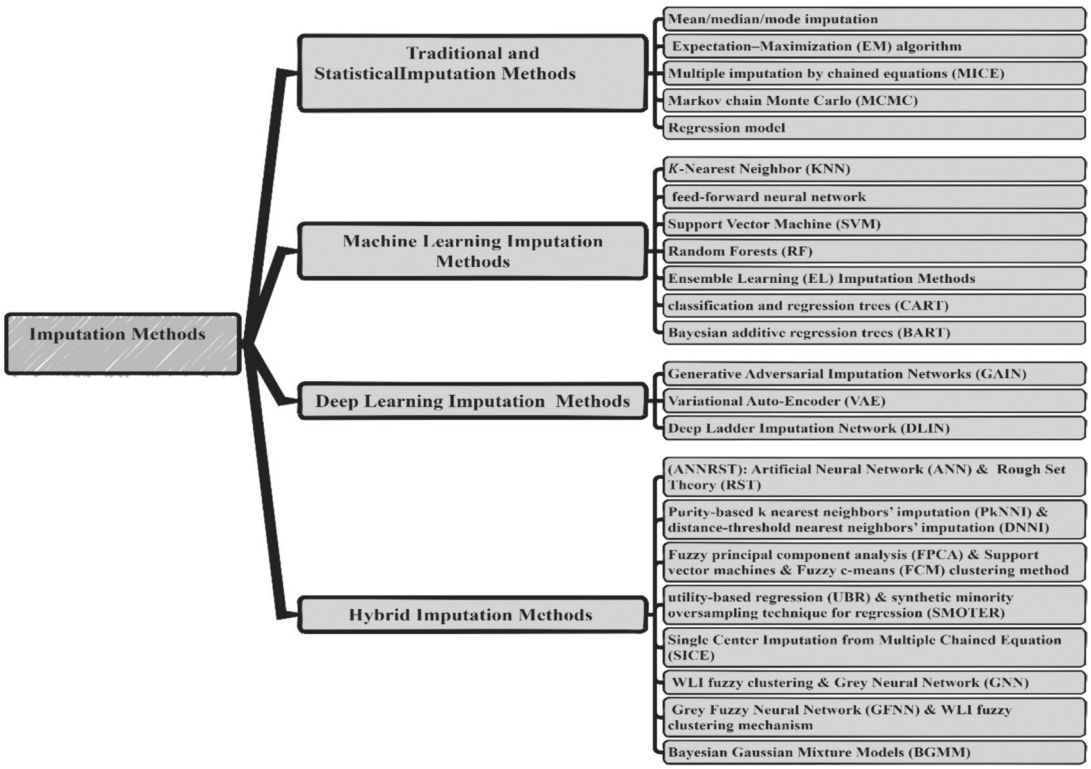


**Fig. 4** Main categorization of imputation methods

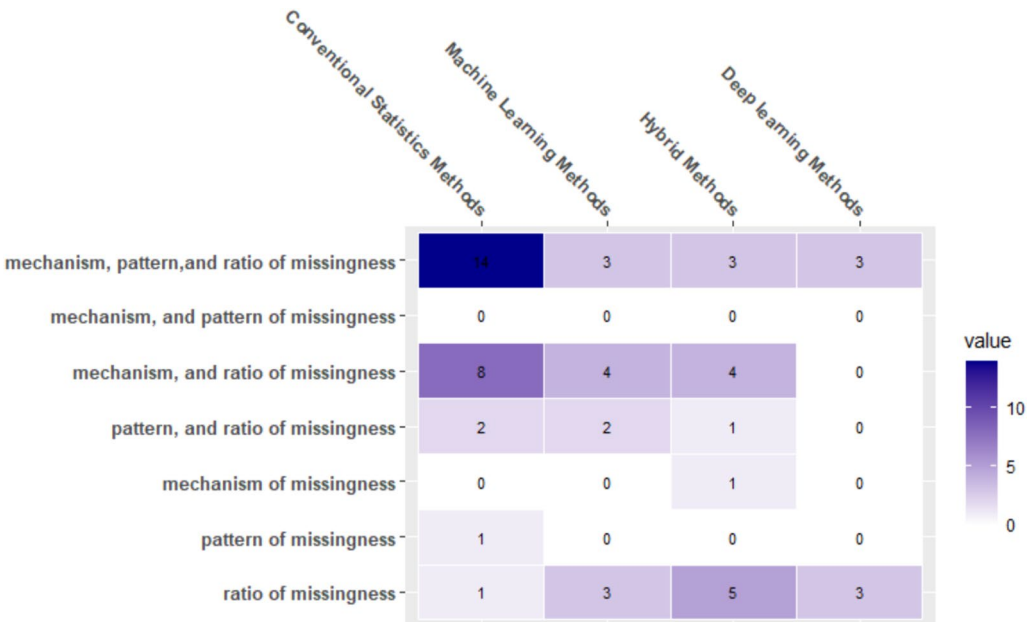


**Fig. 5** Evidence map: The evidence map illustrates the main categorized imputation methods, which are classified into four groups, and various types of structures for missing values assumed in seven cases ((mechanism, pattern, ratio of missing values), (mechanism, pattern), (mechanism, ratio of missing values), (pattern, ratio of missing values), (mechanism), (pattern), (ratio of missing values))

threshold nearest neighbors imputation [49], random forest combined with the expectation maximization (EM) algorithm and (KNN) [84], multilevel multiple correspondence analysis and multilevel factorial analysis (MFA) [62], (KNN), hot-deck, Markov chain Monte Carlo, (MICE), and (EM) [73]. In this category, 75% of the studies employed simulation approaches to implement algorithms for data imputation.

#### *Pattern and Ratio of Missingness*

For this category (8%, 5), studies [52, 60, 63, 72, 76] examined the patterns and ratios of the missingness hypothesis concerning missing values. Within this group, two studies [52, 72] utilized conventional statistical methods, including linear regression techniques [52]. The interpolation method [74] employed, while two studies [60, 63] applied machine-learning algorithms (slap swarm algorithm (SSA) [60] and the instance-based cluster imputation algorithm (INS-CLUS-IMPUTE)) [63] to predict missing values. One study [76] implemented a hybrid method Bayesian-Gaussian mixture models (BGMM) for imputing missing values. In this category, 60% of the studies adopted a simulation approach to implement algorithms for data imputation.

#### *Mechanism Hypothesis*

For this category of missing value structure, only one study [47] examined the hypothesis of the missing data mechanisms in a tabular dataset using a hybrid method for missing value imputation.

#### *Pattern Hypothesis*

For this category of missing value structure, only one study [33] investigated the hypothesis of the missing data patterns using a conventional statistical method (linear regression with half values of random error) for missing value imputation.

