Proceedings of the Federated Conference on Computer Science and Information Systems pp. 129–135
DOI: 10.15439/2018F196
ISSN 2300-5963 ACSIS, Vol. 15

Imputing Missing Values for Improved Statistical Inference Applied to Intrauterine Growth Restriction Problem

Agnieszka Wosiak, Kinga Glinka
Lodz University of Technology
Institute of Information Technology
ul. Wolczanska 215
90-924 Lodz, Poland
Email: agnieszka.wosiak@p.lodz.pl, kinga.glinka@edu.p.lodz.pl

Agata Zamecznik, Katarzyna Niewiadomska-Jarosik
Department of Pediatric
Cardiology and Rheumatology
2nd Chair of Pediatrics
Medical University of Lodz, Poland
Email: agazamek@gmail.com, kasiajarosik@wp.pl

***Abstract*—The paper describes the study on the problem of missing values in medical data collected to discover new dependencies between parameters in children born with intrauterine growth restriction disorder. The aim of the research is to propose a procedure that may be taken to improve the medical inference in the presence of missing data. The approach with use of unconditional mean and k-nearest neighbor imputation has been applied. The experiments proved that application of missing data imputation in original dataset yields more valuable dependencies when compared to original data, maintaining the confidence interval for goodness of fit with the original distribution above 90%. The discovered dependencies in data may establish the basis for new treatment procedures of children with intrauterine growth restriction disorder.**

***Index Terms*—missing values, imputation, medical data analysis, intrauterine growth restriction disorder**

I. Introduction

THE IMPROVEMENT of medical diagnostics and health care is based on scientific studies, which are often based on observations gathered from patients. The reliable analysis of medical dataset usually assumes that subjects of the research were chosen randomly from a greater population at the beginning of the trial. Such an approach is called a randomized controlled trial (RTC) and the analysis of its data is referred to as intention-to-treat (ITT) principle [1].

According to the ITT strategy, all the participants should be included in the analysis regardless whether their outcomes were actually collected [2]. At the same time, the ITT principle requires a complete set of data [3]. The "ideal" ITT analysis is usually not possible to perform, as the problem of missing values commonly occurs [4]. The lacking entries are basically caused by the fact that patient's data are usually gathered as a product of care actions, rather than an organized research protocol [5], [6]. Moreover, in many medical studies, the patients may withdraw or drop out from the trials, which is almost unavoidable and their data may be incomplete [1].

Therefore, appropriate procedures have been created in order to overcome the problem of missing data and estimate a treatment effect.

In most situations, a complete case analysis is considered, and the data of patient with missing vales are discarded. However, in some areas of medical research, more than 50% of missing entries may be encountered [5], [7]. Removing some instances leads to smaller datasets and as a result, to loss of statistical power of the analysis.

As an alternative to a complete case analysis, dropping variables with missing values from the analysis may be also applied [4], [8]. This approach, in turn, neglects valuable observed data and causes less beneficial data analysis.

Another common procedure, that may be a kind of compromise between a complete case analysis and dropping variables, is an available case analysis, where only a piece of patient's data, where no values are provided, is neglected. Despite the complications in analyzing such data due to differences in numbers of instances for various parameters, the method may produce biased estimates of associations [9].

The approach with use of imputation methods is of increasing importance nowadays. Many studies have been conducted on the topic of data imputation techniques [4], [10]–[12], but due to the complexity of the problem of missing data and its close relationship to inner data characteristics, no universal procedure has been discovered and researchers still strive to find standards in data imputation [13].

The aim of this paper is to verify, if appropriate imputation techniques can improve medical inference applied to the problem of intrauterine growth restriction and its relationship with metabolic disorders. There is no universal statistical method that deals with missing data as each study has its own design, measurement characteristics and different assumptions about missing data mechanisms [13]. Therefore, the research constitutes an independent contribution to the relevant literature and also attempts to find a successful way to perform accurate statistical analysis of IUGR in terms of missing data.

The rest of the paper is organized as follows. Section II corresponds to missing values imputation techniques. Section III explains the medical problem of IUGR and is followed by the description of medical data used in the research.

Next, section IV is dedicated to the experiments conducted on sample data and the results. Finally, in Section V, the concluding remarks are discussed.

II. Missing Values Imputation

The imputation methods can be divided into two categories:

- single imputation algorithms,
- multiple imputation algorithms.

A. Single Imputation Methods

In single imputation approach, missing data are imputed by single values. The most popular technique is the mean imputation (MI). The method uses mean of the values of an attribute that contains missing data. The modification of MI technique is using the mode instead of the mean, i.e. the most frequent value in the case of categorical attributes.

