

1 Title:

2 Enhancing data integrity in Electronic Health Records: Review of methods for  
3 handling missing data.

4

5 Amin Vahdati<sup>1</sup>, Sarah Cotterill<sup>1</sup>, Antonia Marsden<sup>1</sup>, Evangelos Kontopantelis<sup>2</sup>

6

7

8

9 <sup>1</sup>Centre for Biostatistics, Division of Population Health, Health Services Research & Primary Care, The  
10 University of Manchester, Manchester, UK

11 <sup>2</sup>Division of Informatics, Imaging and Data Sciences, School of Health Sciences, Faculty of Biology,  
12 Medicine and Health, Manchester

13

14

15 \*Corresponding author:

16 Email: [amin.vahdati@manchester.ac.uk](mailto:amin.vahdati@manchester.ac.uk),

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

## 32 Abstract

### 33 Introduction

34 Electronic Health Records (EHRs) are vital repositories of patient information for medical research,  
35 but the prevalence of missing data presents an obstacle to the validity and reliability of research. This  
36 study aimed to review and categorize methods for handling missing data in EHRs, to help  
37 researchers better understand and address the challenges related to missing data in EHRs.

### 38 Materials and Methods

39 This study employed scoping review methodology. Through systematic searches on EMBASE up to  
40 October 2023, including review articles and original studies, relevant literature was identified. After  
41 removing duplicates, titles and abstracts were screened against inclusion criteria, followed by full-  
42 text assessment. Additional manual searches and reference list screenings were conducted. Data  
43 extraction focused on imputation techniques, dataset characteristics, assumptions about missing  
44 data, and article types. Additionally, we explored the availability of code within widely used software  
45 applications.

### 46 Results

47 We reviewed 101 articles, with two exclusions as duplicates. Of the 99 remaining documents, 21  
48 underwent full-text screening, with nine deemed eligible for data extraction. These articles  
49 introduced 31 imputation approaches classified into ten distinct methods, ranging from simple  
50 techniques like Complete Case Analysis to more complex methods like Multiple Imputation,  
51 Maximum Likelihood, and Expectation-Maximization algorithm. Additionally, machine learning  
52 methods were explored. The different imputation methods, present varying reliability. We identified  
53 a total of 32 packages across the four software platforms (R, Python, SAS, and Stata) for imputation  
54 methods. However, it's significant that machine learning methods for imputation were not found in

55 specific packages for SAS and Stata. Out of the 9 imputation methods we investigated, package  
56 implementations were available for 7 methods in all four software platforms.

## 57 Conclusions

58 Several methods to handle missing data in EHRs are available. These methods range in complexity  
59 and make different assumptions about the missing data mechanisms. Knowledge gaps remain,  
60 notably in handling non-monotone missing data patterns and implementing imputation methods in  
61 real-world healthcare settings under the Missing Not at Random assumption. Future research should  
62 prioritize refining and directly comparing existing methods.

63

64

65

66

67

68

69

## 70 Introduction

71 Electronic Health Records (EHRs) have firmly established themselves as essential repositories of  
72 patient data, serving pivotal roles in medical statistics and diverse research objectives (1). These data  
73 offer a wealth of information, enabling researchers to investigate various health-related phenomena  
74 (2). However, one pressing challenge researchers face when utilising EHRs is the issue of missing  
75 data, which can introduce bias and affect the validity of their findings (3, 4). Real-world EHR data  
76 often have high levels of missingness, presenting a challenge for statistical analyses (5). Imputing  
77 missing data in EHRs is crucial for ensuring data integrity and facilitating reliable healthcare decisions  
78 (3, 4). When it comes to handling missing data in EHR, it is a critical aspect that can significantly  
79 enhance the precision of research and reduce bias. Missing data, if not addressed properly, can lead  
80 to skewed results and unreliable conclusions (3, 6). An Electronic Health Record (EHR) is  
81 characterized as a digital archive of patient information, accessible to authorized users across various  
82 healthcare settings, primarily aimed at facilitating comprehensive healthcare delivery (7). Data that  
83 was not initially gathered for research purposes often exhibits missing values, a common occurrence  
84 in EHR datasets. Consequently, effective management of missing data holds significant importance  
85 within the realm EHRs (8).

86 Missing data is commonly classified into three distinct types, determined by the underlying reasons  
87 for its absence. Missing completely at random (MCAR) is a term used to describe a situation where  
88 the occurrence of missing data is purely random and is independent of both observed and  
89 unobserved variables (9). Missing values in the specific variable don't show any clear differences  
90 compared to the values that were observed (10). E.g. if a blood pressure is missing from a dataset, it  
91 could be because someone accidentally forgot to record it. In such cases, the randomness of the  
92 missing data implies that the calculations made using the available data should ideally be free from  
93 any bias (11). When data are MCAR, the resulting parameter estimates are ideally entirely free from  
94 bias. However, a significant limitation of conducting an analysis solely on complete cases, which

95 essentially overlooks the missing data issue, is the resultant loss of statistical power. This loss is not  
96 negligible in many cases, particularly in multiple regression analyses with a high number of  
97 predictors. Even minimal missing data in individual variables can lead to excluding a significant  
98 portion of the dataset from the study [14].

99 Shifting focus to the missing at random (MAR) category, here, the absence of data is influenced by  
100 the information that is available and recorded in the dataset (12). Unlike MCAR, in MAR, there is a  
101 connection between what we see and what's missing. However, this link is only with the observed  
102 data, not with the missing values themselves (13). For example, in a large dataset, there is  
103 missingness regarding blood pressure data. Upon further analysis, it was found that other variables  
104 systematically differ between observed and unobserved values of blood pressure, particularly when  
105 considering cardiovascular disease and age. Patients without cardiovascular disease and younger  
106 patients had more missing blood pressure data compared to older patients with cardiovascular  
107 disease (11). The term 'informative missingness' is sometimes preferred over MAR, and its presence  
108 can be evaluated using logistic regression to explore the relationship between predictor values and  
109 outcome missingness [14].

110 In the most complex scenario, when data is missing not at random (MNAR), the variable's  
111 unobserved value is associated with its absence's underlying cause. This unobserved value can serve  
112 as a predictor or, more concerning, as an outcome. In this case, producing valid results is challenging  
113 due to the absence of straightforward methods. One approach is to conduct multiple sensitivity  
114 analyses to examine the impact of missing data on the outcomes of the study (14). One example is  
115 how Body Mass Index (BMI) data is recorded. BMI is more often collected for people who are  
116 overweight or obese than for those who are not, because health professionals are aware of the link  
117 between weight and various health conditions. Sometimes, when patients are not obese, their  
118 records might be left blank, and it is hard to tell if the data is missing or if the person just isn't obese.

119 Figuring out this kind of missing data, called Missing Not at Random (MNAR), is tough because it's  
120 hard to know why the data is missing (15, 16).

121 Despite the significance of this methodological concern, there is a noticeable gap in the literature  
122 regarding a comprehensive examination of the various approaches and strategies for handling  
123 missing data in EHR-based observational studies. While there are existing reviews on this topic, they  
124 often focus on narrower aspects or specific methods (16-18). Therefore, there is a compelling need  
125 for a scoping review that can offer a broad overview of the methodological approaches used in  
126 addressing missing data within the context of EHR-based observational studies.

## 127 **Methods**

128 This analysis was conducted using the scoping review methodology defined by Arksey and O'Malley  
129 (19). A scoping review approach is beneficial for identifying existing peer-reviewed research and  
130 pinpointing key evidence. Scoping reviews are designed to chart the evidence landscape rather than  
131 deliver critically evaluated and synthesised findings, eliminating the need for quality assessment of  
132 the included studies. Literature published and accessible in full text on EMBASE by October 2023,  
133 focusing on missing data imputation in EHRs and presented either as a review article or through a  
134 novel approach, was considered eligible for inclusion.