#### *Ratio of missingness hypothesis*

Among 58 primary studies, (20%, 12) studies [16, 34, 40, 66–69, 77, 79, 80, 83, 84] were included in this category. One study [16] employed a conventional statistical method (low-rank approximation-based imputation) to address missing values. Studies [40, 80, 83] utilized machine learning-based methods for imputation. These algorithms include (EL) algorithms in study [40], a cluster-based imputation (CLUSTIMP) method in study [80], and least squares support vector machine (LS-SVM) in study [83]. Studies [34, 67–69, 79] applied hybrid methods for imputing missing values. More details about the algorithms used are as follows: study [34] combined rough set theory (RST) with artificial neural network (ANN) hybridization for missing data imputation; study [67] implemented single center imputation from multiple chained equations (SICE) and MICE; studies [68, 69] employed WLI Fuzzy Clustering and fuzzy grey neural network (FGNN); and study [79] integrated MICE, KNN, and optimal mean and mode imputation. Finally, studies [66, 77, 84] in this category applied a deep learning-based method (auto-encoder to drive a deep learning architecture) for imputing missing values in a tabular dataset. In this category, 58% of the studies utilized a simulation approach to implement algorithms for data imputation.

#### Quality assessment

The assessment process involves assigning a score from zero to one to each primary study based on the Quality Assessment Criteria (QAC). A score of one indicates that the study adequately addresses the QAC question, while a score of 0.75 signifies partial acceptability. Incomplete answers receive a score of 0.5, and a score of zero assigned if the QAC question not addressed. The total score for each study calculated by summing the individual QAC question scores, providing an overall evaluation of the study's quality and adherence to the assessment criteria. Upon completing the quality assessment for each primary study, it noted that the total score of the selected primary studies exceeded 60% [88, 89], representing 84% compliance with each QAC as detailed in Table 2. This observation indicates that the primary studies provide sufficient information to address the research questions.

