

Chapter 13

Missing Data

Cátia M. Salgado, Carlos Azevedo, Hugo Proença
and Susana M. Vieira

Learning Objectives

- What are the different types of missing data, and the sources for missingness.
- What options are available for dealing with missing data.
- What techniques exist to help choose the most appropriate technique for a specific dataset.

13.1 Introduction

Missing data is a problem affecting most databases and electronic medical records (EHR) are no exception. Because most statistical models operate only on complete observations of exposure and outcome variables, it is necessary to deal with missing data, either by deleting incomplete observations or by replacing any missing values with an estimated value based on the other information available, a process called imputation. Both methods can significantly effect the conclusions that can be drawn from the data.

Identifying the source of “missingness” is important, as it influences the choice of the imputation technique. Schematically, several cases are possible: (i) the value is missing because it was forgotten or lost; (ii) the value is missing because it was not applicable to the instance; (iii) the value is missing because it is of no interest to the instance. If we were to put this in a medical context: (i) the variable is measured but for some unidentifiable reason the values are not electronically recorded, e.g. disconnection of sensors, errors in communicating with the database server, accidental human omission, electricity failures, and others; (ii) the variable is not measured during a certain period of time due to an identifiable reason, for instance the patient is disconnected from the ventilator because of a medical decision;

Electronic supplementary material The online version of this chapter (doi:[10.1007/978-3-319-43742-2_13](https://doi.org/10.1007/978-3-319-43742-2_13)) contains supplementary material, which is available to authorized users.

(iii) the variable is not measured because it is unrelated with the patient condition and provides no clinical useful information to the physician [1].

An important distinction must be made between data missing for identifiable or unidentified reasons. In the first case, imputing values can be inadequate and add bias to the dataset, so the data is said to be non-recoverable. On the other hand, when data is missing for unidentifiable reasons it is assumed that values are missing because of random and unintended causes. This type of missing data is classified as recoverable.

The first section of this chapter focuses on describing the theory of some commonly used methods to handle missing data. In order to demonstrate the advantages and disadvantages of the methods, their application is demonstrated in the second part of the chapter on actual datasets that were created to study the relation between mortality and insertion of indwelling arterial catheters (IAC) in the intensive care unit (ICU).

13.2 Part 1—Theoretical Concepts

In knowledge discovery in databases, data preparation is the most crucial and time consuming task, that strongly influences the success of the research. Variable selection consists in identifying a useful subset of potential predictors from a large set of candidates (please refer to Chap. 5—Data Analysis for further information on feature selection). Rejecting variables with an excessive number of missing values (e.g. >50 %) is usually a good rule of thumb, however it is not a risk-free procedure. Rejecting a variable may lead to a loss of predictive power and ability to detect statistically significant differences and it can be a source of bias, affecting the representativeness of the results. For these reasons, variable selection needs to be tailored to the missing data mechanism. Imputation can be done before and/or after variable selection.

The general steps that should be followed for handling missing data are:

- Identify patterns and reasons for missing data;
- Analyse the proportion of missing data;
- Choose the best imputation method.

13.2.1 *Types of Missingness*

The mechanisms by which the data is missing will affect some assumptions supporting our data imputation methods. Three major mechanisms of missingness of the data can be described, depending on the relation between observed (available) and unobserved (missing) data.

For the sake of simplicity, let's consider missingness in the univariate case. To define missingness in mathematical terms, a dataset X can be divided in two parts:

$$X = \{X_o, X_m\} \quad (1)$$

where X_o corresponds to the observed data, and X_m to the missing data, in the dataset.

For each observation we define a binary response whether or not that observation is missing:

$$R = \begin{cases} 1 & \text{if } X \text{ observed} \\ 0 & \text{if } X \text{ missing} \end{cases} \quad (2)$$

The missing value mechanism can be understood in terms of the probability that an observation is missing $\Pr(R)$ given the observed and missing observations, in the form:

$$\Pr(R|x_o, x_m) \quad (3)$$

The three mechanisms are subject to whether the probability of response R depends or not on the observed and/or missing values:

- **Missing Completely at Random (MCAR)**—When the missing observations are dependent on the observed and unobserved measurements. In this case the probability of an observation being missing depends only on itself, and reduces to $\Pr(R|x_o, x_m) = \Pr(R)$. As an example, imagine that a doctor forgets to record the gender of every six patients that enter the ICU. There is no hidden mechanism related to any variable and it does not depend on any characteristic of the patients.
- **Missing at Random (MAR)**—In this case the probability of a value being missing is related only to the observable data, i.e., the observed data is statistically related with the missing variables and it is possible to estimate the missing values from the observed data. This case is not completely ‘random’, but it is the most general case where we can ignore the missing mechanism, as we control the information upon which the missingness depends, the observed data. Said otherwise, the probability that some data is missing for a particular variable does not depend on the values of that variable, after adjusting for observed values. Mathematically the probability of missing reduces to $\Pr(R|x_o, x_m) = \Pr(R|x_o)$. Imagine that if elderly people are less likely to inform the doctor that they had had a pneumonia before, the response rate of the variable pneumonia will depend on the variable age.
- **Missing Not at Random (MNAR)**—This refers to the case when neither MCAR nor MAR hold. The missing data depends on both missing and observed values. Determining the missing mechanism is usually impossible, as it depends on unseen data. From that derives the importance of performing sensitivity analyses and test how the inferences hold under different assumptions. For example, we can imagine that patients with low blood pressure are more likely to have their blood pressure measured less frequently (the missing data for the variable “blood pressure” partially depends on the values of the blood pressure).

