https://doi.org/10.52326/ic-ecco.2022/BME.02

Dealing With Missing Continuous Biomedical Data: a Data Recovery Method for Machine Learning Purposes

Victor Iapascurta [1,2], ORCID: 0000-0002-4540-7045
[1] N. Testemitanu State University of Medicine and Pharmacy, 165, Stefan cel Mare si Sfant Blvd., Chisinau, MD – 2004, Republic of Moldova
[2] Technical University of Moldova, 168, Stefan cel Mare si Sfant Blvd., Chisinau, MD – 2004, Republic of Moldova, victor.iapascurta@doctorat.utm.md

Abstract—**There are different approaches to dealing with missing data. A common one is by deleting observations containing such data, but it is not applicable when the volume of the data is limited. In this case, a number of methods can be applied, such as Last Observation Carried Forward and the like. But these methods are not suitable when all data for a certain parameter are missing. This paper describes a possibility of addressing this issue in the case of time series of biomedical data. Behind the method is the idea of the human body as a complex system in which various parameters are correlated and missing data can be inferred from the available data using the estimated correlation. For this, machine learning-based linear regression models are built and used to recover data describing the sepsis state. Finally, recovered data are used to create a sepsis prediction system.**

Keywords—biomedical data; missing data; data recovery; sepsis; machine learning

I. INTRODUCTION

Missing data are often unavoidable in research, but their potential to influence the research results is rarely discussed. In this respect it is considered of importance the nature of "missingness": at random versus not at random.

Unfortunately, quite often it is not possible to distinguish between missing at random and missing not at random using observed data. Since the data set used in this study is not accompanied by an explanation concerning the nature of missingness, it is assumed that they are missing at random (e.g. during the time the patient is undergoing a medical procedure that requires the sensors used for data collection to be removed, equipment malfunction, etc).

Although data recovery may be possible independent of the nature of missingness, the bias introduced by the recovery procedure is considered to be lower in the case of data missing at random [1].

There are many methods for dealing with missing values in data. Most of these methods are appropriate for static data and the set of tools suitable for continuous data, or time series is of a smaller size. The simplest approach is to delete observations with missing values, known as complete case analysis (CCA) [2], but in many cases, this can be hardly applicable, especially when the volume of available data is limited. This issue is of particular interest when using the data for machine learning purposes, especially when dealing with unbalanced sets, where every observation in the minority class is important.

Methods for data recovery can be conventionally divided in:

a. Single imputation methods - replace a missing data point with a single value, which is usually coming from the observed values from the same subject (Last Observation Carried Forward (LOCF), Baseline Observation Carried Forward (BOCF), and Next Observation Carried Backward (NOCB) or from other sources (e.g., mean value imputation, regression imputation, etc);
b. Multiple imputation methods - by creating several different plausible imputed data sets and appropriately combining results obtained from each of them. There are a number of statistical packages for this purpose (e.g., MICE, Amelia, HMISC in R, etc.)

According to the Guideline on Missing Data in Confirmatory Clinical Trials [3], “if missing values are handled by simply excluding any patients with missing values from the analysis, this will result in a reduction in the number of cases available for analysis and therefore normally result in a reduction of the statistical power. Clearly, the greater the number of missing values, the greater the likely reduction in power. Hence every effort should be made to minimize the amount of missing data [and select an appropriate imputation method].

Unfortunately, there is no methodological approach for handling missing values that is universally accepted in all situations".

As such, how to minimize the amount of missing data and how missing data are going to be handled in the analysis are critical issues that must be considered when planning a trial.

The current work presents a method of data recovery that consists of several steps, including regression imputation, which imputes the predictions from a regression of the missing variables on the observed variables.

One of the behind-the-scene concepts for the current method is the idea of approaching the human body as a complex system in which the various parameters that describe its functioning (in health or disease) are correlated and missing data can be derived from available data using estimated correlation.

Recovered data are finally being used to build a machine learning system, which is part of a larger research with the goal of creating a machine learning-based software application for sepsis prediction.

II. Research Data and Processing Methods

A. *Data*

The data used in this research are from a public database made available by "Early Prediction of Sepsis from Clinical Data: the PhysioNet/Computing in Cardiology Challenge 2019" [4]. The public part of the data comes from two distinct hospitals in the US: Beth Israel Deaconess Medical Center (set A) and Emory University Hospital (set B). These data were collected over the past decade with approval from the appropriate Institutional Review Boards, de-identified, and labeled using Sepsis-3 clinical criteria [5]. They consist o a combination of hourly vital sign summaries, lab values, and static patient descriptions for 40,336 patients, including 8 vital sign variables, 26 laboratory variables, and 6 demographic variables. All patient features were condensed into hourly bins (e.g., multiple heart rate measurements in an hourly time window were summarized as the median heart rate measurement).