135 We employed a customised systematic search strategy that combined keywords pertinent to missing  
136 data imputation and electronic health records (Table. 1) (20, 21). Eligible for inclusion were both  
137 review articles and original studies proposing approaches to managing missing data in EHRs.

138 *Table 1. A search strategy was developed by combining relevant keywords related to imputing missing data and electronic*  
139 *health records, utilizing Boolean operators.*

1. ((patient\* or participant\*) and ("drop-out\*" or "drop out\*" or dropout\* or "loss to follow\*" or "lost to follow\*" or "withdrawal\*") and missing).tw.
2. (missing data).ti.
3. "incomplete data".ti.
4. "missing data".ti.
5. ("sensitivity analysis" and missing).ti.
6. imputation.ti.
7. "full information maximum likelihood".ti.
8. "maximum likelihood".ti.
9. "inverse probability weight".ti.
10. "inverse probability".ti.
11. "multiple imputation".ti.
12. 1 or 2 or 3 or 4 or 5 or 6 or 7 or 8 or 9 or 10 or 11
13. exp medical records systems computerized/
14. exp health records personal/
15. (personal adj (health record\* or medical record\*)).ti,ab,kf.
16. ((electronic\* or online or on-line or digital\*) adj2 (health record\* or medical record\* or personal record\* or patient record\*)).ti,ab,kf.
17. ((web or internet or computer\*) adj3 (health record\* or medical record\* or personal record\* or patient record\*)).ti,ab,kf.
18. (ehr? or phr? or ephr? or emr? or pahr?).ti,ab,kf.
19. (patient adj2 portal\*).ti,ab,kf.
20. 13 or 14 or 15 or 16 or 17 or 18 or 19
21. 12 and 20

140

141 After removing duplicates, we began by systematically reviewing titles and abstracts against the  
142 inclusion criteria to exclude any that were clearly ineligible, followed by screening the remaining full  
143 texts to identify relevant papers. We also searched the publication lists of the final pool of included  
144 papers for relevant references. Data extraction focused on imputation techniques, the characteristics  
145 of datasets, assumptions about missing data and the type of article.

## 146 Results

147 A total of 101 articles were found, and two of them were excluded as duplicates, as indicated in the  
148 PRISMA scoping review flowchart (Figure.1). The remaining 99 documents were screened by title and  
149 abstract. 21 studies were selected for full-text screening. Five studies were eligible for data

150 extraction. Four articles were also found through a manual search, which brings the total number of  
151 studies that will be screened for full text to nine.

152

153 *Figure 1. Prisma flowchart of literature search*

154 Within the nine articles, thirty-one imputation approaches were identified, classified into 10 distinct  
155 imputation methods, each applied under specific assumptions, using simple or more complex  
156 imputation methods (Figure. 2).

157

158 *Figure 2. Imputation methods*

159 The screened articles encompassed various contributions to the field: three papers provided reviews  
160 of existing imputation methods, two engaged in simulation studies, two introduced novel  
161 methodologies, one established a general framework for imputing missing data, and one conducted  
162 a comparative analysis of different autoencoder methods.

163 Certain studies proposed imputation methods suitable for a broad spectrum of EHRs (16, 22, 23),  
164 while others focused on specific data types such as time series (17), marginal structural models (18),  
165 and longitudinal datasets (15, 24, 25). Additionally, a unique study employed a simulation study for  
166 comparing different imputation methods (26). The following synthesis provides a comprehensive  
167 overview of the diverse landscape of imputation methodologies applied to address missing data in  
168 EHRs (Table 2).

169 *Table 2. Characteristics of Included Studies.*

	Authors	Year	Imputation method	Data	Assumptions	Study type
1	Carpenter et al.	2020	Maximum Likelihood,	-	MAR	Practical guideline



			Bayesian approach, Expectation maximisation algorithm, Mean-score estimation, Multiple imputation, IPW.			
2	Kazijevs et al.	2023	Machin learning method	Time series data	MCAR MAR MNAR	Review and benchmarking
3	Leyrat et al.	2020	Complete case analysis, The last observation carried forward, Multiple imputation, Inverse-probability-of-	Simulation study	MCAR MAR Constant Differential	Review and simulation study

			missingness weighting (IPMW),			
4	Tsiampalis et al.	2023	Deep learning unsupervised method, Sub-datasets with complete information, Mean-values of attributes combined autoencoder, Data-driven approach (Denser EHR to Spares EHR), Use of informative observations	-	-	Review
5	Kontopantelis	2017	Longitudinal multiple imputation approaches for	CPRD	MNAR	Method development study

			variables  with very low  individual-level  variability: the  mibmi  command  in Stata			
6	Welch	2013	Two-fold fully  conditional  specification  multiple  imputation	Simulation  study	MCAR	Simulation  study
7	Cesare		Multi-step  approach to  managing  missing data in  time and  patient variant  electronic  health records	AMPATH  Academic Model Providing Access to Healthcare	MAR	Method  development  study
8	Beaulieu-  Jones	2016	Autoencoder,  IterativeSVD  (Singular Value Decomposition),  K-nearest	Pro-Act  dataset	MCAR  MNAR	Compare  imputation  methods

			neighbours imputation (KNN imputation), Soft Impute, Column Mean Filling, Column Median Filling			
9	Blankers	2010	Multiple imputation	Prospective cohort study	MAR, MCAR	Simulation study

170

171 A combined count of 32 software packages dedicated to imputation methods was discovered across  
 172 the four statistical software: R, Python, SAS, and Stata. Notably, machine learning techniques for  
 173 imputation were absent from devoted packages in SAS and Stata. Among the 9 imputation methods  
 174 searched, implementations were accessible for 7 methods across all four software platforms (Table  
 175 3).

176 *Table 3. Imputation methods package*

Category	Method	Software	Package
	CCA	R	tidyr
		Python	Pandas

Simple		SAS	DATA step/ PROC IML
		Stata	rmiss2
	LOCF	R	tidyr/ imputeTS
		Python	pandas
		SAS	MACRO
		Stata	ssc
	Maximum Likelihood	R	maxLik, univariateML
		Python	Mvem, mle, statsmodels
		SAS	sas-twophase-package
		Stata	Cquad
	Expectation maximisation	R	library(dplyr)/ library(ggplot2)
		Python	mixem
		SAS	SAS/IML
		Stata	mi (mi impute mvn)
IPW	R	lpw	
	Python	Causallnference, balance	
	SAS	eAppendix	
Complex			

		Stata	PSWEIGHT
Iterative SVD		R	svd
		Python	Scratch, SciPy, NumPy, scikit-learn, PyTorch
		SAS	IML
		Stata	LAPACK
Multiple imputation		R	rMIDAS, MICE, Amelia
		Python	Autoimpute, MIDASpy
		SAS	MIANALYZE
		Stata	Mi
Soft impute		R	Softimpute, CRAN
		Python	Fancyimpute,
		SAS	-
		Stata	-
Unsupervised deep learning		R	rMIDAS, ruta,
		Python	MIDASpy
		SAS	-
		Stata	-

## 178 A. Simple methods:

### 179 A.1 Complete case analysis

180 Complete-case analysis (CCA), or listwise deletion, exclusively employs the data records that keep no  
181 missing values for any pertinent variable required for analysis, hence excluding observations that  
182 have a missing value for any of the variables included in the analysis (27). CCA can produce estimates  
183 of effect size that are biased, inefficient and/or underpowered. These issues become more  
184 pronounced when there is a high level of missing data, with loss of power and also higher levels of  
185 bias if the data missingness mechanism is MAR or MNAR. (28). The fundamental assumption  
186 underlying CCA is that the data are MCAR.

