Artificial Intelligence in Medicine 58 (2013) 63–72

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Missing data in medical databases: Impute, delete or classify?

Federico Cismondi[a,b,c,*], André S. Fialho[a,b,c], Susana M. Vieira[b], Shane R. Reti[c], João M.C. Sousa[b], Stan N. Finkelstein[a]

[a] Massachusetts Institute of Technology, Engineering Systems Division, 77 Massachusetts Avenue, 02139 Cambridge, MA, USA
[b] Technical University of Lisbon, Instituto Superior Técnico, Department of Mechanical Engineering, CIS/IDMEC – LAETA, Av. Rovisco Pais, 1049-001 Lisbon, Portugal
[c] Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Centre, Harvard Medical School, Boston, MA, USA

ARTICLE INFO

Article history:
Received 18 June 2012
Received in revised form 1 November 2012
Accepted 10 January 2013

Keywords:
Missing data classification
Statistical classifier
Fuzzy systems
Test bed
Intensive care unit

ABSTRACT

Background: The multiplicity of information sources for data acquisition in modern intensive care units (ICUs) makes the resulting databases particularly susceptible to missing data. Missing data can significantly affect the performance of predictive risk modeling, an important technique for developing medical guidelines. The two most commonly used strategies for managing missing data are to impute or delete values, and the former can cause bias, while the later can cause both bias and loss of statistical power.
Objectives: In this paper we present a new approach for managing missing data in ICU databases in order to improve overall modeling performance.
Methods: We use a statistical classifier followed by fuzzy modeling to more accurately determine which missing data should be imputed and which should not. We firstly develop a simulation test bed to evaluate performance, and then translate that knowledge using exactly the same database as previously published work by [13].
Results: In this work, test beds resulted in datasets with missing data ranging 10–50%. Using this new approach to missing data we are able to significantly improve modeling performance parameters such as accuracy of classifications by an 11%, sensitivity by 13%, and specificity by 10%, including also area under the receiver–operator curve (AUC) improvement of up to 13%.
Conclusions: In this work, we improve modeling performance in a simulated test bed, and then confirm improved performance replicating previously published work by using the proposed approach for missing data classification. We offer this new method to other researchers who wish to improve predictive risk modeling performance in the ICU through advanced missing data management.

1. Introduction

Modeling and predicting different clinical outcomes in intensive care units (ICUs) are important for improving health care delivery, and ultimately patient outcomes [1,2]. In many ICUs, multiple patient variables are captured from bedside monitors, laboratory results, clinical progress notes, and admission/discharge data. Advances in computer science and acquisition systems have made it possible to easily collect and store these data in large databases containing long time series from diverse sources [1].

One consequence of large amounts of data from diverse sources is that sampled time series are more prone (a) to be collected in a misaligned uneven fashion [3], as well as (b) to be partially lost or unavailable (missing data) [4]. To manage this, in the first instance, unevenly acquired data usually requires some form of alignment correction into a regular time series template using techniques such as gridding [5]. In doing so, the appearance of missing data is created by the study design solely by virtue of the sampling frequency chosen. For example, blood pressure may be sampled hourly, and lab tests 4 hourly. A gridding template choosing a 1 h sampling frequency will therefore appear to show many periods of missing data for lab tests, when in fact that data is missing because of the choice of sampling frequency rather than the lab tests not being done. In contrast, sometimes missing data is truly missing for reasons not related to study design type features such as sampling frequency; in the example above lab tests may have forgotten to be done, and decisions need to be made on how to handle this form of missing data also.

There are two conventional approaches that are widely used to deal with missing data [6]: (1) delete all variables corresponding to a given sampling time if at least one of them is missing, or (2) impute values for all missing data. A clinical example of approach (1) would be a patient with normal white blood cell counts (WBCs).

* Corresponding author at: Massachusetts Institute of Technology, Engineering Systems Division, 77 Massachusetts Avenue, 02139 Cambridge, MA, USA.
Tel.: +1 617 435 6534; fax: +1 617 278 8188.
E-mail addresses: cismondi@mit.edu, fcismondi@hotmail.com, fcismondi@gmail.com (F. Cismondi).

0933-3657/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2013.01.003

Such a patient might have less frequent WBC sampling, say daily, compared to a patient with a confirmed infection that might be tested more frequently, even hourly [7]. In the case of the normal WBC, if other variables are then deleted when WBC is not measured, information loss can occur with possible bias and loss of statistical power [8,9]. On the other hand, approach (2) can be represented by a patient that has been periodically connected and disconnected from a ventilator; such a patient would only have segments of time series recorded for ventilator acquired variables, but imputing values to replace apparent missing data would create unrealistic information in the dataset, that is, it may suggest the patient was under assisted ventilation all of the time, which in turn could significantly bias predictive results [4,9].

In order to avoid bias and loss of statistical power [4], classifying the pattern of missingness is a useful step that enables the application of appropriate imputation methods to improve database consistency [10]. These pre-processing procedures, together with outlier removal, are estimated to represent 60% of resource effort in predictive risk modeling [11,12].

The objectives of this study include the development of (a) reliable methods to align misaligned unevenly sampled data using gridding and templating, (b) a statistical classification to differentiate absent values resulting from low sampling frequencies from those related to missingness mechanisms and (c) artificial intelligence techniques to classify recoverable and not-recoverable segments of missing data. We classify missing values as recoverable if there is no information in the remaining variables that could explain the missingness, and not-recoverable otherwise. We use the terms recoverable and not-recoverable to identify missing values that have to be imputed or deleted, respectively. The individual performance of each step is assessed using a simulated test dataset that we created specifically for this study. The impact of the misalignment correction and missing data classification steps that we propose are further evaluated by comparing an otherwise identical study design, published using septic shock predictive risk modeling [13].

This article is organized as follows. Section 2 formally describes misaligned unevenly sampled data and missing data phenomena. Section 3 presents the methods used to deal with them. Section 4 details the datasets used to test the proposed methods. Section 5 describes and discusses the results obtained from these methods, after testing them against artificially created test datasets, as well as in predictive risk modeling for septic shock patients. Conclusions are presented in Section 6.

2. Raw medical data issues

2.1. Misaligned unevenly sampled data

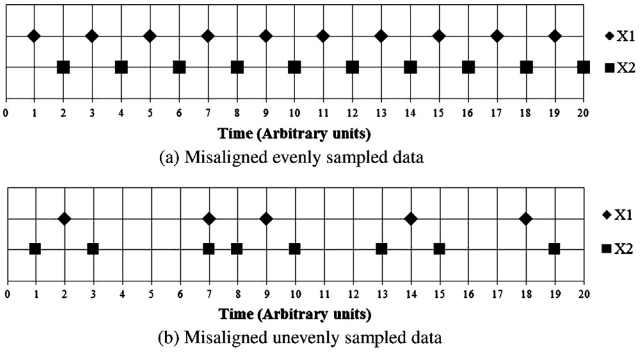
When multiple medical processes from different sources that evolve in time are recorded in databases, the resulting stored data is often in the form of misaligned time series [14]. Technically, time series are said to be misaligned when their samples were not recorded with the same sampling time. This misalignment may occur in two ways: evenly or unevenly. The first scenario, misaligned evenly sampled data, can be formally represented by assuming two discrete random variables X_1 and X_2, both evenly sampled with the same sampling time T [15]:

$$X_1 = X_1(t) = \{x_{1j}(t_j); (t_j - t_{j-1}) = T\} \quad j = 1, 2, \ldots, n \tag{1}$$

$$X_2 = X_2(t) = \{x_{2k}(t_k); (t_k - t_{k-1}) = T\} \quad k = 1, 2, \ldots, n \tag{2}$$

$$t_j - t_k = c_{jk} = c; \quad c \in R \tag{3}$$

where x_{ij} represents the value of a variable X_i at time t_j, and c_{jk} is the absolute time difference between x_{1j} and x_{2k}. As depicted in Fig. 1a, samples of X_1 and X_2 were not acquired at the same times, but they are equally spaced in time, i.e. they are misaligned but evenly sampled.