13.2.2 Proportion of Missing Data

The percentage of missing data for each variable (between patients) and each patient (between variables) must be computed, to help decide which variables and/or patients should be considered candidates for removal or data imputation. A crude example is shown in Table 13.1, where we might want to consider removing patient 1 and the variable “AST” from the analysis, considering that most of their values are missing.

13.2.3 Dealing with Missing Data

Overview of Methods for Handling Missing Data

The methods should be tailored to the dataset of interest, the reasons for missingness and the proportion of missing data. In general, a method is chosen for its simplicity and its ability to introduce as little bias as possible in the dataset.

When data are MCAR or MAR a researcher can ignore the reasons for missing data, which simplifies the choice of the methods to apply. In this case, any method can be applied. Nevertheless it is difficult to obtain empirical evidence about whether or not the data are MCAR or MAR. A valid strategy is to examine the sensitivity of results to the MCAR and MAR assumptions by comparing several analyses, where the differences in results across several analyses may provide some information about what assumptions may be the most relevant.

A significant body of evidence has focused on comparing the performance of missing data handling methods, both in general [2–4] and in context of specific factors such as proportion of missing data and sample size [5–7]. More detailed technical aspects, and application of these methods in various fields can also be found in the works of Jones and Little [8, 9].

In summary, the most widely used methods fall into three main categories, which are described in more detail below.

1. Deletion methods (listwise deletion, i.e. complete-case analysis, pairwise deletion, i.e. available-case analysis)
2. Single Imputation Methods (mean/mode substitution, linear interpolation, Hot deck and cold deck)
3. Model-Based Methods (regression, multiple imputation, k-nearest neighbors)

Table 13.1 Examples of missing data in EHR

	Gender	Glucose	AST	Age
Patient 1	?	120	?	?
Patient 2	M	105	?	68
Patient 3	F	203	45	63
Patient 4	M	145	?	42
Patient 5	M	89	?	80

Deletion Methods

The simplest way to deal with missing data is to discard the cases or observations that have missing values. In general, case deletion methods lead to valid inferences only for MCAR [10]. There are three ways of doing this: complete-case analysis; available-case analysis; and weighting methods.

Complete-Case Analysis (Listwise Deletion)

In complete case analysis, all the observations with at least one missing variable are discarded (Fig. 13.1).

The principal assumption is that the remaining subsample is representative of the population, and will thus not bias the analysis towards a subgroup. This assumption is rather restrictive and assumes a MCAR mechanism. Listwise deletion often produces unbiased regression slope estimates, as long as missingness is not a function of the outcome variable. The biggest advantage of this method is its simplicity, it is always reasonable to use it when the number of discarded observations is relatively small when compared to the total. Its main drawbacks are the reduced statistical power (because it reduces the number of samples n , the estimates will have larger standard errors), waste of information, and possible bias of the analysis specially if data is not MCAR.

Fig. 13.1 Example of complete-case deletion. Cases highlighted in red are discarded

Gender	GLUCOSE	Age
M	?	65
F	120	71
F	99	?
F	140	52
M	88	?
F	85	63
M	170	68
?	153	80
M	115	59
F	103	?

Available-Case Analysis

The available-case method discards data only in the variables that are needed for a specific analysis. For example, if only 4 out of 20 variables are needed for a study, this method would only discard the missing observations of the 4 variables of interest. In Fig. 13.2, imagine that each one of the three represented variables would be used for a different analysis. The analysis is performed using all cases in which the variables of interest are present. Even though this method has the ability to preserve more information, the populations of each analysis would be different and possibly non-comparable.

Weighting-Case Analysis

Weighting is a way of weighting the complete-cases by modelling the missingness in order to reduce the bias introduced in the available-case.

Single-Value Imputation

In single imputation, missing values are filled by some type of “predicted” values [9, 11]. Single imputation ignores uncertainty and almost always underestimates the variance. Multiple imputation overcomes this problem, by taking into account both within—and between—imputation uncertainty.

Fig. 13.2 Example of available-case deletion. If each variable is used for separate analyses, only the cases in which the variable of interest is missing are discarded

Case Study		
S1	S2	S3
Gender	GLUCOSE	Age
M	?	65
F	120	71
F	99	?
F	140	52
M	88	?
F	85	63
M	170	68
?	153	80
M	115	59
F	103	?

Mean and Median

The simplest imputation method is to substitute missing values by the mean or the median of that variable. Using the median is more robust in the presence of outliers in the observed data. The main disadvantages are that (1) it reduces variability, thereby lowering the estimate errors compared to deletion approaches, and (2) it disregards the relationship between variables, decreasing therefore their correlation. While this method diminishes the bias of using a non-representative sample, it introduces other bias.