Two variations of MI can be distinguished: conditional and unconditional. The unconditional mean imputation (UMI) is not conditioned on the values of other parameters that describe the patient's data. Conditional mean approach (CMI) imputes a mean value that depends on the complete attributes for the analyzed record.

The widely applied single imputation technique is the hot deck [5], [14], [15]. The procedure finds the most similar object for the record that contains missing data, and the missing values are imputed from that object. If the most similar object also contains missing data for the same parameters as in the imputed record, then another closest object is found, until all the missing values are successfully imputed. To find the closest object, several distance functions can be used [16].

One of the hot deck techniques used to compensate for missing data is called kNN imputation (kNNI) [17]. It uses k closest complete instances in the dataset for imputing a missing value, assuming that the k most relevant complete objects are the k nearest neighbors of the incomplete instance in the dataset.

Another approach is based on regression (RI) of the missing values using complete data for a given record [18]. Different regression models can be used, usually depending on the types of imputed parameters, e.g. linear for numerical attributes, logistic for binary features or polytomous for discrete values.

B. Multiple Imputation Methods

Multiple imputation methods use several ordered choices for imputing the missing values [9]. The procedure is performed by creating several complete datasets, in which different imputations are based on a random draw from separately estimated underlying distributions.

One of the most popular approach to multiple imputation is multivariate imputation by chained equations (MICE) described in [19]. It provides a full spectrum of conditional distributions and related regression based methods (linear regression, logistic regression and polytomous regression). To make the application of MICE available, a missing data imputation software package was developed [20].

Multiple imputation algorithms also include:

- Markov chains [21],
- machine learning algorithms (e.g EM algorithm) [22],
- genetic algorithms [23].

Results based on those complex methods are increasingly reported, but their use needs to be applied carefully to avoid misleading conclusions. The multiple imputation procedures require modeling the distribution of each attribute with missing values based on the observed data. Therefore, the validity of results performed on the modified datasets depends on the correctness of such modelling [24].

C. Selection of Imputation Methods

The selection of imputation techniques was determined by the assumption that they should be simple and comprehensive, so that human expert could understand the underlying mechanisms. Moreover, the availability of the methods in statistical program packages such as StatSoft Statistica and SPSS facilitates their use [25]. It was also reported that unsupervised imputation methods may provide more accurate imputation for large amounts of missing data [5]. Therefore, in the experimental studies three single imputation methods were applied: unconditional mean imputation (UMI), conditional mean imputation (CMI) and k nearest neighbor with $k = 5$ (5NNI).

III. Data Description

Intrauterine growth restriction (IUGR) is a fetal disorder of growing. It is often related to fetal hypoxia and higher percentage of perinatal mortality. IUGR is a risk factor for many cardiovascular, metabolic, and pulmonologic diseases in adult life [26]. It occurs in about 3-10% of live-born newborns, but in developing countries it concerns up to 20-30% of newborn infants [27]. The comparisons of absolute measurements of the fetuses with reference values, as well as birth weight percentiles, allow detection of deviations between expected and actual fetal growth and identification of newborns being possibly at risk for adverse health events [28]. However, the diagnosis of IUGR is based on non-consistent definitions [29].

The world-wide research studies report that IUGR makes a risk factor for metabolic syndrome [30], [31], however more environmental studies are still needed to put additional treatment in practice [32]–[34].

The research was based on a study group (SG) of 113 children aged 5-10 years (average 8.1 $\pm$ 1.5) born on term with IUGR and birth weight below 10 percentile according to gestational age for the Polish population [35] and a control group (CG) of 39 children aged 4.5 - 12 (average 7.6 $\pm$ 1.2). All patients were selected during prospective studies at the Pediatric Cardiology and Rheumatology Department of Medical University of Lodz in 2010-2013. The study was approved by Medical Ethical Committee of the Health Sciences Faculty of Lodz University (No: RNN/760/10/KB).

The characteristics of all parameters subjected to further analysis included general attributes, cardiovascular parameters, lipids levels and adipocytokines values. Most of the parameters had missing values. The characteristics of the dataset is

presented in Table I, where the first column refers to the type of an attribute, the second is the name, next three columns include the range of values, the mean and standard deviation and the last column holds the percentage of missing values.

IV. Results and Discussion

The purpose of experiments was to find how the missing values imputation methods improve medical inference for the intrauterine growth restriction problem by discovering new significant correlations between attributes.

The experiments were conducted according to the methods introduced in Section II on the dataset described in Section III. Three main procedures were performed:

A. The experimental procedure that includes analysis with original but incomplete data.
B. The experimental procedure that results in choosing the best imputation technique.
C. The experimental procedure that performs the analysis with the imputed data.