### 187 A.2. Last observation carried forward (LOCF)

188 The LOCF method is a historically popular statistical tool in longitudinal studies involving repeated  
189 measurements (29). In this approach, when a value is missing from a later visit, it is filled in with the  
190 previously recorded value of the participant. This leads to an analysis where through a simple form of  
191 imputation, the affected variables can be regarded as complete (18). However, a critical shortcoming  
192 of LOCF is its inclination to provide inaccurate estimates for missing values, notably when the  
193 dependent variable is on an upward (or downward) trajectory over time. LOCF is prone to bias,  
194 potentially leading to either an underestimation or overestimation of the actual effects of the  
195 treatment. The method's suitability is questionable even in situations where the missing data is  
196 purely random (30).

## 197 B. Complex imputation methods

### 198 B.1. Multiple Imputation (MI)

199 Multiple imputation is a commonly used method that replaces missing values by creating plausible  
200 numbers based on the distributions and connections of observed variables in the dataset (31). MI  
201 utilises this technique to estimate the missing values, resulting in multiple datasets considered 'full'.  
202 We use the observed data to evaluate the distribution of the partially observed variables, given the

203 fully known variables (32). Subsequently, we employ this estimation to fill in the missing data. The  
204 rationale behind using multiple imputations is that the imputed data can never possess the identical  
205 characteristics as the observed data. Instead, they are generated from the predicted distribution of  
206 the missing data, given the observed data under the assumption of MAR. This process operates  
207 under a specific Bayes model that accurately represents both the observed data and the mechanism  
208 responsible for missing data (32).

209 Two steps are involved in multiple imputations: 1) Creating replacement values, or "imputations," for  
210 missing data and repeatedly doing so to develop numerous data sets with the missing information  
211 replaced. Multiple Imputation (MI) fills in the missing values using statistical properties of the data,  
212 such as the relationships and distributions of variables in the dataset. 2) Analysing and integrating  
213 the many imputed data sets (33, 34). Following the execution of the planned statistical analysis (such  
214 as regression or t-test) on each imputed data set individually (stage 2), the desired estimates (e.g.,  
215 the average difference in outcome between a treatment group and a control group) from all the  
216 imputed data sets are merged into a single estimate using standard combining methods. The benefit  
217 of using MI in statistical analysis is that it may handle problems other than traditional missing data  
218 problems. MI is a widely used method for managing missing data, and it is offered in many software  
219 programmes (35, 36).

220 Further refinement in data imputation is achieved using Multiple Imputation by Chained Equations  
221 (MICE, also called Fully Conditional Specification (FCS)), a sophisticated technique crucial for datasets  
222 with complex variable relationships (37). MICE works by creating multiple dataset copies, filling  
223 missing values with placeholders, and then employing regression models to predict these values (38).  
224 The pooled predictions from these multiple imputed datasets lead to the selection of final values,  
225 effectively handling both categorical and continuous data. Even after employing MICE, some values,  
226 especially dates, might be inaccurate. The subsequent step involves selecting representative values  
227 for each patient to minimise outlier impacts and manually adjusting date variables to ensure



228 accuracy (25). The primary limitations of this approach involve the time-invariant nature of the MICE  
229 imputation model and the lack of consideration for spatial autocorrelation among nearby clinics.

230 A Bayesian multiple imputation method is introduced to handle left-censored multivariate data  
231 commonly found in environmental and biomedical research (39). This approach addresses the  
232 difficulty of assigning explicit values for observations below the limit of detection (LOD), especially in  
233 longitudinal data settings.

234 Also, k-nearest neighbours (KNN) can be used with bootstrap and sequential imputation to get  
235 multiply imputed data (40). The idea behind utilising KNN for missing values is that a point value may  
236 be estimated by the values of the points nearest to it, depending on other factors. For categorical  
237 variables, the mode of the nearest neighbours is utilised, whereas for numeric data, the mean of the  
238 nearest neighbours is employed. The k-nearest neighbours algorithm requires searching through all  
239 complete cases and selecting the k examples most relevant to a particular missing information  
240 (Nearest neighbour selection for iteratively KNN imputation). The KNN imputation technique  
241 encounters two significant hurdles: firstly, determining the optimal value of k in advance, and  
242 secondly, selecting the most appropriate k nearest neighbours [53].

## 243 **B.2. Maximum likelihood**

244 Maximum likelihood imputation is a theoretically robust approach for estimating parameters in  
245 regression models when dealing with missing data (41). This technique includes estimating a set of  
246 parameters that maximise the chance of obtaining the observed data (42). The method is  
247 characterised by the explicit articulation of the likelihood of the intended data, integration over  
248 missing values to ascertain the likelihood of observed data, and the subsequent maximisation of this  
249 likelihood to derive maximum likelihood estimates. This methodological rigour ensures a principled  
250 and statistically sound treatment of missing data, especially in complex and unbalanced study  
251 designs (22, 43). A limitation of maximum likelihood methods is the need for relatively large data sets  
252 (44).

253 Maximum likelihood and MI procedures have been methodically developed with a foundational  
254 assumption of MAR (45). Note that if the missingness mechanism strays from the MAR assumption,  
255 as observed in MNAR scenarios, methods such as maximum likelihood and multiple imputation  
256 might produce biased results, especially in small longitudinal samples exhibiting non-normality (43).  
257 In instances of varying population distributions, Maximum Likelihood is preferred over MI when the  
258 distribution is non-normal (46).

### 259 B.3. Expectation-Maximisation (EM) algorithm

260 Expectation maximisation offers an iterative approach to maximising the likelihood estimation by  
261 using latent variables (47). This algorithm is designed to maximise the likelihood of observed data by  
262 iteratively refining parameter estimates. The EM algorithm consists of two distinct phases:  
263 Expectation and Maximization. In the Expectation phase, the algorithm begins by imposing a value  
264 for the missing or latent variables based on the current estimates of the parameters. This step  
265 involves calculating the expected values of the missing data given the observed data and the current  
266 parameter estimates. Essentially, it estimates the posterior distribution of the missing variables. This  
267 imputation process is crucial for establishing a foundation for subsequent parameter refinement.  
268 Following the Expectation step, the Maximization phase comes into play. In this step, the algorithm  
269 evaluates the parameters that maximise the expected log-likelihood obtained from the first step. It  
270 involves adjusting the parameter values to enhance the fit between the observed and imputed data.  
271 The Maximisation step essentially serves as a parameter update based on the newly imputed values,  
272 optimising the likelihood function. The iterative nature of the EM algorithm involves repeating these  
273 two steps until a convergence criterion is met, indicating that the algorithm has reached a stable  
274 solution. Convergence is often achieved when there is minimal change in the parameter estimates  
275 between successive iterations. One notable advantage of the EM algorithm is its ability to provide  
276 consistent estimates of means and covariance matrices, which are essential for characterising the  
277 underlying distribution of the data. However, EM may be computationally intensive and requires a  
278 sufficiently large sample size to ensure the reliability of the parameter estimates (48-50).

279 **B.4. Inverse probability weighting (IPW)**

280 In this method, complete cases are weighted by the inverse of their probability of being complete  
 281 cases. Table four illustrates a basic idea. Because the likelihood of missing data being dependent on  
 282 its value introduces bias, the observed mean is 13/6 instead of the whole dataset's mean of 2 (i.e.,  
 283 the actual values of the nine observations). The third row of table four displays the estimated chance  
 284 of observing the data when considering the group variable and assuming MAR given the group. This  
 285 assumption was validated by cross-referencing with the entire data (22).