Linear Interpolation

This method is particularly suitable for time-series. In linear interpolation, a missing value is computed by interpolating the values of the previous and next available measurements for the patient. For example, if the natremia changes from 132 to 136 mEq/L in 8 h, one can reasonably assume that its value was close to 134 mEq/L at midpoint.

Hot Deck and Cold Deck

In the hot deck method, a missing attribute value is replaced with a value from an estimated distribution of the current data. It is especially used in survey research [9]. Hot deck is typically implemented in two stages. First, the data is partitioned into clusters, and then each instance with missing data is associated with one cluster. The complete cases in a cluster are used to fill in the missing values. This can be done by calculating the mean or mode of the attribute within a cluster. Cold deck imputation is similar to hot deck, except that the data source is different from the current dataset. Hot-deck imputation replaces the missing data by realistic values that preserve the variable distribution. However it underestimates the standard errors and the variability [12].

Last Observation Carried Forward

Sometimes called “sample-and-hold” method [13]. The last value carried forward method is specific to longitudinal designs. This technique imputes the missing value with the last available observation of the individual. This method makes the assumption that the observation of the individual has not changed at all since the last measured observation, which is often unrealistic [14].

Model-Based Imputation

In model-based imputation, a predictive model is created to estimate values that will substitute the missing data. In this case, the dataset is divided into two subsets: one with no missing values for the variable under evaluation (used for training the model) and one containing missing values, that we want to estimate. Several modeling methods can be used such as: regression, logistic regression, neural networks and other parametric and non-parametric modeling techniques. There are two main drawbacks in this approach: the model estimates values are usually more well-behaved than the true values, and the models perform poorly if the observed and missing variables are independent.

Linear Regression

In this model, all the available variables are used to create a linear regression model using the available observations of the variable of interest as output. The advantages of this method is that it takes into account the relationship between variables, unlike the mean/median imputation. The disadvantages are that it overestimates the model fit and the correlation between the variables, as it does not take into account the uncertainty in the missing data and underestimates variances and covariances. A method that was created to introduce uncertainty is the stochastic linear regression (see below).

The case of multivariate imputation is more complex as missing values exist for several variables, which do not follow the same pattern of missingness through the observations. The method used is a multivariate extension of the linear model and relies on an iterative process carried until convergence.

Stochastic Regression

Stochastic regression imputation aims to reduce the bias by an extra step of augmenting each predicted score with a residual term. This residual term is normally distributed with a mean of zero and a variance equal to the residual variance from the regression of the predictor on the target. This method allows to preserve the variability in the data and unbiased parameter estimates with MAR data. However, the standard error tends to be underestimated, because the uncertainty about the imputed values is not included, which increases the risk of type I error [15].

Multiple-Value Imputation

Multiple Imputation (MI) is a powerful statistical technique developed by Rubin in the 1970s for analysing datasets containing missing values [7, 16]. It is a Monte Carlo technique that requires 3 steps (Fig. 13.3).

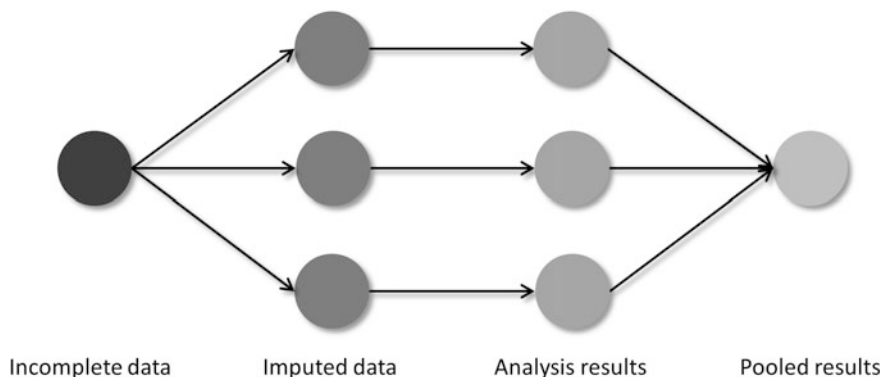


Fig. 13.3 The concept of multiple imputation, with $M = 3$

- Imputation, where the missing values are filled in using any method of choice, leading to $M \geq 2$ completed datasets (5–10 is generally sufficient) [10]. In these M multiply-imputed datasets, all the observed values are the same, but the imputed values are different, reflecting the uncertainty about imputation [10].
- Analysis: each of the M completed datasets is analysed (e.g. a logistic regression classifier for mortality prediction is built), which gives M analyses.
- Pooling: the M analyses are integrated into a final result, for example by computing the mean (and 95 % CI) of the M analyses.

K-Nearest Neighbors

K-nearest neighbors (kNN) can be used for handling missing values. Here, they will be filled with the mean of the k values coming from the k most similar complete observations. The similarity of two observations is determined, after normalization of the dataset, using a distance function which can be Euclidean, Manhattan, Mahalanobis, Pearson, etc. The main advantage of the kNN algorithm is that given enough data it can predict with a reasonable accuracy the conditional probability distribution around a point and thus make well informed estimations. It can predict qualitative and quantitative (discrete and continuous) attributes. Another advantage of this method is that the correlation structure of the data is taken into consideration. The choice of the k -value is very critical. A higher value of k would include attributes which are significantly different from our target observation, while lower value of k implies missing out of significant attributes.