A. Experimental Procedure that Includes Analysis with Original but Incomplete Data

The procedure was performed to discover the characteristics of all parameters and to perform their comparison between the control and study groups. The intention was to confirm by the epidemiological studies the hypothesis that:

- IUGR enhances the susceptibility to metabolic syndrome, and
- there is a correlation between levels of lipids and adipocytokines in IUGR group.

The results of the statistical analysis for the original dataset are presented in Table II.

The first hypothesis was successfully confirmed only for total cholesterol and triglicerides. The significant differences for the rest of parameters were not possible to obtain, mostly due to the numerous missing values and interrelated low significance level, as the level of missing values for adipocytokines was almost 50% in the study group and over 80% in the control group. Therefore, the presence of relationship between lipids and adipocytokines was not possible to be confirmed by statistical analysis as well.

B. Experimental Procedure that Results in Choosing the Best Imputation Technique

In literature the imputation methods are usually related to machine learning problems, mainly to the classification [11], [36]–[38]. Then, the validation can be based on comparisons of imputed datasets to the results obtained for the complete datasets with use of the standard classification metrics, e.g. accuracy or TP rate.

The IUGR problem described in the paper did not refer to the classification, and no class labels were available. Therefore, the choice of the best imputation technique was based on the differences between distributions for the complete sets of data and the sets with randomly dropped and artificially imputed values. The procedure involved five steps:

1) Choose the parameters that were originally complete.
2) Randomly introduce missing data into each parameter in the amounts of: 5%, 10%, 20%, 30%, 40% and 50%.
3) Impute the missing values in each dataset using three imputation methods: mean, conditional mean and kNN with k=5.
4) Compare the distributions of original and modified data for each parameter.
5) For each amount of data imputed, choose the method that built the distribution closest to the original distribution.

In our dataset only 5 parameters out of 18 were originally complete and those data were further used for verification the best suitable imputation technique.

We used missing completely at random (MCAR) approach to drop the data. Values were dropped in the amounts of 5%, 10%, 20%, 30%, 40% and 50%. Each type of missing values' generation was repeated 10 times. As a result we performed 150 experiments (5 attributes x 10 draws x 3 imputation methods). The percentage of cases, where the particular imputation technique was the closest to original distribution, taking into account the amounts of imputed data, are presented in Table III.

The results of comparison for distributions revealed that either the simplest imputation technique by mean values, or more complex with 5NN, can be used for imputation in term of our IUGR datasets. However, it can be also noticed that the imputation by mean gives better results when only small amounts of data are missing. Moreover, the experiments revealed that for smaller amount of missing data, the confidence intervals for goodness of fit with original distribution were above 95%, and at least 80% for the highest amounts of missing values.

C. Experimental procedure that performs the analysis with the imputed data

The final procedure was performed in three steps:

1) Impute the missing data with the method that resulted best for a specified amount of missing values.
2) Perform comparison of characteristics of all parameters between datasets with original and imputed values.
3) Verify the dependencies between lipids and adipocytokines with use of correlation analysis.

As the indication of the best imputation method was not clear enough (although there was a slightly higher recommendation of 5NNI), we decided to use two approaches for further analysis: unconditioned mean imputation and 5NN imputation.

The results of the analysis that includes unconditional mean imputation were presented in Table IV, whereas Table V presents the results for 5-nearest neighbor imputation.

When comparing results of statistical analysis for original dataset (Table II) and for the dataset imputed by unconditional mean (Table IV), one can notice that the statistically significant differences were attained for the parameters with rather small amount of missing data (12% for glucose - Fig. 1 and HDL - Fig. 2), whereas no additional medical conclusions could be drown for the parameters with higher levels of missing values.

TABLE I: Characteristics of attributes for the dataset

Type	Name	Range	Mean	Standard Deviation	Missing values
General	Age (years)	4.5 - 12.0	7.97	1.68	0%
	Body mass (kg)	14.5 - 73.0	25.08	7.73	0%
	BMI	10.8 - 31.6	15.72	2.60	0%
	Birth mass (g)	1800 - 4700	2808	472	0%
	Gestational age	38 - 42	39	0.89	0%
Cardiovascular	Average heart rate	70 - 120	88	10	14%
	SBP	11 - 129	103	13	14%
	DBP	40 - 85	62	8	14%
	SBP load	0 - 96	21.79	17.33	5%
	DBP load	0 - 60	10.54	10.81	5%
Lipids	Glucose	66 - 133	85.90	8.83	12%
	Total cholesterol (mg/dl)	81 - 214	155.11	25.22	12%
	HDL (mg/dl)	27.9 - 100.6	60.91	15.76	12%
	LDL (mg/dl)	24.0 - 133.7	82.53	19.20	12%
	Triglicerides (mg/dl)	24 - 236	70.93	32.99	12%
Adipocytokines	Leptin (ng/ml)	0.48 - 30.79	6.52	6.61	60%
	Adiponectin (μg/dl)	7.33 - 36.70	19.92	6.22	60%
	Resistin (ng/ml)	1.23 - 9.73	2.45	1.50	60%