(a) Misaligned evenly sampled data

(b) Misaligned unevenly sampled data

Fig. 1. Simplified representation of real time series after the acquisition process.

The second scenario, misaligned unevenly sampled data, can be also formally represented by assuming two discrete random variables X_1 and X_2, unevenly sampled [15]:

$$X_1 = X_1(t) = \{x_{1j}(t_j); (t_j - t_{j-1}) = a_j \in R\} \quad j = 1, 2, \ldots, n \tag{4}$$

$$X_2 = X_2(t) = \{x_{2k}(t_k); (t_k - t_{k-1}) = b_k \in R\} \quad k = 1, 2, \ldots, n \tag{5}$$

$$t_j - t_k = c_{jk}; \quad c_{jk} \in R \tag{6}$$

where a_j and b_k are the times between samples of X_1 and X_2, at times t_j, t_{j-1}, and t_k, t_{k-1}, respectively. In this case, as depicted in Fig. 1b, X_1 and X_2 samples neither occur at the same times nor are equally spaced in time, i.e. they are misaligned and unevenly sampled; both the time interval between two samples of the same variable, a_j or b_k, and the time interval between two samples of different variables, c_{jk}, are random, which represents the worst case sampling scenario [16].

2.2. Missing data

The discussion presented in the previous section demonstrates how missing gaps of information between samples can effectively be a result of the sampling method used, and not of the process being sampled. In the rest of this study, we consider as *true missing data* those gaps in the time series that occur because of the process being sampled.

True missing data can be formally presented in a dataset D, which may consist of D_i variables, with $i = 1, \ldots, n$. Each variable D_i may consist of k_i samples, with $k_i \in N$. Let us denote as Y the variable D_i for which the missingness is to be assessed, and X the remaining variable(s) in the dataset D [10,17]. The variable Y consists of observed and missing values, i.e. $Y = \{Y_{obs}, Y_{miss}\}$. Similarly for the remaining variables, $X = \{X_{obs}, X_{miss}\}$. Data is said to be missing at random (MAR) if the probability of Y_{miss} is independent of Y. Formally,

$$P(Y_{miss}|Y, X) = P(Y_{miss}|X) \Rightarrow Y_{miss} \perp\!\!\!\perp Y \tag{7}$$

This means missingness may depend on X, but to be MAR it must be independent of present and/or absent values of Y. If the probability of Y_{miss} is dependent of Y, the missingness is named missing not at random (MNAR). Formally,

$$P(Y_{miss}|Y, X) = P(Y_{miss}|Y, X) \Rightarrow Y_{miss} \top\!\!\!\top Y \tag{8}$$

This means missingness may depend on X, but to be MNAR it must be dependent of present and/or absent values of Y. Missing completely at random (MCAR) is a special case of MAR, where

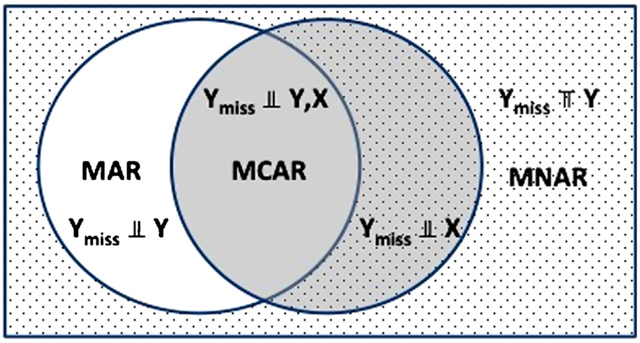

$Y_{miss} \perp\!\!\!\perp Y,X$
$Y_{miss} \top\!\!\!\top Y$
MAR
MCAR
MNAR
$Y_{miss} \perp\!\!\!\perp Y$
$Y_{miss} \perp\!\!\!\perp X$

Fig. 2. Missing data classification depicting Y_{miss} dependence on X and Y.

missing values are independent of the information in both Y and X. Formally,

$$P(Y_{miss}|Y,X) = P(Y_{miss}) \Rightarrow Y_{miss} \perp\!\!\!\perp Y, X \tag{9}$$

Since the values of Y_{miss} are unknown, it is impossible to use them to assess any dependence. Thus, it is impossible to prove the MAR/MNAR condition. For a better visualization and understanding, the logical relations between the three sets of missing data are depicted in Fig. 2. In Fig. 2 we can see that there are two types of MAR: those dependent of X, and those independent of X (MCAR). The same situation is found for MNAR.

Focusing exclusively on the dependence of Y_{miss} on X, two main interactions can be pointed out (Fig. 3). If the probability of Y_{miss} depends on X $((Y_{miss} \top\!\!\!\top X))$, those missing data are called observed not at random (ONAR) [6]. This interaction includes fractions of MNAR and MAR present in a dataset. If there are no systematic differences on the fully observed variables X between Y_{obs} and Y_{miss} segments ($Y_{miss} \perp\!\!\!\perp X$), then those missing data are called observed at random (OAR) [6]. This interaction includes all MCAR, and the remaining fraction of MNAR not classified as ONAR. This dependence will be further exploited in the explanation of the methods to classify missing data, since the dependence of Y on X can be assessed.

3. Methods

3.1. Alignment of misaligned unevenly sampled data

A strategy to align samples in a dataset is needed since most modeling tools need all variables to have samples acquired/recorded at the same sampling times [6,5]. We utilize two alignment methods, gridding and templating.

Gridding consists of aligning all variables according to a fixed sampling rate which becomes the grid with time series on the x axis and data points on the y axis. All samples of every variable are shifted to occupy the nodes of the grid [13]. This approach offers the advantage of allowing all variables to be resampled uniformly, which is highly desirable for most modeling techniques [5].

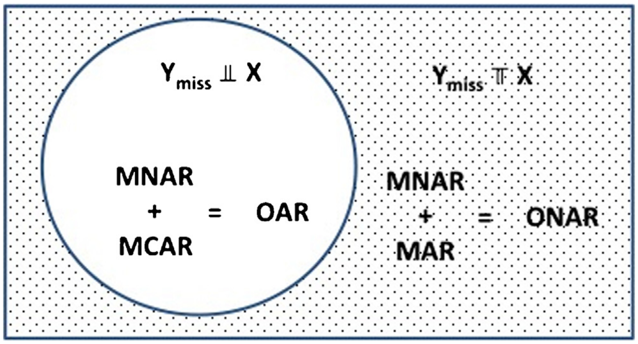


Fig. 3. Missing data classification depicting Y_{miss} dependence exclusively on X.