The data contain more than 80% of missing values (e.g., set B). Since set A contains fewer missing values (i.e., 79,4%) and the prevalence of sepsis is higher (i.e., 8,80% vs 5,71% in set B) this set is used for further research.

There are 20336 patients/subsets in the set, including 1790 septic subsets, of which 502 subsets contain all missing values for at least one parameter (out of 6 parameters of interest). After applying initial selection criteria (e.g., the presence of at least 7 hourly observations before sepsis is diagnosed, absence of artifacts, etc.) the number of subsets that contain missing values but can potentially be recovered is 211. These subsets are the focus of the current research.

Table 1 shows the appearance of an original sepsis file with all-missing values (NA) for one parameter (i.e. Temperature). It describes observations on seven parameters of interest selected for further research (i.e. heart rate (HR), arterial blood oxygen saturation (O_2Sat), temperature (Temp), systolic blood pressure (SBP), diastolic blood pressure (DBP), respiratory rate (Resp), the age and labeling (0 – for non-sepsis cases and 1 – for sepsis).

TABLE I. Original appearance of a sepsis file

HR	SaO_2	Temp	SBP	DBP	Resp	Age	Sepsis label
83	100	**NA**	129	50	17	77.3	0
80	99	**NA**	89	41	18	77.3	0
79.5	100	**NA**	143	52.5	19	77.3	0
85	100	**NA**	161	56	18	77.3	0
69	95	**NA**	91	43	15	77.3	0
66	98	**NA**	116	40	20	77.3	0
68	99	**NA**	148	50	17	77.3	0
73	97	**NA**	117	44	14	77.3	1

B. *Methods*

The algorithm used here for data recovery purposes is a Generalized Linear Model (GLM) [6] provided by the H2O platform (www.h2o.ai) and is described below.

Gaussian approach (behind GLM) models the dependency between a response y and a covariates vector x as a linear function:

$$y = x^T\beta + \beta_0 + \epsilon, \quad (1)$$

where, β is the parameter vector, β_0 represents the intercept term and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a gaussian random variable which is the noise in the model.

The estimation of the model is obtained by maximizing the log-likelihood over the parameter vector β for the observed data. The GLM [6] used in this research fits the model by solving the following likelihood optimization with parameter regularization:

$$\max_{\beta,\beta_0} (GLM\ Log \cdot likelihood - \ Regularization\ Penalty). \quad (2)$$

The regularization penalty is the weighted sum of the ℓ_1 (least absolute shrinkage parameter) and ℓ_2 (ridge regression) norms of the coefficients vector and is defined as:

$$\lambda P_\alpha(\beta) = \lambda\left(\alpha\|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_2^2\right) \quad (3)$$

with no penalty for the intercept term, where α is the elastic net parameter, $\alpha \in [0,1]$ and λ is a tuning parameter.

The optimization over *N* observations is performed as follows:

$$\max_{\beta\beta_0} \sum_{i=1}^{N} log\, f\,(y_i;\, \beta,\beta_0) - \lambda\left(\alpha\|\beta\|_1 + \frac{1}{2}(1-\alpha)\|\beta\|_2^2\right). (4)$$

At the final machine learning (ML) stage, the Gradient Boosting Machine (GBM) provided by the same H2O platform [7] proved to be the best-performing algorithm.

The programming language used for current research is R [8], including a number of packages coming from the same environment and used for various tasks throughout the research. The same language/environment is used for interacting with the H2O ML platform and plotting.

III. Data Processing and Results

Data processing flow in this research consists of a number of steps, including missing value recovery and aims to generate datasets suitable for machine learning purposes. The following is a description of the main processing steps.

A. Preprocessing stage

Initially, there are 502 sepsis files with all-missing values for at least one parameter of interest (i.e. heart rate, arterial blood oxygen saturation, temperature, systolic and diastolic blood pressure respiratory rate) and 211 files where hourly missing values can be reconstructed. In order to create a more or less balanced dataset to be subsequently used for ML there was randomly sampled a commensurate number of non-sepsis subsets (i.e. 349 files/cases, or to be not larger than 40% compared to the sepsis subset).

B. Sepsis Data Reconstruction

The first step for sepsis data reconstruction includes applying LOCF (Last Observation Carried Forward, by "DescTools" package, R). This will recover missing values in columns in which some of the values are missing, but will not work for columns with all missing values.

In order to address the all-missing values cases there was examined the correlation between the 6 parameters of interest for the sepsis cases with no missing values described previously (Fig. 1). The plot includes an additional parameter – the age, which shows a moderate correlation with some parameters of interest and also has no missing values.