TABLE II: Characteristics of lipids and adipocytokines levels in the original dataset

Parameter	Study group (SG) Mean $\pm$SD (*)	Control group (CG) Mean $\pm$SD (*)	p-value (**)
Glucose (mg/dl)	86.50 $\pm$ 9.55	84.46 $\pm$ 6.67	0.228
Total cholesterol (mg/dl)	159.08 $\pm$ 25.37	145.44 $\pm$ 22.32	0.004
HDL (mg/dl)	62.14 $\pm$ 14.27	57.93 $\pm$ 18.80	0.160
LDL (mg/dl)	81.77 $\pm$ 20.23	84.39 $\pm$ 16.57	0.475
Triglicerides (mg/dl)	75.99 $\pm$ 36.54	58.62 $\pm$ 16.97	0.005
Leptin (ng/ml)	6.68 $\pm$ 6.78	4.35 $\pm$ 3.56	0.500
Adiponectin (μg/dl)	19.94 $\pm$ 6.21	19.83 $\pm$ 7.42	0.974
Resistin (ng/ml)	2.48 $\pm$ 1.55	2.01 $\pm$ 0.37	0.551

(*) described as average values $\pm$ standard deviations
(**) p-value $<$0.05 defines statistical significance

TABLE III: Summary of evaluation for imputation techniques

% of missing values	UMI	CMI	5NNI
5%	80%	10%	10%
10%	40%	20%	40%
20%	50%	10%	40%
30%	40%	0%	60%
40%	40%	20%	40%
50%	30%	20%	50%

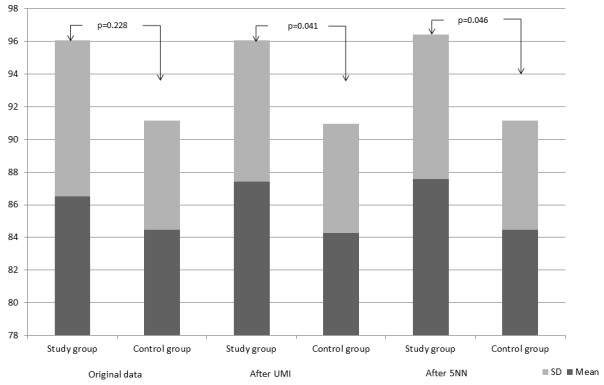

Fig. 1: Differences for glucose in study and control groups before and after performing imputation

When data were imputed with 5NN (Table V), new differences were discovered between levels of leptin and resistin, for which the percentage of missing values equaled 60%.

The correlations between lipids and adipocytokines are presented in Tables VI and VII, for datasets after imputation by unconditioned mean and 5NN respectively.

Unconditioned mean imputation enabled discovering statistically significant correlations between glucose and resistin, total cholesterol and leptin, HDL and leptin, and triglicerids and resistin. The 5NN imputation additionally revealed depen-

TABLE IV: Characteristics of lipids and adipocytokines levels in the dataset after unconditional mean imputation

	Study group (SG)	Control group (CG)	
Parameter	**Mean ±SD (*)**	**Mean ±SD (*)**	**p-value (**)**
Glucose (mg/dl)	84.26 ± 6.68	84.40 ± 8.65	0.041
Total cholesterol (mg/dl)	145.44 ± 22.32	158.45 ± 23.29	0.003
HDL (mg/dl)	57.53 ± 16.20	62.95 ± 13.08	0.038
LDL (mg/dl)	84.39 ± 16.57	81.89 ± 18.53	0.457
Triglicerides (mg/dl)	58.61 ± 16.97	75.18 ± 13.52	0.004
Leptin (ng/ml)	6.30 ± 1.20	6.60 ± 4.66	0.692
Adiponectin (μg/dl)	19.92 ± 2.09	19.93 ± 4.27	0.984
Resistin (ng/ml)	2.41 ± 0.17	2.46 ± 1.07	0.726

(*) described as average values ± standard deviations
(**) p-value <0.05 defines statistical significance

TABLE V: Characteristics of lipids and adipocytokines levels in the dataset after imputation with use of 5-nearest neighbor