## Discussion

This systematic review aims to identify the most suitable method for imputing missing data in a static tabular dataset. This study designed around three research questions. The first research question seeks to provide an overview of recent trends in various imputation methods that address missing values in structured clinical datasets. The second research question focuses on identifying common imputation methods and categorizing them based on the combination of different modes derived from three underlying assumptions: mechanism, pattern, and ratio of missing values in a structured clinical dataset. Third research question determines the most appropriate imputation method according to the characteristics of the missing values. To answer the first research question, reference information—including title, publication year, publication source, study purpose, and applied method—extracted from each primary study. The response to this question illustrated in Fig. 2. In addressing the second question, information such as dataset source, data types, missingness mechanism, missingness pattern, the ratio of missing values, study design, the role of missing values in the variables, and the imputation method extracted from each study. Details related to this question presented in

**Table 2** The quality assessment of primary studies

| Research Questions | Quality Assessment Criteria | Answering Score (0, 0.5, 0.75, 1) | Total Score |
|---|---|---|---|
| **RQ1**<br>What is the utilization pattern of different data imputation techniques? | **QAC1**<br>Introduce imputation methods given some hypothesis for missing values: **58** | 58*1 | 58 studies (100%) |
| **RQ2**<br>What are common techniques used for data imputation, considering the mechanism of missing values, the pattern of missing values, and the missing ratios? | **QAC2**<br>(Mechanism, pattern, ratio of missing values): **23**<br>((Mechanism and pattern), (Mechanism and ratio of missing values), (Pattern and ratio of missing values)): **21**<br>((Mechanism), (Pattern), (Ratio of missing values)): **14** | 23*1<br>21*0.75<br>14*0.5 | 45 studies (77%) |
| **RQ3**<br>What is the most suitable data imputation method, considering the nature and characteristics of the missing values? | **QAC3**<br>Simulation study: **36**<br>Prediction model: **16**<br>Other approaches: **6** | 36*1<br>16*0.5<br>6*0 | 44 studies (75%) |

"Primary studies analysis" Section. Finally, to answer the third research question regarding the most suitable imputation method, 58 studies examined to determine whether the introduced imputation method based on a simulation approach. In the simulation method, all scenarios verified and evaluated, and then the most appropriate method proposed based on the problem conditions. Table 2 presents the evaluation results for each of the research questions in this study, based on Quality Assessment Criteria relevant to each question. Therefore, it is essential to focus on the mechanism and pattern of missing values in accurately diagnosing the data assignment method. The analysis of 58 studies highlights the significance of considering these factors when selecting an imputation method, resulting in improved data quality. Several recent review studies have examined various data imputation methods [35, 89–93] with some concentrating on specific categories such as machine learning [89–91] or deep learning [18, 92]. In contrast, this review encompasses a wide range of imputation methods, including conventional statistical methods, machine learning, and deep learning methods. It serves as a comprehensive guide for researchers in selecting a suitable imputation method during data preprocessing by considering the structural attributes of missing values.

We created an evidence map by integrating assumptions about the structure of missing data and various imputation methods to identify the most appropriate approach. The assumptions considered three primary factors: mechanism, pattern, and ratio of missingness. Several imputation techniques, including conventional statistical, machine learning, hybrid, and deep learning, used to construct the evidence map. Figure 3 organizes the resulting matrix (7×4) as follows: The assumptions (mechanism, pattern, and ratio of missingness) presented in the first row. The second, third, and fourth rows display the binary combinations of structural features (mechanism, pattern), (mechanism, ratio of missingness), and (pattern, ratio of missingness), respectively. The fifth row corresponds to the mechanism, the sixth row corresponds to the pattern, and the seventh row corresponds to the ratio of missingness. Conventional statistical methods, machine learning methods, hybrid methods, and deep learning methods are included in the columns of the evidence map matrix. This structured approach enables a comprehensive examination of various combinations of missing data attributes and imputation methods. It provides valuable insights for researchers in selecting the most suitable strategy for imputing missing data in their datasets.