13.2.4 Choice of the Best Imputation Method

Different imputation methods are expected to perform differently on various datasets. We describe here a generic and simple method that can be used to evaluate the performance of various imputation methods on your own dataset, in order to help selecting the most appropriate method. Of note, this simple approach does not test the effect of deletion methods. A more complex approach is described in the case study below, in which the performance of a predictive model is tested on the dataset completed by various imputation methods.

Here is how to proceed:

1. Use a sample of your own dataset that does not contain any missing data (will serve as ground truth).
2. Introduce increasing proportions of missing data at random (e.g. 5–50 % in 5 % increments).
3. Reconstruct the missing data using the various methods.
4. Compute the sum of squared errors between the reconstructed and the original data, for each method and each proportion of missing data.
5. Repeat steps 1–4 a number of times (10 times for example) and compute the average performance of each method (average SSE).
6. Plot the average SSE versus proportion of missing data (1 plot per imputation method), similarly to the example shown in Fig. 13.4.

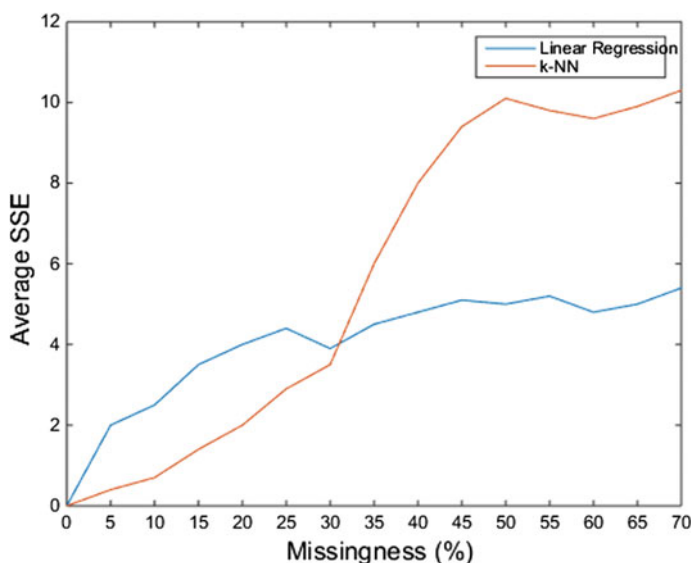


Fig. 13.4 Average SSE between original and reconstructed data, for various levels of missingness and 2 imputation methods (data only for illustrative purposes)

7. Choose the method that performs best at the level of missing data in your dataset. E.g. if your data had 10 % of missing data, you would want to pick k-NN; at 40 % linear regression performs better (made-up data, for illustrative purpose only).

13.3 Part 2—Case Study

In this section, various imputation methods will be applied to two “real world” clinical datasets used in a study that investigated the effect of inserting an indwelling arterial catheter (IAC) in patients with respiratory failure. Two datasets are used, and include patients that received an IAC (IAC group) and patients that did not (non-IAC). Each dataset is subdivided into 2 classes, with class 1 corresponding to patients that died within 28 days and class 0 to survivors. The proportion of missing data and potential reasons for missingness are discussed first. The following analyses were then carried out:

1. Various proportions of missing data at random were inserted into the variable “age”, then imputed using the various methods described above. The distribution of the imputed observations was compared to the original distribution for all the methods.
2. The performance of imputed datasets with different degrees of missingness was tested on a predictive model (logistic regression to predict mortality), first for univariate missing data (the variable age), then for all the variables (multivariate).

The code used to generate the analyses and the figures is provided in the in the accompanying R functions document.

13.3.1 Proportion of Missing Data and Possible Reasons for Missingness

Table 13.2 shows the proportion of missing data in some of the variables of the datasets. 26 variables represent the subset that was considered for testing the different imputation methods, and were selected based on the assumption that missing data occurring in these variables is recoverable.

Since IAC are mainly used for continuous hemodynamic monitoring and for arterial blood sampling for blood gas analysis, we can expect a higher percentage of missing data in blood gas-related variables in the non-IAC group. We can also expect that patient diagnoses are often able to provide an explanation for the lack of specific laboratory results: if a certain test is not ordered because it will most likely provide no clinical insight, a missing value will occur; it is fair to estimate that such

Table 13.2 Missing data in some of the variables of the IAC and non-IAC datasets

	IAC		Non-IAC	
	# points	%	# points	%
Arterial line time day	0	0	792	100
Hospital length of stay	0	0	0	0
Age	0	0	0	0
Gender	0	0	0	0
Weight first	39	3.96	71	8.96
SOFA first	2	0.20	4	0.51
Hemoglobin first	2	0.20	5	0.63
Bilirubin first	418	42.48	365	46.09
...				

value lies within a normal range. In both cases, the fact that data is missing contains information about the response, thus it is MNAR. Body mass index (BMI) has a relatively high percentage of missing data. Assuming that this variable is calculated automatically from the weight and height of patients, we can conclude that this data is MAR: because the height and/or weight are missing, BMI cannot be calculated. If the weight is missing because someone forgot to introduce it into the system then it is MCAR. Besides the missing data mechanism, it is also important to consider the sample distribution in each variable, as some imputation methods assume specific data distributions, usually the normal distribution.