286 *Table 4. Example for inverse probability weighting imputation technique*

Group	A			B			C		
Full data	1	1	1	2	2	2	3	3	3
Observed data	1	?	?	2	2	2	?	3	3
Probability of observation given group	1/3	-	-	1	1	1	-	2/3	2/3

287

288 The weights, which are the inverses of the observation probabilities, are used to compute a weighted  
 289 mean using this estimate. Therefore, to make up for the three observations in group A, we give each  
 290 observation a weight of 3. Next, we compute the weighted mean:

$$291 \frac{(1 \times \frac{3}{1}) + (2+2+2) \times 1 + (3+3 \times \frac{3}{2})}{\frac{3}{1} + 1 + 1 + 1 + \frac{3}{2} + \frac{3}{2}} = 2$$

292 Since the IPW analysis relies on estimated weights from observed data, it only applies under MAR  
 293 (51). Statistical analysis is complicated by IPW. To begin, it considers only complete records when re-  
 294 weighting them, which results in the elimination of any data point that is lacking values. As a result,  
 295 the study may not be as comprehensive or representative as it could have been because data from  
 296 these missing variables cannot be recovered (52). If covariates are not MAR given the dependent

297 variable, a complete records analysis might be more suitable, necessitating the inclusion of the  
298 dependent variable in the weight model. It can be challenging to estimate weights precisely when  
299 relevant variables in the weight model have missing values, further complicating the inclusion of  
300 these variables. Lastly, IPW results can be affected by big weights and the weight model chosen,  
301 which can make it hard to choose. Finding a suitable weight model becomes much more of a  
302 challenge when there is no explicit documentation on handling these problems (22).

303 There is a modified method of IPW called Inverse Probability of Missingness Weighting (IPMW). In  
304 studies with time-varying confounding and data gaps, IPMW is used. While effective in monotone  
305 missing data scenarios, like participant dropout, IPMW becomes complex with non-monotone  
306 missing patterns, where data intermittently lacks in various variables. This complexity increases with  
307 multiple time points and incomplete variables. Additionally, the broader inverse probability  
308 weighting (IPW) method, which combines weights for each missing variable, can lead to variable  
309 weights and less accurate treatment effect estimates. "IPW" refers to the general approach, with  
310 "IPMW" denoting its specific uses for missing data (53). IPMW is used in Marginal Structural Model  
311 (MSM) for dealing with missing confounder in non-randomised longitudinal studies (18). Marginal  
312 structural models (MSMs) are commonly used to estimate causal intervention effects in longitudinal  
313 nonrandomized studies (54).

## 314 B.5. Iterative SVD (Singular Value Decomposition)

315 To understand Iterative SVD, it is essential first to grasp the concept of standard SVD. Singular Value  
316 Decomposition is a matrix factorisation technique used in many fields, including signal processing,  
317 statistics, and machine learning (55). SVD is a mathematical technique that decomposes a matrix into  
318 its constituent elements, providing insight into its structure. It decomposes a matrix  $\bar{A}$  into three  
319 other matrices  $U$ ,  $\Sigma$  and  $V^T$  where  $U$  and  $\Sigma$  and  $V^T$  are orthogonal matrices, and  $\Sigma$  is a diagonal matrix  
320 containing singular values. This decomposition can capture the underlying structure of the

$$321 \quad A_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T.$$

322 Iterative Singular Value Decomposition (SVD) addresses missing values in the initial data matrix.  
323 Given that SVD necessitates an entirely populated matrix, an initial imputation is performed by  
324 substituting missing values with mean or median values or utilising row/column averages. After the  
325 initial imputation, standard SVD is performed on the modified data matrix. This yields the matrices  $U$   
326 and  $\Sigma$  and  $V^T$  , providing a lower-dimensional representation of the data. The missing values are  
327 subsequently recalculated utilising the knowledge acquired from the SVD decomposition. This entails  
328 utilising the singular value decomposition (SVD) components to restore the data matrix and replacing  
329 the previously imputed values with new estimations. The final two steps are performed iteratively.  
330 During each iteration, the Singular Value Decomposition (SVD) is performed on the matrix that has  
331 been updated with imputed values. Subsequently, these values are adjusted depending on the new  
332 decomposition (56, 57).

333 Singular Value Decomposition (SVD) offers notable advantages, including significantly higher  
334 accuracy compared to simple row averages across diverse datasets (58). It excels in analysing time-  
335 series data with low noise levels, effectively estimating gene expression based on temporal  
336 regulation patterns (57). However, SVD has limitations, requiring complete matrices for operation.  
337 Imputation strategies, such as substituting row averages for missing values, are necessary.  
338 Additionally, SVD's performance is sensitive to data type, exhibiting potential challenges in non-time  
339 series datasets lacking clear expression patterns (59). Its linear regression nature in lower-  
340 dimensional space may result in diminished performance for non-time series data, where expression  
341 patterns are often less distinct (57). Despite these considerations, SVD remains a powerful tool,  
342 particularly suited for specific data characteristics and applications.

#### 343 **B.6. Longitudinal imputation approach:**

344 The two-fold fully conditional specification (FCS) technique is an adaptation of FCS as multiple  
345 imputation technique, that takes account of the temporal structure of longitudinal data (60, 61).

346 The method commences by selecting a time window width, for instance, 1-, 2-, or 3-time blocks,  
347 establishing the range of time blocks around the target time point for imputation. During this stage,  
348 the typical FCS imputation technique is utilised to fill in missing values for each variable at a  
349 particular time (61). This is done by employing regression models incorporating values from adjacent  
350 time points within the designated time window (61, 62). This method enables the inclusion of  
351 temporal dependencies within the imputation model. Once the within-time imputation for a time  
352 block is finished, the same process is done for all time blocks in the dataset. This step guarantees  
353 that the imputation model considers the longitudinal evolution of the data (61, 62).

354 The two-fold FCS method improves the collection of temporal correlations in longitudinal EHR data  
355 by considering adjacent time blocks throughout the imputation process. This simplifies the process  
356 of imputation models and minimises the likelihood of collinearity and overfitting (61). The system  
357 enables the utilisation of several modelling approaches tailored to specific variable types, thus  
358 accommodating the wide range of data types commonly encountered in EHRs. The two-fold  
359 technique can result in more accurate estimates compared to traditional FCS/MICE when there are  
360 longitudinal and time-dependent patterns present, reducing bias.

361 Like other multiple imputation approaches, the two-fold FCS functions assume that data are MAR.  
362 The algorithm can be computationally intensive, especially with large datasets and comprehensive  
363 time windows (24). Bespoke longitudinal imputation algorithms, like the `mibmi` code in STATA, are  
364 customized for specific parameters such as BMI over time. These algorithms start with outlier  
365 detection and employ standard and regression-based cleaning methods for BMI computation. While  
366 effective in generating multiple imputed datasets, they come with drawbacks including high  
367 computational costs, exclusion of patients with limited BMI records, and the necessity for careful  
368 extrapolation to avoid inaccuracies.