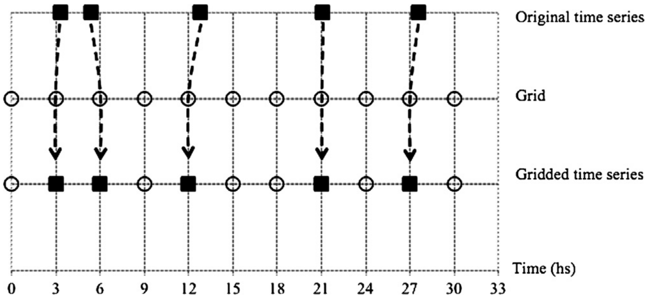


Fig. 4. Simplified representation of the gridding process, with the *shifted* and *unshifted* versions of a variable.

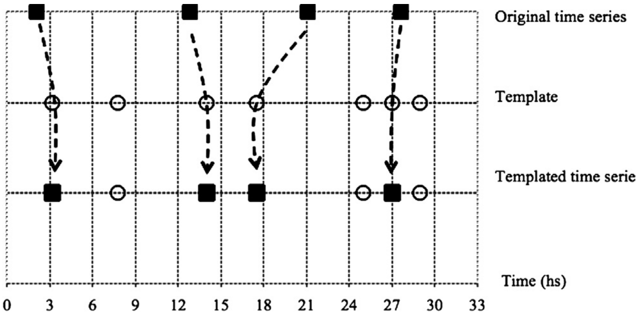


Fig. 5. Simplified representation of a *template variable*, together with the *shifted* and *unshifted* versions of a second variable.

Fig. 4 demonstrates this approach. Here we can see how the time series on the top axis before gridding, is aggregated to the closest node on the bottom axis grid below. Where several data points all aggregate to the same node, cubic interpolation determines the value that is placed on the grid [18]. No new values are created at this stage of the process, but blank values before gridding are effectively blank values after gridding.

Templating is a similar technique for correcting misaligned time series, except in templating the fixed-sampling-period grid is replaced with the sampling times of one of the variables in the dataset. The same principles of cubic-interpolating multiple values that aggregate to the same nearest node are still applied. In this study, heart rate is chosen as the templating variable because it has the highest sampling frequency, it is the time series with the most information about the original sampling, and it minimizes displacement and loss of information [18]. An example of templating is demonstrated in Fig. 5.

At the end of the alignment correction process, a non-rectangular matrix, similar to Table 1a, is obtained. Using only the columns with complete data in the nonrectangular matrix (columns 2, 5 and 6) creates the previously mentioned problem of loss of possibly useful information, which turns into very few points

Table 1
Examples of (a) non-rectangular and (b) rectangular matrix fully populated with values and NaNs.

	1	2	3	4	5	6	7
(a)							
Complete variable	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}
Aligned variable		x_{21}			x_{22}	x_{23}	
(b)							
Complete variable	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}
Aligned variable	NaN	x_{21}	NaN	NaN	x_{22}	x_{23}	NaN

to train predictive models at the risk of bias and poor statistics. On the other hand, imputing values in all the empty bins would create a significant amount of unrealistic information, which in itself introduces bias.

Since we did not know how much bias and subjective data manipulation we could introduce by choosing arbitrarily gridding periods, we used values ranging from 1 to 24 h in order to evaluate the impact in missing data classification and posterior predictive modeling.

Nearly all standard medical predictive risk methods require complete information in all variables presented to the models [6]. In other words, they assume a complete dataset, in a rectangular matrix shape (Table 1b) to perform decision-making inference [19]. In this study, in order to meet this requirement and create a rectangular matrix fully populated, all empty bins are filled with Not-a-Number (NaN) designations (as exemplified in Table 1b). Not-a-Number is commonly used in modeling simply as a placeholder for undefined values. It is important to note that not all NaNs in a rectangular matrix have the same derivation, that is, they may be related to each variable's sampling frequency, or to different underlying mechanisms of missing data.

3.2. Missing data classification

There is little published work addressing missing data as a classification problem. Instead of analyzing if data should be imputed or not, previous studies have accepted that it should, and then mainly focused on defining the correct imputation strategy according to the type of missing data [20–22]. Since missing data is ubiquitous, a correct strategy must be found in order to avoid loss of information [4,6] and potentially very serious bias due to systematic differences between observed and not observed data [23]. A generally accepted guideline for managing missing data is that when it represents more than 10% of the total information expected to be present, all records with missing values can be deleted without a significant loss of statistical power in the modeling results based on such dataset; if missing data is more than 10%, then an imputation strategy can be used since deleting would result in a significant loss of statistical power [10]. However, in some databases missing data can range up to 50% or more and in these circumstances, imputing the data that is genuinely absent (missing not at random, as will be described later) would seem inherently incorrect [18,24]; imputing values might create unrealistic states of the process being modeled.

Several approaches have been proposed to examine patterns of missing ness, including the creation of missing dummy variables that are treated as covariates in logistic regression models [10], computation of *t*-tests to compare respondents and non-respondents on an item [25,26], and incorporating missing data as independent dummy variables into a multiple regression models [27]. Again, although these methods deal with the classification of missing data according to their dependence on data, none of them identifies which data should then be actually recovered or left missing.

In this study we propose the use of a two-step classification of missing data during preprocessing:

- 1st step: classification of lack of sampling vs. true missing data.
- 2nd step: classification of recoverable and not-recoverable true missing data.

These steps will be detailed in the following subsections.

3.2.1. 1st step: classification of lack of sampling vs. true missing data

The sampling frequency of a medical variable may depend on several medical and technological issues [28]:

- how often clinicians need new results to evaluate patients' condition,
- how fast the variable measured changes its values,
- how easily the variable can be measured, considering a balance between benefit vs. cost/risk of measuring it,
- how practical is to measure a variable.

In intensive care, small changes in the values of variables could lead to significant changes in clinical status, which in turn could warrant more or less sampling ($f_{sampling}$). In fact the sampling frequency for most variables follows an approximated normal distribution centered on a mean sampling frequency with a reasonably small standard deviation [29]. In this study, and following the method proposed by [18], we use twice the upper limit of the 95% confidence interval ($95\%CI_{upperlimit}$) of the sampling period (the inverse of $f_{sampling}$) for a given variable as an acceptable threshold to separate the usual sampling process from a missingness event; we defined as *lack of sampling* those missing values resulting from sampling processes, and as *true missing data* those resulting from other events. That is, if the value for a variable is not present within twice the $95\%CI_{upperlimit}$ for its usual sampling frequency, then we define that data point as true missing data. For example, hematocrit may have a sampling period with a 3 ± 1 h 95% confidence interval, giving a $95\%CI_{upperlimit}$ of 4 h, and at twice the $95\%CI_{upperlimit}$ equaling a total time gap of 8 h. If the gridding time gap between any two hematocrit samples is less than 8 h, then the hourly nodes of the grid between the two hematocrit samples will be classified as lack of sampling, imputed, and not classified as true missing data. As already pointed out, the data is missing due to our choice of sampling frequency. On the other hand, if the time gap between two samples is greater than two times the $95\%CI_{upperlimit}$ (greater than 8 h in the hematocrit example) the nodes in between the two samples are classified as true missing data, and are then processed for further classification in terms of their recoverability. It should be noted that the choice of twice the $95\%CI_{upperlimit}$ is an arbitrary decision used in previous studies, and fundamentally seeking to adopt a very conservative approach to delayed data acquisition, e.g. a late sampling of hematocrit.

