Quality Control, Data Cleaning, Imputation

Dawei Liu, Hanne I. Oberman, Johanna Muñoz, Jeroen Hoogland and Thomas P. A. Debray

Abstract

This chapter addresses important steps during the quality assurance and control of RWD, with particular emphasis on the identification and handling of missing values. A gentle introduction is provided on common statistical and machine learning methods for imputation. We discuss the main strengths and weaknesses of each method, and compare their performance in a literature review. We motivate why the imputation of RWD may require additional efforts to avoid bias, and highlight recent advances that account for informative missingness and repeated observations. Finally, we introduce alternative methods to address incomplete data without the need for imputation.

Keywords

Missing data · Imputation · Missing at random · Missing not at random · Missing completely at random · Informative missingness · Sporadically missing · Systematically missing · Joint modelling imputation · Conditional modelling imputation · Machine learning imputation · Nearest neighbor · Matrix completion · Tree-based ensembles · Support vector machines · Neural networks · Rubin's rules · Pattern submodels · Surrogate splits · Missing indicator · Heckman selection model

1 Introduction

1.1 Quality Control

Increasingly often, researchers have access to data collected from the routine clinical practice with information on patient health or the delivery of health care from a variety of sources other than traditional clinical trials [1, 2]. These data are also known as Real World Data (RWD).

D. Liu
Biogen Digital Health, Biogen, 225 Binney Street, Cambridge, MA 02142, USA

H. I. Oberman
Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands

J. Muñoz · J. Hoogland · T. P. A. Debray
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, Utrecht, The Netherlands

T. P. A. Debray
Health Data Research UK and Institute of Health Informatics, University College London, Gibbs Building, 215 Euston Road, London NW1 2BE, UK

T. P. A. Debray (✉)
Smart Data Analysis and Statistics, Utrecht, The Netherlands
e-mail: tdebray@fromdatatowisdom.com

F. W. Asselbergs et al. (eds.), *Clinical Applications of Artificial Intelligence in Real-World Data*,
https://doi.org/10.1007/978-3-031-36678-9_2

Some examples of RWD include administrative databases or clinical registries with electronic healthcare records (EHR), which contain information on patient characteristics, admission details, treatment procedures and clinical outcomes [3].

The generation and collection of RWD is often pragmatic, and limited efforts are made to control the data collection scheme or information flow. The quality of RWD thus can vary dramatically across clinical domains and individual databases [4–8]. For example, health care records are often incomplete and may contain information that is inaccurate or even inconsistent with other data sources [9, 10]. It is therefore imperative that studies involving RWD investigate the nature of recorded information to improve their quality, raise awareness on their strengths and weaknesses, and take these into account to facilitate valid inference on the research question at hand.

Although there is no formal framework to assess the quality of RWD, it is common to focus on at least three domains: accuracy, timeliness and completeness [11]. Data accuracy relates to the validity of individual data entries [12]. It is typically assessed by examining distributional properties of the observed data (e.g., mean, standard deviation, range) and comparing this information with other sources (e.g., previously published population characteristics). Timeliness refers to the degree to which the available data represent reality from the required point in time. Problems can arise when recorded observations (e.g. taken after surgery) do not adequately reflect the patient's health state at the intended measurement time (e.g. before surgery). Finally, completeness represents the existence and amount of missing data.

In this chapter, we first briefly discuss important preprocessing steps in data quality assurance and quality control (QA/QC). Subsequently, we focus on the handling of missing data. As RWD is typically incomplete when missing values are not handled properly, straightforward analysis will very likely lead to misleading conclusions. As such, there is a strong justification to consider and select appropriate analytical methods for handling missing data.

1.2 Data Preparation

The analysis of RWD often necessitates multiple preprocessing steps to create a meaningful and analyzable dataset from the raw data. In general, we can distinguish between three types of preprocessing steps: data integration, data cleaning, and data transformation.

The first step is to identify and integrate relevant sources of data (e.g. hospital registries, administrative databases) such that all information of interest becomes available for the studied individuals. These data may, for instance, include information on signs and symptoms, diseases, test results, diagnoses, referrals, and mortality. Sometimes, it is also possible to retrieve information from unstructured data sources including texts, audio recordings, and/or images (Ref Chap. 8 on text mining). When multiple sources of data are available, it is possible to check for duplicate or inconsistent information across data sources, and thus the accuracy of the data can be assessed. Strategies for data integration are discussed in Ref Chap. 7 on data integration. Once all relevant data sources have been integrated, it is important to select those individuals that are eligible for the intended analysis. The selection requires the identification of the target population, and is often based on disease status or combinations of information (e.g. morbidity code with relevant prescription or results from a diagnostic test). In addition, it is helpful to define relevant time points, including the starting time (also known as index date or baseline) and endpoint (e.g., the outcome of interest) of the study. Although measurements at other time points can be discarded from the dataset, this information can sometimes be used to facilitate risk prediction or missing data imputation (Sect. 6.2). When repeated measurements are available for one or more variables, they can be formatted using two approaches [13]. One approach is to code observations made at different time points as separate columns, leading to a so-called “wide format”. This approach works well when the repeated measurements occur at regular time intervals, which is rather uncommon for RWD. A second approach is to record repeated information as separate rows, and to include a “time” variable that indicates when the measurements were taken. This approach is also known as the “long format”.

As a second step in data preprocessing, it is recommended to inspect the constructed dataset and to generate descriptive summaries such as the mean, standard deviation, range and amount of missing values for each variable [14]. This information can be used to assess completeness of the data and to identify outliers with impossible or extreme values. When invalid measurements or recordings are detected, corresponding values can be treated as missing data and subsequently be recovered using imputation methods. Alternatively, in case of extreme but valid values, the analysis may be rendered more robust to outliers by windsorizing (i.e., observations are transformed by limiting extreme values) or trimming (i.e., simply discarding extreme observations). Such methods always cause a loss of information, and their use should be guided by good reasons to reduce the influence of such observations. This will heavily depends on the analysis of interest. For instance, mean and variance measures are heavily affected by outliers, but the median is not affected at all. Unfortunately, it is often difficult to assess the validity of individual measurements. For this reason, researchers may sometimes consider analysis methods that directly account for the (potential) presence of measurement error in the entire dataset during model estimation (Ref Chap. 9 on measurement error).

Finally, in the last step, data transformations can be performed. For instance, it is sometimes helpful to transform continuous variables (e.g., in line with model assumptions or to improve numerical stability), to re-code categorical variables (e.g., dummy coding to allow unordered and non-equidistant steps between categories), or to collapse multiple variables into an aggregate measure (i.e., data reduction). Further, when the focus of a study is on the development of a prediction model, it is necessary to set up a training and validation set. Although it is common to randomly split the data into two parts, resampling methods have been recommended to make better use of the data in terms of bias and efficiency (Ref Chap. 15 on model evaluation).

2 Missing Data

Pre-processing often brings to light that records in some data fields are missing. This requires careful consideration since it may indicate loss of information and almost surely affects the analysis and the subsequent interpretation of findings. The degree to which this is the case primarily relates to the type of missing data. Therefore, first and foremost, it is important to try to understand why data are missing, as this will guide any further processing.

2.1 Types of Missing Data Mechanisms

In the broadest sense, there two large groups of missing data mechanisms.

The first group relates to situations where data cannot or should not be measured. For example, it is not possible to assess tumor characteristics or disease severity for healthy patients. Although the absence of any measurements could here be identified and treated as a missing data problem, this strategy should be avoided because it fails to address the fact that no information is actually missing.

The second group arises when variables could have been measured but were not recorded (i.e., information is actually missing). It is, for instance, possible that observations are missing because no measurements were taken or because available measurements were considered invalid or not correctly recorded. Alternatively, it is possible that data collection is complete for individual patients. However, when data are combined across patients or clinical centers, key variables may become incomplete. When trying to understand the consequences of these missing data and to guide the best way forward, it is helpful to distinguish between three mechanisms by which missing data can arise [15]: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR).

Briefly, MCAR occurs when the probability that a certain type of measurement is missing does not depend on the values of either observed or missing data. This directly implies that missingness is not related to any of the recorded data and that records with missing data do not form any special group. As an example, physical examination records can be lost due to an administrative computer error. There are no measures, either observed or unobserved, that explain missingness for these particular cases: missingness is said to be *completely* at random.[1]

In MAR, the probability that a variable is missing differs across records based on the values of observed data. For example, a particular type of diagnostic measure may be ordered more often upon certain blood sample deviations. If these data are indeed missing at random (MAR), this means that the probability that a value is missing is again completely at random *within* subgroups with the same blood sample analysis. That is, after taking observed blood sample measures into account, there is no further information that predicts missingness.

Lastly, MNAR describes the situation that the probability that a certain type of measurement is missing is associated with *un*observed data. For instance, if certain measures are more often performed in those with a high suspicion of an unfavorable outcome, but this suspicion cannot be derived from other measures that were observed and recorded in the database. An important particular case is where missingness depends on the value of the measure being missing itself. For instance, alcoholics might be less likely to respond to a questionnaire on alcohol intake.

The distinction between these types of missing data mechanisms is helpful when thinking about the inferences one can make based on the observed data only, without modelling the missing data mechanism itself. As it turns out, several methods can obtain unbiased inference when the MAR assumption holds without explicitly modelling the missing data mechanism[2] (See Sect. 4). Although MAR is often a useful and sometimes convenient assumption from a statistical point of view, the analysis of incomplete data will often need to be supplemented by sensitivity analyses that allow for a more complex missingness mechanism [16]. Methods for this purpose are discussed in more detail in Sect. 6.

2.2 Types of Missing Data Patterns

The manifestation of missing values (regardless of their cause) can be classified into different patterns, each of which requires a different analysis approach. We here focus on common patterns that arise when analyzing RWD.

Real world data are often collected over a period of time and may therefore contain multiple observations for one or more variables. When data are incomplete, it is helpful to distinguish between monotone (e.g., dropout) and non-monotone (intermittent) patterns of missingness (Fig. 1). The dropout pattern occurs when a variable is observed up to a certain

[1] The notion of 'completely at random' is intended to mean: not depending on any observed or missing values out of the measures analyzed. Therefore, is does not have to imply that the missing data pattern is totally unsystematic; it may for instance relate to a measure that is not measured and not of interest for the final analysis. Therefore, the definition of MCAR (and equivalently MAR and MNAR) depends on the set of variables of interest.

[2] In the likelihood and Bayesian paradigm, and when mild regularity conditions are satisfied, the MCAR and MAR mechanisms are ignorable, in the sense that inferences an proceed by analyzing the observed data only, without explicitly addressing the missing data mechanism. In this situation, MNAR mechanisms are nonignorable. Note that in frequentist inference the missingness is generally ignorable only under MCAR [92].

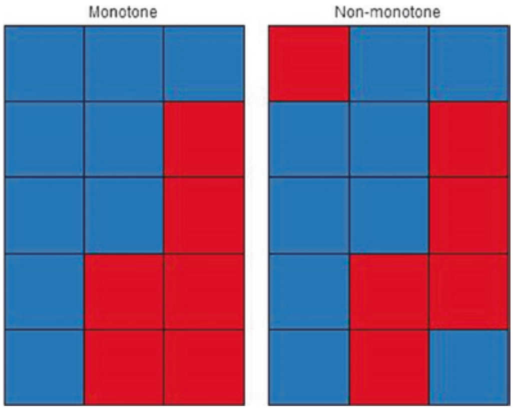


Fig. 1 Illustration of missing data patterns in multivariable data. Each row represents the measurements for a unique patient or timepoint. Columns represent individual variables. Missing values are displayed in red, observed values are displayed in blue

time-point, and missing thereafter [17]. This situation may, for instance, occur when an individual leaves the study prematurely or dies. More generally, a missing data pattern is said to be monotone if the variables can be sorted conveniently according to the percentage of missing data [18]. Univariate missing data form a special monotone pattern. The presence of monotone missingness offers important computational savings and can sometimes be addressed using likelihood-based methods (Sect. 5.2). Conversely, the intermittent pattern occurs when an observed value occurs after a missing value. Because the collection of RWD is often driven by local healthcare demands, measurements tend to be unavailable for time points that are of primary interest to researchers. Intermittent patterns of missingness are therefore relatively common for variables that were measured at multiple occasions. In Sect. 6.2, we discuss dedicated imputation methods to address these non-monotone patterns of missingness.

Real-world data originating from multiple sources (e.g., hospitals, or even countries) tend to be clustered, with distributions and effects that may differ between clusters. In this context, one can distinguish between data values that are sporadically missing (at least some values available in each cluster) and those that are systematically missing (not measured at all in a particular cluster) [18–20]. Systematically missing data are more common when combining routinely collected data from multiple different sources, such as in claims databases. Also, in a pharmacoepidemiologic multi-database studies, there is a high likelihood of missing data because the multiple databases involved may record different variables [21, 22]. Sporadically missing values often occur and are just the within cluster counterpart of usual missing data. This also leads to the main advantage of dealing with just sporadically missing data. Since at least some information on the joint distribution of the data is available in each cluster, regular missing data methods can be implemented *within* clusters if they have sufficient size. In contrast, more evolved missing data methods that accommodate the clustered nature of the data are necessary to handle systematically missing data. A detailed account of missing data methods designed for clustered data is available elsewhere [18–20].