## 369 C. Machine learning methods

370 Machine learning methods for imputing missing data operate by dividing datasets into training and  
371 test sets and learning from observed variables. There are several methods, using supervised, semi-  
372 supervised, or unsupervised machine learning techniques [59]. Like other imputation methods,  
373 machine learning methods do not work with all kinds of data [60]. Their performance could change  
374 depending on the value type, which include numeric, non-numeric, text, graph, etc. So, it is essential  
375 to fully understand the underlying data patterns before doing data imputation to ensure that the  
376 best machine learning method is chosen for handling missing data situations correctly and effectively  
377 [61]. Selecting the proper machine learning imputation method depends on the dataset and analysis  
378 aim [61]. When compared to standard imputation methods, the use of machine learning for  
379 imputing missing values has shown improved performance in prediction and data analysis [62]. Non  
380 machine learning techniques may result in a smaller sample size and more bias since they limit  
381 variability [63]. Nevertheless, due to recent technological progress, machine learning capitalises on  
382 substantial computational resources to tackle these obstacles efficiently through precise estimation  
383 of absent values, thus enhancing the performance of data analysis [64]. Current method proposals  
384 that use machine learning techniques can have their crucial improvements in accuracy, performance,  
385 and time consumption brought to light through research and analysis [24]. Different machine  
386 learning methods have been used in EHRs for imputing missing data in the screened articles. We will  
387 introduce them more.

### 388 C.1. Deep learning unsupervised method

389 Deep learning unsupervised is an imputation technique specifically tailored to handle longitudinal  
390 patient data, defined by the progression of the disease over time (63). This strategy primarily  
391 involves collecting data by grouping essential clinical variables that an expert physician has  
392 determined. The data, characterised by missing values denoted as "NA," is structured into records  
393 associated with patients via unique identifiers. The methodology entails an initial preprocessing  
394 stage in which the data is converted into a matrix structure. The matrix has n rows (records) and m

395 columns (variables) and is subjected to z-score normalisation for numeric variables and one-hot  
396 encoding for categorical variables. The approach's critical element resides in utilising a deep  
397 autoencoder framework. This system consists of two distinct encoders: Encoder1, which is dedicated  
398 to creating embeddings at the record level, and Encoder2, which captures the heterogeneity at the  
399 patient level. These encoders operate by projecting the input into a hidden space, and then a  
400 decoder combines these embeddings and processes them through numerous layers that include  
401 nonlinear transformations. The parameters are adjusted in the last stage of the process, and the  
402 patient data is re-entered to provide comprehensive datasets. The utilisation of a deep learning  
403 architecture in this method enables the modelling of intricate inter-variable interactions, which is  
404 crucial for effectively filling in missing values in cardiovascular patient data. The strategy effectively  
405 minimises imputation mistakes and biases in cardiovascular patient care by utilising patient-level  
406 heterogeneity and temporal patterns, representing a significant development in data processing (64).  
407 The deep autoencoder framework, with multiple layers and non-linear transformations, can be  
408 computationally intensive (65). Training and fine-tuning such models may require significant  
409 computing resources and time, which could be a practical limitation in certain healthcare settings.  
410 Deep learning models, particularly intricate ones such as autoencoders, frequently suffer from a lack  
411 of transparency and interpretability (66, 67). The comprehension of the model's process in  
412 determining imputed values or the rationale behind its conclusions might be difficult, which can  
413 cause apprehension in clinical environments where interpretability is essential.

## 414 C.2. Soft-Impute

415 This technique is especially relevant in fields such as machine learning and data science, where  
416 dealing with incomplete data is a common challenge. Soft-Impute is grounded in spectral  
417 regularisation, using convex relaxation techniques to fill in missing values in large matrices. The core  
418 of this method lies in its use of the nuclear norm as a regulariser. The nuclear norm, essentially the  
419 sum of the singular values of a matrix, helps maintain the low-rank structure of the solution and  
420 avoids overfitting, which is crucial when dealing with large datasets. The process of Soft-Impute is



421 iterative. It employs a soft-threshold Singular Value Decomposition (SVD) in each iteration to update  
422 the missing elements of the matrix (68). Through iterative updates using SVD, Soft-Impute efficiently  
423 replaces missing values, gradually improving the matrix's completeness. One of the critical strengths  
424 of Soft-Impute is its ability to compute a regularisation path, offering a series of solutions  
425 corresponding to different values of the regularisation parameter. This capability allows the method  
426 to find the optimal balance between the complexity of the model and its fit to the data, making it  
427 highly effective for practical applications. In terms of scalability and efficiency, Soft-Impute stands  
428 out. With high computational efficiency, it can handle huge matrices, such as those encountered in  
429 the Netflix challenge for predicting user movie ratings. This efficiency stems from its fast  
430 computation of a low-rank SVD of a dense matrix. Furthermore, the method has shown impressive  
431 performance, achieving good training and test errors compared to other state-of-the-art techniques  
432 [70].

## 433 Discussion

434 The comprehensive review conducted in this study provides a valuable summary of imputation  
435 methods for addressing missing data in EHRs. With the increasing digitization of healthcare data, the  
436 issue of missing data has become a significant concern, as it can impact the reliability and validity of  
437 analyses derived from such data. In this study, we identified a total of 101 articles and chose nine  
438 studies for data extraction, which described 31 imputation approaches. These methods were divided  
439 into two primary groups: simple imputation methods and complex imputation methods. Simple  
440 methods included CCA and LOCF, while complex methods comprised MI, Maximum Likelihood, EM  
441 algorithm, IPW, Iterative SVD, longitudinal imputation approach, deep learning unsupervised  
442 method, and Soft-Impute. Researchers should carefully consider the characteristics of their data, the  
443 assumptions of each method, and the specific context of their study when selecting an imputation  
444 method. The choice of imputation method should align with the goals of the analysis and the nature  
445 of the missing data mechanism.

446 The performance of different imputation methods has been assessed in various studies. A  
447 comparison of eight imputation methods under different missing data mechanisms (MI, CCA, mean  
448 imputation, LOCF, HOT deck imputation, regression imputation, KNN, EM algorithm) suggested that  
449 MI exhibits the least standard errors compared to other methods, indicating its effectiveness in  
450 handling missing data (69). Simulation studies offer valuable insights, but it is crucial to validate  
451 these findings with real-world datasets to determine the practical applicability and reliability of  
452 multiple imputation (MI) in comparison to other imputation methods.

453 Comparison of multiple imputation methods, including chained equations, random forests, and  
454 denoising autoencoders, revealed distinct patterns across different missing data mechanisms.  
455 Chained equations and random forests showed reduced bias and comparable standard errors under  
456 MCAR. However, denoising autoencoders exhibited elevated bias in cases of MAR. Furthermore, all  
457 methods demonstrated increased bias proportional to the extent of missing data under MNAR  
458 conditions (70).

459 Assessing the performance of FCS and Two-Fold FCS methods in managing missing variables within  
460 longitudinal studies, particularly in estimating regression coefficients via a linear regression model,  
461 reveals that Two-fold FCS methods yield estimates that are slightly more biased and less precise  
462 compared to FCS (71, 72). Two-Fold FCS was specially adapted from FCS to deal with missing data in  
463 longitudinal studies, so it seems surprising that FCS would perform better than two-fold FCS in  
464 longitudinal data. This observed bias likely stems from the inherent limitation of these approaches in  
465 restricting variables within the univariate imputation models, consequently risking the omission of  
466 crucial information of the missing data (71). One study suggests that, in longitudinal data, imputation  
467 of data from patients with a similar pattern of data may outperform traditional MI methods (73).  
468 Multiple imputation (MI) is an effective method for dealing with bias caused by missing data in  
469 longitudinal studies, resulting in accurate parameter estimates.

470 A comparison of eight common statistical and machine learning imputation methods (simple  
471 imputation, regression, EM, MICE, KNN, clustering imputation, random forest, and decision tree)  
472 showed that KNN and RF were the most effective imputation methods in the cohort study dataset  
473 under the MAR assumption (74). Machine learning methods attempt to explore the relationship  
474 between variables and predict missing data more precisely. This superior performance of machine  
475 learning methods suggests their potential for handling complex data patterns and nonlinear  
476 relationships, thereby offering more accurate imputations in challenging datasets.