13.3.2 Univariate Missingness Analysis

In this section, the specific influence of each imputation method will be explored for the variable age, using all the other variables. Two different levels of missingness (20 and 40 %) were artificially introduced in the datasets. The original dataset represents the ground truth, to which the imputed datasets were compared using frequency histograms.

Complete-Case Analysis

The complete-case analysis method discards all the incomplete observations with at least one missing value. The distribution of the “imputed” dataset is going to be equal to the original dataset minus the observations that have a missing value in variable age. Figure 13.5 shows an example of the distribution of the variable age in the IAC group.

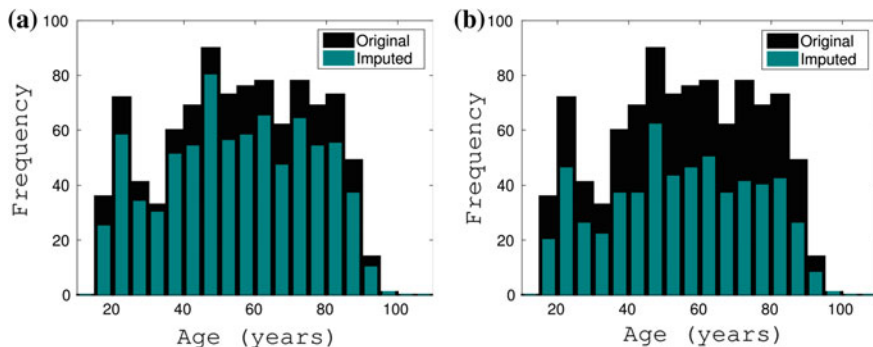


Fig. 13.5 Histogram of variable age in the IAC group before and after univariate complete case method

This method is only exploitable when there is a small percentage of missing data. This method does not require any assumption in the distribution of the missing data, besides that the complete cases should be representative of the original population, which is difficult to prove.

Single Value Imputation

Mean and Median Imputation

Mean and median methods are very crude imputation techniques, which ignore the relationship between age and the other variables and introduce a heavy bias towards the mean/median values. These simple methods allow us to better understand the biasing effect, something that is obvious in the examples Fig. 13.6.

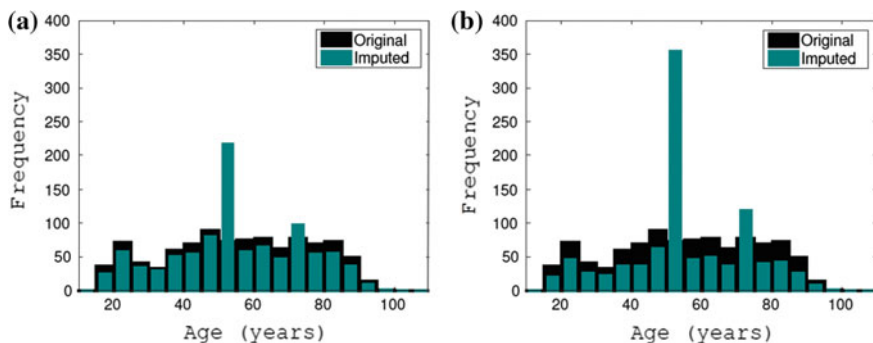


Fig. 13.6 Histogram of variable age in the IAC group before (original) and after (imputed) mean for univariate imputation

Linear Regression Imputation

The linear regression method imputes most of the data at the center of the distribution (example in Fig. 13.7). The extremities of the distribution are not well modeled and are easily ignored. This is due to two features of this technique: first, the assumption that the linear regression is a good fit to the data, and second, the assumption that the missing data lays over the regression line, bending the reality to fit the deterministic nature of the model. Compared to the mean/median imputation, the linear regression assumes a relation between the variables, however it overestimates this relation by assuming that the missing points are over the regression line. The model assumes that the percentage of variance explained is 100 %, thus it underestimates variability.

Stochastic Linear Regression Imputation

The stochastic linear regression is an attempt to loosen the deterministic assumption of the linear regression. In this case, the distribution of the imputed data fits better the original data than previous methods (Fig. 13.8). This method can introduce impossible values, such as negative age. It is a first step to model the uncertainty present in the dataset that represents a trade-off between the precision of the values and the uncertainty introduced by the missing data.

K-Nearest Neighbors

We limit the demonstration to the case where $k = 1$. In the extreme case where all neighbors are used without weights, this method converges to the mean imputation.