2.3 A Bird's Eye View on Missing Data Methods

Datasets that are collected in real-world settings are typically large and complex. They are large not only in the sense of the number of individuals, but also in terms of the number of collected variables. At the same time, the structure of RWD also tends to be very complicated. It generally has mixed variable types, containing continuous, categorical and time-to-event variables, some of which could have very sophisticated relationships. It is also common that many variables have missing values and that some variables are incomplete for most individuals. Moreover, when missingness occurs, it is often difficult to determine whether the missing data mechanism is MCAR, MAR or MNAR. Instead, it is very likely that all three missing data mechanisms co-exist in the dataset. The validity of analyses involving RWD will therefore often depend highly on whether missing data were handled appropriately.

Fortunately, several strategies exist to address the presence of missing data. In this chapter, we focus on imputation methods which can address many of the aforementioned challenges. These methods replace the missing values by one (single imputation, see Sect. 3) or more (multiple imputation, see Sect. 4) plausible values. Imputation avoids the need to discard patient records and separates the missing data problem from the substantive analysis problem (e.g., estimation of a causal effect or predictive model). This implies that imputed data can be analysed using standard methods and software, and as such be directly available for inference (e.g., parameter estimation or hypothesis testing) and the generation of risk predictions. However, as we discuss later in this chapter, single imputation methods are best avoided in most settings because they are not capable of preserving uncertainty about the missing values and their imputation [23]. We therefore recommend more advanced approaches that are based on multiple imputation (Sect. 4) or avoid imputation altogether (Sect. 5). These methods can mainly be applied when data are MCAR or MAR. When the missingness mechanism is MNAR or unknown, additional methods need to be employed (Sect. 6.1).

Traditional methods for multiple imputation have been studied extensively in the literature, and are briefly summarized in Sects. 4.1 and 4.2. More recently, numerous imputation methods have also been proposed in the field of machine learning [24]. Although these methods tend to be relatively data hungry, they offer increased flexibility and may therefore improve the quality of subsequent analyses (Sect. 4.3). To evaluate the potential merit of advanced imputation methods, we embarked on a literature review and focused on imputation methods that are well-suited to handle mixed data types, a large number of both cases and variables, and different types of missing data mechanisms. Briefly, we searched relevant publications on PubMed and ArXiv that describe quantitative evaluations of missing data methods. Initially, we identified 15 relevant papers based on our own experience in the field. These papers compared several statistical and machine learning imputation techniques and were used to inform an active learning literature review. To this purpose, we used the software ASReview, a machine-learning framework that facilitates the screening of titles and abstracts [25, 26]. To achieve full merit of the framework, a 'stopping criterion' is required–in our case when the software had selected all 15 priory identified publications. A flow diagram of the review methods is presented in Fig. 2. We made use of the following eligibility criteria:

- Inclusion criteria: the paper concerns an evaluation of missing data methods through simulation; the paper matches the search query "(simulation[Title/Abstract]) AND ((missing[Title/Abstract]) OR (incomplete[Title/Abstract]))"; the paper is selected by ASReview before the stopping criterion is reached.
- Exclusion criteria during abstract screening: the paper does not concern an evaluation of missing data methods through simulation; the paper concerns a datatype that deviates from typical EHR data (e.g., imaging data, free text data, traffic sensor data); the paper only concerns (variations of) the *analysis* model, not the *imputation* model; the paper only concerns (variations of) one missing data method.
- Exclusion criteria during full text screening (all of the above, plus): the paper only concerns two missing data methods, one of which is complete case analysis; the paper only concerns single-patient data; the paper only concerns a MCAR missingness mechanism (equivalently, the paper does not concern MAR, MNAR or empirical missingness mechanisms).

After omitting duplicates and removing papers that did not meet the eligibility criteria, we obtained 67 publications. These are listed on zotero.org/groups/4418459/clinical-applications-of-ai/library.

Based on the aforementioned considerations, we decided to focus on five types of machine learning methods that can be used for imputation: nearest neighbour methods, matrix

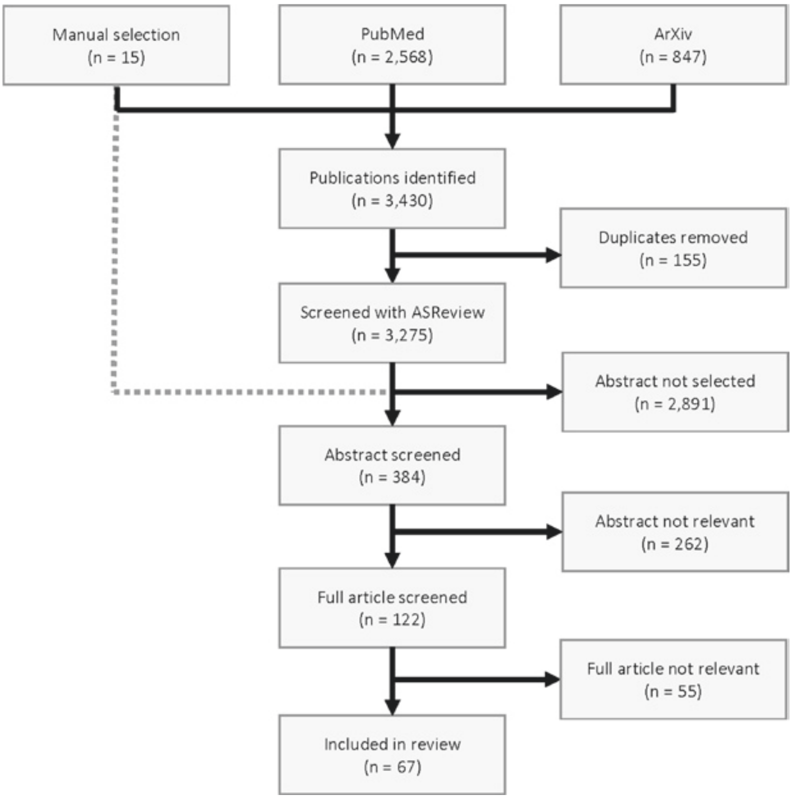


Fig. 2 Flow chart of the literature review to identify quantitative evaluations of missing data methods

completion, support vector machines, tree-based ensembles, and neural networks. In the following sections, we briefly introduce each method, discuss its strengths and weaknesses, and provide software implementations. We summarize the main findings from our review in Sect. 6.3, offering also a list of recommendations.

2.4 Introduction of Case Study Data (MIMIC-III)

The Medical Information Mart for Intensive Care (MIMIC)-III database contains information on 38,597 adults and 7870 neonates that were admitted to critical care units at Beth Israel Deaconess Medical Center [27, 28]. Various types of patient-level data are available, including vital signs, laboratory measurements, imaging reports, received treatments, hospital length of stay, and survival. Although many variables were only measured upon admission, temporal data are also available. For instance, there are 753 types of laboratory measurements in MIMIC-III, each with on average 8.13 observations per patient. As illustrated in Fig. 3, the missingness rate in MIMIC-III greatly varies between variables and can be as high as 96%.

3 Single Imputation Methods

A common approach to address the presence of missing values is to simply replace them by a plausible value or prediction [30]. This approach is adopted by many software packages that implement contemporary machine learning methods. Below, we outline and illustrate three single imputation methods to recover missing systolic blood pressure levels in MIMIC-III.

In single value imputation (SVI), it is widespread to replace missing values of a variable

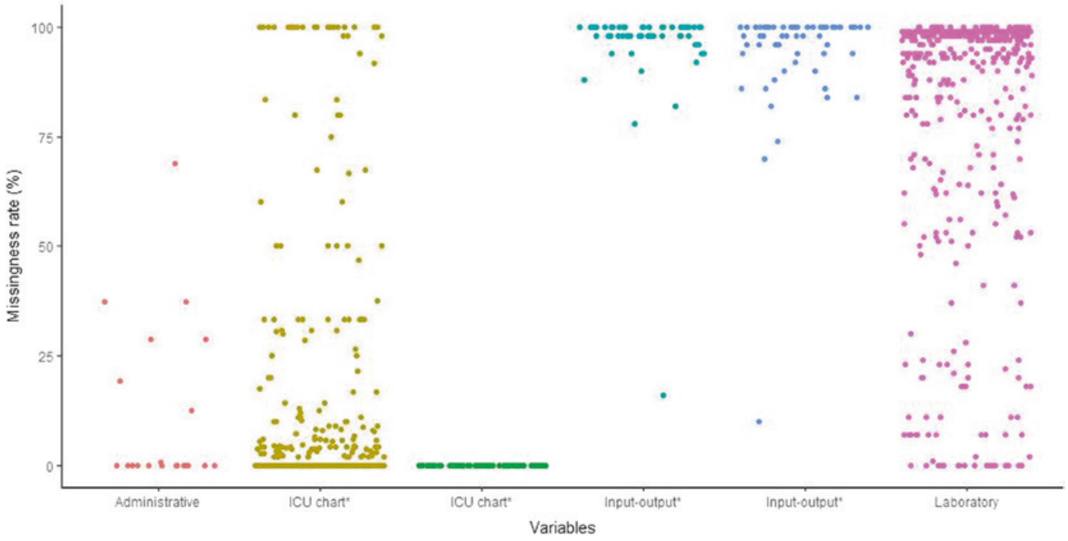


Fig. 3 Visualization of missing data in MIMIC-III [29]. Missingness rate is calculated as the proportion of individuals that do not have any observation for a given variable. Administrative variables include demographic data and were not much affected by missing values (e.g., missingness rate for date of birth = 0%). Intensive Care Unit (ICU) chart variables include patient monitoring variables. Input–output variables relate to intake substances (e.g., liquids, medication) and excretions (e.g., urine, fluid from the lungs).Finally, laboratory variables include microbiology results. * Two different critical care information systems were in place over the data collection period. For this reason, missingness rates for ICU chart and input–output variables are presented as separate categories

by a convenient summary statistic, such as the mean, median, or mode of the corresponding variable. For example, patients without follow-up data are sometimes assumed to be alive. Similarly, when blood oxygenation levels are incomplete, it is possible to assume that corresponding patients are in perfect health and simply impute a constant that reflects this condition (e.g., 100%). Alternatively, when the health conditions of included patients are suboptimal, it is possible to impute the average of the observed blood oxygenation levels (left graph in Fig. 4).

A more advanced approach to generate imputations is to adopt multivariable (e.g., regression or machine learning) models that replace each missing value by a prediction [18]. For instance, it is possible to predict blood oxygenation levels in the MIMIC-III database using information on patient age by adopting a regression model (middle graph in Fig. 4). As more (auxiliary) variables are used to predict the missing values, the accuracy of imputed values tends to increase [31].

Unfortunately, single imputation methods tend to distort the data distribution because they do not account for sampling variability and model uncertainty [17, 18]. Because this usually leads to biased inference, single imputation methods are best avoided [30]. Their implementation can, however, be acceptable in some circumstances [18]. For example, it is possible to add noise to imputed values in order to account for sampling variability (right graph in Fig. 4). Also, when applying a prediction model in clinical practice, single imputation methods can greatly facilitate real-time handling of missing values on a case-by-case basis [31, 32].

4 Multiple Imputation Methods

In general, the preferred approach to address the presence of missing data is to adopt multiple imputation [18, 30]. In this approach, each missing value in the original dataset is replaced by a set of $m > 1$ simulated values, leading to

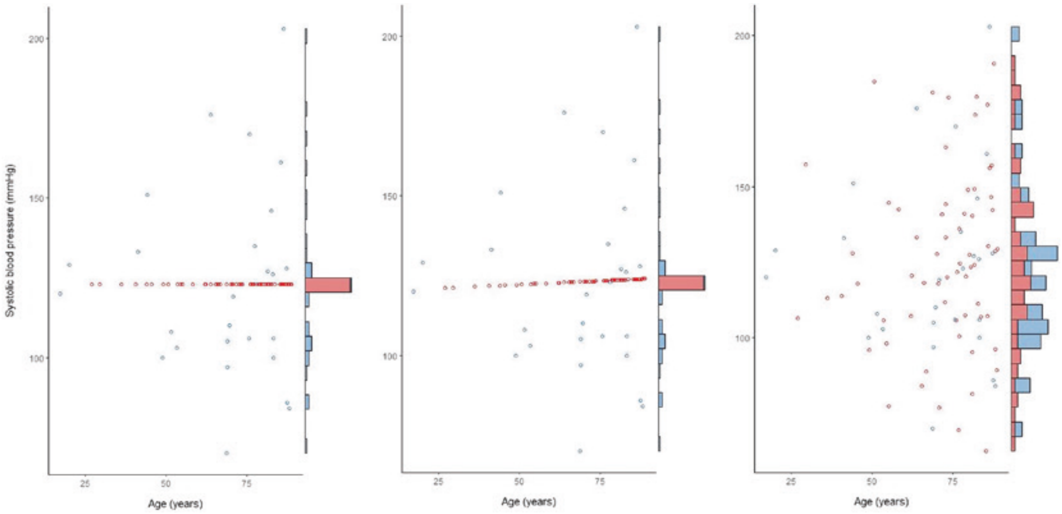


Fig. 4 Illustration of imputation strategies using 100 patients from MIMIC-III. The observed data are displayed in blue and represent the first available measurement for systolic blood pressure after hospital admission. Imputed data are displayed in red, and were generated using mean imputation (left), regression imputation (middle), stochastic regression imputation (right)

multiple completed datasets. The entire procedure is illustrated in Fig. 5.