Based on this evidence map, it concluded that Case (i) (mechanism, pattern, and ratio of missingness) exhibited the highest application of statistical methods for imputing missing data. Cases (ii) (mechanism and pattern), (v) (mechanism), and (vi) (pattern) demonstrated the lowest utilization in missing data imputation. The default case (vii) (ratio of missingness) alone attracted more interest than the case (v) mechanism and the case (vii) pattern. Case (iii) (mechanism and ratio of missingness) garnered more attention than cases (ii) and (iv) in the combination of two cases of the structural features of missing values. Among the seven structural features assumed for missing values, case (i), case (iii), and case (vii) received greater focus than others did. Conventional statistical methods predominantly employed for imputing missing values related to the case (i). Statistical, machine learning, and hybrid methods used to address missing values associated with the case (iii). Case (vii) applied machine learning, deep learning, and hybrid methods.

Statistical methods considered sensitive and essential specific prerequisites for modeling, whereas learning-based methods recognized for their robustness and

reduced sensitivity to such prerequisites. In traditional statistical methods, the focus is on the mechanism, pattern, and ratio of missingness. Machine learning methods emphasize the mechanism and ratio of missingness, while deep learning and hybrid methods prioritize the missingness ratio. As a result, findings derived from the analysis of 58 studies well supported by the distinction between statistical and learning-based methods in addressing missing data. Statistical methods deemed sensitive and require specific prerequisites for modeling, while learning-based methods, including machine learning and deep learning, known for their robustness and less influenced by the prerequisites typically needed for modeling. Moreover, in traditional statistical methods, the key structural feature of interest encompasses a case (i) (mechanism, pattern, and ratio of missingness). In contrast, machine-learning methods concentrate on the case (iii) (mechanism and ratio of missingness). Deep learning and hybrid methods emphasize the case (vii) (missingness ratio) structural feature over other assumed structural features. This distinction underscores the strengths and considerations of various approaches when managing missing data in a dataset. Researchers can gain from understanding these nuances to choose the most suitable method based on their data characteristics and research objectives.

Among the 58 studies reviewed, the simulation approach used by a notable percentage: 62% of studies used conventional statistical methods, 66% of studies employing machine-learning methods, 64% of studies incorporating hybrid methods, and 66% of studies applying deep learning methods. Based on this information, it is advisable to conduct a simulation study to confirm that the proposed imputation method is indeed the most appropriate approach, considering the structural characteristics of the missing values. A simulation study can enhance the robustness of the proposed imputation method in practical applications.

Consequently, this review emphasizes that knowing about the structure of missing values is essential when applying conventional statistical techniques. This involves examining the mechanism, pattern, and ratio of missing values, as these factors significantly affect the model's performance in estimating missing values because of their high sensitivity. Conversely, machine learning and deep learning techniques, recognized for their robustness, may require less focus on the structural characteristics of missing values compared to the default assumptions regarding the mechanism, pattern, and ratio of missing data. Neglecting these structural aspects may not considerably affect the model's performance in predicting missing values.

## Conclusions

This systematic review emphasizes the importance of understanding the structure and characteristics of missing data in clinical datasets when selecting appropriate imputation methods. The analysis of 58 studies revealed that conventional statistical methods are most effective when considering the mechanisms, patterns, and ratios of missing values, while machine learning and deep learning techniques are more robust and often focus primarily on the missingness ratio.

The review created an evidence map that links missing data assumptions to suitable imputation methods, highlighting the prevalence of simulation-based approaches in validating these techniques. The findings suggest that researchers should tailor their imputation strategies based on the specific characteristics of their datasets, particularly when using conventional statistical methods. Ultimately, employing simulation methods can enhance data quality and improve the accuracy of medical decision-making, underscoring the need for methodologically sound imputation practices in clinical research. A conceptual framework as a road map on how to choose the most appropriate imputation method according to characteristics of missing values is essential in future literature.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02310-6.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

**Acknowledgements**
Not applicable.

**Authors' contributions**
M.A. conducted data extraction, analysis, and interpretation of data, and drafted and edited the manuscript. E.H. screened relevant studies and performed data extraction. H.T. collaborated on designing the data extraction checklist, conducted a pilot check on data extraction for some papers, offered feedback on all manuscript versions, substantively revised the manuscript, and supervised this project. All authors read and approved the final manuscript.