477 Understanding the assumptions behind each imputation method is essential for choosing the best  
478 one based on the dataset and missing data mechanism. Also, one should evaluate the computational  
479 complexity and resource needs when choosing imputation algorithms. Deep learning unsupervised  
480 methods can handle complex data patterns better but take a lot of processing power and time to  
481 train and fine-tune models. Although simpler algorithms like CCA and LOCF are more  
482 computationally efficient, they can be biased and inefficient. Existing methods address monotone  
483 missing patterns, temporal correlations in longitudinal data, and complex model interpretability, but  
484 they can be improved. The software available to researchers may influence their choice of method as  
485 not every method is available in each software.

486 Regarding knowledge gaps, firstly there is a lack of consensus regarding the superiority of certain  
487 imputation methods over others, particularly in scenarios involving complex missing data patterns.  
488 This discrepancy underscores the need for further comparative studies to evaluate the strengths and  
489 limitations of different imputation techniques comprehensively. Secondly, existing imputation  
490 methods may not adequately address the challenges posed by non-monotone missing data patterns,  
491 where missingness occurs sporadically or irregularly across the dataset. While methods like MI and  
492 machine learning algorithms have been proposed for handling monotone missing data, there is  
493 limited research on techniques specifically designed for non-monotone missingness. Developing  
494 robust imputation strategies tailored to these complex missing data patterns is crucial for improving

495 the accuracy and reliability of analyses conducted on longitudinal datasets. Furthermore, there is a  
496 need for more research on the practical implementation of imputation methods in real-world  
497 healthcare settings. Many existing studies focus on theoretical comparisons of imputation techniques  
498 using simulated datasets or controlled experiments. While current approaches such as MI and other  
499 techniques have shown promise in dealing with missing data, further research is required to build  
500 viable solutions suited particularly to MNAR circumstances.

501 Imputation strategies for addressing missing data are often underutilized due to a prevalent practice  
502 of not explicitly reporting missing data (75, 76). This reporting failure adds to the knowledge gap on  
503 the kind and amount of missing data in epidemiological studies. Because of this, chances to use  
504 imputation techniques efficiently to solve problems with missing data are often missed or  
505 underutilized. The absence of imputation procedures in this study might potentially undermine the  
506 validity and reliability of the results. This highlights the need of thorough reporting and the use of  
507 suitable approaches to address missing data in research projects.

## 508 Conclusion

509 Addressing missing data is an important aspect of the analysis of EHRs and various imputation  
510 methods are available to researchers. These methods range from simple to complex, each relying on  
511 assumptions that have been discussed in this paper.

512 When selecting imputation methods, key considerations include understanding the assumptions  
513 underlying each method, evaluating computational complexity, and effectively addressing monotone  
514 missing data patterns. Further research is needed to establish method superiority, especially in  
515 complex missing data scenarios, and to develop robust strategies for real-world healthcare  
516 applications.

517

## 518 References

- 519 1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research  
520 applications and clinical care. *Nature Reviews Genetics*. 2012;13(6):395-405.
- 521 2. Annis A, Reaves C, Sender J, Bumpus S. Health-Related Data Sources Accessible to Health  
522 Researchers From the US Government: Mapping Review. *Journal of Medical Internet Research*.  
523 2023;25:e43802.
- 524 3. Zhou Y, Shi J, Stein R, Liu X, Baldassano RN, Forrest CB, et al. Missing data matter: an  
525 empirical evaluation of the impacts of missing EHR data in comparative effectiveness research.  
526 *Journal of the American Medical Informatics Association*. 2023:ocad066.
- 527 4. Beesley LJ, Salvatore M, Fritsche LG, Pandit A, Rao A, Brummett C, et al. The emerging  
528 landscape of health research based on biobanks linked to electronic health records: Existing  
529 resources, statistical challenges, and potential opportunities. *Statistics in medicine*. 2020;39(6):773-  
530 800.
- 531 5. Sondhi A, Weberpals J, Yerram P, Jiang C, Taylor M, Samant M, et al. A systematic approach  
532 towards missing lab data in electronic health records: A case study in non-small cell lung cancer and  
533 multiple myeloma. *CPT: Pharmacometrics & Systems Pharmacology*. 2023.
- 534 6. Al-Ghraiyyah T, Sim J, Fernandez R, Lago L. Managing missing and erroneous data in nurse  
535 staffing surveys. *Nurse Researcher*. 2023;31(1).
- 536 7. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of  
537 electronic health records: a review of the research literature. *International journal of medical  
538 informatics*. 2008;77(5):291-304.
- 539 8. Psychogyios K, Ilias L, Ntanos C, Askounis D. Missing value imputation methods for electronic  
540 health records. *IEEE Access*. 2023;11:21562-74.
- 541 9. Pham TM, Pandis N, White IR. Missing data, part 2. Missing data mechanisms: Missing  
542 completely at random, missing at random, missing not at random, and why they matter. *American  
543 journal of orthodontics and dentofacial orthopedics*. 2022;162(1):138-9.
- 544 10. Nakai M, Chen D-G, Nishimura K, Miyamoto Y. Comparative study of four methods in missing  
545 value imputations under missing completely at random mechanism. *Open Journal of Statistics*.  
546 2014;2014.
- 547 11. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and  
548 missing at random? *Int J Epidemiol*. 2014;43(4):1336-9.
- 549 12. Curnow E, Tilling K, Heron JE, Cornish RP, Carpenter JR. Multiple imputation of missing data  
550 under missing at random: including a collider as an auxiliary variable in the imputation model can  
551 induce bias. *Frontiers in epidemiology*. 2023;3:1237447.
- 552 13. Lee KJ, Carlin JB, Simpson JA, Moreno-Betancur M. Assumptions and analysis planning in  
553 studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification.  
554 *International Journal of Epidemiology*. 2023:dyad008.
- 555 14. Heymans MW, Twisk JW. Handling missing data in clinical research. *Journal of clinical  
556 epidemiology*. 2022;151:185-8.
- 557 15. Kontopantelis E, Parisi R, Springate DA, Reeves D. Longitudinal multiple imputation  
558 approaches for body mass index or other variables with very low individual-level variability: the  
559 mibmi command in Stata. *BMC research notes*. 2017;10(1):41.
- 560 16. Tsiampalis T, Panagiotakos D. Methodological issues of the electronic health records' use in  
561 the context of epidemiological investigations, in light of missing data: a review of the recent  
562 literature. *BMC medical research methodology*. 2023;23(1):180.
- 563 17. Kazijevs M, Samad MD. Deep imputation of missing values in time series health data: A  
564 review with benchmarking. *Journal of biomedical informatics*. 2023;144:104440.
- 565 18. Leyrat C, Carpenter JR, Bailly S, Williamson EJ. Common Methods for Handling Missing Data  
566 in Marginal Structural Models: What Works and Why. *American journal of epidemiology*.  
567 2021;190(4):663-72.