The generation of plausible values typically involves modelling the observed data distribution and imposing corresponding parameters on the missing data. A major advantage of multiple imputation is that the extent to which the missing values can accurately be recovered becomes more transparent. The variability of imputed values will be large for variables that cannot adequately be retrieved from the observed data (and vice versa). For example, when temperature measurements are missing for a patient diagnosed with COVID-19 and having symptoms that often coexist with fever, imputed values will have a high probability to indicate the presence of fever. In contrast, fever imputations for a patient with a positive COVID-19 test and only mild disease can be expected to be more variable.

A key challenge in multiple imputation is to generate random samples that are plausible and

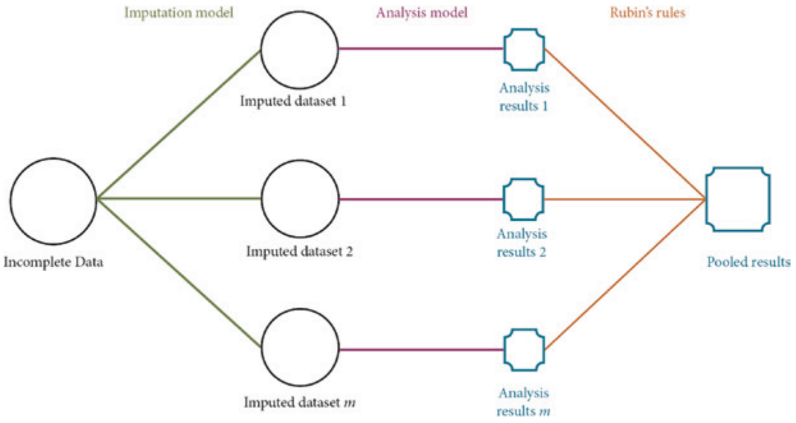
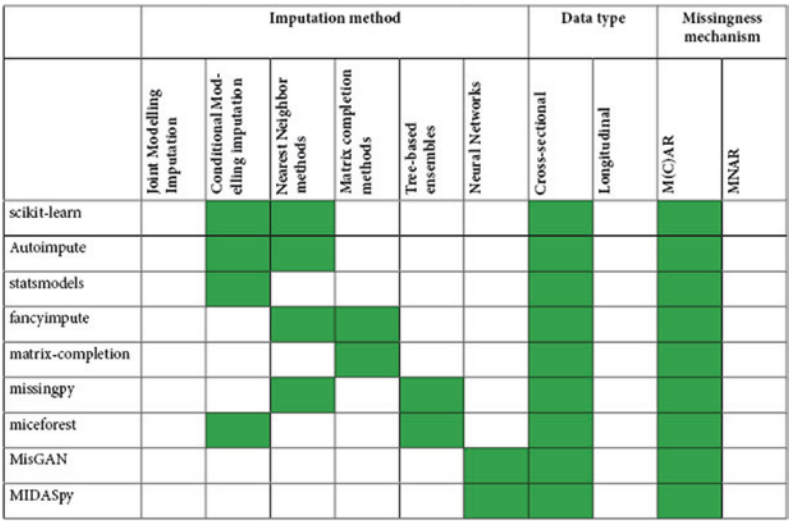


Fig. 5 Scheme of main steps in multiple imputation, adapted from [18]



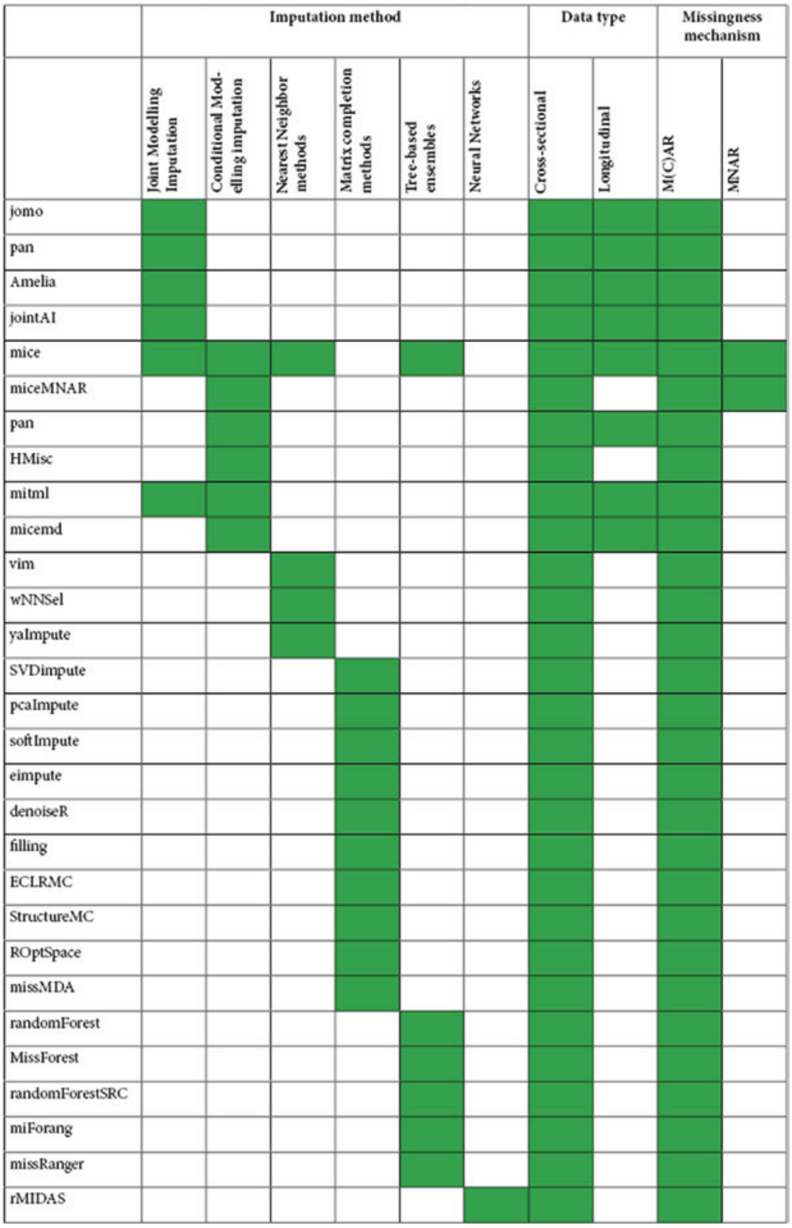
	Imputation method						Data type		Missingness mechanism	
	Joint Modelling Imputation	Conditional Modelling imputation	Nearest Neighbor methods	Matrix completion methods	Tree-based ensembles	Neural Networks	Cross-sectional	Longitudinal	M(C)AR	MNAR
scikit-learn		🟩	🟩				🟩		🟩	
Autoimpute		🟩	🟩				🟩		🟩	
statsmodels		🟩					🟩		🟩	
fancyimpute			🟩	🟩			🟩		🟩	
matrix-completion				🟩			🟩		🟩	
missingpy			🟩		🟩		🟩		🟩	
miceforest		🟩			🟩		🟩		🟩	
MisGAN						🟩	🟩		🟩	
MIDASpy						🟩	🟩		🟩	

Fig. 6 Python modules for multiple imputation. If an analyst decides to use SVM for imputation, they may need to manually incorporate the algorithm into the imputation procedure. In Python, SVM can be implemented using the scikit-learn library, or using GitHub repositories such as SVMAlgorithm and SupportVectorMachine

exhibit an appropriate amount of variability. Conceptually, this can be achieved by generating imputations from a probability distribution. For instance, consider that some patients in MIMIC-III have missing values for age. A simple solution is to approximate the empirical (observed) age distribution, which has a mean value of 65.8 years and a standard deviation of 18.5 years, with a suitable well-known distribution. New values for patient age could then be generated from a normal distribution with the aforementioned characteristics. It may be clear that the aforementioned (univariate) approach does not account for any relation with other variables in the dataset, and thus leads to imputations that are not very plausible. A better approach is to consider the entire (multivariate) distribution of the available data and draw imputations tailored to each patient [33]. Here, we discuss two broad strategies to generate personalized imputations: joint modelling imputation and conditional modelling imputation. For the latter, both statistical and machine learning methods can be used. Software implementations are summarized in Fig. 6 (Python) and Fig. 7 (R).

4.1 Joint Modelling Imputation

A direct approach to consider the entire data distribution is to explicitly specify a parametric joint model for the observed data [34]. The parameters of this (imputation) model are estimated from the observed data, and subsequently used to generate imputed values. It is, for instance, common to assume that the observed patient characteristics arise from a multivariate normal model. The mean and covariance can be estimated using Markov Chain Monte Carlo (MCMC) methods and directly be used to draw imputed values that account for individual patient characteristics [32]. This approach is also known as multivariate normal imputation [35]. Recent work shows that multiple imputation based on more flexible joint models of the data (e.g. allowing for variables of different types, hierarchical structure of the data, or interaction effects) can also be achieved within the Bayesian framework [36, 37]. Often, it is difficult to identify an appropriate joint model that describes the observed data. Many datasets contain a combination of binary, continuous,



	Imputation method						Data type		Missingness mechanism	
	Joint Modelling Imputation	Conditional Modelling imputation	Nearest Neighbor methods	Matrix completion methods	Tree-based ensembles	Neural Networks	Cross-sectional	Longitudinal	M(C)AR	MNAR
jomo	🟩						🟩	🟩	🟩	
pan	🟩						🟩	🟩	🟩	
Amelia	🟩						🟩	🟩	🟩	
jointAI	🟩						🟩	🟩	🟩	
mice	🟩	🟩	🟩		🟩		🟩	🟩	🟩	🟩
miceMNAR		🟩					🟩		🟩	🟩
pan		🟩					🟩	🟩	🟩	
HMisc		🟩					🟩		🟩	
mitml	🟩	🟩					🟩	🟩	🟩	
micemd		🟩					🟩	🟩	🟩	
vim			🟩				🟩		🟩	
wNNSel			🟩				🟩		🟩	
yaImpute			🟩				🟩		🟩	
SVDimpute				🟩			🟩		🟩	
pcaImpute				🟩			🟩		🟩	
softImpute				🟩			🟩		🟩	
eimpute				🟩			🟩		🟩	
denoiseR				🟩			🟩		🟩	
filling				🟩			🟩		🟩	
ECLRMC				🟩			🟩		🟩	
StructureMC				🟩			🟩		🟩	
ROptSpace				🟩			🟩		🟩	
missMDA				🟩			🟩		🟩	
randomForest					🟩		🟩		🟩	
MissForest					🟩		🟩		🟩	
randomForestSRC					🟩		🟩		🟩	
miForang					🟩		🟩		🟩	
missRanger					🟩		🟩		🟩	
rMIDAS						🟩	🟩		🟩	

Fig. 7 Software packages in R for multiple imputation. More detailed information for R packages is available from https://cran.r-project.org/web/views/MissingData.html

categorical, and other data types. These mixed data types usually cannot be described using a multivariate distribution with a well-known density. A common strategy to relax this limitation is to approximate the (multivariate) data distribution by a series of conditional (univariate) distributions which is the focus of the next section.

4.2 Conditional Modelling Imputation

Conditional modelling imputation implies that a separate imputation model is estimated for each incomplete variable [38]. For instance, a logistic regression model can be used to describe the conditional distribution of a binary variable (e.g., current smoker). Conversely, a linear regression model can be used to describe the conditional distribution of a continuous variable (e.g., systolic blood pressure). As discussed in Sect. 4.3, it is also possible to adopt machine learning models to describe these conditional distributions. Imputed values are then generated by sampling successively from each of the conditional models, which requires an iterative Monte Carlo procedure. This approach is also known as conditional modelling imputation [32], chained equations imputation [39], or fully conditional specification.