Based on the correlation coefficients there were selected 3 most correlated parameters for each of the 6 parameters (e.g., for temperature the most correlated parameters are HR, SBP, and Age; for respiration the most correlated are HR, O_2Sat, and temperature, etc.).

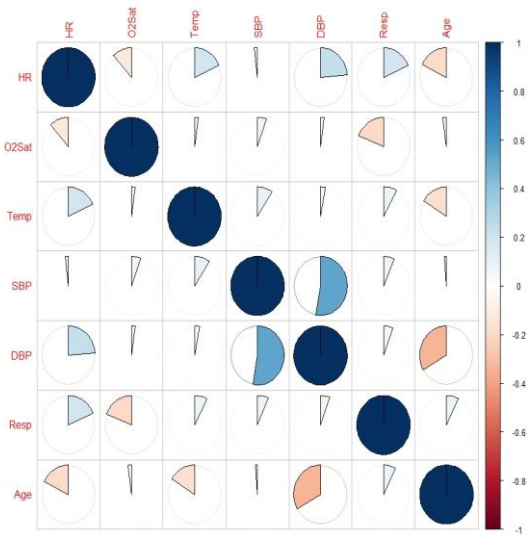

Figure 1. Correlation plot

Table 2 presents the correlation matrix for 7 parameters. Bold font denotes the highest correlation coefficients.

TABLE II. Correlation Matrix

	HR	O_2Sat	Temp	SBP	DBP	Resp	Age
HR	1.00	-.12	.18	-.02	**.24**	**.18**	**-.18**
O_2Sat	**-.12**	1.00	.02	**.06**	.02	**-.19**	-.02
Temp	**.18**	.02	1.00	**.09**	.03	.08	**-.16**
SBP	-.02	.06	**.09**	1.00	**.53**	**.07**	-.01
DBP	**.24**	.02	.03	**.53**	1.00	.06	**-.33**
Resp	**.18**	**-.19**	**.08**	.07	.06	1.00	.07
Age	-.18	-.02	-.16	-.01	-.33	.07	1.00

Using this correlation information a number of Generalized Linear Models (GLM) were trained and the best-performing models were selected for further search. Table 3 shows the main characteristics of these models.

These models are integrated into the data recovery pipeline and used to reconstruct the sepsis files with missing values.

TABLE III. LOGISTIC REGRESSION PARAMETERS (NORMALIZED)

	Intercept	Parameter/Coefficient		
HR	89.4392	*DBP* 3.2658	*Resp* 3.2980	*Age* -2.3973
O_2Sat	97.4894	*HR* -0.2305	*SBP* 0.1840	*Resp* - 0.4954
Temp	37.2479	*HR* 0.1233	*SBP* 0.0720	*Age* -0.1001
SBP	123.6418	*Temp* 1.6593	*DBP* 11.7255	*Resp* 0.6840
DBP	61.4999	*HR* 2.5176	*SBP* 6.8117	*Age* -3.7355
Resp	20.3789	*HR* 0.9001	*O_2Sat* -1.0138	*Temp* 0.3220

Table 4 shows the appearance of a recovered file. Recovered values are in bold. This is the same file as in Table 1 above.

TABLE IV. APPEARANCE OF A SEPSIS FILE AFTER RECOVERY

HR	SaO_2	Temp	SBP	DBP	Resp	Age	Sepsis label
83	100	**37.13**	129	50	17	77.3	0
80	99	**36.98**	89	41	18	77.3	0
79.5	100	**37.15**	143	52.5	19	77.3	0
85	100	**37.25**	161	56	18	77.3	0
69	95	**36.91**	91	43	15	77.3	0
66	98	**36.97**	116	40	20	77.3	0
68	99	**37.09**	148	50	17	77.3	0
73	97	**37.02**	117	44	14	77.3	1

C. Preparing Data Sets for Machine Learning

Once reconstructed, the data are split into training (393 files/subsets) and test (167 files/subsets) sets. On each file/subset there is applied a sliding window approach that groups observations in chunks of length three. Finally, the difference between the parameter's values in three consecutive hourly samples is estimated as well as the algorithmic complexity (by the Block Decomposition Method) on each of the two 3x3 matrices [9]. The resulting 14L vector generated for each sample represents the format of data to be passed to the ML algorithm. Since each file contains at least 7 hourly observations of the 6 parameters of interest on which the sliding window approach is applied, the size of the final data sets is larger than the number of initially selected files/subsets.

D. Machine Learning Stage

The training set for ML consists of 3126 samples (1330 sepsis and 1796 non-sepsis). This set is used to train a number of ML models using the H2O platform with 10-fold cross-validation.