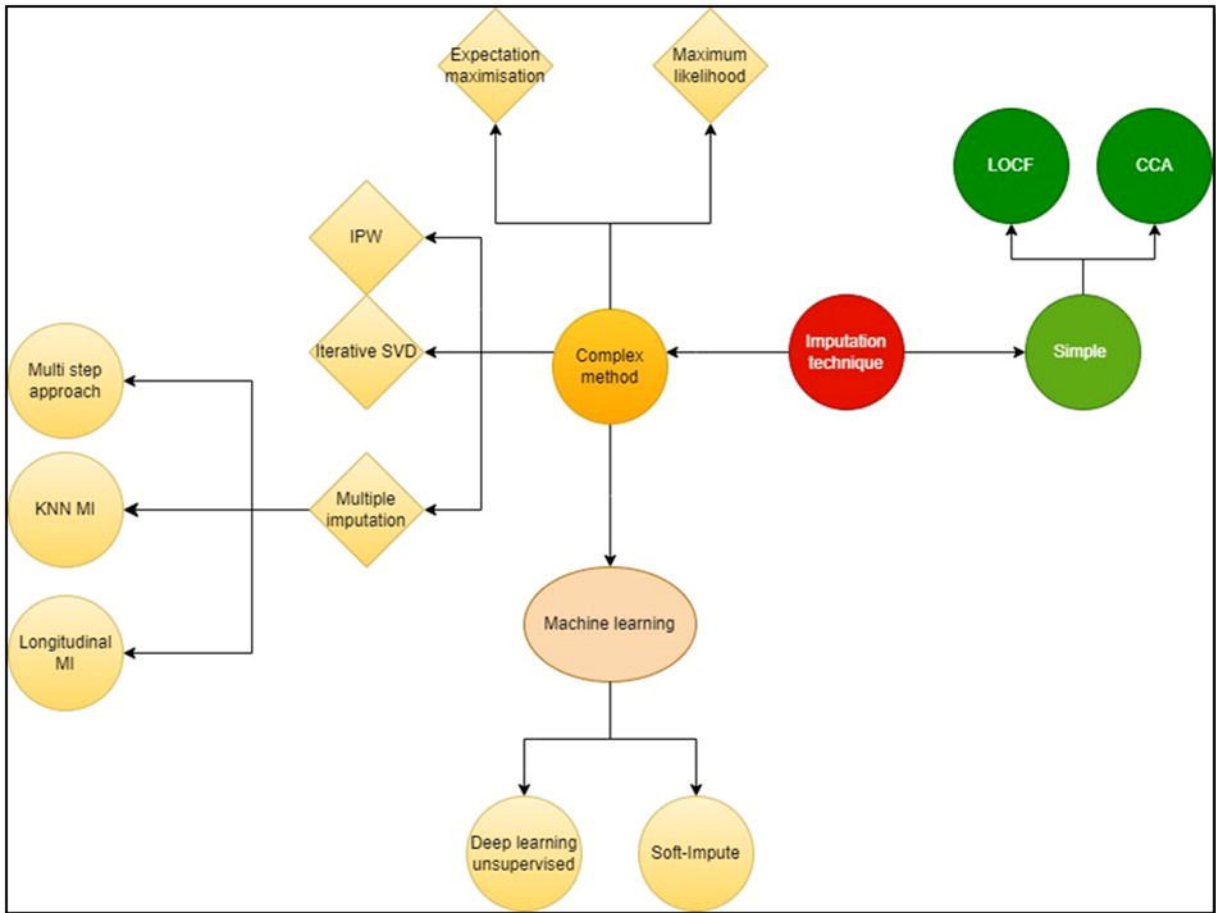
- 568 19. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *International*  
569 *journal of social research methodology*. 2005;8(1):19-32.
- 570 20. Ammenwerth E, Neyer S, Hörbst A, Mueller G, Siebert U, Schnell-Inderst P. Adult patient  
571 access to electronic health records. *Cochrane Database of Systematic Reviews*. 2021(2).
- 572 21. Medcalf E, Turner R, Espinoza D, Bell K. Methods for dealing with missing outcome data in  
573 randomised controlled trials: a methodological scoping review. 2022.
- 574 22. Carpenter JR, Smuk M. Missing data: A statistical framework for practice. *Biometrical Journal*.  
575 2021;63(5):915-47.
- 576 23. Beaulieu-Jones BK, Moore JH. MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH  
577 RECORD USING DEEPLY LEARNED AUTOENCODERS. *Pacific Symposium on Biocomputing Pacific*  
578 *Symposium on Biocomputing*. 2017;22:207-18.
- 579 24. Welch CA, Petersen I, Bartlett JW, White IR, Marston L, Morris RW, et al. Evaluation of two-  
580 fold fully conditional specification multiple imputation for longitudinal electronic health record data.  
581 *Statistics in medicine*. 2014;33(21):3725-37.
- 582 25. Cesare N, Were LPO. A multi-step approach to managing missing data in time and patient  
583 variant electronic health records. *BMC research notes*. 2022;15(1):64.
- 584 26. Blankers M, Koeter MWJ, Schippers GM. Missing data approaches in eHealth research:  
585 simulation study and a tutorial for nonmathematically inclined researchers. *Journal of medical*  
586 *Internet research*. 2010;12(5):e54.
- 587 27. Ross RK, Breskin A, Westreich D. When Is a Complete-Case Approach to Missing Data Valid?  
588 The Importance of Effect-Measure Modification. *Am J Epidemiol*. 2020;189(12):1583-9.
- 589 28. Mukaka M, White SA, Terlouw DJ, Mwapasa V, Kalilani-Phiri L, Faragher EB. Is using multiple  
590 imputation better than complete case analysis for estimating a prevalence (risk) difference in  
591 randomized controlled trials when binary outcome observations are missing? *Trials*. 2016;17(1):1-12.
- 592 29. Lachin JM. Fallacies of last observation carried forward analyses. *Clin Trials*. 2016;13(2):161-  
593 8.
- 594 30. Lydersen S. Last observation carried forward. *Tidsskrift for Den norske legeforening*. 2019.
- 595 31. Li P, Stuart EA, Allison DB. Multiple imputation: a flexible tool for handling missing data.  
596 *Jama*. 2015;314(18):1966-7.
- 597 32. Liu X. Chapter 14 - Methods for handling missing data. In: Liu X, editor. *Methods and*  
598 *Applications of Longitudinal Data Analysis*. Oxford: Academic Press; 2016. p. 441-73.
- 599 33. Muñoz J, Efthimiou O, Audigier V, de Jong VM, Debray TP. Multiple imputation of incomplete  
600 multilevel data using Heckman selection models. *Statistics in medicine*. 2023.
- 601 34. Umar N, Gray A. Comparing single and multiple imputation approaches for missing values in  
602 univariate and multivariate water level data. *Water*. 2023;15(8):1519.
- 603 35. Yenduri S. An empirical study of imputation techniques for software data sets: Louisiana  
604 State University and Agricultural & Mechanical College; 2005.
- 605 36. Yadav ML, Roychoudhury B. Handling missing values: A study of popular imputation packages  
606 in R. *Knowledge-Based Systems*. 2018;160:104-18.
- 607 37. Mbona SV, Mwambi H, Ramroop S. Multiple imputation using chained equations for missing  
608 data in survival models: applied to multidrug-resistant tuberculosis and HIV data. *Journal of Public*  
609 *Health in Africa*. 2023;14(8).
- 610 38. Guguloth S, Telu A, Sairam U, Voruganti S, editors. *Activity Recognition in Missing Data*  
611 *Scenario Using MICE Algorithm*. International Conference on Soft Computing and Pattern  
612 Recognition; 2022: Springer.
- 613 39. Chen H, Quandt SA, Grzywacz JG, Arcury TA. A Bayesian multiple imputation method for  
614 handling longitudinal pesticide data with values below the limit of detection. *Environmetrics*.  
615 2013;24(2):132-42.
- 616 40. Faisal S, Tutz G. Multiple imputation using nearest neighbor methods. *Information Sciences*.  
617 2021;570:500-16.
- 618 41. Han J, Lee Y, Kim JK. Maximum Likelihood Imputation. arXiv preprint arXiv:220709891. 2022.