4.3 Machine Learning Imputation

Multiple imputation methods often require explicit assumptions about the distribution(s) of the data, including consideration of the potential presence of interactive and non-linear effects. If the imputation model(s) are based on invalid distributional assumptions or fail to incorporate important covariate effects, subsequent analyses can lead to substantial bias [40]. For instance, consider that an interaction exists between the age of a patient and their blood test results (which contains missing values). If this interaction is not explicitly accommodated during imputation, its magnitude will be attenuated in the imputed data. Thus, constructing an appropriate imputation model requires consideration of how the imputed data will eventually be used [23]. Unfortunately, it is often difficult to predetermine how data will be analyzed, especially when the available data sources were not designed for the intended analysis. It is therefore helpful for imputation models to anticipate certain features of the data (such as interactions, nonlinearities, and complex distributions) without making any specific commitments. Such flexibility can be realized by non-parametric (e.g., nearest neighbor) or semi-parametric models (e.g., neural networks, random forests, or support vector machines) that avoid making distributional assumptions about the observed data. Below, we discuss a selection of common approaches that yield multiple imputed datasets. In general, machine learning methods can be used in two different contexts. One approach is to embed machine learning models in conditional modelling imputation to describe the conditional distribution of a certain variable. For example, missing blood pressure levels could be imputed using a random forest. A second approach is to generate imputed values directly using a dedicated machine learning method, such as matrix completion or adversarial networks.

4.3.1 Nearest Neighbor Methods

Nearest neighbor (NN) methods offer a non-parametric approach to generate imputations without making distributional assumptions. To this purpose, a distance metric is used to determine the relatedness between any two individuals and to identify *neighbors* with complete information for each individual with one or more missing values. Imputation is then achieved by simply copying the observed values from the nearest neighbor (1-NN) or by combining the observed values from *k* nearest neighbors (kNN) into a weighted average [41]. Since NN methods generate imputations by (re)sampling from observed data, no special efforts are required to address complex data types. Accordingly, they are often used with incomplete variables that are restricted to a certain range (e.g., due to truncation), skewed, or semi-continuous. To allow for multiple imputed values, NN methods typically determine the distance between two individuals using a random subset of variables, rather than all observed variables [42]. Although NN methods can directly be used as a non-parametric imputation approach, they can also serve as an intermediate step in semi-parametric imputation procedures [43]. For instance, predictive mean matching combines conditional modelling imputation with NN methods to draw imputations from the observed data [44]. It has been demonstrated that NN methods perform well when data are MCAR or MAR [45–47]. Although NN methods are simple and easy to implement [48], they strongly depend on the specification of a suitable multivariate distance measure and a reasonably small dimension (since there are fewer near neighbors in high dimensional space). Consequently, the performance of NN methods tends to suffer from high dimensionality problems [49] and declines when k is too small or too large. Finally, NN methods do not facilitate the incorporation of MNAR mechanisms, and therefore appear less suitable in RWD.

4.3.2 Matrix Completion Methods

Matrix completion methods aim to recover an intact matrix from the dataset with incomplete observations. To this purpose, they decompose the original (high-dimensional) matrix into a product of lower dimensional matrices [50]. Missing data are then imputed by identifying an appropriate low-rank approximation to the original data matrix.

For instance, singular value decomposition (SVD) can be used to describe a dataset X with n rows (e.g., patients) and k columns (e.g., variables) by a matrix product $X = UDV\prime$. In this expression, D is a diagonal matrix with k singular values, U is an $n \times k$ matrix of left singular vectors, and V is an $k \times k$ matrix of right singular vectors. The entries of D are used to scale U and V, and therefore describe how much information each singular vector provides to the original data matrix. Recall that the rank of a matrix is the maximal number of linearly independent column vectors or row vectors in the matrix, which is also equal to the number of non-zero singular values of the matrix. By omitting singular values that are close to 0 from D (and omitting the corresponding vectors from U and V), the rank of a matrix can be reduced without much loss of information. This, in turn, gives a lower-rank approximation to the original matrix. In case of missing data, the key idea is to find a low-rank approximation that closely fits the observed entries in X from a lower-rank approximation, with the rank sufficiently reduced to fill in the missing parts of X.

Other methods that apply matrix completion include (robust) principle component analysis (PCA) and nuclear-norm regularization [50, 51]. In the latter, the singular values are summarized into a nuclear norm that is optimized using expectation maximization. Matrix completion methods do not make any assumptions about the distribution of the observed data, and can handle high-dimensional data in a straightforward manner. Although their implementation is mainly justified when data are MCAR or MAR, several extensions exist for MNAR situations [52, 53]. Unfortunately, matrix completion is primarily used for numerical data. For categorical data, mode imputation is generally used. Another limitation is related to the implicit linearity assumption. As rank is a concept for the linear relationship between rows or columns of a matrix, the method does not preserve nonlinear relationships between rows or columns.

4.3.3 Tree-Based Ensembles

Tree-based ensemble methods estimate multiple decision trees on the available data and adopt boosting (e.g., XGBoost) or bagging (e.g., random forests) to combine their predictions. Tree-based ensembles can be applied to mixed data types, do not require distributional assumptions, and naturally allow for variable selection. Moreover, their recursive partitioning operation predisposes to capture nonlinear effects and interactions between variables. Several simulation studies have shown that tree-based ensemble methods can outperform commonly used multiple imputation methods [54–56]. We here focus on the use of random forests to generate imputed values, for which at least four different implementations are available [57]. In Sect. 5.3, we discuss additional approaches for developing random forests without the need for imputation.

The first tree-based approach to handle missing data was proposed by Breiman and is implemented in the R package *randomForest* with the function “rfImpute” [58]. It relies on the concept of “proximity” for missing data imputation. Missing values are initially replaced by a simple summary such as their mean or mode, then a forest is constructed and the proximity matrix is calculated. The proximity matrix is a square matrix where each row and column represents a specific individual. Each matrix entry then quantifies the probability that the individuals from the corresponding row and column fall in the same leaf node. The missing value of a particular variable for a specific individual is imputed using an average over the non-missing values of the variable or the most frequent non-missing value where the average or frequency is weighted by the proximities between the case and the non-missing value cases. The process is repeated for each imputed dataset [58].

A second approach termed "on-the-fly-imputation method" was proposed by Ishwara et al. and is implemented in the R package *randomForestSRC* [59]. In this method, only observed values are used to calculate the split-statistic when growing a tree. At each node of a tree, when a split decision needs to be made, missing values will be replaced by random observed values within the corresponding subtree. After each node split, imputed values are set back to missing and the process continues until no more splits can be made. Missing data in terminal nodes are then imputed using the mean or mode of out-of-bag non-missing terminal node data from all the trees.

A third approach was proposed by Stekhoven and Buehlmann and is implemented in the R packages *MissForest and missRanger* [54]. In this method, missing values are initially imputed using simple methods such as mean or mode. The completed data is then used to construct a forest, which in turn is used to predict the missing values. In contrast to the approach proposed by Breiman, this process of training and predicting iterates until a stopping criterion is met, or until a maximum number of user-specified iterations is reached.

Finally, a fourth approach is to use random forests to approximate the conditional (univariate) distribution of the observed data [60]. The chained equations framework is then used to iteratively replace the missing values for each incomplete variable (Sect 4.2). Conditional modeling imputation using random forests has, for instance, been implemented by the function *mice.impute.rf* in the R package *mice* and tends to yield better performance than the three approaches mentioned above [55, 61]. A major advantage of this approach is that imputed data can be analyzed using any method of choice.

4.3.4 Support Vector Machines

Support Vector Machines (SVM) were developed more than thirty years ago [62, 63] and have been successfully used in many real-world applications focusing on classification or prediction. A key building block and also the driving force behind SVM's success is the employment of a kernel function. The kernel function implicitly defines a high-dimensional, or even infinite dimensional feature space (hyperplane), in which data points from different classes could be linearly separated or a continuous response variable could be linearly related to the feature vector. The kernel function needs to be carefully selected, and often takes the form of a Gaussian or polynomial (e.g., when the model should allow for non-linear relations). The most typical scenario for the application of SVM is when all predictors are continuous and when the outcome is binary or continuous. When a predictor variable is categorical, dummy coding needs to be applied. Extensions of SVM are available that can handle categorical or survival outcome data. After the completion of the training process, an SVM generally depends only on a small subset of the original data points, called "support vectors". Although SVM are very powerful in handling high-dimensional data, they are not commonly used for missing data imputation. Possibly, this is because SVM algorithms are very sensitive to noise and less suitable when the sample size is large. For the application of SVM for missing data imputation, no formal statistical software packages were found.

4.3.5 Neural Networks

Neural networks are emerging methods in the field of machine learning and are commonly applied for data generation, feature extraction and dimension reduction. We here discuss two main categories of neural networks that can be used for missing data imputation: autoencoders (AEs) and generative adversarial nets (GANs).

An AE is an artificial neural network specifically designed to learn a representation of the observed data. It typically contains an encoder and a decoder. The encoder maps the original input data to a lower-dimensional representation through successive hidden layers of a neural network. The final layer of an encoder is the output layer, which simply describes the original input layer in a lower dimension [64, 65]. The decoder then maps the output from the encoder to reconstruct the original input, again through successive hidden layers of a neural network. Unfortunately, standard implementations of AEs require data to be complete, and they may end up learning an identity map (hence perfectly reconstructing the input data when an identity map is used instead of successfully reducing the complexity). To address these problems, several AE variants have been proposed. One approach is to adopt denoising autoencoders (DAE) that corrupt the input data with noise [66]. The most common way of adding noise is to randomly set some of the observed input values to zero. This approach can also be applied to incomplete input data, by simply replacing missing values by zero. To facilitate multiple imputation, missing values can be replaced by random samples [67]. Further, it is also possible to treat missing values as an additional type of corrupted data, and to draw imputations from an AE trained to minimize the reconstruction error on the originally observed data. This approach has, for instance, been implemented by Multiple Imputation with Denoising Autoencoders (MIDAS) [68]. A second extension of AE is to adopt variational autoencoders (VAEs) that learn to encode the input using a latent vector from a probabilistic distribution [69–71]. The original data can then be imputed by sampling from the latent posterior distribution.

GANs are another type of neural network that consists of two parts; a generator and a discriminator [72]. In an adversarial process, the generator learns to generate samples that resemble the original data distribution, and the discriminator learns to distinguish whether a presented example is original or artificial. The GAN procedure can be extended to allow for the imputation of missing data [73–75]. Generative Adversarial Imputation Nets (GAIN) adapt the original GAN architecture as follows [75]. The generator learns to model the distribution of the data and to impute missing values accurately. The discriminator then learns to distinguish which values were observed or imputed. The generator's input combines the original input data and a mask matrix that indicates the presence of missing values. Conversely, the input of the discriminator is given by the output of the generator and a hint matrix, which reveals partial information about the missingness of the original data. The discriminator then learns to reconstruct the mask matrix.

4.4 Analyzing and Combining the Imputed Datasets

Once multiple imputed datasets have been generated, they can be analyzed separately using the procedure that would have been followed if all data were complete (Fig. 5). For example, studies aiming to evaluate a relative treatment effect can perform a regression analysis in the imputed data to estimate an odds ratio adjusted for confounders. From each analysis, one or more parameter estimates (and corresponding estimates of uncertainty) are then obtained and need to be combined. Pooling results across multiple imputed datasets is not trivial and typically requires to consider three sources of uncertainty. In particular, there is estimation error within each imputed dataset (e.g., reflected by the estimated standard errors in each completed dataset), variation due to missing data (reflected by the between-imputation variance of parameter estimates), and uncertainty arising from a finite number of imputations. Although point estimates (e.g., regression coefficients) can simply be averaged across the imputed datasets, the pooling of standard errors requires adopting a series of equations that account for aforementioned sources of uncertainty. These equations are also known as Rubin's rules [33, 76, 77] and have been implemented in most contemporary software packages.

If pooling is done appropriately, multiple imputation methods yield valid parameter estimates with appropriate confidence intervals. In some situations, however, the implementation of Rubin's rules cannot be justified. For example, an exception arises when data are available for the entire population [78]. The application of Rubin's rules also becomes more complicated when imputed datasets are analyzed using non-parametric approaches (e.g., recursive partitioning) or approaches that do not result in the same number of parameters across imputations (e.g., variable selection algorithms) [79–81]. In such situations, it may be helpful to avoid imputation altogether.

5 Non-imputation Methods

5.1 Complete Case Analysis

A simple approach to address missing data is to simply remove incomplete records from the dataset. This approach, also known as complete case analysis (CCA), is generally valid but needlessly inefficient under the usually unrealistic MCAR assumption. The adoption of CCA is therefore more appealing when conducting likelihood-based inference under MAR conditions or in datasets where only the outcome is missing (Sect. 5.2). Unfortunately, CCA does not offer a solution when estimated models (e.g., for risk prediction or classification) are applied to new patients with incomplete data.