**Funding**
The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

**Availability of data and materials**
Not applicable.

**Data availability**
No datasets were generated or analysed during the current study.

## Declarations

### Ethics approval and consent to participate

This study was assessed by the research council of Mashhad University of Medical Sciences (Reference Number: IR.MUMS.REC.1402.069). The study was approved because no identifying data have been reported.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.

Received: 6 April 2024 Accepted: 19 August 2024
Published online: 28 August 2024

### References

1. Little RJ, Rubin DB. Statistical Analysis with Missing Data, vol. 793. Hoboken, NJ, USA: Wiley; 2019.
2. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92.
3. Galimard JE, Chevret S, Protopopescu C, Resche-Rigon M. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. Stat Med. 2016;35(17):2907–20.
4. Miettinen OS. Theoretical epidemiology: principles of occurrence research in medicine. In Theoretical epidemiology: principles of occurrence research in medicine 1985 (pp. xxii-359).
5. Humphries M. Missing Data & How to Deal: an overview of missing data. Popul Res Cent. 2013; 45.
6. Li T, Hutfless S, Scharfstein DO, Daniels MJ, Hogan JW, Little RJA, et al. Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus. J Clin Epidemiol. 2014;67:15–32. https://doi.org/10.1016/j.jclinepi.2013.08.013.
7. Suthar B, Patel H, Goswami A. A survey: classification of imputation methods in data mining. Int J Emerg Technol Adv Eng. 2012;2(1):309–12.
8. Graham JW, Cumsille PE, Elek-Fisk E. Methods for handling missing data. Handbook of psychology. 2003:87–114.
9. Buuren SV. Flexible Imputation of Missing Data. Chapman & Hall CRC. 2018. https://doi.org/10.1201/9780429492259.
10. Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014;1(2):293–314.
11. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009;338.
12. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. Am J Epidemiol 2014; 179:764–74? https://doi.org/10.1093/aje/kwt312.
13. Palanivinayagam A, Damaševičius R. Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods. Information. 2023;14(2):92.
14. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17:520–5.
15. Luis J, Gomez S, Vidal ARF, Verleysen M. K nearest neighbors with mutual information for simultaneous classification and missing data imputation. Neurocomputing. 2009;72(7–9):1483–93.
16. Khan SI, Hoque AS. SICE: an improved missing data imputation technique. Journal of Big Data. 2020;7(1):1–21.
17. Jain R, Xu W. Dynamic model updating (DMU) approach for statistical learning model building with missing data. BMC Bioinformatics. 2021;22(1):1–5.
18. Sun Y, Li J, Xu Y, Zhang T, Wang X. Deep learning versus conventional methods for missing data imputation: A review and comparative study. Expert Systems with Applications. 2023:120201
19. Sherwood B, Wang L, Zhou XH. Weighted quantile regression for analyzing health care cost data with missing covariates. Stat Med. 2013;32(28):4967–79.
20. Crambes C, Henchiri Y. Regression imputation in the functional linear model with missing values in the response. Journal of Statistical Planning and Inference. 2019;201:103–19.
21. Andridge RR, Little RJ. A review of hot deck imputation for survey non-response. Int Stat Rev. 2010;78(1):40–64.
22. Sullivan D, Andridge R. A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. Comput Stat Data Anal. 2015;82:173–85.