- 619 42. Williams R. Missing data part II: Multiple imputation & maximum likelihood. 2017.
- 620 43. Lu P, Shelley M. Testing the missingness mechanism in longitudinal surveys: A case study  
621 using the health and retirement study. *International Journal of Social Research Methodology*.  
622 2023;26(4):439-52.
- 623 44. Myrtveit I, Stensrud E, Olsson UH. Analyzing data sets with missing data: An empirical  
624 evaluation of imputation methods and likelihood-based methods. *IEEE Transactions on Software  
625 Engineering*. 2001;27(11):999-1013.
- 626 45. Allison PD, editor *Handling missing data by maximum likelihood*. SAS global forum; 2012: San  
627 Diego, CA, USA:.
- 628 46. Yuan K-H, Yang-Wallentin F, Bentler PM. ML versus MI for missing data with violation of  
629 distribution conditions. *Sociological methods & research*. 2012;41(4):598-629.
- 630 47. Brusa L, Bartolucci F, Pennoni F. Tempered expectation-maximization algorithm for the  
631 estimation of discrete latent variable models. *Computational Statistics*. 2023;38(3):1391-424.
- 632 48. Le TD, Beuran R, Tan Y, editors. *Comparison of the Most Influential Missing Data Imputation  
633 Algorithms for Healthcare*. 2018 10th International Conference on Knowledge and Systems  
634 Engineering (KSE); 2018 1-3 Nov. 2018.
- 635 49. Aljuaid T, Sasi S, editors. *Proper imputation techniques for missing values in data sets*. 2016  
636 International Conference on Data Science and Engineering (ICDSE); 2016 23-25 Aug. 2016.
- 637 50. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing  
638 data in machine learning. *Journal of Big data*. 2021;8:1-37.
- 639 51. Li P, Qin J, Liu Y. Instability of inverse probability weighting methods and a remedy for  
640 nonignorable missing data. *Biometrics*. 2023.
- 641 52. Kalpourtzi N, Carpenter JR, Touloumi G. Handling missing values in surveys with complex  
642 study design: A simulation study. *Journal of Survey Statistics and Methodology*. 2024;12(1):105-29.
- 643 53. Segura-Buisan J, Leyrat C, Gomes M. Addressing missing data in the estimation of  
644 time-varying treatments in comparative effectiveness research. *Statistics in Medicine*.  
645 2023;42(27):5025-38.
- 646 54. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in  
647 epidemiology. *Epidemiology*. 2000:550-60.
- 648 55. Sidiropoulos ND, De Lathauwer L, Fu X, Huang K, Papalexakis EE, Faloutsos C. Tensor  
649 decomposition for signal processing and machine learning. *IEEE Transactions on signal processing*.  
650 2017;65(13):3551-82.
- 651 56. Beaulieu-Jones BK, Moore JH. MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH  
652 RECORD USING DEEPLY LEARNED AUTOENCODERS. *Pac Symp Biocomput*. 2017;22:207-18.
- 653 57. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value  
654 estimation methods for DNA microarrays. *Bioinformatics*. 2001;17(6):520-5.
- 655 58. Menon AK, Elkan C. Fast algorithms for approximating the singular value decomposition.  
656 *ACM Transactions on Knowledge Discovery from Data (TKDD)*. 2011;5(2):1-36.
- 657 59. Hourani Ma, El EIM. Microarray missing values imputation methods: Critical analysis review.  
658 *Computer Science and Information Systems*. 2009;6(2):165-90.
- 659 60. Cai M, van Buuren S, Vink G. Joint distribution properties of fully conditional specification  
660 under the normal linear model with normal inverse-gamma priors. *Scientific Reports*.  
661 2023;13(1):644.
- 662 61. Welch C, Bartlett J, Petersen I. Application of multiple imputation using the two-fold fully  
663 conditional specification algorithm in longitudinal clinical data. *The Stata Journal*. 2014;14(2):418-31.
- 664 62. Nevalainen J, Kenward MG, Virtanen SM. Missing values in longitudinal dietary data: a  
665 multiple imputation approach based on a fully conditional specification. *Statistics in medicine*.  
666 2009;28(29):3657-69.
- 667 63. Xu D, Hu PJ-H, Huang T-S, Fang X, Hsu C-C. A deep learning-based, unsupervised method to  
668 impute missing values in electronic health records for improved patient management. *Journal of  
669 Biomedical Informatics*. 2020;111:103576.

- 670 64. Xu D, Sheng JQ, Hu PJH, Huang TS, Hsu CC. A Deep Learning–Based Unsupervised Method to  
671 Impute Missing Values in Patient Records for Improved Management of Cardiovascular Patients. *IEEE*  
672 *Journal of Biomedical and Health Informatics*. 2021;25(6):2260-72.
- 673 65. Charte D, Charte F, García S, del Jesus MJ, Herrera F. A practical tutorial on autoencoders for  
674 nonlinear feature fusion: Taxonomy, models, software and guidelines. *Information Fusion*.  
675 2018;44:78-96.
- 676 66. Esser-Skala W, Fortelny N. Reliable interpretability of biology-inspired deep neural networks.  
677 *NPJ Systems Biology and Applications*. 2023;9(1):50.
- 678 67. Avelar PHdC, Wu M, Tsoka S. Incorporating Prior Knowledge in Deep Learning Models via  
679 Pathway Activity Autoencoders. *arXiv preprint arXiv:230605813*. 2023.
- 680 68. Mazumder R, Hastie T, Tibshirani R. Spectral regularization algorithms for learning large  
681 incomplete matrices. *The Journal of Machine Learning Research*. 2010;11:2287-322.
- 682 69. Gad AM, Abdelkhalek RHM. Imputation methods for longitudinal data: A comparative study.  
683 *International Journal of Statistical Distributions and Applications*. 2017;3(4):72.
- 684 70. Getz K, Hubbard RA, Linn KA. Performance of Multiple Imputation Using Modern Machine  
685 Learning Methods in Electronic Health Records Data. *Epidemiology*. 2023;34(2):206-15.
- 686 71. Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for  
687 missing data in longitudinal studies. *BMC medical research methodology*. 2018;18:1-16.
- 688 72. De Silva AP, Moreno-Betancur M, De Livera AM, Lee KJ, Simpson JA. A comparison of multiple  
689 imputation methods for handling missing values in longitudinal data in the presence of a time-  
690 varying covariate with a non-linear association with time: a simulation study. *BMC medical research*  
691 *methodology*. 2017;17:1-11.
- 692 73. Jazayeri A, Liang OS, Yang CC. Imputation of missing data in electronic health records based  
693 on patients' similarities. *Journal of Healthcare Informatics Research*. 2020;4(3):295-307.
- 694 74. Li J, Guo S, Ma R, He J, Zhang X, Rui D, et al. Comparison of the effects of imputation  
695 methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research*  
696 *Methodology*. 2024;24(1):41.
- 697 75. Yu H, Perumean-Chaney SE, Kaiser KA. What is Missing in Missing Data Handling? An  
698 Evaluation of Missingness in and Potential Remedies for Doctoral Dissertations and Subsequent  
699 Publications that Use NHANES Data. *Journal of Statistics and Data Science Education*. 2024;32(1):3-  
700 10.
- 701 76. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and  
702 handling of missing data in cohort studies with repeated assessment of exposure measures. *BMC*  
703 *medical research methodology*. 2012;12:1-10.

704





## Identification of studies via databases and registers

Identification

Records identified from:  
Embase (n = 101 )

Records removed *before*  
*screening*:  
Duplicate records removed  
(n = 2 )

Screening

**Title and abstract review:**  
Records screened  
(n = 99)

Records excluded\*\*  
(n = 78)

**Full text review:**  
Reports assessed for eligibility  
(n = 21 )

Reports excluded:  
(n = 16)

Included

Studies included in review  
(n = 5 )  
Reports of included by manual  
search (n = 4)