5.2 Likelihood-Based Methods

More advanced approaches to address missing values define a model for the observed data only. For example, survival models can be used to analyze binary outcome variables that are affected by censoring (e.g., due to dropout). Similarly, multilevel models can be used to analyze repeated outcomes that were measured at arbitrary follow-up times. A special situation arises when missing values only occur for the outcome, as multiple imputation then requires auxiliary variables that are not part of the analysis model to offer an advantage over likelihood-based methods. The adoption of likelihood-based methods is therefore particularly appealing when missingness only depends on covariates that are included in the analysis model (such that missingness is ignorable) [82].

Likelihood-based methods can also be used to address missing covariate values, and often require advanced procedures for parameter estimation [83, 84]. Although likelihood-based methods tend to be much faster and produce more accurate results than multiple imputation, their applicability is limited to very specific analytical scenarios. Likelihood-based methods may therefore have limited usefulness in RWD, where patterns of missingness can be very complex and additional adjustments may be required to account for other sources of bias (e.g., time-varying confounding).

5.3 Pattern Submodels

A straightforward alternative to imputation methods is to develop separate models for each missingness pattern. For instance, those individuals for which c-reactive protein (CRP) has been observed contribute to a different model than those individuals for which CRP was not observed. This idea has also been referred to as a pattern submodel approach [85]. This type of approach is particularly helpful when the number of missingness patterns is fairly limited with respect to the number of observations, since model development occurs in partitions of the original data. Nonetheless, this is a setting that can be expected to occur quite often RWD. For instance, a whole array of venous blood results, genetics, or imaging data will often be entirely missing or entirely observed. Key benefits of patterns submodels include ease of use (both during development and application) and the fact that it does not rely on assumptions about the missingness pattern. Clear costs include loss of information due to partitioning of the data into missingness patterns (this can be relaxed to allow borrowing of information between patterns, but this invokes the MAR assumption across the patterns for which it is relaxed), and the fact that many models are developed instead of just one. As already noted by Mercaldo and Blume [85], different methods can be envisioned to allow borrowing of information between missingness patterns while retaining some of the robustness with respect to missing data mechanisms, but this is still ongoing research.

5.4 Surrogate Splits

Surrogate splits is a missing data method that is specific to tree-based methods and was proposed in the context of classification and regression trees [86]. The key idea is to not only find the optimal split point when building a tree, but also find second best (or more) split points on variables other than the one providing the optimal split point. This allows using an alternative (surrogate) split variable when the optimal variable is missing. Similar ideas have been proposed throughout tree-based methods research. For instance, instead of finding surrogate splits, the popular XGBoost method [87] finds a default direction for each split point in case the variable to split on is missing. While these methods are easy to apply on any data set with missing values, they have important limitations. For instance, surrogate splits are not able to use information from observed data to infer something about the missing variable. Instead, imputed values are generated conditionally on their position in the tree, which roughly correspond to conditional mean imputation. A more robust approach would be to apply the tree-based methods in multiple imputed data based on flexible methods that preserve more of the data complexities, and subsequently bag the results.

5.5 Missing Indicator

The indicator method replaces missing values by a fixed value (zero or the mean value for the variable) and the indicators are used as dummy variables in analytical models to indicate that a value was missing. The procedure is applied to each incomplete variable, and can be implemented in any analysis method (e.g., regression, decision trees, neural network). The indicator method allows for systematic differences between the observed and the unobserved data by including the response indicator, and thus to address MNAR. However, its implementation usually leads to biased model parameters and can create peculiar feedback mechanisms between the user of the model (e.g. a clinician) and the model itself [88]. For this reason, it is generally discouraged to adopt the missing indicator method for addressing missing data.[3]

6 Imputation of Real-World Data

Although the principles and methods outlined in Sect. 4 are primarily designed for imputing missing data in medical studies a clear sampling or data collection design (e.g., an observational cohort study or clinical trial), they can also be applied to incomplete sources of RWD that were not generated under a specific research design. In this section, we discuss two common characteristics of RWD that require more advanced imputation methods and software packages that were discussed in Sect. 4. A first challenge is the presence of informative missingness and typically arises when missing data mechanisms are complex and partially unknown. A second challenge is the presence of repeated observations, which occurs when patients are followed for a period of time. Below, we discuss methods that are well suited to address these challenges.

6.1 Informative Missingness

It is often difficult to determine the exact mechanisms by which missing values occur in RWD.

[3] While the details are beyond the scope of this chapter, Mercaldo and Blume [85] describe the implementation of missing indicator methodology in the context of multiple imputation, which does provide unbiased inference and has an interesting relation to the pattern submodels described above.

In fact, the distinction between MCAR, MAR and MNAR is a theoretical exercise and all these missingness mechanisms could co-exist in RWD. It is not uncommon that important causes of missingness are not recorded, and missingness in routine healthcare data is often informative [9, 21]. Unfortunately, traditional imputation methods are not well equipped to address this situation, as they do not distinguish between the observed and missing data distribution.

For example, the CRP test is often ordered when there is suspicion of an infection or an inflammation. Lab results may therefore be missing when elevated levels are deemed unlikely. Although multiple imputation could be used to recover these missing test results from information recorded in the EHR database, this approach is problematic when data on signs and symptoms are unavailable. Similar problems arise when test results are directly linked to their missingness. For instance, it is possible that some patients were referred from another hospital based on their lab results, and therefore did not undergo further testing. In general, when missing data mechanisms depend on unobserved information, the presence of missing values becomes informative about the patient, their physician or even the health care center [89, 90].

The plausibility of the MAR assumption (and thus the validity of "traditional" imputation methods) can often be increased by implementing imputation models with auxiliary variables that explain the reasons of missingness during imputation [91]. As more patient characteristics are recorded, it becomes less likely that the presence of missing values depends on unobserved information. For instance, when hospital registries only record information on patient age, sex, and blood test results, CRP levels are highly likely to be MNAR when unavailable. Conversely, when information on signs, symptoms, diagnostic suspicions, and other laboratory markers are also recorded, it becomes more likely that these observations explain why CRP is missing. At the very least, it will decrease the influence of MNAR mechanisms.

Unfortunately, the use of auxiliary variables becomes problematic when they are substantially affected by missing values or when they do not strongly predict the presence of missingness. Unfortunately, EHR databases are notoriously prone to prominent levels of missingness, often caused by complex recording processes. For this reason, the imputation of RWD may benefit from more advanced imputation methods that explicitly account for different missing data mechanisms [92]. When data are MNAR, it is necessary to model the joint distribution of the data and the missingness through selection, pattern-mixture or shared parameter models [93, 94]. Selection models factorize the joint distribution into the marginal distribution of the complete data and the distribution of the missingness. As an example, we discuss the Heckman selection model in more detail below [95, 96]. Conversely, pattern-mixture models separate the marginal distribution for the missingness mechanism and the data distribution conditional on the type of missingness. Essentially, this requires to estimate separate (pattern sub) models for each missingness pattern and to combine their inferences by means of integration. Finally, shared parameter models assume that the data distribution and the missingness indicator are conditionally independent after conditioning on a set of shared parameters or latent variables. This type of model has been successfully applied in settings where the missingness mechanism is related to an underlying process that changes over time. These so-called joint models[4] combine information from a mixed model for a longitudinal outcome and a temporal event model for censoring events with a set of latent variables or random effects.

A common strategy for informative missingness is to directly model the relationship between the risk of a variable being missing and its unseen value [96–98]. This strategy is based on the Heckman selection model [95], and can

[4] In this context, 'joint' is used to describe models that share a parameter, and is not to be confused with joint models that fully describe a multivariate distribution.

be used to assess and correct potential non-random missingness of outcome data. Briefly, the selection model approach involves two equations to predict the missing value and their availability. Both equations are linked together through their residual error terms, which are modelled using a bivariate (e.g., normal) distribution. The correlation of this distribution is estimated from the available data and indicates to what extent the magnitude of the missing values affects their probability of missingness (i.e., presence of MNAR). A special situation arises when there is no correlation between the error terms, as the Heckman model then generates imputations under the MAR assumption. An important requirement for the implementation of Heckman-type imputation models is the availability of exclusion restriction variables. These variables are related to the probability of missingness, but not to the missing value itself. For example, if younger physicians are more motivated to routinely record data into EHR systems, the age of the treating healthcare professional could be treated as an exclusion restriction variable. Similarly, it is possible that CRP tests are ordered more frequently for patients with a certain healthcare insurance program or socioeconomic background. As discussed, information on missingness mechanisms could also be addressed using traditional imputation methods that adopt auxiliary variables, especially if their inclusion converts MNAR situations into MAR. Indeed, it has been demonstrated that Heckman-selection models perform comparably to traditional imputation methods when missing values do not depend on unobserved information [99]. However, Heckman-selection models do not require the MAR assumption and therefore appear more suitable when the missing data mechanisms are unclear. Several simulation studies have demonstrated that Heckman-selection models can greatly decrease bias, even when the proportion of missing data is substantial [96, 98, 99].

6.2 Longitudinal and Sequence Data

RWD are often collected over a period of time and may therefore contain multiple observations for one or more variables. Traditionally, these data are collected at frequent and regular time intervals. The recorded observations then describe a smooth trajectory that strongly resembles the underlying time process. In RWD, however, there are many challenges as compared to traditional longitudinal data. First, a large number of variables in the dataset are measured over time. For example, the MIMIC-III dataset contains patient medical records from 2001 to 2012 and includes thousands of variables with repeated measurements [100]. For standard longitudinal or sequence data, the number of variables is generally very small. Second, each variable generally has its own scheme of measurement times, and the measurement interval can be irregular and may even vary across individuals. As illustrated in Fig. 3, many clinical variables in the MIMIC-III dataset are affected by irregular measurement times. For standard longitudinal data, all variables typically follow the same scheme of measurement schedule, and for time series data, the measurement interval is fixed and remains the same for the entire series. Third, complex relationships can exist between measurements of different variables at different time points. Finally, missing data can be confounded with the irregularity of measurement schedule, and when missing data do exist, they tend to be informative and the missing rate can be very high for some variables. Due to these challenges, RWD are highly prone to MNAR mechanisms and intermittent patterns of missingness (Sect. 2.2). Traditional imputation methods are not capable of handling missing data in longitudinal datasets like EHR. In this section, we therefore discuss advanced imputation methods that are dedicated to longitudinal data. These methods can be used to reconstruct the entire trajectory of longitudinal variables for each distinct individual, but also to recover single observations at particular points in time (e.g., at the startpoint or endpoint of the study).

One approach to address the presence of missing values in longitudinal data is to recover each trajectory separately, using methods designed for time series (TS) reconstruction. Although TS methods were originally designed for the analysis of evenly spaced observations, some methods could also be used when measurement times are irregular [101]. It is, for instance, possible to replace the missing values by their respective mean or mode of the repeated measurements. These univariate algorithms are best suited for stationary series (i.e., when statistical properties of the data generation process do not change over time) and should generally be avoided because they tend to introduce bias for non-stationary series. More advanced univariate algorithms for TS imputation may account for trend (i.e., the long-term direction of the data), seasonality (i.e., systematic patterns that repeat periodically), or even certain irregularities (i.e., distribution of the residuals) of the repeated observations [102]. These algorithms often rely on moving averages or interpolation methods, and can be satisfactory when the stretches of missing data are short and if the TS is not much affected by noise [103]. Last observation carried forward (LOCF) is a special type of interpolation, where the last observed value replaces the next missing observations. Another common example is the use of autoregressive integrated moving average (ARIMA) models, which eliminate autoregressive parts from the TS and can also adjust for seasonality. However, because their implementation can distort the data distribution and their relation with other variables, univariate TS algorithms should be used with caution. Instead, multivariate TS algorithms could be used to create time lagged and lead data, and to include smooth basis functions over time in the imputation model [104]. Simulation studies found that this strategy tends to outperform simple univariate TS algorithms [103]. There are several R packages available for missing data imputation in time series. Due to space limitations, we will not list the packages individually, and refer the reader to https://CRAN.R-project.org/view=TimeSeries.

A different class of methods allows borrowing of information across individuals. When repeated measurements are structured in the wide format, time-related variables can be imputed using the methods discussed in Sect. 4 without the need for further adjustment. For instance, the Sequential Organ Failure Assessment (SOFA) score is widely employed in the daily monitoring of acute morbidity in critical care units [105]. It is calculated using information on the patient's respiratory, cardiovascular, hepatic, coagulation, renal and neurological systems, and prone to missing values when some test results are unavailable. For example, most predictors of the SOFA score were affected by missing values in MIMIC-III, with missingness rates ranging from 58.88–99.98% (Fig. 3). When repeated measurements of the SOFA score are formatted into separate columns with daily observations, corresponding variables can be imputed with joint modelling methods such as JM-MVN [30] or with conditional modelling methods such as FCS-fold [106]. Alternatively, recurrent neural networks can be used to capture long-term temporal dependencies without the need for distributional assumptions [29, 102, 107–109]. Also other machine learning methods have been customized to allow for imputation of longitudinal data, including matrix completion and nearest neighbor methods [102]. Unfortunately, these methods are not well suited to recover irregularly spaced observations (Figs. 8 and 9).