This statistical classification was done through the Statistics Toolbox® of the MATLAB® software. The code with the specific details can be requested to the author by email.

After this classification, missing segments that are classified as a lack of sampling are imputed by averaging the existing samples on each side of the data gap of the same variable, as in [13].

3.2.2. 2nd step: classification of recoverable and not-recoverable true missing data

Before applying any missing data treatment method, it is important to understand the underlying mechanisms of missingness, which may vary [30]. Missing medical data can be divided into two general groups [31]:

- The variable is not measured during a certain period of time because of an identifiable reason. For example, in an intensive care unit, a patient could be disconnected from the ventilator for several hours because of a medical decision, and variables registered by that machine would not be recorded during that time. It would be erroneous to replace and impute values during that period since they were not supposed to exist, i.e. it would be like saying the patient was under assisted ventilation when he was not, which can significantly bias posteriors predictions. We define this group as not-recoverable missing data.

- The variable is measured but, for some unidentifiable reason, the values are not recorded. Medical examples include accidental disconnection of sensors, errors in the communication with the server/storing facility, accidental human omission of data registration, electricity failures, and unlabeled samples that cannot be associated to any specific experiment. These variables are supposed to have values but they are not recorded in the database. We define this group as recoverable missing data.

In order to identify recoverable missing data using a data-driven technique, we propose to relate missing data observed at random (OAR) and observed not at random (ONAR) [6], presented in Section 2.2, with recoverable and not-recoverable missing data. This last relation is based on proving that variable(s) X, having information about Y_{miss}, imposed the decision of not measuring Y during the acquisition process. In other words, this means that X was used as a constraining rule to record Y, e.g. there was a reason not to record those values that can be explained with information in other variables. Previous analyses have shown the consistency of the relation between OAR/ONAR and recoverable/not-recoverable missing data [18,32].

3.2.3. *Fuzzy modeling for classification of true missing data*

Fuzzy modeling is a tool that allows an approximation of non-linear systems when there is little or no previous knowledge of the system to be modeled. A detailed description of fuzzy logic and modeling can be found in [33]. Briefly, fuzzy models use rules and logical connectives to establish relations between the features defined to derive the model. A fuzzy classifier contains a rule base consisting of a set of fuzzy if–then rules together with a fuzzy inference mechanism.

Since the relations between variable(s) X could have a non-linear nature, fuzzy systems were used in this work to binary classify recoverable/not-recoverable true missing data as follows: Y_{miss} (the true missing data in a given variable) is analyzed by using variable Y as the output and variable(s) X as input(s) in a fuzzy model; the input(s) go through a forward selection of features [34,35] to obtain the subset of X that better classifies Y_{miss}. To avoid magnitude effects in the classification process [36], variable(s) X were normalized as follows:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{10}$$

where X_{norm} is the normalized version of X, while X_{min} and X_{max} represent the minimum and maximum values of X, respectively. The minimum–maximum normalization method is commonly used in engineering applications to normalize the data due to its linear transforming form [36]. Additionally, Y was normalized by setting Y_{miss} to 1 and Y_{obs} to 0.

In this work, Takagi–Sugeno (TS) fuzzy models were used [37], which consist of fuzzy rules where each rule describes a local input–output relation. When TS fuzzy systems are used, each discriminant function consists of rules of the type

$$\begin{aligned} & \text{If } x_1 \text{ is } A^c_{i1} \text{ and } \ldots \text{ and } x_M \text{ is } A^c_{iM} \\ \textit{Rule}\, R^c_i: \ & \text{Then } d^c_i(X_{norm}) = f^c_i(X_{norm}), \\ & i = 1, \ldots, K \end{aligned} \tag{11}$$

where $x_1, \ldots, x_M$ are the values of each feature of the vector X_{norm}, and f_{ic} is the consequent function for rule R_{ic}. In these rules, the index c indicates that the rule is associated with the output class c. Therefore, the output of each discriminant function $d_c(X_{norm})$ can be interpreted as a score (or evidence) for the associated class c given the input feature vector. The degree of activation of the ith rule for class c is given by:

$$\beta_i = \prod_{j=1}^{M} \mu_{A^c_{ij}}(\mathbf{x}), \tag{12}$$

where $\mu_{A^c_{ij}}(\mathbf{x}) : \mathbb{R} \rightarrow [0, 1]$. The discriminant output for each class c, with $c = 1, \ldots, C$, is computed by aggregating the individual rules contribution:

$$d_c(\mathbf{x}) = \frac{\sum_{i=1}^{K} \beta_i f^c_i(\mathbf{X}_{norm})}{\sum_{i=1}^{K} \beta_i} \tag{13}$$

The classifier assigns the class label corresponding to the maximum value of the discriminant functions, i.e.

$$Y_{miss} = \max_{c} d_c(\mathbf{X}_{norm}) \tag{14}$$

When the fuzzy model classifies Y_{miss} as 1, those Y_{miss} points are considered not-recoverable. On the other hand, if Y_{miss} are classified as 0, those Y_{miss} points are considered recoverable, meaning they are found to have the same behavior than Y_{obs} [18], and hence, they can be imputed. Training, test and validation sets for the fuzzy models were defined by randomly selecting 30%, 40% and 30% of the dataset, respectively. The variables with predictive value for the missing data classification were selected during the training/testing process, while the performance of the classifier was evaluated over the unseen validation set through a 10-fold leave-one-out cross-validation (LOOCV) process.

For example, let us assume a dataset consisting of 3 variables X_1, X_2, and X_3. If we want to analyze the missingness in X_3, we create a dummy variable Y that has ones when X_3 is missing and zeros when X_3 is present. Then, we select a random training subset of instances of X_1, X_2 (collectively called X) and Y, and we train the fuzzy classifier in order for it to learn how to recognize both classes in Y. After the training, we test the model with the remaining subset of X, we obtain the output of the model, and we compare it to what Y is supposed to be in order to assess the performance of the model and choose the subset of X variables useful to classify the missingness in Y. When the model correctly classifies Y as missing is because it found enough information in other variables to explain the missingness, i.e. the missing value is not recoverable; when the model mistakes a missing value as an existing one is because there is not enough information in the other variables to explain the missingness, i.e. the missing value is recoverable.

Summarizing, a fuzzy system is the modeling algorithm we used to determine if a missing value has a relationship and is dependent on other variables or not. If the potential value for a missing variable is dependent on other variables, and those other variables are all present, then the values are truly missing, are not recoverable, and should not be imputed. If the potential value for a variable is independent of other variables, then the values are falsely missing, are recoverable, and should be imputed by averaging the existing samples on each side of the data gap of the same variable, as in [13].

In this paper, the fuzzy classification of missing data was performed using the Fuzzy Toolbox®, a component of the MATLAB® suite, using *Genfis3*. The code with the specific details can be requested to the author by email.

4. Databases description

4.1. *Creation of the test set*

In order to assess the performance of the methods that we propose, we follow the testing framework format proposed in [9]. This framework allows us to test misalignment, uneven sampling and missingness mechanisms (Algorithm 1). We create the test set as follows.

