doi: 10.22100/ijhs.v4i3.523
Original Article

IJHS 2018;4(3):21-26
ijhs.shmu.ac.ir

IJHS
International Journal of Health Studies

A Predictive Model for Mortality of Patients with Thalassemia using Logistic Regression Model and Genetic Algorithm

Mahmoud Hajipour[1], Kobra Etminani[2], Zahra Rahmatinejad[2], Maryam Soltani[3], Koorosh Etemad[4], Saeid Eslami[2], Amin Golabpour[5*]

[1] Student Research Committee, Dept. of Epidemiology, School of Public Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[2] Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
[3] Razi Clinical Research Development Unit (RCRDU), Birjand University of Medical Sciences (BUMS), Birjand.
[4] Department of Epidemiology, Environmental and Occupational Hazards Control Research Center, School of Public Health, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[5] School of Medicine, Shahroud University of Medical Sciences, Shahroud, Iran.

Received: 24 March 2019
Accepted: 7 July 2019

Abstract

Background: *Due to the thalassemia severe complications, prediction of mortality or patients survival has a great importance in early treatment phases. This study purpose was to predict the mortality rate of patients with thalassemia major and thalassemia intermedia, by the use of the binary logistic regression algorithm and genetic algorithm combination.*

Methods: *This retrospective cohort study was conducted on 909 thalassemia patients by using a questionnaire during 2004-2014. The data of all patients referring to Imam Reza Hospital from 2004 to 2014 have been considered. This study predictive variable is considered to be death or survival of the patient. In this research, we embedded the missing data by the use of the proposed data mining model and MICE algorithm. Totally, 100 patients were excluded from this research, due to the missing or out-of-range data. Death was considered as dependent variable. Also, a predictive model was designed in order to predict the patient mortality using MATLAB language.*

Results: *Mean age of the thalassemia patients was 25.7±9.04 years old and at the end of the study death was reported in 185 subjects. Additionally, there were also 26 independent variables. Moreover, the missing variables mean for each patient was 1.8±0.81. The combined predictive model was able to predict the patient survival rate with 94.35% accuracy. In this research, it was found out that 26 independent variables, which were collected from 12 variables were patient mortality predictors. Also, missing data imputation is an important method for increasing the data mining algorithms efficiency.*

Conclusions: *According to this study results, the use of missing algorithm with the data analysis aid yielded more accurate results, in comparison with the MICE algorithm. Furthermore, 12 parameters affected the patient mortality prediction, which were extracted by the genetic algorithm. Accuracy of the predictive model for the patient death detection was favorable. Consequently, it is recommended to use this model in order to predict the patient mortality.*

Keywords: Thalassemia, Regression, Missing data, Data mining.
***Corresponding to:** A Golabpour, **Email:** a.golabpour@shmu.ac.ir
Please cite this paper as: Hajipour M, Etminani K, Rahmatinezhad Z, Soltani M, Etemad K, Eslami S, Golabpour A. A predictive model for mortality of patients with thalassemia using logistic regression model and genetic algorithm. Int J Health Stud 2018;4(3):21-26.

Introduction

Thalassemia is considered as a serious health hazard, which is responsible for more than 1.5 billion individuals death.[1] In general, about 200 million people are suffering from thalassemia in all over the world.[2] Iran with more than three million individuals suffering from thalassemia minor and fifteen thousand thalassemia major patients is in the global thalassemia belt. Unfortunately, this disease could be associated with severe complications and many physical problems, including chronic and severe anemia, stunted growth, splenomegaly, hepatomegaly, bone diseases, especially ostensible alterations in the skull and facial bones, delayed puberty, heart failure and also cardiac and endocrine dysfunction.

Some of this disease major complications are economic issues arising from the hospitalization, regular blood transfusion and iron therapy cost. In addition, extra amount of iron leads to heart disease, which is the leading cause of the patients mortality.[3,4] Management of chronic, non-communicable diseases is one of the healthcare systems top priorities. In this regard, timely prognosis could be considered as extremely effective.[4,5]

Genetic algorithm (GA) applies Darwin's natural selection principle in order to find the optimal formula for predicting or matching the pattern. This type of algorithm is identified as an appropriate alternative for regression-based prediction methods.

GA is a programming method in artificial intelligence, which uses genetic evolution as a problem-solving pattern. The problem to be solved has the inputs, which are turned into solutions during the modeling process. Afterwards, solutions are evaluated by the fitness function as candidates, and if the exit condition is provided the algorithm will be ended. Generally, GA is an iterative algorithm, which the most parts of it are selected by random processes.[6] In this research, an algorithm was presented in order to remove missing data. Moreover, some models for prediction of patients with thalassemia death were presented by using GA and logistic regression. This research purpose is to provide a predictive model for the patients with thalassemia death and also to determine those variables that affect patients' death or survival.