23. Delalleau O, Courville A, Bengio Y. Efficient EM training of Gaussian mixtures with missing data. arXiv preprint arXiv:1209.0521. 2012 Sep 4.
24. Pelckmans K, De Brabanter J, Suykens JA, De Moor B. Handling missing values in support vector machine classifiers. Neural Netw. 2005;18(5–6):684–92.
25. Twala B. An empirical comparison of techniques for handling incomplete data using decision trees. Appl Artif Intell. 2009;23(5):373–405.
26. Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Mach Learn. 1999;36:105–39.
27. Whitehead M, Yaeger L. Sentiment mining using ensemble classification models. InInnovations and advances in computer sciences and engineering 2010 (pp. 509–514). Springer Netherlands.
28. Gupta A, Lam MS. Estimating missing values using neural networks. Journal of the Operational Research Society. 1996;47:229–38.
29. Sharpe PK, Solly RJ. Dealing with missing values in neural network-based diagnostic systems. Neural Comput Appl. 1995;3:73–7.
30. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group* T. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of internal medicine. 2009; 151(4):264–9.
31. Liu N, Chee ML, Niu C, Pek PP, Siddiqui FJ, Ansah JP, Matchar DB, Lam SS, Abdullah HR, Chan A, Malhotra R. Coronavirus disease 2019 (COVID-19): an evidence map of medical literature. BMC Med Res Methodol. 2020;20:1–1.
32. Abassi RA, Msengwa AS. Classification of breast cancer recurrence based on imputed data: a simulation study. BioData Mining. 2022;15(1):30.
33. Ahmad A, Mohamed HH. The enhancement of linear regression algorithm in handling missing data for medical data set.
34. Setiawan NA, Venkatachalam PA, Ahmad Fadzil MH. A knowledge discovery from incomplete coronary artery disease datasets using a rough set. International Journal of Medical Engineering and Informatics. 2011;3(1):60–77.
35. Alabadla M, Sidi F, Ishak I, H, Affendey L, Hamdan H. A. Extralmpute: A Novel Machine Learning Method for Missing Data Imputation. Journal of Advances in Information Technology. 2022; 13(5): 470–476. https://doi.org/10.12720/jait.13.5.470-476
36. Alade OA, Selamat A, Sallehuddin R. The Effects of Missing Data Characteristics on the Choice of Imputation Techniques. Vietnam Journal of Computer Science. 2020;7(02):161–77.
37. Algarni A, Ragab M, Alamri W, Mostafa SM. Towards Improving Predictive Statistical Learning Model Accuracy by Enhancing Learning Technique. Comput Syst Sci Eng. 2022;42(1):303–18.
38. Almasinejad P, Golabpour A, Mollakhalili Meybodi MR, Mirzaie K, Khosravi A. A dynamic model for imputing missing medical data: a multiobjective particle swarm optimization algorithm. J Healthcare Eng. 2021; 2021.
39. Alsaber A, Al-Herz A, Pan J, AL-Sultan AT, Mishra D, KRRD Group. Handling missing data in a rheumatoid arthritis registry using a random forest approach. Int J Rheumatic Dis. 2021;24(10):1282–93.
40. Batra S, Khurana R, Khan MZ, Boulila W, Koubaa A, Srivastava P. A Pragmatic Ensemble Strategy for Missing Values Imputation in Health Records. Entropy. 2022;24(4):533.
41. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: data analysis. JMIR Med Inform. 2018;6(1): e8960.
42. Beesley LJ, Taylor JM. Accounting for not-at-random missingness through imputation stacking. Stat Med. 2021;40(27):6118–32.
43. Bernardini M, Doinychko A, Romeo L, Frontoni E, Amini MR. a novel missing data imputation approach based on clinical conditional Generative