Figure 13.9 demonstrates that this method introduces in our particular dataset a huge bias towards the central value. The reason for this arises from the fact that

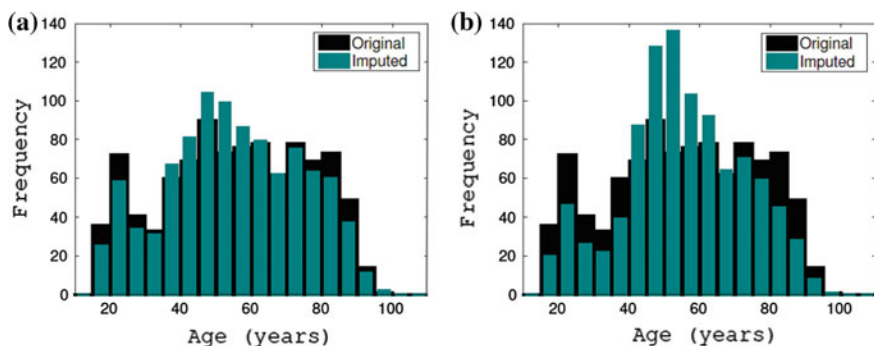


Fig. 13.7 Histogram of the variable age in the IAC group before (original) and after (imputed) linear for univariate imputation

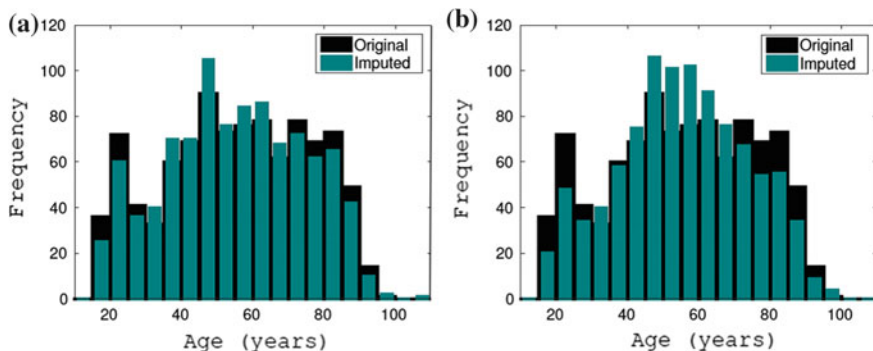


Fig. 13.8 Histogram of variable age in the IAC group before (original) and after (imputed) stochastic linear for univariate imputation

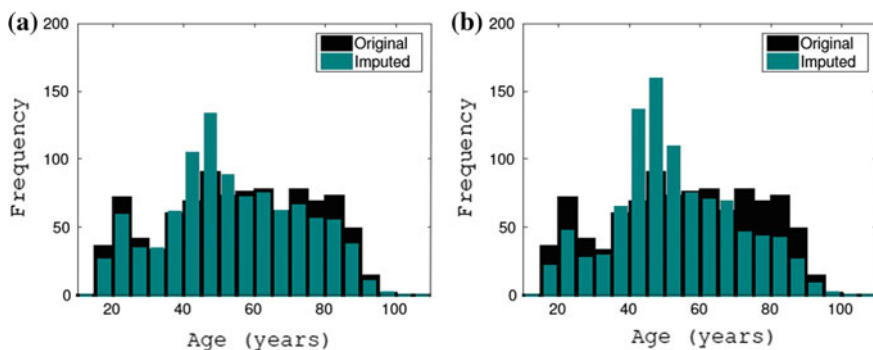


Fig. 13.9 Histogram of variable age in the IAC group before (original) and after (imputed) KNN for univariate imputation

almost half of the variables are binary, which end up having a much higher weight on the distances than continuous variables (which are always less than 1, due to the unitary normalization performed in data pre-processing). Computations with kNN increase in quality with the number of observations in the dataset, and indeed this method is very powerful given the right conditions.

Multiple Imputation

Multiple imputation with linear regression and multivariate normal regression are extensions of the single imputation methods of the same name and use sampling to create multiple different datasets, that represent different possibilities of what might be the original dataset. These methods allow a better modeling of the uncertainty present in the missing values and are, usually, more solid in terms of statistical

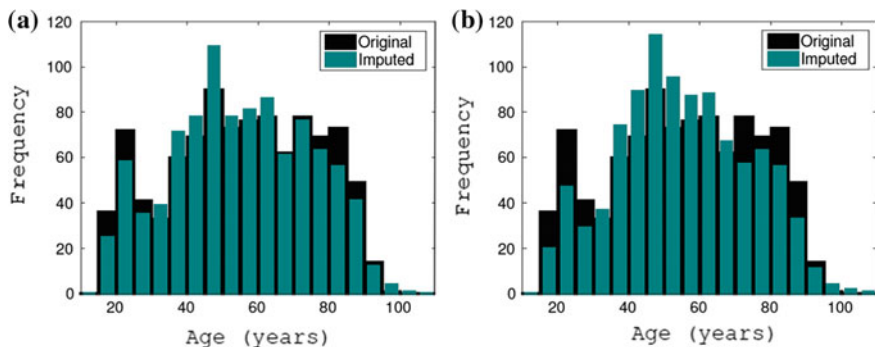


Fig. 13.10 Histogram of variable age in the IAC group before (original) and after (imputed) multiple imputation multivariate normal regression for univariate imputation

properties and results. We chose to work with 10 datasets, which were averaged so that the graphical representation would look similar to the previous methods.