Materials and Methods

This retrospective cohort study was conducted on 909 thalassemia patients from 2004 to 2014. In this study, all patients with thalassemia data were collected from patients' records in 10 years went to Mashhad Imam Reza Hospital. Patients were divided into two groups: the dead and survived, and the 26 variables were also extracted from the patient. This work was a part of the last author's PhD dissertation in supported by a grant [grant #931034] from Mashhad University of Medical Sciences Research Council, one dependent and 26 independent variables were applied, which are displayed in table 1 with details. We preprocessed the data in order to prepare them for modelling. Afterwards, the proposed model was run on the data and the final model was assessed.

Table 1. The independent variables characteristics

2.1. Data preprocessing

2.1.1. Missing data

MICE algorithm and the proposed optimized algorithm, which is an altered form of MICE algorithm were applied in this study, in order to find the missing data. Therefore, a modified version of MICE algorithm was used in the research.[7]

2.2.2. Outliers

The outliers were detected both in row and column-wise. The rows (record of each patient), which contained outlier data, were eliminated in the first stage. During the time of this process, also the Df Beta Algorithm was used in order to identify the outlier data, which its structure is estimated by using Equation 1.[8]

$$DfBeta_i = \hat{\beta} - \hat{\beta}_i = \frac{(X^T X)^{-1} x_i e_i}{1 - h_{ii}} \quad (1)$$

Where x_i of the ith row from the X matrix. ie_i is the eth remainder, and ih_ii is hth member of H diagonal matrix, which is calculated by Equation 2.

$$H = X(X^T X)^{-1} X^T \quad (2)$$

DFBDTA was estimated for all of the rows. If its amount was $\geq \frac{2}{\sqrt{n}}$, which row would be selected as outlier data and also omitted.

Equation 3 was used in order to eliminate the outliers from each row in the next phase (features) (9).

$$min = mean - 3\sigma, \quad max = mean + 3\sigma \quad (3)$$

Mean is considered as the average amount and σ is variance. If the obtained amounts were less than minimum or greater than maximum, they were regarded as outlier and any row that contained these data must be eliminated.

2.2. Model construction

Three methods were used for designing the models. The logistic regression algorithm was applied on all features in the first method and a model was designed based on the obtained results. In the second method, dimensionality reduction was carried out by the use of the genetic algorithm on the data, followed by designing and evaluating the model. In this regard, the MICE algorithm was used for missing data imputation in both methods. The third method is similar to the second one and the only difference is that the MICE algorithm optimized version was used in this method. All of the three methods are explained in this research following section.

2.1. Model design using logistic regression algorithm

In this method, missing data imputation was carried out by using MICE algorithm. Data was divided into two series of training and test. Afterwards, the regression model was applied on the training data and also the test data was used in order to evaluate the results. Totally, 70% of the data was used in order to create the model and 30% was applied in order to test the model. Logistic regression algorithm was applied on all of the data, and also regression coefficients were determined. The final model was evaluated by parameters assessment classification.

2.2.2. Model design using logistic regression and genetic algorithms with MICE imputes

After the missing data imputation, the MICE algorithm was used and dimensionality reduction was applied. In order to achieve that, a features subset was selected using genetic algorithm, so that this subset could produce a more efficient model. The features subset production structure is indicated in figure 1.

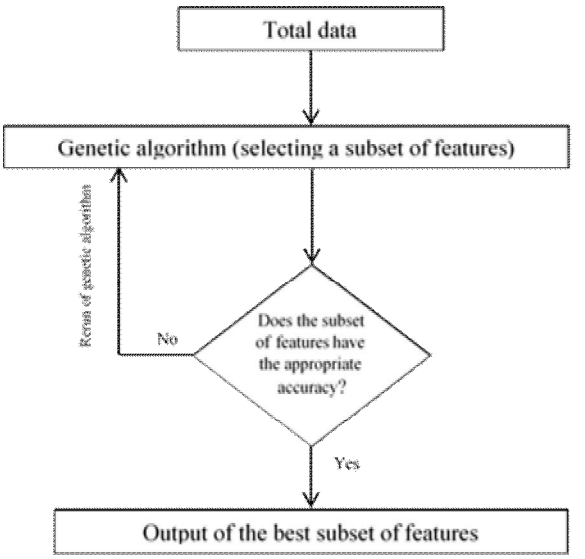

Figure 1. Structure of the genetic algorithm for dimensionality reduction

Figure 1. The genetic algorithm structure for dimensionality reduction

After dimensionality reduction, data was divided into two sections of training and test again and logistic regression algorithm was designed only on the training data. Afterwards, the test data assessed the model.