Adversarial Networks applied to EHR datasets. Comput Biol Med. 2023;163: 107188.
44. Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. Am J Epidemiol. 2010;172(9):1070–6.
45. Carreras G, Miccinesi G, Wilcock A, Preston N, Nieboer D, Deliens L, Groenvold M, Lunder U, van der Heide A, Baccini M. Missing not at random in end-of-life care studies: multiple imputation and sensitivity analysis on data from the ACTION study. BMC Med Res Methodol. 2021;21:1–2.
46. Casiraghi E, Wong R, Hall M, Coleman B, Notaro M, Evans MD, Tronieri JS, Blau H, Laraway B, Callahan TJ, Chan LE. A method for comparing multiple imputation techniques: A case study on the US national COVID cohort collaborative. J Biomed Inform. 2023;139: 104295.
47. Chen J, Hunter S, Kisfalvi K, Lirio RA. A hybrid approach of handling missing data under different missing data mechanisms: VISIBLE 1 and VARSITY trials for ulcerative colitis. Contemp Clin Trials. 2021;100: 106226.
48. Cheng CH, Chang JR, Huang HH. A novel weighted distance threshold method for handling medical missing values. Comput Biol Med. 2020;122: 103824.
49. Cheng CH, Huang SF. A novel clustering-based purity and distance imputation for handling medical data with missing values. Soft Comput. 2021;25(17):11781–801.
50. Choi YJ, Nam CM, Kwak MJ. Multiple imputation techniques applied to appropriateness ratings in cataract surgery. Yonsei Med J. 2004;45(5):829–37.
51. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol. 2003;56(1):28–37.
52. Cleophas EP, Cleophas TJ. Clinical research: A novel approach to regression substitution for handling missing data. Am J Ther. 2013;20(5):514–9.
53. Curioso I, Santos R, Ribeiro B, Carreiro A, Coelho P, Fragata J, Gamboa H. Addressing the curse of missing data in clinical contexts: A novel approach to correlation-based imputation. Journal of King Saud University-Computer and Information Sciences. 2023;35(6): 101562.
54. Dekermanjian JP, Shaddox E, Nandy D, Ghosh D, Kechris K. Mechanism-aware imputation: a two-step approach in handling missing values in metabolomics. BMC Bioinformatics. 2022;23(1):179.
55. DiazOrdaz K, Kenward MG, Gomes M, Grieve R. Multiple imputation methods for bivariate outcomes in cluster randomized trials. Stat Med. 2016;35(20):3482–96.
56. Dong W, Fong DY, Yoon JS, Wan EY, Bedford LE, Tang EH, Lam CL. Generative adversarial networks for imputing missing data for big data clinical research. BMC Med Res Methodol. 2021;21:1.
57. Dzulkalnine MF, Sallehuddin R. Missing data imputation with fuzzy feature selection for diabetes dataset. SN Applied Sciences. 2019;1(4):362.
58. Ferri P, Romero-Garcia N, Badenes R, Lora-Pablos D, Morales TG, de la Cámara AG, García-Gómez JM, Sáez C. Extremely missing numerical data in Electronic Health Records for machine learning can be managed through simple imputation methods considering informative missingness: A comparative of solutions in a COVID-19 mortality case study. Comput Methods Programs Biomed. 2023;242: 107803.
59. Haliduola HN, Bretz F, Mansmann U. Missing data imputation using utility-based regression and sampling approaches. Comput Methods Programs Biomed. 2022;226: 107172.
60. Hassan GS, Ali NJ, Abdulsahib AK, Mohammed FJ, Gheni HM. A missing data imputation method based on the Salp swarm algorithm for diabetes disease. Bulletin of Electrical Engineering and Informatics. 2023;12(3):1700–10.
61. Hegde H, Shimpi N, Panny A, Glurich I, Christie P, Acharya A. MICE vs PPCA: Missing data imputation in healthcare. Inform Med Unlocked. 2019;17: 100275.
62. Husson F, Josse J, Narasimhan B, Robin G. Imputation of mixed data with multilevel singular value decomposition. J Comput Graph Stat. 2019;28(3):552–66.
63. Ilango P, Vijayakumar K, Rajasekhara BM. Instance-driven clustering for the imputation of missing data in KDD. International Journal of Communication Networks and Distributed Systems. 2014;12(1):69–81.
64. Jafrasteh B, Hernández-Lobato D, Lubián-López SP, Benavente-Fernández I. Gaussian processes for missing value imputation. Knowl-Based Syst. 2023;273: 110603.
65. Jain R, Xu W. Dynamic model updating (DMU) approach for statistical learning model building with missing data. BMC Bioinformatics. 2021;22(1):221.
66. Jolani S. Hierarchical imputation of systematically and sporadically missing data: an approximate Bayesian approach using chained equations. Biom J. 2018;60(2):333–51.
67. Kabir S, Farrokhvar L. Non-linear missing data imputation for healthcare data via index-aware autoencoders. Health Care Manag Sci. 2022;25(3):484–97.
68. Kim KH, Kim KJ. Missing-data handling methods for lifelong-based wellness index estimation: Comparative analysis with panel data. JMIR Med Inform. 2020;8(12): e20597.