The algorithms used include Gradient Boosting Machine (GBM), Generalized Linear Model (GLM), Distributed Random Forest (DRF), Stacked Ensemble (SE), and Deep Learning (DL). Although GBM and SE showed the best performance, because of explainability reasons the GBM model was chosen for further research.

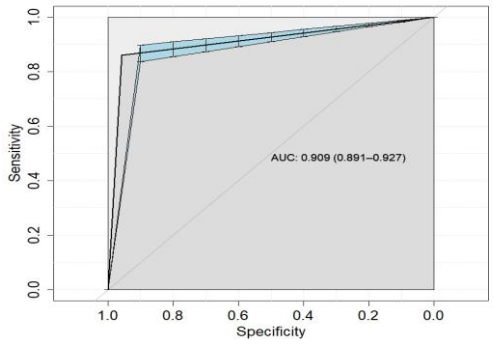

Figure 2. GBM model performance on the test set (by AUC).

The classification efficiency (sepsis vs non-sepsis) of this model (as measured by AUC) on the test set consisting of 167 cases/1065 samples that did not participate in model training is 0.91 (95% CI: 0.89 – 0.93) as shown in Fig. 2.

More detailed statistics concerning the best-performing GBM are summarized in Table 5.

TABLE V. CONFUZION MATRIX AND STATISTICS

		Reference	
		0	1
Prediction	0	598	62
	1	26	379
Accuracy		0.9174	
95% CI		0.8992 - 0.9332	
P-Value		< 2.2e-16	
Cohen's Kappa		0.8277	
Mcnemar's Test P-Value		0.0001907	

IV. DISCUSSION AND CONCLUSIONS

With careful planning, it is possible to reduce the amount of data that are missing to a certain extent. This is important because missing data are a potential source of bias when analyzing data. Handling missing data during model building is a challenge that this study addresses using a known perspective. But as far as we know it is the first time the method is used for such or similar datasets. The core of the method consists of a number of GLMs (one for each missing parameter of interest to be imputed/recovered) combined with LOCF [10] at an earlier stage in the data processing.

The proposed method has certain limitations. First, the method is not tested on different datasets, such as datasets containing categorical features, datasets generated by

another disease (non-septic), etc. It also does not consider data in which there is no correlation between features and does not take into account the type of the variable distribution (normal, logarithmic, etc.).

Thus, our approach could be considered for similar datasets with continuous features and outcomes, and with similar correlations between features. Possible future studies will have the scope to determine the robustness of the method in different datasets.

One of the future research directions would be using the approach described in this paper on a larger data set (e.g. the full set A, which includes a total of 20366 patients and 1760 septic cases). Under these conditions, measures of the method adequacy and reliability can serve the performance of classification ML models in discriminating septic and non-septic cases together with sensitivity analysis results, especially when evaluating a number of methods used for missing data imputation (e.g. comparing the results of the full set analysis to those of the complete case analysis).

REFERENCES

[1] J. Sterne et. al. "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, 2009, 338:b2393

[2] R. Jain and W. Xu, Dynamic model updating (DMU) approach for statistical learning model building with missing data, *BMC Bioinformatics* (2021) 22:221

[3] *Guideline on Missing Data in Confirmatory Clinical Trials*, 2010 EMA/CPMP/EWP/1776/99 Rev. 1 Committee for Medicinal Products for Human Use (CHMP), https://www.ema.europa.eu/en

[4] M. Reyna et al. (2019) "Early Prediction of Sepsis from Clinical Data: the PhysioNet Computing in cardiology Challenge 2019" (version 1.0.0), *PhysioNet*. https://doi.org/10.13026/v64v-d857

[5] M. Singer et al. "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *Journal of the American Medical Association", vol*. 315(8), pp.801–810, 2016.

[6] T. Nykodym, T. Kraljevic, A. Wang and W. Wong, *Generalized Linear Modeling with H2O*, 6ht ed., H2O.ai, Inc.: Mountain View, CA, 2022, pp. 10-17

[7] M. Malohlava and A. Candel, *Gradient Boosting Machine with H2O*, 7th ed., H2O.ai, Inc.: Mountain View, CA, 2022, pp. 8-14
R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. R version 4.0.5, 2021

[8] V. Iapascurta, "A less traditional approach to biomedical signal processing for sepsis prediction", *5th International Conference on Nanotechnologies and Biomedical Engineering*, 2021, *Springer IFMBE Proceedings Series*, vol. 87 pp. 215-222

[9] *Single Imputation Methods for Missing Data: LOCF, BOCF, LRCF (Last Rank Carried Forward), and NOCB (Next Observation Carried Backward)*, January, 2021, available: http://onbiostatistics.blogspot.com/2021/01/single-imputation-methods-for-missing.html