Because RWD are rarely collected at regular time intervals, it is often more helpful to structure sequential observations in the long format. Imputation methods then need to adjust for the time of measurement and the non-independence of observations. This requires to adopt hierarchical (also known as multilevel) models that group related observations, which can be achieved using joint modelling or conditional modelling imputation. A detailed overview of imputation methods for longitudinal data is provided by Huque et al. [110]. When adopting multilevel imputation methods, the longitudinal relation

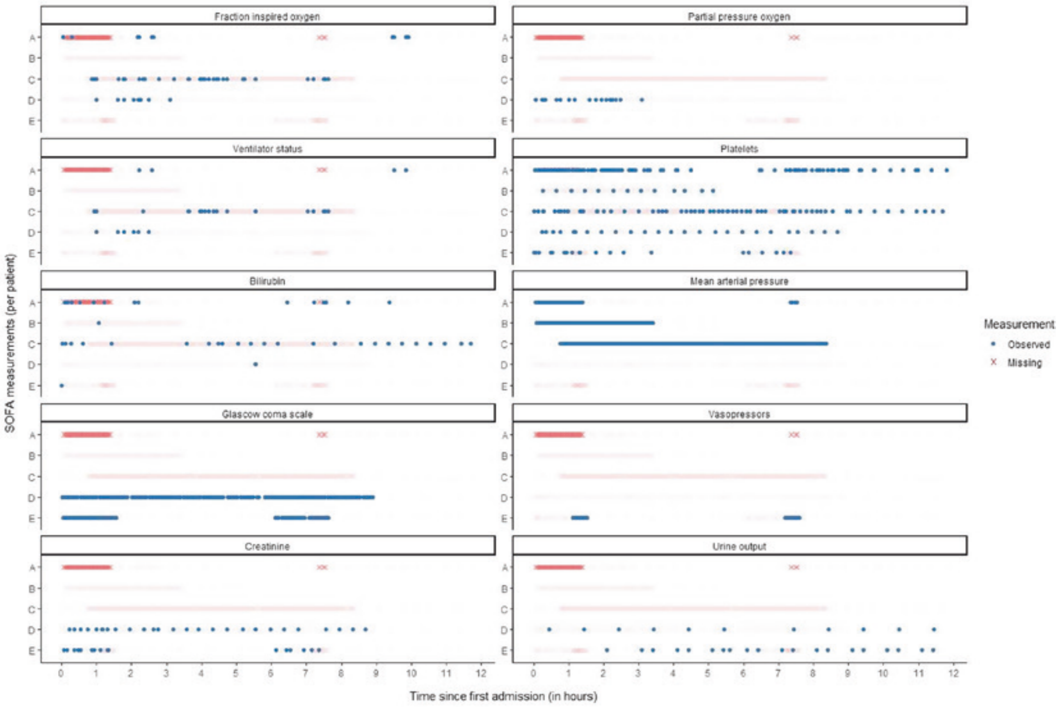


Fig. 8 Illustration of longitudinal data for five patients from MIMIC-III. Repeated measurements are presented for all predictors of the SOFA score. Each point represents a contact moment between the patient and healthcare provider

of repeated observations can be preserved by including measurement time as an explanatory (possibly random) variable. It is then common to assume a linear relationship for the effect of the time variable. Unfortunately, this approach may become problematic when there is no linear association between the incomplete variable and its predictors, when there is no compatibility between the joint distribution and the full conditional model, or when there is a lack of congeniality between the imputation model and the analysis model. Therefore, it has been proposed to adopt Bayesian substantive-model-compatible methods in which the joint distribution of the variables in the imputation model is specified by a substantive analysis model and an incomplete explanatory variable model [36, 111, 112]. Alternatively, van Buuren proposed the time raster imputation method [13] to convert irregular observations into a set of regular measurements using a piecewise linear mixed model. Initially, the user must specify an ordered set of *k* break times. Next, a B-spline model is used to represent each subject's time points with knots that are given by *k*. This approach then yields a k-column matrix *X*. Finally, the incomplete time-dependent variables are imputed using chained equations with a clustering method, using the reference variables, other time-dependent variables and *X* as predictors in the imputation method for each incomplete variable. More recently, Debray et al. developed conditional modelling imputation methods that adjust for clustering and autocorrelation. These methods were implemented using chained equations and can be used to recover missing observations at arbitrary time points [113]. Simulations showed that this approach substantially outperforms simpler imputation methods such as LOCF or rounding, and can also yield valid inferences when longitudinal data are MNAR.

It is also possible to impute longitudinal data using machine learning methods such as recurrent neural networks (RNNs). Although RNNs have been described since 1986 [114], they have rarely been used for longitudinal data analysis until the past decade. Standard RNNs bear many similarities to traditional feedforward neural networks, and can use the output from previous time steps as input for the next time step. In this manner, RNNs offer the ability to handle sequential or time series data. Traditional implementations of RNN cannot process information across many time steps and therefore have a short-term memory. This limitation can be addressed by adopting gated architectures that control the flow of information in the RNN [115]. The long short-term memory (LSTM) [116] and the Gated Recurrent Unit (GRU) are common examples of this architecture [117].

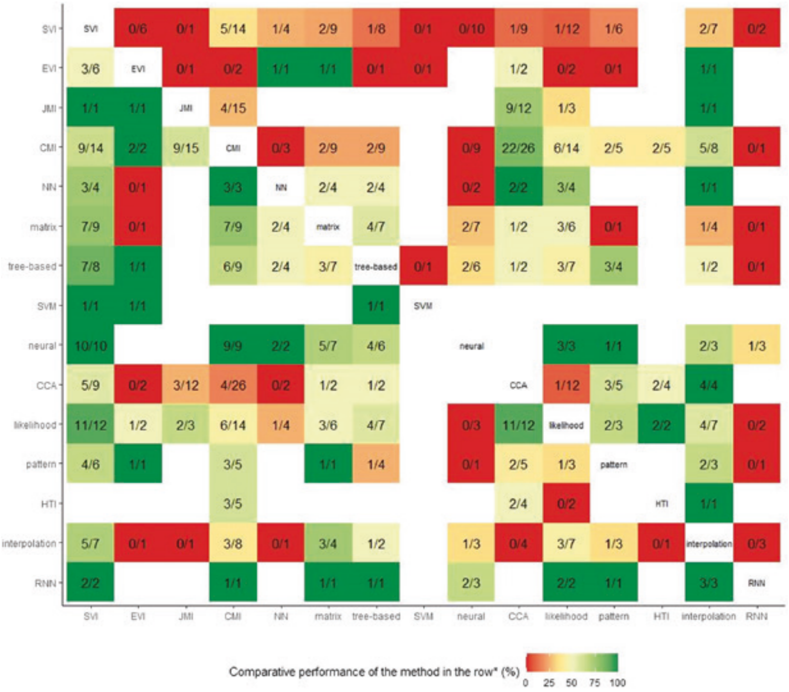
Traditional RNNs require that all variables have the same measurement schedule. For this reason, they are not well suited for the imputation and analysis of RWD. In the past few years, there have been tremendous research developments to facilitate the analysis of multivariate sequence data collected with irregular measurement schedules [102]. RNN methos can, for instance, be enhanced by adopting adversarial training, attention mechanisms, or multidirectional structures. We here distinguish between three common types of RNN for imputation of longitudinal data. A first type of RNN methods generate multiple imputed datasets, and include Bidirectional Recurrent Imputation for Time Series [118], multi-directional recurrent neural networks [119], and residual neural networks [120]. A second, similar type of RNN methods adopt generative adversarial networks to learn the overall distribution of a multivariate time series data and to generate imputed datasets [121]. Finally, a third type of RNN methods do not yield imputed datasets, but offer an integrated solution to the analysis of incomplete longitudinal data. To this purpose, they adopt missing indicators ("masks") and/or the time interval between the observed values as input values of the network [29, 122–124]. To increase the ability to capture long-term relations in the data, these non-imputation methods often adopt a GRU or LSTM architecture. Estimation of aforementioned RNNs is not straightforward and often requires dedicated software packages, which may not always be readily available or easy to use.

6.3 Choosing an Appropriate Imputation Method

The selection of an appropriate imputation method will often depend on the ultimate goal of the data analysis. If the goal is to make statistical inferences, such as estimating regression parameters or testing certain hypotheses, it is important that the imputation method provides not only unbiased estimates of parameters of interest, but also unbiased estimates of their associated (co)variance. On the other hand, if the goal of data analysis is to make predictions or classification, a suitable imputation method should be able to maintain the desired prediction or classification accuracy.

As discussed in this chapter, multiple imputation offers a generic solution to handle the presence of missing data. Multiple imputation can be used in both "inference-focused" and "prediction-focused" studies, and can also be used on a case-by-case basis (e.g., when calculating predictions in clinical practice). Multiple imputation methods that have widely been studied approximate the observed data using a well-known multivariate probability distribution (Sect. 4.1) or approximate this distribution through a series of conditional (often regression-based) models (Sect 4.2). Although these methods can greatly differ in operationalization and underlying assumptions, simulation studies have demonstrated that they generally achieve similar performance [19, 38, 125]. Overall, (semi-)parametric imputation methods can reliably be used for inference and prediction, and tend to perform well in datasets with a limited number of variables. Caution is, however, warranted when complex relations exist in the data (e.g., presence of treatment-covariate interactions), when observations are not independent (e.g., presence of repeated measurements) or when mechanisms of missingness are complex (e.g., presence of MNAR). In these situations, the required complexity of imputation methods drastically increases and manual configuration is often necessary to avoid bias (e.g., see Sects. 6.1 and 6.2) [36, 40]. In this regard, non-parametric methods offer several important advantages. First, there is no need to specify the functional form of the outcome relationship. Instead, non-linear effects and interactions are directly derived from the observed data. Second, there is no need to distinguish between different data types, as most machine learning methods can easily handle discrete, continuous and other data types. Third, because variable selection and dimensionality reduction are integrated into many machine learning procedures, they are well capable of dealing with high-dimensional datasets. Finally, because machine learning methods are extremely flexible, they are well suited to avoid incompatibilities between the imputation and substantive analysis model [40]. This is an important issue when pursuing statistical inference and is often overlooked. Machine learning methods are therefore particularly appealing when there is limited understanding about likely sources of variability in the data, as data-driven procedures are used to determine how the imputations should be generated.

Results from the literature review are summarized in Fig. 9. Each row in this figure represents



	SVI	EVI	JMI	CMI	NN	matrix	tree-based	SVM	neural	CCA	likelihood	pattern	HTI	interpolation	RNN
SVI	SVI	0/6	0/1	5/14	1/4	2/9	1/8	0/1	0/10	1/9	1/12	1/6		2/7	0/2
EVI	3/6	EVI	0/1	0/2	1/1	1/1	0/1	0/1		1/2	0/2	0/1		1/1	
JMI	1/1	1/1	JMI	4/15						9/12	1/3			1/1	
CMI	9/14	2/2	9/15	CMI	0/3	2/9	2/9		0/9	22/26	6/14	2/5	2/5	5/8	0/1
NN	3/4	0/1		3/3	NN	2/4	2/4		0/2	2/2	3/4			1/1	
matrix	7/9	0/1		7/9	2/4	matrix	4/7		2/7	1/2	3/6	0/1		1/4	0/1
tree-based	7/8	1/1		6/9	2/4	3/7	tree-based	0/1	2/6	1/2	3/7	3/4		1/2	0/1
SVM	1/1	1/1					1/1	SVM							
neural	10/10			9/9	2/2	5/7	4/6		neural		3/3	1/1		2/3	1/3
CCA	5/9	0/2	3/12	4/26	0/2	1/2	1/2			CCA	1/12	3/5	2/4	4/4	
likelihood	11/12	1/2	2/3	6/14	1/4	3/6	4/7		0/3	11/12	likelihood	2/3	2/2	4/7	0/2
pattern	4/6	1/1		3/5		1/1	1/4		0/1	2/5	1/3	pattern		2/3	0/1
HTI				3/5						2/4	0/2		HTI	1/1	
interpolation	5/7	0/1	0/1	3/8	0/1	3/4	1/2		1/3	0/4	3/7	1/3	0/1	interpolation	0/3
RNN	2/2			1/1		1/1	1/1		2/3		2/2	1/1		3/3	RNN

Fig. 9 Comparative performance of imputation methods as identified through a literature review. The color indicates the fraction of simulation studies in which the method in the row outperforms the method in the column. **Single imputation methods**: SVI = single value imputation; EVI = expected value imputation, **Multiple imputation methods**: JMI = joint modelling imputation, CMI = conditional modelling imputation, NN = nearest neighbor imputation, matrix = matrix factorization, tree-based = tree-based ensembles, SVM = support vector machine imputation, generative = neural network-based imputation, **Non-imputation methods**: CCA = complete case analysis, likelihood = likelihood-based approaches, pattern = missing data pattern methods, **Imputation of MNAR**: HTI = Heckman-type imputation, **Imputation of longitudinal data**: interpolation = interpolation methods (incl. last observation carried forward), RNN = recurrent neural networks

one method of accommodating missing data, in order of appearance in this chapter. In particular, we highlight single imputation (single value imputation, expected value imputation), joint modelling imputation, conditional modelling imputation, non-parametric imputation, non-imputation methods, and methods dedicated for informative missingness and longitudinal data. Each cell displays the total number of simulation studies in which the method in the row outperforms the method in the column. Methods that work comparatively well have a higher percentage of papers in which they outperform other methods, signified by rows with many green cells. Most studies evaluated performance by quantifying the mean squared error of imputed values.