	Study group (SG)	Control group (CG)	
Parameter	**Mean ±SD (*)**	**Mean ±SD (*)**	**p-value (**)**
Glucose (mg/dl)	84.46 ± 6.68	87.58 ± 8.83	0.046
Total cholesterol (mg/dl)	145.44 ± 22.32	160.62 ± 23.60	0.001
HDL (mg/dl)	57.93 ± 18.80	63.38 ± 13.17	0.049
LDL (mg/dl)	84.39 ± 16.57	83.67 ± 19.12	0.834
Triglicerides (mg/dl)	58.62 ± 16.97	76.31 ± 13.61	0.001
Leptin (ng/ml)	3.67 ± 1.25	5.55 ± 4.82	0.017
Adiponectin (μg/dl)	20.47 ± 2.32	20.17 ± 4.38	0.691
Resistin (ng/ml)	2.36 ± 0.59	2.74 ± 1.13	0.048

(*) described as average values ± standard deviations
(**) p-value <0.05 defines statistical significance

TABLE VI: Correlations between lipids and adipocytokines in the dataset after unconditional mean imputation

Parameter	**Leptin**		**Adiponectin**		**Resistin**	
	r	**p-value (*)**	**r**	**p-value (*)**	**r**	**p-value (*)**
Glucose	0.1034	0.276	0.0871	0.359	-0.4631	0.013
Total cholesterol	0.3616	0.036	0.0673	0.479	-0.0262	0.783
HDL	0.2405	0.020	0.0531	0.576	0.0036	0.970
LDL	0.1165	0.219	0.0479	0.615	-0.0876	0.356
Triglicerides	0.1104	0.244	-0.0335	0.725	0.2861	0.023

(*) p-value <0.05 defines statistical significance

TABLE VII: Correlations between lipids and adipocytokines in the dataset after 5NN imputation

Parameter	**Leptin**		**Adiponectin**		**Resistin**	
	r	**p-value (*)**	**r**	**p-value (*)**	**r**	**p-value (*)**
Glucose	0.2626	0.046	0.0762	0.423	-0.3647	0.050
Total cholesterol	0.3632	0.042	-0.0013	0.989	-0.0096	0.919
HDL	0.2809	0.037	0.0425	0655.	-0.0014	0.988
LDL	0.1004	0.290	-0.0162	0.865	-0.0624	0.512
Triglicerides	0.0963	0.310	-0.0413	0.664	0.2806	0.037

(*) p-value <0.05 defines statistical significance

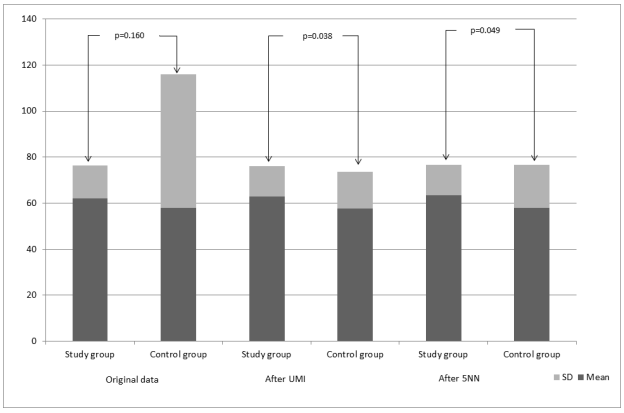

Fig. 2: Differences for HDL in study and control groups before and after performing imputation

dencies between glucose and leptin. These relationships can build the basis for further medical diagnosis and new treatment procedures.

V. Conclusions

Missing values make one of the most common problems for real data collection and extraction in medicine. It is mainly due to the fact, that their presence excludes the intention-to-treat principle and interferes with statistically significant inference. Incomplete data may also refer to other measurements e.g. derived from modern textronic structures used in medicine or protective clothing. Their correct interpretation and analysis ensures reliable operation of intelligent sensors, and as result the entire control system of life signs of the body [41].

Each medical study has its own design, measurement characteristics and different assumptions about missing data mechanisms. Therefore, there is no universal statistical method that deals with missing data, and new investigations should be performed. In this paper, a procedure to improve medical reasoning applied to the problem of discovering new dependencies in the presence of intrauterine growth restriction in children is proposed.

The procedure consists of selecting the imputation technique that results best as applied to the characteristics of data considered and uses the chosen method to impute missing values in data subjected for further analysis. In the empirical test two imputation methods were chosen: unconditional mean and k-nearest neighbor. The statistical analysis of imputed dataset proved to yields more valuable dependencies when compared to original data, maintaining the confidence interval for goodness of fit with the original distribution above 90%. The discovered dependencies in data may establish the basis for new treatment procedures of children with intrauterine growth restriction disorder.