2.2.3. Model design using logistic regression and genetic algorithms with optimized MICE algorithm

This method is similar to the second one; however, the optimized MICE algorithm is used in the missing MICE algorithm lieu. The optimized MICE algorithm performance for finding the missing data includes some phases; first, the variables are arranged based on the missingness level. Afterwards, variables with minimum level of missingness were set at the beginning and those with the maximum level of missingness were put at the end. The classification model was developed by using data mining algorithms, which its independent variables had no missing data. Meanwhile, this model dependent variables were those with the minimum amount of missing data. In the next stage, the missing data imputation of the dependent variable was carried out by using data mining algorithm. Afterwards, this dependent variable was added to the independent variables set. The next dependent variable was selected according to the missing data number, and data mining algorithm was applied once again. This process was sequentially continued until the time that no variable existed with missing data.

The above mentioned process was performed repeatedly until no change was made in the data. At the end of the process, all of the missing data were completed and no one of them was left. After completing the data imputation, the second method was applied and also genetic algorithm was used for dimensionality reduction; after that, the reduced dimensions were designed by the logistic regression model.

Results

According to this research results, there were 659 empty fields in the collected data, indicating that 1.89% of the data were missing. Additionally, the missing variables mean for each patient was 1.8±0.81. Although there was a small number of missing data, but it should not be neglected because of its presence in all of the patient files.

54 records were eliminated from the information in the first stage by the DfBeta algorithm in order to detect outliers, followed by removal of 46 records in the second stage and totally, 100 records were removed. Moreover, 809 records were used for the model creation and evaluation. In the model creation first stage, logistic regression algorithm was applied on all of the 43 variables. The algorithm was performed 1000 times.

As explained in the 1-3 section of this study, the data was divided into two sections of training and test. Three indices of sensitivity, specificity, and accuracy were also applied in order to evaluate the algorithm. The output of the algorithm performing is displayed in table 2.

Table 1. Characteristics of the independent variables

Row	Variable	Qualitative		Quantitative		Type		Definition Scientific-practical
		Rank	Name	Discrete	Continuous	Dependent	Independent	
1	Hepatitis		*			*		Hepatitis B: hepatitis B virus detection using laboratory tests in humans. Hepatitis C: hepatitis C virus detection using laboratory tests in humans.
2	Splenectomy		*			*		Splenectomy: surgical removal of the spleen
3	Heart failure		*			*		Heart failure: the inability of the heart to pump sufficient amount of blood
4	Hypogonadism		*			*		Hypogonadism in males: when the body does not secrete enough testosterone, a hormone that plays an important role in the growth and development of manhood during puberty or impaired sperm production or both. Hypogonadism in females: inactivity, progressive weakness or absence of ovaries leading to reduced levels of female sex hormones.
5	Hypothyroidism		*			*		Hypothyroidism: when the thyroid gland fails to produce sufficient amount of thyroxine hormone that the body needs.
6	Hypoparathyroidism		*			*		Hypoparathyroidism: reduced production of parathyroid hormone (PTH) from the parathyroid glands.
7	Age at initiation blood transfusion			*			*	The time between birth and the onset of blood transfusion per month.
8	Age at desferal initiation			*			*	The time between birth and the onset of desferal per month
9	Type of medication used for blood sampling		*				*	The type of medication used for blood sampling: desferal, L1, Desfonak, Deferoxamine mesylate, Osveral, Ex jade
10	The number of desferal injections			*			*	Number of desferal injections per month
11	of the amount of desferal at each injection				*		*	The amount of desferal that is injected every time.
12	Method of desferal injection		*				*	Method of desferal injection by injection or pumping
13	Type of used blood		*				*	Type of used blood, which is regulated or cleaned with filters and is healthy.
14	Hemoglobin level				*		*	Hemoglobin is one of the main ingredients of red blood cells, which combines with oxygen and carries it in the blood. Measurement of the amount of hemoglobin is indicative of total red blood cell count.
15	Ferritin level				*		*	Ferritin is a protein that has been linked to excess iron body and could be released if necessary. Measurement of serum ferritin is one of the criteria to determine iron deficiency anemia. Increased amount of ferritin (more than 1000 ng/ml) indicates the increased iron storage in the body.
16	Age			*			*	The period between birth and the time of enrollment in the study
17	Gender		*				*	The phenotypic manifestations of the patient's gender based on the opinion of the questionnaire
18	Patient blood type		*				*	The blood groups contain AB^+,AB^-,B^+,B^-,A^+,A^-,O^+, and O^- cells.
19	Received blood group		*				*	The blood groups contain AB^+,AB^-,B^+,B^-,A^+,A^-,O^+, and O^- cells.
20	Patient occupational status		*				*	The total activities of the patient in order to gain profit or wage.
21	Patient education level		*				*	The degree of scientific interest patient that is recorded in the file.
22	Father's occupational status		*				*	The total activities of patient's father to gain profit or wage.
23	Father's educational level		*				*	The educational level of patient's father as recorded in the file.
24	Mother's occupational status		*				*	Work or a series of activities of patient's mother to gain profit or wage.
25	Mother's educational level		*				*	The educational level of the patient's mother as recorded in patient file.
26	Marital status		*				*	Marital status of patients