Our literature review confirms that missing data is an important problem in RWD and requires dedicated methods. In particular, it is rarely justifiable to delete incomplete records, and to perform a so-called complete case analysis. Although missing values can accurately be recovered by adopting single imputation methods, simulation studies showed that their implementation usually leads to bias when estimating model parameters. For this reason, single imputation methods should be reserved for situations where imputations are needed on a case-by-case basis (e.g., when implementing a prediction model in clinical practice). Conversely, methods that perform consistently well are often based on multiple imputation using neural networks or other non-parametric approaches. As discussed, most of these methods can address mixed data types under various missingness mechanisms, and do not require user input to inform variable selection. Recurrent neural networks appear particularly useful because they can manage informative missingness and incomplete longitudinal data. However, when repeated measurements are relatively sparse, (semi-)parametric approaches that explicitly model their relatedness (e.g., through random effects) may be more suitable. Unfortunately, the implementation of multiple imputation methods can be very demanding w.r.t. available resources and may therefore not always be desirable. As discussed in Sect. 5, it is possible to avoid the need for imputation in some circumstances. For instance, when adopting statistical models for prediction, the presence of missing data can simply be addressed by estimating pattern submodels [92]. These models require fewer assumptions about the missing data mechanisms, and can perform well even when data are MNAR.

Finally, our review highlights several gaps in the published literature. Methods that appear promising but have not extensively been studied are based on SVM, or parametric models that estimate the joint distribution of the data and the missingness. Further, there is little consensus on appropriate strategies to evaluate missing data methods. For example, many simulation studies focus on situations where data are MCAR, or do not consider the validity of statistical inference that is based on imputed datasets. For this reason, it would be helpful to develop guidelines for the conduct and reporting of simulation studies focusing on missing data imputation, to facilitate fair comparisons between methods. For reasons of brevity, our review did not distinguish between different implementations of similar methods, such as the tree-based methods implemented within the chained equations framework. Uniting statistical and machine learning methods holds a promise to obtain imputations that are both accurate and confidence valid.

7 Summary

The analysis of RWD often requires extensive efforts to address data quality issues. In this chapter, we primarily focused on the presence of missing data and discussed several imputation methods. Although these methods are no panacea for poor quality RWD, their implementation may help address situations where RWD are incomplete or require recovery due to temporality or accuracy issues.

1. Assess whether missing data can be handled using non-imputation methods. For example, when the goal is to develop a prediction model, it is possible to avoid the need for imputation by adopting pattern submodels or built-in algorithms for dealing with missing values
2. When pursuing imputation strategies, multiple imputed values should be generated to preserve uncertainty (and thus allow for inference)
3. Include the covariates and outcome from the substantive (analysis) model [78]
4. Include as many variables as possible, especially (auxiliary) variables that are related to the variables of interest or the presence of missingness [24, 78]
5. Consider imputation methods that allow for informative missingness when missing data mechanisms cannot be ignored
6. Especially in very large data sets with many cases and variables (RWD): use flexible imputation models [24]. This can be achieved by adopting machine learning methods that have built-in procedures for variable selection and dimensionality reduction such as neural networks
7. Evaluate the quality of imputed data by inspecting trace plots and distribution of imputed values [125].

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under ReCoDID grant agreement No 825746.

References

1. Cave A, Kurz X, Arlett P. Real-world data for regulatory decision making: challenges and possible solutions for Europe. Clin Pharmacol Ther. 2019;106(1):36–9.
2. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W. What is real-world data (RWD)? A review of definitions based on literature and stakeholder interviews. Value in Health [Internet]. 2017 May [cited 2017 Jun 12]; Available from: http://linkinghub.elsevier.com/retrieve/pii/S1098301517301717.
3. Cook JA, Collins GS. The rise of big clinical databases. Br J Surg. 2015;102(2):e93–101.
4. Michaels JA. Use of mortality rate after aortic surgery as a performance indicator. Br J Surg. 2003;90(7):827–31.
5. Black N, Payne M. Directory of clinical databases: improving and promoting their use. Qual Saf Health Care. 2003;12(5):348–52.
6. Aylin P, Lees T, Baker S, Prytherch D, Ashley S. Descriptive study comparing routine hospital administrative data with the Vascular Society of Great Britain and Ireland's National Vascular Database. Eur J Vasc Endovasc Surg. 2007;33(4):461–5; discussion 466.
7. Kelly M, Lamah M. Evaluating the accuracy of data entry in a regional colorectal cancer database: implications for national audit. Colorectal Dis. 2007;9(4):337–9.
8. Stey AM, Ko CY, Hall BL, Louie R, Lawson EH, Gibbons MM, et al. Are procedures codes in claims data a reliable indicator of intraoperative splenic injury compared with clinical registry data? J Am Coll Surg. 2014;219(2):237-244.e1.
9. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit on Translat Bioinforma. 2010;1(2010):1–5.
10. Peek N, Rodrigues PP. Three controversies in health data science. Int J Data Sci Anal [Internet]. 2018 [cited 2018 Mar 12]; Available from: https://doi.org/10.1007/s41060-018-0109-y.
11. Ehrenstein V, Kharrazi H, Lehmann H, Taylor CO. Obtaining data from electronic health records [Internet]. Tools and technologies for registry interoperability, registries for evaluating patient outcomes: A user's guide, 3rd ed., Addendum 2 [Internet]. Agency for Healthcare Research and Quality (US); 2019 [cited 2021 Aug 27]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK551878/.
12. Feder SL. Data quality in electronic health records research: quality domains and assessment methods. West J Nurs Res. 2018;40(5):753–66.
13. van Buuren S. Longitudinal data. In: Flexible imputation of missing data, 2nd edn. Boca Raton: Chapman and Hall/CRC; 2018. (Chapman & Hall/CRC Interdisciplinary Statistics).
14. Diehl J. Preprocessing and visualization. Aachen, Germany: RWTH Aachen University; 2004 Jan. Report No.: 235087.
15. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92.
16. Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Methods Med Res. 2007;16(3):259–75.
17. Little RJA, Rubin DB. Statistical analysis with missing data, 2nd edn. Hoboken, NJ: Wiley; 2002. 381 p. (Wiley series in probability and statistics).

18. van Buuren S. Flexible imputation of missing data [Internet], 2nd edn. Boca Raton: CRC Press, Taylor & Francis Group; 2018 [cited 2018 Nov 8]. 415 p. (Chapman & Hall/CRC Interdisciplinary Statistics). Available from: https://stefvanbuuren.name/fimd/.
19. Audigier V, White IR, Jolani S, Debray TPA, Quartagno M, Carpenter JR, et al. Multiple imputation for multilevel data with continuous and binary variables. Stat Sci. 2018;33(2):160–83.
20. Debray TPA, Snell KIE, Quartagno M, Jolani S, Moons KGM, Riley RD. Dealing with missing data in an IPD meta-analysis. In: Individual participant data meta-analysis: a handbook for healthcare research. Hoboken, NJ: Wiley; 2021. (Wiley series in statistics in practice).
21. Hunt NB, Gardarsdottir H, Bazelier MT, Klungel OH, Pajouheshnia R. A systematic review of how missing data are handled and reported in multi-database pharmacoepidemiologic studies. Pharmacoepidemiol Drug Saf. 2021;pds.5245.
22. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–5.
23. Murray JS. Multiple imputation: a review of practical and theoretical findings. Statist Sci [Internet]. 2018 [cited 2021 May 7];33(2). Available from: https://projecteuclid.org/journals/statistical-science/volume-33/issue-2/Multiple-Imputation-A-Review-of-Practical-and-Theoretical-Findings/10.1214/18-STS644.full.
24. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15(90):3133–81.
25. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021;3(2):125–33.
26. Van de Schoot R, De Bruin J, Schram R, Zahedi P, De Boer J, Weijdema F, et al. ASReview: active learning for systematic reviews [Internet]. Zenodo; 2021 [cited 2021 Sep 8]. Available from: https://zenodo.org/record/5126631.
27. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;24(3): 160035.
28. Johnson A, Pollard T, Mark R. MIMIC-III Clinical Database [Internet]. PhysioNet; 2019 [cited 2021 Sep 24]. Available from: https://physionet.org/content/mimiciii-demo.
29. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8(1):6085.
30. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7(2):147–77.
31. Nijman SWJ, Hoogland J, Groenhof TKJ, Brandjes M, Jacobs JJL, Bots ML, et al. Real-time imputation of missing predictor values in clinical practice. Eur Heart J Digital Health. 2020;2(1):154–64.
32. Nijman SWJ, Groenhof TKJ, Hoogland J, Bots ML, Brandjes M, Jacobs JJL, et al. Real-time imputation of missing predictor values improved the application of prediction models in daily practice. J Clin Epidemiol. 2021;19(134):22–34.
33. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
34. Harel O, Mitchell EM, Perkins NJ, Cole SR, Tchetgen Tchetgen EJ, Sun B, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):576–84.
35. Carpenter JR, Kenward MG. Multiple imputation and its application [Internet]. 1st ed. John Wiley & Sons, Ltd; 2013 [cited 2014 Dec 18]. (Statistics in Practice). Available from: https://doi.org/10.1002/9781119942283.
36. Erler NS, Rizopoulos D, Jaddoe VW, Franco OH, Lesaffre EM. Bayesian imputation of time-varying covariates in linear mixed models. Stat Methods Med Res. 2019;28(2):555–68.
37. Erler NS, Rizopoulos D, Rosmalen J van, Jaddoe VWV, Franco OH, Lesaffre EMEH. Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. Stat Med. (2016).
38. Hughes RA, White IR, Seaman SR, Carpenter JR, Tilling K, Sterne JAC. Joint modelling rationale for chained equations. BMC Med Res Methodol. 2014;14:28.
39. Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Statistical Software [Internet]. 2011;45(3). Available from: http://doc.utwente.nl/78938/.
40. Meng X-L. Multiple-imputation inferences with uncongenial sources of input. Stat Sci. 1994;9(4):538–58.
41. Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. Comput Stat Data Anal. 2015;1(90):84–99.
42. Bay SD. Combining nearest neighbor classifiers through multiple feature subsets. In: Proceedings of the fifteenth international conference on machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1998. p. 37–45. (ICML '98).
43. Ding Y, Ross A. A comparison of imputation methods for handling missing scores in biometric fusion. Pattern Recogn. 2012;45(3):919–33.
44. Vink G, Frank LE, Pannekoek J, van Buuren S. Predictive mean matching imputation of semicontinuous variables: PMM imputation of semicontinuous variables. Stat Neerl. 2014;68(1):61–90.
45. Faisal S, Tutz G. Multiple imputation using nearest neighbor methods. Inf Sci. 2021;570:500–16.

46. Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. Appl Artif Intell. 2019;33(10):913–33.
47. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010;50(2):105–15.
48. Thomas T, Rajabi E. A systematic review of machine learning-based missing value imputation techniques. Data Tech Appl. 2021;55(4):558–85.
49. Marimont RB, Shapiro MB. Nearest neighbour searches and the curse of dimensionality. IMA J Appl Math. 1979;24(1):59–70.
50. Davenport MA, Romberg J. An overview of low-rank matrix recovery from incomplete observations. IEEE J Sel Top Sig Proc. 2016;10(4):608–22.
51. Li XP, Huang L, So HC, Zhao B. A survey on matrix completion: Perspective of Signal Processing. arXiv:190110885 [eess] [Internet]. 2019 May 7 [cited 2021 Aug 20]; Available from: http://arxiv.org/abs/1901.10885.
52. Sportisse A, Boyer C, Josse J. Imputation and low-rank estimation with Missing Not At Random data. arXiv:181211409 [cs, stat] [Internet]. 2020 Jan 29 [cited 2021 Aug 20]; Available from: http://arxiv.org/abs/1812.11409.
53. Hernandez-Lobato JM, Houlsby N, Ghahramani Z. Probabilistic Matrix Factorization with non-random missing data. In: International conference on machine learning [Internet]. PMLR; 2014 [cited 2021 Aug 20]. p. 1512–20. Available from: https://proceedings.mlr.press/v32/hernandez-lobatob14.html.
54. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–8.
55. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a caliber study. Am J Epidemiol. 2014;179(6):764–74.
56. Ramosaj B, Pauly M. Who wins the miss contest for imputation methods? Our vote for miss BooPF. arXiv: 171111394 [stat] [Internet]. 2017 Nov 30 [cited 2021 Aug 24]; Available from: http://arxiv.org/abs/1711.11394.
57. Tang F, Ishwaran H. Random forest missing data algorithms. Stat Anal Data Min. 2017;10(6):363–77.
58. Breiman L. Manual for setting up, using, and understanding random forest V4.0 [Internet]. 2003. Available from: https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.
59. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The Ann Appl Stat. 2008;2(3):841–60.
60. Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. Am J Epidemiol. 2010;172(9):1070–6.
61. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Methodol. 2020;20(1):199.
62. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
63. Vapnik V. The nature of statistical learning theory [Internet], 2nd edn. New York: Springer-Verlag; 2000 [cited 2021 Aug 24]. (Information Science and Statistics). Available from: https://www.springer.com/gp/book/9780387987804.
64. Pereira RC, Santos MS, Rodrigues PP, Abreu PH. Reviewing autoencoders for missing data imputation: technical trends, applications and outcomes. J Artif Intell Res. 2020;14(69):1255–85.
65. Beaulieu-Jones BK, Moore JH. Missing data imputation in the electronic health record using deeply learned autoencoders. Pac Symp Biocomput. 2017;22:207–18.
66. Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on machine learning [Internet]. New York, NY, USA: Association for Computing Machinery; 2008 [cited 2021 Aug 25]. p. 1096–103. (ICML '08). Available from: https://doi.org/10.1145/1390156.1390294.
67. Gondara L, Wang K. MIDA: multiple Imputation using denoising autoencoders. arXiv: 170502737 [cs, stat] [Internet]. 2018 Feb 17 [cited 2021 Aug 25]; Available from: http://arxiv.org/abs/1705.02737.
68. Lall R, Robinson T. The MIDAS touch: accurate and scalable missing-data imputation with deep learning. Polit Anal. 2021;26:1–18.
69. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv:13126114 [cs, stat] [Internet]. 2014 May 1 [cited 2021 Aug 25]; Available from: http://arxiv.org/abs/1312.6114.
70. Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. In: Proceedings of the 31st international conference on international conference on machine learning, Vol. 32. Beijing, China: JMLR.org; 2014. p. II-1278-II–1286. (ICML'14).
71. Ma C, Tschiatschek S, Turner R, Hernández-Lobato JM, Zhang C. VAEM: a deep generative model for heterogeneous mixed type data. In: Advances in neural information processing systems [Internet]. Curran Associates, Inc.; 2020 [cited 2021 Aug 25]. p. 11237–47. Available from: https://papers.nips.cc/paper/2020/hash/8171ac2c5544a5cb54ac0f38bf477af4-Abstract.html.

72. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. arXiv:14062661 [cs, stat] [Internet]. 2014 Jun 10 [cited 2021 Aug 25]; Available from: http://arxiv.org/abs/1406.2661.
73. Li SC-X, Jiang B, Marlin B. MisGAN: learning from incomplete data with generative adversarial networks. arXiv:190209599 [cs, stat] [Internet]. 2019 Feb 25 [cited 2021 Aug 25]; Available from: http://arxiv.org/abs/1902.09599.
74. Shang C, Palmer A, Sun J, Chen K-S, Lu J, Bi J. VIGAN: missing view imputation with generative adversarial networks. arXiv:170806724 [cs, stat] [Internet]. 2017 Nov 1 [cited 2021 Aug 25]; Available from: http://arxiv.org/abs/1708.06724.
75. Yoon J, Jordon J, van der Schaar M. GAIN: missing data imputation using generative adversarial nets. arXiv:180602920 [cs, stat] [Internet]. 2018 Jun 7 [cited 2021 Aug 25]; Available from: http://arxiv.org/abs/1806.02920.
76. van Buuren S. Rubin's rules. In: Flexible imputation of missing data, 2nd edn. Boca Raton: CRC Press, Taylor & Francis Group; 2018. (Chapman & Hall/CRC Interdisciplinary Statistics).
77. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30(4):377–99.
78. Vink G, van Buuren S. Pooling multiple imputations when the sample happens to be the population. arXiv:14098542 [math, stat] [Internet]. 2014 Sep 30 [cited 2021 Aug 27]; Available from: http://arxiv.org/abs/1409.8542..
79. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Stat Med. 2008;27(17):3227–46.
80. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol. 2009;28(9):57.
81. Zhao Y, Long Q. Variable selection in the presence of missing data: imputation-based methods. Wiley Interdiscip Rev Comput Stat. 2017;9(5): e1402.
82. Little RJA. Regression with missing X's: a review. J Am Stat Assoc. 1992;87(420):1227–37.
83. Herring AH, Ibrahim JG. Likelihood-based methods for missing covariates in the cox proportional hazards model. J Am Stat Assoc. 2001;96(453):292–302.
84. Xie Y, Zhang B. Empirical Likelihood in Nonignorable covariate-missing data problems. Int J Biostat. [Internet]. 2017 [cited 2021 Sep 21];13(1). Available from: https://www.degruyter.com/document/doi/10.1515/ijb-2016-0053/html.
85. Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. Biostatistics [Internet]. 2018 [cited 2018 Sep 27]; Available from: https://academic.oup.com/biostatistics/advance-article/doi/10.1093/biostatistics/kxy040/5092384.
86. Breiman L. Classification and regression trees. Wadsworth International Group; 1984. 376 p.
87. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining - KDD '16. 2016;785–94.
88. van Smeden M, Groenwold RHH, Moons KG. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. J Clin Epidemiol. 2020.
89. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc. 2016.
90. Haneuse S, Arterburn D, Daniels MJ. Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task. JAMA Netw Open. 2021;4(2): e210184.
91. Hardt J, Herke M, Leonhart R. Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. BMC Med Res Methodol. 2012;12:184.
92. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test (Madr). 2009;18(1):1–43.
93. Michiels B, Molenberghs G, Bijnens L, Vangeneugden T, Thijs H. Selection models and pattern-mixture models to analyse longitudinal quality of life data subject to drop-out. Stat Med. 2002;21(8):1023–41.
94. Creemers A, Hens N, Aerts M, Molenberghs G, Verbeke G, Kenward MG. Generalized shared-parameter models and missingness at random. Stat Model. 2011;11(4):279–310.
95. Heckman JJ. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann Econ Soc Meas. 1976;5(4):475–92.
96. Koné S, Bonfoh B, Dao D, Koné I, Fink G. Heckman-type selection models to obtain unbiased estimates with missing measures outcome: theoretical considerations and an application to missing birth weight data. BMC Med Res Methodol. 2019;19(1):231.
97. Muñoz J, Hufstedler H, Gustafson P, Bärnighausen T, De Jong VMT, Debray TPA (2023) Dealing with missing data using the Heckman selection model: methods primer for epidemiologists. Int J Epidemiol 2(1):5–13
98. Galimard J-E, Chevret S, Curis E, Resche-Rigon M. Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. BMC Med Res Methodol. 2018;18(1):90.
99. Holmes FW. A comparison of the heckman selection model, ibrahim, and lipsitz methods for dealing with nonignorable missing data. J Psychiatry Behav Sci. 2021;4(1):1045.

100. Deasy J, Liò P, Ercole A. Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or curation. Sci Rep. 2020;10(1):22129.
101. Eckner A. A Framework for the analysis of unevenly spaced time series data [Internet]. 2014 [cited 2021 Sep 24]. Available from: https://www.semanticscholar.org/paper/A-Framework-for-the-Analysis-of-Unevenly-Spaced-Eckner/bb307aa6671a5a65314d3a26fffa6c7ef48a3c86.
102. Fang C, Wang C. Time series data imputation: a survey on deep learning approaches. arXiv: 201111347 [cs] [Internet]. 2020 Nov 23 [cited 2021 Aug 25]; Available from: http://arxiv.org/abs/2011.11347.
103. Bauer J, Angelini O, Denev A. Imputation of multivariate time series data - performance benchmarks for multiple imputation and spectral techniques. SSRN J [Internet]. 2017 [cited 2021 Aug 27]; Available from: https://www.ssrn.com/abstract=2996611.
104. Zhang Z. Multiple imputation for time series data with Amelia package. Ann Transl Med. 2016;4(3):56.
105. Lambden S, Laterre PF, Levy MM, Francois B. The SOFA score—development, utility and challenges of accurate assessment in clinical trials. Crit Care. 2019;23(1):374.
106. Nevalainen J, Kenward MG, Virtanen SM. Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification. Stat Med. 2009;28(29):3657–69.
107. Guo Y, Liu Z, Krishnswamy P, Ramasamy S. Bayesian recurrent framework for missing data imputation and prediction with clinical time series. arXiv: 191107572 [cs, stat] [Internet]. 2019 [cited 2021 May 7]; Available from: http://arxiv.org/abs/1911.07572..
108. Yu K, Zhang M, Cui T, Hauskrecht M. Monitoring ICU mortality risk with a long short-term memory recurrent neural network. Pac Symp Biocomput. 2020;25:103–14.
109. Li Q, Xu Y. VS-GRU: a variable sensitive gated recurrent neural network for multivariate time series with massive missing values. Appl Sci. 2019;9(15):3041.
110. Huque MH, Carlin JB, Simpson JA, Lee KJ. A comparison of multiple imputation methods for missing data in longitudinal studies. BMC Med Res Methodol. 2018;18(1):168.
111. Enders CK, Du H, Keller BT. A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. Psychol Methods. 2020;25(1):88–112.
112. Goldstein H, Carpenter JR, Browne WJ. Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. J R Stat Soc A Stat Soc. 2014;177(2):553–64.
113. Debray TP, Simoneau G, Copetti M, Platt RW, Shen C, Pellegrini F et al (2023) Methods for comparative effectiveness based on time to confirmed disability progression with irregular observations in multiple sclerosis. Stat Methods Med Res. https://doi.org/10.1177/09622802231172032
114. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6.
115. Weerakody PB, Wong KW, Wang G, Ela W. A review of irregular time series data handling with gated recurrent neural networks. Neurocomputing. 2021;21(441):161–78.
116. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
117. Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) [Internet]. Doha, Qatar: Association for Computational Linguistics; 2014 [cited 2021 Sep 22]. p. 1724–34. Available from: https://aclanthology.org/D14-1179.
118. Cao W, Wang D, Li J, Zhou H, Li L, Li Y. BRITS: Bidirectional recurrent imputation for time series. In: Advances in neural information processing systems [Internet]. Curran Associates, Inc.; 2018 [cited 2021 Sep 22]. Available from: https://proceedings.neurips.cc/paper/2018/hash/734e6bfcd358e25ac1db0a4241b95651-Abstract.html.
119. Yoon J, Zame WR, van der Schaar M. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. IEEE Trans Biomed Eng. 2019;66(5):1477–90.
120. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). 2016. p. 770–8.
121. Luo Y, Cai X, ZHANG Y, Xu J, Xiaojie Y. Multivariate time series imputation with generative adversarial networks. In: Advances in neural information processing systems [Internet]. Curran Associates, Inc.; 2018 [cited 2021 Sep 22]. Available from: https://papers.nips.cc/paper/2018/hash/96b9bff013acedfb1d140579e2fbeb63-Abstract.html.
122. Lipton ZC, Kale DC, Wetzel R. Modeling Missing Data in Clinical time series with RNNs. Proc Mach Learn Healthc. 2016;2016:17.
123. Baytas IM, Xiao C, Zhang XS, Wang F, Jain AK, Zhou J. Patient subtyping via time-aware LSTM networks. KDD. 2017.

124. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting clinical events via recurrent neural networks. In: Proceedings of the 1st machine learning for healthcare conference [Internet]. PMLR; 2016 [cited 2021 Sep 22]. p. 301–18. Available from: https://proceedings.mlr.press/v56/Choi16.html
125. Quartagno M, Carpenter JR. Multiple imputation for discrete data: Evaluation of the joint latent normal model. Biom J. 2019;61(4):1003–19.
126. Raghunathan T, Bondarenko I. Diagnostics for multiple imputations [Internet]. Rochester, NY: Social Science Research Network; 2007 Nov [cited 2021 Sep 24]. Report No.: ID 1031750. Available from: https://papers.ssrn.com/abstract=1031750.