Algorithm 1. Creation of the test bed

```
Create a rectangular matrix $D$ with 16 columns and 5000 rows.
for $column = 1 \rightarrow$ no. of columns do
  Assign a random value to $row$=1, according to phisiological ranges of the variable that column represents.
  for $row = 2 \rightarrow$ no. of rows do
    Assign a value, following a normal distribution function centered on the last value added.
  end for
end for
Create a time vector $T$, with each value randomly and incrementally defined.
Simulate data removal processes:
for$column = 1 \rightarrow$ no. of columns do
    - Remove segments of data with random length from $column$, following a normal distribution, to mimick uneven sampling frequencies.
    - Remove random samples from $column$ to mimick MCAR mechanisms.
end for
for $column = 1 \rightarrow$ no. of columns do
    - $Y = column$ ;
    - $X = D_r$ ; where $D_r$ is a random subset of columns of $D$ without considering $column$
  for $X_i = 1 \rightarrow$ no. of columns in $X$ do
    Define random upper and lower limits for $X_i$; the randomness is constrained to reach the desired percentage of MAR.
    Remove samples of $Y$ for which $X_i$ values are in the range previously defined.
  end for
end for
for$column = 1 \rightarrow$ no. of columns do
    - $Y = column$ ;
    - Define random upper and lower limits for $Y$; the randomness is constrained to reach the desired percentage of MNAR.
    - Remove samples of $Y$ with values in the range previously defined.
end for
Simulate misalignment:
for$column = 1 \rightarrow$ no. of columns
    - Create a new time vector $T_{column}$ with values shifted according to normal distributions centered in the original sampling times of $T$.
end for
```

First, a rectangular matrix is generated, composed of 16 variables with 5000 samples each. In order to emulate the time series representation of a real database (see Section 4.2), random values were assigned to each variable, following a normal distribution function centered on the last value added. Simultaneously, a vector was created, indicating the time when each sample was acquired. Each value of the vector was randomly and incrementally defined. The previously defined rectangular matrix is subjected to a 3 step data removal process emulating a real dataset acquisition process and missing data mechanisms:

1 Uneven sampling frequencies of data acquisition were mimicked by removing segments of data with random length, following a normal distribution.

2 MCAR, MAR and MNAR were individually simulated in equal amounts, representing up to 10%, 30% and 50% of the original data present in the dataset. MCAR was introduced by randomly removing samples within the rectangular matrix. MAR was introduced by removing samples from each variable using the following procedure:
   - one of the 16 variables was considered as Y (the variable in which the missingness will be created),
   - a random number of the remaining 15 was defined as X (variables with information about Y_{miss}),
   - a range of values for each variable of X was defined, with random upper and lower limits; the randomness was constrained to reach the desired percentage of missingness,
   - samples in Y, for which variables X had values in the ranges previously defined, were removed.

   MNAR was introduced in each variable by two different means: dependent only on values of Y, and dependent on values of both Y and X. These two types of MNAR were created in such a way that each of them accounted for half of the total MNAR introduced in the dataset.

3 Finally, misalignment was simulated by shifting sampling times. In other words, for each variable, a new time vector was created with values shifted according to normal distributions centered in the original sampling times.

At the end of this process we obtain a framework for constructing 16 misaligned unevenly sampled time series containing all three types of missing data (MCAR, MAR and MNAR).

Each of the steps previously defined to create a test bed is schematically shown in Fig. 6. Special attention should be given to the simulation of MNAR and MAR. The dashed bins in V2 mean those values determine the absence in V1, i.e. V1 is missing only because of V2 according to the MAR definition. The dashed bins in V1 determine the absence in V3, but since values of both V1 and V3 are in between certain thresholds, V3 is missing because of values of V1 and V3, according to the definition of one of the possible MNAR.

4.2. MEDAN database

In order to evaluate the effect of classifying missing data on predictive risk modeling, a real world, publicly available ICU database named MEDAN was used [29,38]. The MEDAN database contains personal records, physiological parameters, procedures, diagnoses/therapies, medications, and respective outcomes (survived or deceased) for a set of 139 septic shock patients. In [13], the survival of septic shock patients was predicted by using the values of 16 input variables.

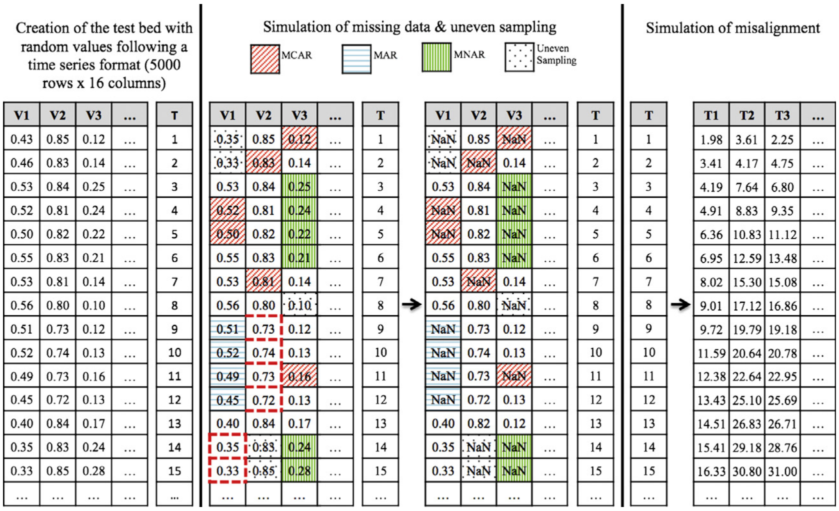
The outcome to predict, deceased or survived, was the binary result at the end of the ICU admission of each patient. All the instances of 16 input variables were fed into a neuro-fuzzy model, and the outcome was predicted for each of those instances and for each patient. In this work, we use the same neuro-fuzzy classifier than [13] in order to compare results and specifically evaluate the impact of missing data classification in the prediction performance.

4.3. Raw data issues in the test set

We undertook the gridding process with four different gridding periods as proposed in [13] (m = 1, 3, 12 and 24 h). On the other hand, we used heart rate as the highest sampling frequency variable for templating. We also controlled the amount of missing data in order to emulate the quantities existing in MEDAN.

Table 2 shows the percentages of missing values for each variable after templating and gridding periods. From Table 2 it can be seen that the total percentage of missing data is greater than 10% of the total dataset for templating and for gridding periods of 1, 3 and 12 h. As previously discussed, generally accepted missing data approaches would accept the test set as crossing the 10% threshold and being worthy of imputation. For our purposes however we will first apply our missing data correction and then recoverability process before making imputation decisions.

Contrary to most retrospective studies based on real world databases, by using the test dataset described here, it is possible to know the underlying mechanisms of misalignment, uneven sampling, and type and location of missing data present within the data. Thus, the accuracy of the proposed approaches when dealing with those topics can be assessed.