Further studies will involve other medical domains, e.g. monosymptomatic nocturnal enuresis in children where the problem of missing data was encountered [42]. They will also focus on investigating the impact of amounts of missing data on the validity of an imputation technique. Some other methods for dealing with missing values based on rough sets will be used, as proposed by J. Grzymala-Busse et al. [43], [44]. Moreover, the problem of high-dimensional data and feature selection techniques should be considered. More and more data are collected either by interviews, equipment [39] or extraction from text [45], speech [46] or images [47], including medical imaging [48]. In high-dimensional datasets missing data may be more frequent [49] and appropriate feature selection technique [50], [51] may improve the imputation accuracy [10]. Novel solutions of outlier detection based on linguistically quantified statements may be also considered to remove impurities from the data [52].

References

[1] Armijo-Olivo S., Warren S., Magee D. (2009). *Intention to treat analysis, compliance, drop-outs and how to deal with missing data in clinical research: a review*. Physical Therapy Reviews, Vol. 14(1), pp. 36-49, DOI: 10.1179/174328809X405928.

[2] Higgins J. P., Green S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

[3] Lachin J. M. (2000). *Statistical considerations in the intent-to-treat principle*. Contemporary Clinical Trials, Vol. 21(3), pp. 167-189.

[4] Janssen K. J., Donders A. R. T., Harrell F. E., Vergouwe Y., Chen Q., Grobbee D. E., Moons K. G. (2010). *Missing covariate data in medical research: to impute is better than to ignore*. Journal of Clinical Epidemiology, Vol. 63(7), pp. 721-727, DOI: 10.1016/j.jclinepi.2009.12.008.

[5] Farhangfar A., Kurgan L. A., Pedrycz W. (2007). *A novel framework for imputation of missing values in databases*. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, Vol. 37(5), pp. 692-709, DOI: 10.1109/TSMCA.2007.902631.

[6] Cios K. J., Moore G. W. (2002). *Uniqueness of medical data mining*. Artificial Intelligence in Medicine, Vol. 26(1-2), pp. 1-24.

[7] Kurgan L. A., Cios K. J., Sontag, M., Accurso F. J. (2005). *Mining the cystic fibrosis data*. In: Next generation of data-mining applications, IEEE Press, pp. 415-444.

[8] Klebanoff M. A., Cole S. R. (2008). *Use of multiple imputation in the epidemiologic literature*. American Journal of Epidemiology, Vol. 168(4), pp. 355-357, DOI: 10.1093/aje/kwn071.

[9] Donders A. R. T., Van Der Heijden G. J., Stijnen T., Moons K. G. (2006). *A gentle introduction to imputation of missing values*. Journal of Clinical Epidemiology, Vol. 59(10), pp. 1087-1091, DOI: 10.1016/j.jclinepi.2006.01.014.

[10] Aydilek I. B., Arslan, A. (2013). *A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm*. Information Sciences, Vol. 233, pp. 25-35, DOI: 10.1016/j.ins.2013.01.021.

[11] Farhangfar A., Kurgan L., Dy, J. (2008). *Impact of imputation of missing values on classification error for discrete data. Pattern Recognition*, Vol. 41(12), pp. 3692-3705, DOI: 10.1016/j.patcog.2008.05.019.

[12] Moons K. G., Donders R. A., Stijnen T., Harrell F. E. (2006). *Using the outcome for imputation of missing predictor values was preferred*. Journal of Clinical Epidemiology, Vol. 59(10), pp. 1092-1101, DOI:10.1016/j.jclinepi.2006.01.009.

[13] Li T., Hutfless S., Scharfstein D. O., Daniels M. J., Hogan J. W., Little R. J., Royh J. A., Law A.H., Dickersin K. (2014). *Standards should be applied in the prevention and handling of missing data for patient-centered outcomes research: a systematic review and expert consensus*. Journal of Clinical Epidemiology, Vol. 67(1), pp. 15-32, DOI: 10.1016/j.jclinepi.2013.08.013.

[14] Andridge R. R., Little, R. J. (2010). *A review of hot deck imputation for survey non-response*. International Statistical Review, Vol. 78(1), pp. 40-64, DOI:10.1111/j.1751-5823.2010.00103.x.

[15] Myers T. A. (2011). *Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data*. Communication Methods and Measures, Vol. 5(4), pp. 297-310, DOI:10.1080/19312458.2011.624490.

[16] Joenssen D. W., Bankhofer U. (2012). *Hot deck methods for imputing missing data*. In International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 63-75, DOI:10.1007/978-3-642-31537-4_6.

[17] Zhang S. (2011). *Shell-neighbor method and its application in missing data imputation*. Applied Intelligence, Vol. 35(1), pp. 123-133, DOI:10.1007/s10489-009-0207-6.

[18] Yu Q., Miche Y., Eirola E., Van Heeswijk M., SeVerin E., Lendasse A. (2013). *Regularized extreme learning machine for regression with missing data*. Neurocomputing, Vol. 102, pp. 45-51, DOI:10.1016/j.neucom.2012.02.040.