Multivariate normal regression

Multiple imputation multivariate normal distribution gave more importance to the values of the center of the distribution (Fig. 13.10). The main assumption of this method is that the data follows a multivariate normal distribution, something that is not completely true for this dataset, which contains numerous binary variables. Nonetheless, even in the presence of categorical variables and distributions that are not strictly normal, it should perform reasonably well [10, 19]. The multiple imputation method enhances the modeling of uncertainty by adding a bootstrap sampling to the expectation maximization algorithm, giving raise to better predictions of the possible missing data by considering multiple possibilities of the original data. Obviously, when averaging the data for histogram representation, some of that richness is lost. Nonetheless, the quality of the regression is obvious when compared to the previous methods.

Linear regression

The multiple imputation linear regression method uses all the variables except the target variable (age) to estimate the missing data of this last variable. The data is modelled using linear regression and Gibbs sampling. Figure 13.11 demonstrates that this represents by far the most accurate imputation method in this particular dataset.

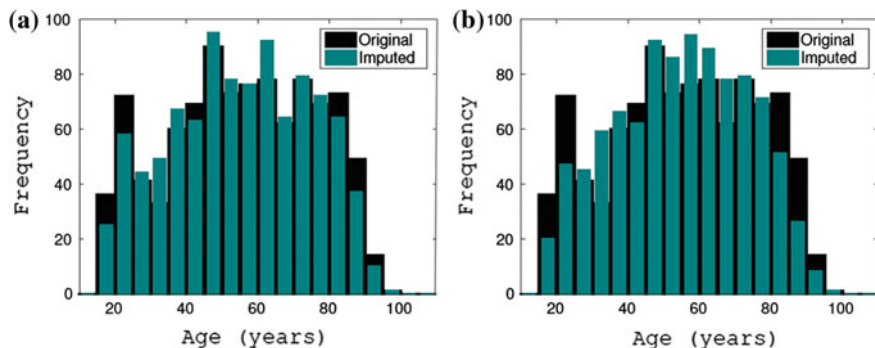


Fig. 13.11 Histogram of variable age in the IAC group before (original) and after (imputed) multiple imputation generalized regression for univariate imputation

13.3.3 *Evaluating the Performance of Imputation Methods on Mortality Prediction*

This test aims to assess the generalization capabilities of the models constructed using imputed data, and check their performance by comparing them to the original data. All the methods described previously were used to reconstruct a sample of both IAC and non-IAC datasets, with increasing proportions of missing data at random, first only on the variable age (univariate), then on all the variables in the dataset (multivariate). A logistic regression model was built on the reconstructed data and tested on a sample of the original data (that does not contain imputations or missing data).

The performance of the models is evaluated in terms of area under the receiver operating characteristic curve (AUC), accuracy (correct classification rate), sensitivity (true positive classification rate—TPR, also known as recall), specificity (true negative classification rate—TNR) and Cohen’s kappa. All the methods were compared against a reference logistic regression that was fitted with the original data without missingness. The results were averaged over a 10-fold cross validation and the AUC results are presented graphically.

The influence of one variable has a limited effect, even if age is the variable most correlated with mortality (Fig. 13.12). At most, the AUC decreased from 0.84 to 0.81 for IAC and from 0.90 to 0.87 for the non-IAC case, if we exclude the complete-case analysis method that performs poorly from the beginning. For lower values of missingness (less than 50 %), all the other models perform similarly. Among univariate techniques, the methods that performed the best on both datasets are the two multiple imputation methods, namely the linear regression and the multivariate normal distribution, and the one-nearest neighbors algorithm. In the case of univariate missingness, the nearest neighbors reveals to be a good estimator if several complete observations exist, as it is the case. With increasing of the

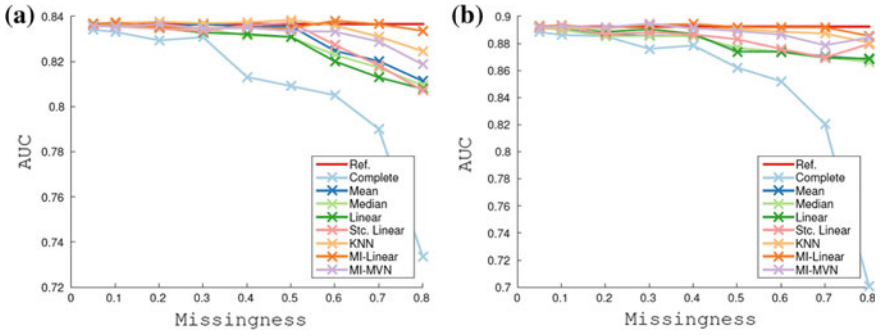


Fig. 13.12 Mean AUC performance of the logistic regression models modelled with different imputation methods for different degrees of univariate missingness of the Age variable

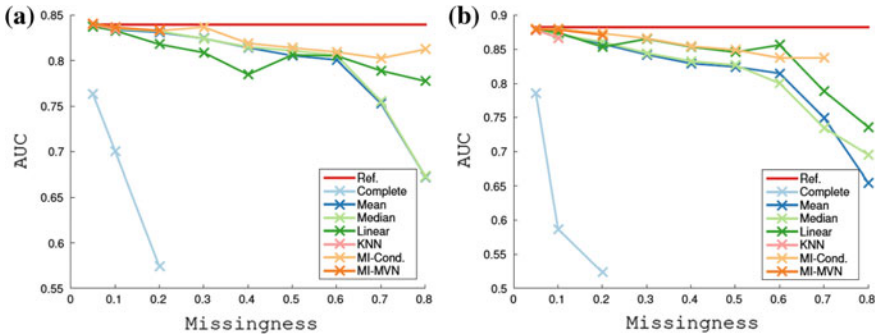


Fig. 13.13 Mean AUC of the logistic regression models for different degrees of multivariate missingness

missingness, the simpler methods introduced more bias in the modeling of the datasets.