Creation of the test bed with random values following a time series format (5000 rows x 16 columns)
Simulation of missing data & uneven sampling
MCAR
MAR
MNAR
Uneven Sampling
Simulation of misalignment

V1	V2	V3	…	T
0.43	0.85	0.12	…	1
0.46	0.83	0.14	…	2
0.53	0.84	0.25	…	3
0.52	0.81	0.24	…	4
0.50	0.82	0.22	…	5
0.55	0.83	0.21	…	6
0.53	0.81	0.14	…	7
0.56	0.80	0.10	…	8
0.51	0.73	0.12	…	9
0.52	0.74	0.13	…	10
0.49	0.73	0.16	…	11
0.45	0.72	0.13	…	12
0.40	0.84	0.17	…	13
0.35	0.83	0.24	…	14
0.33	0.85	0.28	…	15
…	…	…	…	…

V1	V2	V3	…	T
0.35	0.85	0.12	…	1
0.33	0.83	0.14	…	2
0.53	0.84	0.25	…	3
0.52	0.81	0.24	…	4
0.50	0.82	0.22	…	5
0.55	0.83	0.21	…	6
0.53	0.81	0.14	…	7
0.56	0.80	0.10	…	8
0.51	0.73	0.12	…	9
0.52	0.74	0.13	…	10
0.49	0.73	0.16	…	11
0.45	0.72	0.13	…	12
0.40	0.84	0.17	…	13
0.35	0.83	0.24	…	14
0.33	0.85	0.28	…	15
…	…	…	…	…

→

V1	V2	V3	…	T
NaN	0.85	NaN	…	1
NaN	NaN	0.14	…	2
0.53	0.84	NaN	…	3
NaN	0.81	NaN	…	4
NaN	0.82	NaN	…	5
0.55	0.83	NaN	…	6
0.53	NaN	0.14	…	7
0.56	0.80	NaN	…	8
NaN	0.73	0.12	…	9
NaN	0.74	0.13	…	10
NaN	0.73	NaN	…	11
NaN	0.72	0.13	…	12
0.40	0.82	0.12	…	13
0.35	NaN	NaN	…	14
0.33	NaN	NaN	…	15
…	…	…	…	…

→

T	T1	T2	T3	…
1	1.98	3.61	2.25	…
2	3.41	4.17	4.75	…
3	4.19	7.64	6.80	…
4	4.91	8.83	9.35	…
5	6.36	10.83	11.12	…
6	6.95	12.59	13.48	…
7	8.02	15.30	15.08	…
8	9.01	17.12	16.86	…
9	9.72	19.79	19.18	…
10	11.59	20.64	20.78	…
11	12.38	22.64	22.95	…
12	13.43	25.10	25.69	…
13	14.51	26.83	26.71	…
14	15.41	29.18	28.76	…
15	16.33	30.80	31.00	…
…	…	…	…	…

Fig. 6. Schematic representation of the creation of the test bed (NaN: missing value). *Note*: all tables were populated with arbitrary values and units.

Table 2
Total amounts of missing data and individual contributions of each variable for the templating, and each gridding period *m* used in the test bed.

Variable name	Templating	% of missing data for each gridding period			
		$m = 1$ h	$m = 3$ h	$m = 12$ h	$m = 24$ h
V1/heart rate	0	18.1	6.5	3.4	2.7
V2/systolic blood pressure	1.2	19.2	6.1	2.7	1.8
V3/diastolic blood pressure	1.1	19.4	6.1	2.7	1.7
V4/temperature	34.3	55.0	25.5	6.2	4.2
V5/central venous pressure	63.8	82.1	53.6	18.6	14.0
V6/partial saturation of O2	31.9	54.8	32.3	16.5	11.6
V7/white blood count	79.6	93.7	81.5	38.7	9.8
V8/hemoglobin	65.1	88.8	71.0	31.5	8.6
V9/hematocrit	76.3	92.1	77.8	35.7	9.5
V10/platelet count	77.9	93.8	81.7	39.4	11.2
V11/pro-thrombin time	80.1	94.6	84.0	46.7	19.2
V12/serum sodium	75.2	91.0	75.9	35.6	12.3
V13/serum potassium	69.9	89.6	72.4	32.4	11.4
V14/creatinine	80.8	95.4	86.3	49.6	16.7
V15/blood sugar	61.7	83.9	57.4	21.9	9.3
V16/urine volume	13.9	30.9	15.4	7.3	5.9
Average percentage of missing data	50.8	68.9	52.1	24.3	9.4
Total amount of records	19,695	63,330	21,070	5221	2577

5. Results

5.1. Alignment of misaligned unevenly sampled data

Table 2 demonstrates that, depending on the method used to align the 16 variables, the amount of missing data in a rectangular matrix can vary significantly. If reducing the amount of missing data is a goal, then a sampling frequency of $m = 24$ h would be the right choice for aligning variables. However, the last row in Table 2 shows the high cost of achieving this goal, namely the amount of records that can be used for modeling is low, and indeed significantly lower compared to templating and to lower values of *m*. This approach would affect statistical power of predictive models and the frequency of predictions that can be made (in this case, one prediction every 24 h).

If templating or lower values of *m* are used, the amount of available records for modeling grows significantly, but so does the amount of missing data. It should also be noted that by using a templating process, the variable used as the template has zero missing data, which means that all the original information for that variable can be used for modeling. Table 2 also shows that, for templating columns V2 and V3, the closer the sampling frequency between a given variable and the template, the lower the amount of missing data that variable will have. To summarize the alignment processes, we would note that the alignment method that best suits the modeling purposes should be based on the frequency of predictions needed, based on the outcome to be predicted out of that dataset, and a balance between bias, loss of statistical power and complexity of data imputation.

5.2. Classification of lack of sampling vs. true missing data

The results of the statistical classifier differentiating lack of sampling from true missing data is presented in Table 3. Across

Table 3
Classification of lack of sampling vs. true missing data. Sensitivity: true missing data hit rate; specificity: lack of sampling hit rate.

	Templating	Gridding periods				
		$m = 1$ h	$m = 3$ h	$m = 8$ h 33 min	$m = 12$ h	$m = 24$ h
% of true missing data	22.12	27.89	24.40	20.21	16.25	9.36
ACC (%)	77.6	75.4	77.1	78.2	78.4	78.4
Sensitivity	73.4	66.2	70.6	75.7	77.4	77.4
Specificity	96.8	95.5	96.6	96.9	97.1	97.1

Table 4
Accuracy of classification for recoverable/not-recoverable true missing data using fuzzy models. ACC_{NR}: accuracy of classification of not-recoverable missing data. ACC_R: accuracy of classification of recoverable missing data; ACC_{Total}: accuracy of classification of total true missing data.

%	Templating	Gridding periods				
		$m = 1$ h	$m = 3$ h	$m = 8$ h 33 min	$m = 12$ h	$m = 24$ h
ACC_{NR}	76.5	68.5	74.3	78.6	80.5	83.2
ACC_R	47.1	36.8	39.4	53.1	56.2	62.3
ACC_{Total}	63.9	55.3	60.4	65.85	74.7	76.6

the whole table, ACC ranges from 75.4% to 78.4%, which is clearly better than a majority or random classifier ($ACC_{majorityclassifier} = 50\%$). The greatest accuracy (78.4%) not surprisingly occurs with the least amount of missing data (9.36%), and this value decreases as missing data goes up/gridding periods go down (p-value ≤ 0.05 in all cases). As a general rule of thumb [39], ACC $\leq$ 75% is considered marginal from an accuracy perspective, and so one notes that as the percentage of missing data approaches 30%, so the ACC approaches this 75% threshold. The practical implication for this technique is that when more than 30% of the information is missing, the assumptions underlying the statistical classifying approach may fail in distinguishing lack of sampling from true missing data.