[19] Van Buuren S., Oudshoorn K. (1999). *Flexible multivariate imputation by MICE*. Leiden, The Netherlands: TNO Prevention Center.

[20] Horton N. J., Lipsitz S. R. (2001). *Multiple imputation in practice: comparison of software packages for regression models with missing variables*. The American Statistician, Vol. 55(3), pp. 244-254.

[21] Zhang P. (2003). *Multiple imputation: theory and method. International Statistical Review*, vol. 71(3), pp. 581-592, DOI:10.1111/j.1751-5823.2003.tb00213.x

[22] Fichman M., Cummings J. N. (2003). *Multiple imputation for missing data: Making the most of what you know*. Organizational Research Methods, vol. 6(3), pp. 282-308.

[23] Zhong M., Sharma S., Lingras P. (2004). *Genetically designed models for accurate imputation of missing traffic counts*. Transportation Research Record: Journal of the Transportation Research Board, vol. 1879, pp. 71-79, DOI:10.3141/1879-09.

[24] Sterne J. A., White I. R., Carlin J. B., Spratt M., Royston P., Kenward M. G., Carpenter J. R. (2009). *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls*. BMJ, vol. 338(b2393), DOI: 10.1136/bmj.b2393.

[25] Gadbury G. L., Coffey C. S., Allison D. B. (2003). *Modern statistical methods for handling missing repeated measurements in obesity trial data: beyond LOCF*. Obesity Reviews, Vol. 4(3), pp. 175-184, DOI:10.1046/j.1467-789X.2003.00109.x.

[26] Mahajan, S.D. and Aalinkeel, R. and Singh, S. and Shah, P. and Gupta, N. and Kochupillai, N.: "Endocrine regulation in asymmetric intrauterine fetal growth retardation", Journal of Maternal-Fetal and Neonatal Medicine, 2006, vol. 19(10), pp. 615–623, DOI: 10.1080/14767050600799901

[27] Black, R.E. and Victora, C.G. and Walker, S.P. and Bhutta, Z.A. and Christian, P. and de Onis, M. and et al.: "Maternal and child undernutrition and overweight in low-income and middle-income countries", Lancet, 2013, vol. 382, pp. 427–451, DOI: 10.1016/S0140-6736(13)60937-X

[28] Gürgen, F. and Zeynep, Z. and Füsun, V.: "Intrauterine growth restriction (IUGR) risk decision based on support vector machines", Expert Systems with Applications, 2012, vol.39(3), pp. 2872–2876, DOI: 10.1016/j.eswa.2011.08.147

[29] Bagi, K.S. and Shreedhara, K.S.: "Biometric measurement and classification of IUGR using neural networks", Proceedings of the International Conference on Contemporary Computing and Informatics (IC3I 2014), 2014, pp. 157–161, DOI: 10.1109/IC3I.2014.7019613

[30] Dessi A., Atzori L., Noto A., Visser A. G. H., Gazzolo D., Zanardo V., Magistris A. D. (2011). *Metabolomics in newborns with intrauterine growth retardation (IUGR): urine reveals markers of metabolic syndrome*. The Journal of Maternal-Fetal & Neonatal Medicine, Vol. 24(sup2), pp. 35-39 DOI:10.3109/14767058.2011.605868.

[31] Neitzke U. T. A., Harder T., Plagemann A. (2011). *Intrauterine growth restriction and developmental programming of the metabolic syndrome: a critical appraisal*. Microcirculation, Vol. 18(4), pp. 304-311, DOI:10.1111/j.1549-8719.2011.00089.x .

[32] Zamecznik, A. and Niewiadomska-Jarosik, K. and Wosiak, A. and Zamojska, J. and Moll, J. and Stańczyk, J.: *Intra-uterine growth restriction as a risk factor for hypertension in children six to 10 years old*, Cardiovascular Journal of Africa, 2014, pp.73–77, DOI: 10.5830/CVJA-2014-009

[33] Niewiadomska-Jarosik K., Zamojska J., Zamecznik A., Stańczyk J., Wosiak A., Jarosik P. (2017). *Myocardial dysfunction in children with intrauterine growth restriction: an echocardiographic study*. Cardiovascular Journal of Africa, Vol. 28(1), pp. 36-39, DOI:10.5830/CVJA-2016-053.

[34] Zamecznik A., Stańczyk J., Wosiak A., Niewiadomska-Jarosik K. (2017). *Time domain parameters of heart rate variability in children born as small-for-gestational age*. Cardiology in the Young, Vol. 27(4), pp. 663-670, DOI:10.1017/S1047951116001001.