The quality of the imputation methods was also evaluated in the presence of multivariate missingness with an uniform probability in all variables (Fig. 13.13). It has to be noted that obtaining results for more than 40 % of missingness in all the variables is quite infeasible in most cases, and there are no assurances of good performances with any of the methods. Some methods were not able to perform complete imputations over a certain degree of missingness (e.g. the complete-case analysis stopped having enough observations after 20 % of missingness).

Overall, and quite surprisingly, the methods had a reasonable performance even for 80 % of missingness in every variable. The reason behind this is that almost half of the variables are binary, and because of their relation with the output, reconstructing them from frequent values in each class is usually the best guess. The decrease in AUC was due to a decrease in the sensitivity, as the specificity values remained more or less unchanged with the increase in missingness. The method that performed the best overall in terms of AUC was the multiple imputation linear

regression. In IAC it achieved a minimum value of AUC of 0.81 at 70 % of missingness, corresponding to a reference AUC of 0.84 and in non-IAC it achieved an AUC of 0.85 at 70 % of missingness, close to the reference AUC of 0.89.

13.4 Conclusion

Missing data is a widespread problem in EHR due to the nature of medical information itself, the massive amounts of data collected, the heterogeneity of data standards and recording devices, data transfers and conversions, and finally Human errors and omissions. When dealing with the problem of missing data, just like in many other domains of data mining, there is no one-size-fits-all approach, and the data scientist should ultimately rely on robust evaluation tools when choosing an imputation method to handle missing values in a particular dataset.

Take-Home Messages

- Always evaluate the reasons for missingness: is it MCAR/MAR/MNAR?
- What is the proportion of missing data per variable and per record?
- Multiple imputation approaches generally perform better than other methods.
- Evaluation tools must be used to tailor the imputation methods to a particular dataset.

Open Access This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, duplication, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, a link is provided to the Creative Commons license and any changes made are indicated.

The images or other third party material in this chapter are included in the work's Creative Commons license, unless indicated otherwise in the credit line; if such material is not included in the work's Creative Commons license and the respective action is not permitted by statutory regulation, users will need to obtain permission from the license holder to duplicate, adapt or reproduce the material.

References

1. Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN (2013) Missing data in medical databases: impute, delete or classify? *Artif Intell Med* 58(1):63–72
2. Peng CY, Harwell MR, Liou SM, Ehman LH (2006) Advances in missing data methods and implications for educational research
3. Peugh JL, Enders CK (2004) Missing data in educational research: a review of reporting practices and suggestions for improvement. *Rev Educ Res* 74(4):525–556
4. Young W, Weckman G, Holland W (2011) A survey of methodologies for the treatment of missing values within datasets: limitations and benefits. *Theor Issues Ergon Sci* 12(1):15–43

5. Alesh M (2009) The impact of missing data in a generalized integer-valued autoregression model for count data. *J Biopharm Stat* 19(6):1039–1054
6. Knol MJ, Janssen KJM, Donders ART, Egberts ACG, Heerdink ER, Grobbee DE, Moons KGM, Geerlings MI (2010) Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J Clin Epidemiol* 63(7):728–736
7. Little RJA, Rubin DB (2002) Missing data in experiments. In: *Statistical analysis with missing data*. Wiley, pp 24–40
8. Jones MP (1996) Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J Am Stat Assoc* 91(433):222–230
9. Little RJA (2016) *Statistical analysis with missing data*. Wiley, New York
10. Schafer JL (1999) Multiple imputation: a primer. *Stat Methods Med Res* 8(1):3–15
11. de Waal T, Pannekoek J, Scholtus S (2011) *Handbook of statistical data editing and imputation*. Wiley, New York
12. Roth PL (1994) Missing data: a conceptual review for applied psychologists. *Pers Psychol* 47(3):537–560
13. Hug CW (2009) *Detecting hazardous intensive care patient episodes using real-time mortality models*. Thesis, Massachusetts Institute of Technology
14. Wood AM, White IR, Thompson SG (2004) Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials* 1(4):368–376
15. Enders CK (2010) *Applied missing data analysis*, 1st edn. The Guilford Press, New York
16. Rubin DB (1988) An overview of multiple imputation. In: *Proceedings of the survey research section, American Statistical Association*, pp 79–84
17. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. *Crit Care Med* 39(5):952–960
18. Scott DJ, Lee J, Silva I, Park S, Moody GB, Celi LA, Mark RG (2013) Accessing the public MIMIC-II intensive care relational database for clinical research. *BMC Med Inform Decis Mak* 13(1):9
19. Schafer JL, Olsen MK (1998) Multiple imputation for multivariate missing-data problems: a data analyst’s perspective. *Multivar Behav Res* 33(4):545–571