A further observation is that when the gridding period m is greater than the average sampling frequency of all variables, which for this MEDAN database was 8 h 33 min, we see that ACC did not improve beyond 78.4%. This implies that a good balance can be reached by choosing the gridding period to be equal to the average sampling frequency. This information is new knowledge and we believe will substantively inform study designs of this type.

In the final stage following classification, missing segments that are classified as sampling related are imputed by averaging the existing samples on each side of the data gap of the same variable, as in [13].

5.3. Classification of recoverable and not-recoverable true missing data

As previously detailed, our aim is to determine if the true missing data are recoverable or not-recoverable and for this purpose we use fuzzy models. Table 4 shows the performance of our fuzzy models from an accuracy perspective, as they were applied to the simulated dataset. This table shows that: (1) the total accuracy (ACC_{Total}) is greater in each instance compared to a default majority classifier (50%) and (2) ACC_{Total} significantly decreases (p-value ≤ 0.05 in all cases) as the gridding period becomes shorter and, consequently, missing data increases. Table 4 also shows for gridding with $m = 8$ h 33 min,that reasonably good amounts of missing data, both recoverable and not-recoverable are correctly classified by the fuzzy models. This last finding, together with the optimal gridding period discussed in the previous subsection, would suggest using the average of the $95\%CI_{upperlimit}$ for all variables as the gridding period, or to the use of templating for misalignment correction.

Further analysis of the fuzzy models is shown in Table 5 which details each classification result as: correctly classified as recoverable (RCC), incorrectly classified as recoverable (RIC), correctly classified as not-recoverable (NCC), and incorrectly classified as not-recoverable (NIC).

By observing the discriminated classification results it can be seen that: (1) the sum of RCC and NCC (the correct classifications) is higher than the sum of RIC and NIC (the incorrect classifications) for all values of m; (2) RCC is lower than NCC for all values of m. The implications from these findings are that these models better classify not-recoverable than recoverable true missing data, and that true missing data were more often classified as not-recoverable despite the test bed being created with equal amounts of recoverable and not-recoverable values. In other words, these models tend to better identify those missing values that are not supposed to be recovered which in itself protects the posterior imputation algorithm from the mistake of creating false information in the dataset.

Table 5
Discriminated results for the classification of true missing data using fuzzy models. NCC: not-recoverable correctly classified; NIC: not-recoverable incorrectly classified; RCC: recoverable correctly classified; RIC: recoverable incorrectly classified.

%	Templating	Gridding periods			
		$m = 1$ h	$m = 3$ h	$m = 12$ h	$m = 24$ h
NCC	38.60	34.25	37.15	40.25	41.6
NIC	25.48	31.6	30.3	21.9	18.85
Missing data classified as not-recoverable	64.08	65.85	67.45	62.15	60.45
RCC	25.43	18.4	19.7	28.1	31.15
RIC	10.49	15.75	12.85	9.75	8.4
Missing data classified as recoverable	35.92	34.15	32.55	37.85	39.55
Sum of NCC and RCC	64.03	52.65	56.85	68.35	72.75
Sum of NIC and RIC	35.97	47.35	43.15	31.65	27.25

After the classification of true missing data, missing segments that are classified as recoverable are imputed by averaging the existing samples on each side of the data gap of the same variable, as in [13].

5.4. Application to patients' survival prediction using MEDAN database

In order to evaluate the impact of the missing data classification method proposed in this paper, we used this process to recast previously published work exploring patient mortality from septic shock in a real world database (MEDAN) [13].

Table 6 shows the comparative effects of listwise deletion, not classifying missing data at all [13], using only the 1st step proposed in this work, and using the 1st and the 2nd steps of missing data classification together.

As shown in Table 2, the *average percentage of missing data* ranges from 69% to 9% of the total information that should be present in a complete dataset. Since listwise deletion proceeds to delete all records with at least one missing value, the total amount of records available for modeling after the deletion is very low (approximately 0.5% for $m = 1$ h), which considerably affects the statistical power of the results shown in Table 6. However, it is possible to see that the all performance measures improve for this method when m is increased, showing that the less records are deleted, the best the prediction results obtained.

From Table 6 it can be seen that, applying only the 1st classification step, there are improvements in AUC, sensitivity, specificity and ACC when compared to [13]. It is also noted that any improvements in AUC are maintained over increasing gridding periods. Templating was not considered at this point as [13] did not use that technique.

Table 6 also shows further improvements in AUC when both the 1st and 2nd missing data classification steps are used together. Using this combination, true missing data classified as recoverable are imputed as per the method proposed in [13]. This imputation involves imputing random noise that follows a normal distribution around existing values of the variable being recovered [40].

Table 6
Comparison of patient survival prediction between Paetz et al. published work, and statistical analysis, and combined statistical and fuzzy logic analysis corrections for missing data.

Sampling	Method	AUC	Sensitivity	Specificity	ACC (%)	Paetz cf (*p*-values)
$m = 1$ h	Listwise deletion	0.55	0.60	0.53	55.3	
	Paetz [13]	0.57	0.60	0.55	56.7	
	1st step	0.59	0.62	0.56	59.7	<0.01
	1st and 2nd step	0.65	0.70	0.59	65.4	<0.01
$m = 3$ h	Listwise deletion	0.59	0.62	0.57	59.2	
	Paetz [13]	0.62	0.64	0.58	62.1	
	1st step	0.64	0.67	0.61	65.6	0.02
	1st and 2nd step	0.71	0.75	0.65	73.2	0.01
$m = 8$ h 33 min	Listwise deletion	0.62	0.64	0.59	62.3	
	Paetz [13]	0.69	0.73	0.65	67.7	
	1st step	0.73	0.77	0.69	71.1	<0.01
	1st and 2nd step	0.82	0.86	0.75	78.3	0.03
$m = 12$ h	Listwise deletion	0.64	0.67	0.62	65.1	
	Paetz [13]	0.68	0.71	0.63	65.8	
	1st step	0.73	0.77	0.69	71.1	0.01
	1st and 2nd step	0.78	0.82	0.72	76.1	<0.01
$m = 24$ h	Listwise deletion	0.66	0.69	0.63	65.1	
	Paetz [13]	0.66	0.68	0.61	64.3	
	1st step	0.73	0.77	0.69	71.1	0.02
	1st and 2nd step	0.76	0.79	0.70	75.6	<0.01

Those records with missing values classified as not-recoverable were deleted from the dataset.

Table 6 also demonstrates that a gridding period of 8 h 33 min gives the best classification results. This supports the previously suggested findings of the average of the various sampling periods for all variables, as the AUC maximizing gridding period.

This work extends the results presented in [18] by: (1) showing the impact of missing data classification on posterior predictive modeling, (2) introducing an application on a real database, and (3) explaining in detail the methods used both for missing data classification and for the creation of the test bed.

5.5. Limitations

The method proposed in this paper assumes that when the missingness in one variable can be explained by others, those missing values should not be recovered/imputed. This is not always true, and this might lead to biased predictive results. More experimentation needs to be carried out, and a deeper understanding of the missing data mechanisms for each specific data subset is crucial in order to avoid misleading results.