[35] Malinowski, A. and Chlebna-Sokół, D.: "Dziecko łódzkie-metody badań i normy rozwoju biologicznego", Ankal, 1998, (In Polish)

[36] Baneshi M. R., Talei A. R. (2010). *Impact of imputation of missing data on estimation of survival rates: an example in breast cancer*. Iranian Journal of Cancer Prevention, Vol 3(3), pp. 127-131.

[37] Luengo J., Garcia S., Herrera F. (2012). *On the choice of the best imputation methods for missing values considering three groups of classification methods*. Knowledge and information systems, Vol. 32(1), pp. 77-108, DOI: 10.1007/s10115-011-0424-2.

[38] Tran C. T., Andreae P., Zhang M. (2015). *Impact of imputation of missing values on genetic programming based multiple feature construction for classification*. In Evolutionary Computation (CEC), 2015 IEEE Congress on, pp. 2398-2405, DOI: 10.1109/CEC.2015.7257182.

[39] Ridgway G. R., Lehmann M., Barnes J., Rohrer J. D., Warren J. D., Crutch S. J., Fox N. C. (2012). *Early-onset Alzheimer disease clinical variants multivariate analyses of cortical thickness*. Neurology, vol. 79(1), pp. 80-84, DOI:10.1212/WNL.0b013e31825dce28.

[40] Pawlak R., Korzeniewska E., Koneczny C., Halgas, B. (2017). *Properties Of Thin Metal Layers Deposited On Textile Composites By Using The Pvd Method For Textronic Applications*. Autex Research Journal. Vol. 17(3). pp. 229-237 DOI: 10.1515/aut-2017-0015.

[41] Korzeniewska E., Walczak M., Rymaszewski J. (2017). *Elements of elastic electronics created on textile substrate*. Proceedings of The 24th International Conference Mixed Design of Integrated Circuits and Systems - MIXDES 2017. pp. 447-450.

[42] Tkaczyk M., Maternik M., Krakowska A., Wosiak A., Miklaszewska M., Zachwieja K., Runowski D., Jander A., Ratajczak D., Korzeniecka-Kozyrska A., Mader-Wolynska I., Kilis-Pstrusinska K. (2017). *Evaluation of the effect of 3-month bladder basic advice in children with monosymptomatic nocturnal enuresis*. Journal of Pediatric Urology. Vol. 13. pp. 615.e1-e615.e6. DOI: 10.1016/j.jpurol.2017.03.039.

[43] Grzymala-Busse J. W., Clark P. G., Kuehnhausen M. (2014). *Generalized probabilistic approximations of incomplete data*. International Journal of Approximate Reasoning. Vol. 55(1). pp. 180-196. DOI: 10.1016/j.ijar.2013.04.007.

[44] Clark P. G., Grzymala-Busse J. W., Rzasa W. (2014). *Mining incomplete data with singleton, subset and concept probabilistic approximations*. Information Sciences. Vol. 280. pp. 368-384. DOI: 10.1016/j.ins.2014.05.007.

[45] Komenda M., Karolyi M., Vyskovsky R., Jezova K., Scavnicky J.(2017). *Towards a Keyword Extraction in Medical and Healthcare Education*. Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, pp. 173-176, DOI: 10.15439/2017F351.

[46] Bhaskar J., Sruthi K., Nedungadi P. (2015). *Hybrid approach for emotion classification of audio conversation based on text and speech mining*. Procedia Computer Science, vol. 46, pp. 635-643, DOI:10.1016/j.procs.2015.02.112.

[47] Wojciechowski A., Staniucha R. (2016). *Mouth features extraction for emotion classification*. Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pp. 1685-1692, DOI: 10.15439/2016F390.

[48] Tomczyk, A. (2014). Detection of line segments. Journal of Applied Computer Science. Vol. 22 No. 2 (2014), pp. 81-90, URL: http://it.p.lodz.pl/file.php/12/2014-2/jacs-2014-2-Tomczyk.pdf

[49] Zaitseva E., Levashenko V., Kvassay M., Deserno T.M. (2016). *Reliability Estimation of Healthcare Systems using Fuzzy Decision Trees*. Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 8, pp. 331-340, DOI:10.15439/2016F150.

[50] Paja W. (2015). *Medical diagnosis support and accuracy improvement by application of total scoring from feature selection approach*. Proceedings of the 2015 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 5, pp. 281-286, DOI: 10.15439/2015F361.

[51] Paja W, Pancerz K. (2017). *Feature Selection Methods Applied to Severe Brain Damages Data*. Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki (eds). ACSIS, Vol. 11, pp. 199-202, DOI: 10.15439/2017F382.

[52] Duraj A., Niewiadomski A., Szczepaniak P. S. (2018) *Outlier detection using linguistically quantified statements*. International Journal of Intelligent Systems. DOI: 10.1002/int.21924