69. Kuppusamy V, Paramasivam I. Integrating WLI fuzzy clustering with grey neural network for missing data imputation. International Journal of Intelligent Enterprise. 2017;4(1–2):103–27.
70. Kuppusamy V, Paramasivam I. Grey Fuzzy Neural Network-Based Hybrid Model for Missing Data Imputation in Mixed Database. International Journal of Intelligent Engineering & Systems. 2017; 10(2).
71. Lee JH, Huber JC Jr. Evaluation of multiple imputations with large proportions of missing data: how much is too much? Iran J Public Health. 2021;50(7):1372.
72. Ma Y, Zhang W, Lyman S, Huang Y. The HCUP SID imputation project: improving statistical inferences for health disparities research by imputing missing race data. Health Serv Res. 2018;53(3):1870–89.
73. Miao SD, Li SQ, Zheng XY, Wang RT, Li J, Ding SS, Ma JF. Missing data interpolation of Alzheimer's disease based on column-by-column mixed mode. Complexity. 2021;2021:1–6.
74. Nadimi-Shahraki MH, Mohammadi S, Zamani H, Gandomi M, Gandomi AH. A hybrid imputation method for multi-pattern missing data: A case study on type II diabetes diagnosis. Electronics. 2021;10(24):3167.
75. Nijman SW, Groenhof TK, Hoogland J, Bots ML, Brandjes M, Jacobs JJ, Asselbergs FW, Moons KG, Debray TP. Real-time imputation of missing predictor values improved the application of prediction models in daily practice. J Clin Epidemiol. 2021;134:22–34.
76. Pereira RC, Abreu PH, Rodrigues PP. Partial multiple imputations with variational autoencoders: tackling not at randomness in healthcare data. IEEE J Biomed Health Inform. 2022;26(8):4218–27.
77. Pezoulas VC, Tachos NS, Olivotto I, Barlocco F, Fotiadis DI. A "smart" Imputation Approach for Effective Quality Control across Complex Clinical Data Structures. In2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) 2022. (pp. 1049–1052). IEEE.
78. Phung S, Kumar A, Kim J. A deep learning technique for imputing missing healthcare data. In2019 41st annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2019. (pp. 6513–6516). IEEE.
79. Quartagno M, Carpenter JR. Multiple imputation for discrete data: Evaluation of the joint latent normal model. Biom J. 2019;61(4):1003–19.
80. Rani P, Kumar R, Jain A. HIOC: a hybrid imputation method to predict missing values in medical datasets. International Journal of Intelligent Computing and Cybernetics. 2021;14(4):598–616.
81. Shobha K, Savarimuthu N. Clustering-based imputation algorithm using unsupervised neural network for enhancing the quality of healthcare data. J Ambient Intell Humaniz Comput. 2021;12(2):1771–81.
82. Sportisse A, Boyer C, Josse J. Imputation and low-rank estimation with missing not at random data. Stat Comput. 2020;30(6):1629–43.
83. Tomita H, Fujisawa H, Henmi M. A bias-corrected estimator in multiple imputation for missing data. Stat Med. 2018;37(23):3373–86.
84. Wang G, Lu J, Choi KS, Zhang G. A transfer-based additive LS-SVM classifier for handling missing data. IEEE transactions on cybernetics. 2018;50(2):739–52.
85. Xu D, Hu PJ, Huang TS, Fang X, Hsu CC. A deep learning–based, unsupervised method to impute missing values in electronic health records for improved patient management. J Biomed Inform. 2020;111: 103576.
86. Xu D, Daniels MJ, Winterstein AG. Sequential BART for imputation of missing covariates. Biostatistics. 2016;17(3):589–602.
87. Zang H, Kim HJ, Huang B, Szczesniak R. Bayesian causal inference for observational studies with missingness in covariates and outcomes. Biometrics. 2023;79(4):3624–36.

88. Yang L, Zhang H, Shen H, Huang X, Zhou X, Rong G, Shao D. Quality assessment in systematic literature reviews: A software engineering perspective. Inf Softw Technol. 2021;130: 106397.
89. Alabadla M, Sidi F, Ishak I, Ibrahim H, Affendey LS, Ani ZC, Jabar MA, Bukar UA, Devaraj NK, Muda AS, Tharek A. Systematic review of using machine learning in imputing missing values. IEEE Access. 2022;10:44483–502.
90. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. Journal of Big Data. 2021;8:1–37.
91. Thomas T, Rajabi E. A systematic review of machine learning-based missing value imputation techniques. Data Technologies and Applications. 2021;55(4):558–85.
92. Liu M, Li S, Yuan H, Ong ME, Ning Y, Xie F, Saffari SE, Shang Y, Volovici V, Chakraborty B, Liu N. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. Art Intel Med. 2023:102587.
93. Setiawan I, Gernowo R, Warsito B. A Systematic Literature Review on Missing Values: Research Trends, Datasets, Methods, and Frameworks. In E3S Web of Conferences 2023. (Vol. 448, p. 02020). EDP Sciences.

## Publisher's Note