In this paper, after the classification and imputation of recoverable missing data, those instances that still contain not-recoverable missing values are deleted following a listwise deletion process. Since this becomes a selective deletion (after a classification), this method could introduce significant bias depending on the amount of not-recoverable missing data. To avoid this, the method proposed in this paper should be used combined with a modeling approach that handles missing data in its inputs, so no selective deletion would take place.

6. Conclusions

In this paper we present a new approach for managing missing data that improves overall modeling performance. We describe gridding and templating as alignment techniques, and we then present an argument for statistically classifying and managing missing data into sampling related or true missing data. In the next step, true missing data is further classified and managed using fuzzy models into recoverable or not-recoverable. We demonstrate improved modeling performance in a simulated test bed, and then confirm improved performance replicating previously published work. We believe more detailed management of missing data is a useful and important step for improving predictive risk modeling, and we present here combined statistical and fuzzy modeling techniques to support that goal.

Acknowledgments

The authors would like to acknowledge the help and space provided by the Division of Clinical Informatics of the Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology. Both human and technical resources were available through them, and were critical for the development of this work. This work is supported by the Portuguese Government under the programs: project PTDC/SEM-ENR/100063/2008, Fundação para a Ciência e Tecnologia (FCT), and by the MIT-Portugal Program and FCT grants SFRH/BPD/65215/2009, SFRH/43043/2008 and SFRH/43081/2008, Ministério do Ensino Superior, da Ciência e da Tecnologia, Portugal.

References

[1] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isnt. British Medical Journal 1996;312:71–2.

[2] Brilli RJ, Spevetz A, Branson RD, Campbell GM, Cohen H, Dasta JF, et al. Critical care delivery in the intensive care unit: defining clinical roles and the best practice model. Critical Care Medicine 2001;29(10):2007–19.

[3] Aigner W, Miksch S, Müller W, Schumann H, Tominski C. Visual methods for analyzing time-oriented data. IEEE Transactions on Visualization and Computer Graphics 2008;14(1):47–60.

[4] Heitjan DF. Annotation: what can be done about missing data? American Journal of Public Health 1997;87(4):548–50.

[5] Erdogan E, Ma S, Beygelzimer A, Rish I. Statistical models for unequally spaced time series. SIAM 2004.

[6] Allison PD. Missing data. Sage University papers series on quantitative applications in the social sciences. Thousand Oaks, CA: SAGE Publications; 2001.

[7] Mehari SM, Havill JH. Written guidelines for laboratory testing in intensive care-still effective after 3 years***. Critical Care Resucitation 2001;3:158–62.

[8] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. British Journal of Medicine 2009:2393b.

[9] Gorelick MH. Bias arising from missing data in predictive model. Journal of Clinical Epidemiology 2006;59(10):1115–23.

[10] Little RJA, Rubin DB. Statistical analysis with missing data. Wiley series in probability and mathematical statistics. Probability and mathematical statistics Hoboken, New Jersey: Wiley; 2002.

[11] Cios KJ, Kurgan LA. Trends in data mining and knowledge discovery. In: Pal NR, Jain LC, Teoderesku N, editors. Knowledge discovery in advanced information systems. Heidelberg: Springer; 2005. p. 200–2.

[12] Pyle D. Data preparation for data mining. San Francisco, CA: Morgan Kaufmann; 1999.

[13] Paetz J. Knowledge-based approach to septic shock patient data using a neural network with trapezoidal activation functions. Artificial Intelligence in Medicine 2003;28(6):207–30.

[14] Kalyadin NI, Lemenkov VA, Losev IR, Kantor SI. Problems of medical monitoring of patients and the requirements for development of computer monitoring systems. Biomedical Engineering 1996;30(2):81–5.

[15] Hamilton JD. Time series analysis. NJ, USA: Princenton University Press; 1994.

[16] Xia H. Bayesian hierarchical model for combining two-resolution metrology data. Ph.D. dissertation, Texas A&M University; 2008.

[17] Honaker J, King G. What to do about missing values in time-series cross-section data. American Journal of Political Science 2010;54(2):561–81.

[18] Cismondi F, Fialho AS, Vieira SM, Sousa JMC, Reti SR, Howell MD, et al. Computational intelligence methods for processing misaligned, unevenly sampled time series containing missing data. In: CIDM. 2011. p. 224–31.

[19] Shih WJ. Problems in dealing with missing data and informative censoring in clinical trials. Current Controlled Trials in Cardiovascular Medicine 2002;3(1):4.

[20] Wayman JC. Multiple imputation for missing data: what is it and how can I use it? Annual meeting of the American Education Research Association. Chicago: AERA; 2003. p. 1–16.

[21] 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. Artificial Intelligence in Medicine 2010;50(2):105–15.

[22] Liew AW, Law NF, Yan H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics 2010;12(5):498–513.

[23] Barnard J, Meng XL. Applications of multiple imputation in medical studies: from *AIDS* to *NHANES*. Statistical Methods in Medical Research 1999;8(1):17–36.

[24] Ramoni M, Sebastiani P, Dybowski R. Robust outcome prediction for intensive-care patients. Methods of Information in Medicine 2001;40(1): 39–45.

[25] Barnard J, Meng XL. Working with missing data. Family Science Review 1997;1(10):76–102.

[26] Huisman M. Missing data in behavioural science research: investigation of a collection of data sets. Kuantitieve Methoden 1998;87(4):548–50.

[27] Orme RG, Reis J. Multiple regression with missing data. Journal of Social Service Research 1991;15:61–91.

[28] Clifford GD, Scott DJ, Villarroel M. User guide and documentation for the mimic ii database, version 2.1; 2009.

[29] Hanisch E, Brause R, Arlt B, Paetz J, Holzer K. The MEDAN database. Computer Methods and Programs in Biomedicine 2003;75(1):23–30.

[30] Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychological Methods 2002;7:147–77.

[31] Guobing L, Copas JB. Missing at random, likelihood ignorability and model completeness. Annals of Statistics 2004;32:754–65.

[32] Vach W, Blettner M. Missing values in epidemiological studies; 1997.

[33] Sousa JMC, Kaymak U. Fuzzy decision making in modeling and control. Singapore: World Scientific Pub. Co.; 2002.

[34] Liu H, Hiroshi M. Feature selection for knowledge discovery and data mining. Norwell, MA, USA: Kluwer Academic Publishers; 1998.

[35] Mendonça LF, Vieira SM, Sousa JMC. Decision tree search methods in fuzzy modeling and classification. International Journal of Approximate Reasoning 2007;44:106–23.

[36] Milligan GW, Cooper MC. A study of standardization of variables in cluster analysis. Journal of Classification 1988;5:181–204.

[37] Takagi T, Sugeno M. Fuzzy identification of systems and its applications to modelling and control. IEEE Transactions on Systems, Man and Cybernetics 1985;15:116–32.

[38] Paetz J, Arlt B, Erz K, Holzer K, Brause R, Hanisch E. Data quality aspects of a database for abdominal septic shock patients. Computer Methods and Programs in Biomedicine 2004;75:23–30.

[39] Lemeshow S, Gall JRL. Modelling the severity of illness of ICU patients: a systems update. Journal of American Medical Association 1994;272: 1049–55.

[40] Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. Analytical methods for social research. Cambridge:Cambridge University Press; 2007.