Table 2. Performing logistic regression algorithm on all the variables

Parameters	Sensitivity			Specificity			Accuracy		
	Minimum	Average	Maximum	Minimum	Average	Maximum	Minimum	Average	Maximum
Training Data	96.52%	98.14%	100%	85.59%	91.19%	100%	94.35%	96.56%	100%
Test Data	87.96%	95.23%	99.47%	67.27%	83.89%	100%	86.42%	92.62%	97.53%

Table 2. Performing logistic regression algorithm on all of the variables

As shown in table 3, test data accuracy was reported to be as 92.62%. In the second method, which was model creation with MICE algorithm and dimensionality reduction with genetic algorithm, the variables number reduced from 43 to 10. The reduced variables are displayed in table 3.

Table 3. Dimensionality reduction using logistic regression algorithm and genetic algorithm with MICE imputation

Table 3. Dimensionality reduction using logistic regression algorithm and genetic algorithm with imputation of MICE

Row	Variable
1	Age
2	Blood Sampling Interval
3	Defroksaminash
4	Desfnak
5	Heart Failur State
6	Hypo Age
7	Method of Desferal Injection
8	Osteo Age
9	Splenectomy
10	Weight

Logistic regression algorithm was applied after the features reduction to 10 and its output is presented in table 4. Also, algorithm accuracy for the test data was 93.72%, according to the information of this table.

Table 4. Performing algorithm on 10 reduced variables with genetic algorithm and imputation with MICE algorithm

In the third method, the model was created using the optimized MICE algorithm for the missing data and genetic algorithm imputation for dimensionality reduction. According to the results, the variables number decreased from 43 to 12, which its output is indicated in table 5.

Table 5. Dimensionality reduction using logistic regression algorithm and genetic algorithm with MICE optimized imputation

Logistic regression algorithm was applied on the remaining 12 variables. This algorithm was performed for 1000 times repeatedly. Evaluation output is shown in Table 6, accordingly, algorithm accuracy for the test data was 94.35%.

Table 6. Performing algorithm on 12 reduced variables by using the genetic algorithm

In figure 2, the three methods were compared for model creation in terms of the three parameters of sensitivity, specificity and accuracy. The comparison criterion shown in figure 2 is these parameters mean.

Figure 2. The three methods comparison in terms of the three parameters of sensitivity, specificity and accuracy

Table 4. Performing algorithm on 10 reduced variables with genetic algorithm and imputation with MICE algorithm

Parameters	Sensitivity			Specificity			Accuracy		
	Minimum	Average	Maximum	Minimum	Average	Maximum	Minimum	Average	Maximum
Training Data	95.33%	96.93%	98.19%	78.69%	85.88%	90.7%	92.4%	94.41%	96.29%
Test Data	91.85%	95.54%	100%	62.96%	84.3%	96.3%	89.3%	93.72%	98.35%

Table 5. Dimensionality reduction using logistic regression algorithm and genetic algorithm with optimized imputation of MICE

Row	Variable
1	Age
2	Blood Sampling Interval
3	Defroksaminash
4	Desfnak
5	Exjade
6	Heart Failur State
7	Hypo age
8	Marital Status
9	Method of Desferal Injection
10	Osteo age
11	Splenectomy
12	Weight

Table 6. Performing algorithm on 12 reduced variables using genetic algorithm

Parameters	Sensitivity			Specificity			Accuracy		
	Minimum	Average	Maximum	Minimum	Average	Maximum	Minimum	Average	Maximum
Training Data	96.07%	97.33%	98.44%	82.54%	87.56%	90.31%	93.46%	95.1%	96.82%
Test Data	92.55%	96.99%	100%	70%	85.66%	98.24%	90.53%	94.35%	98.35%

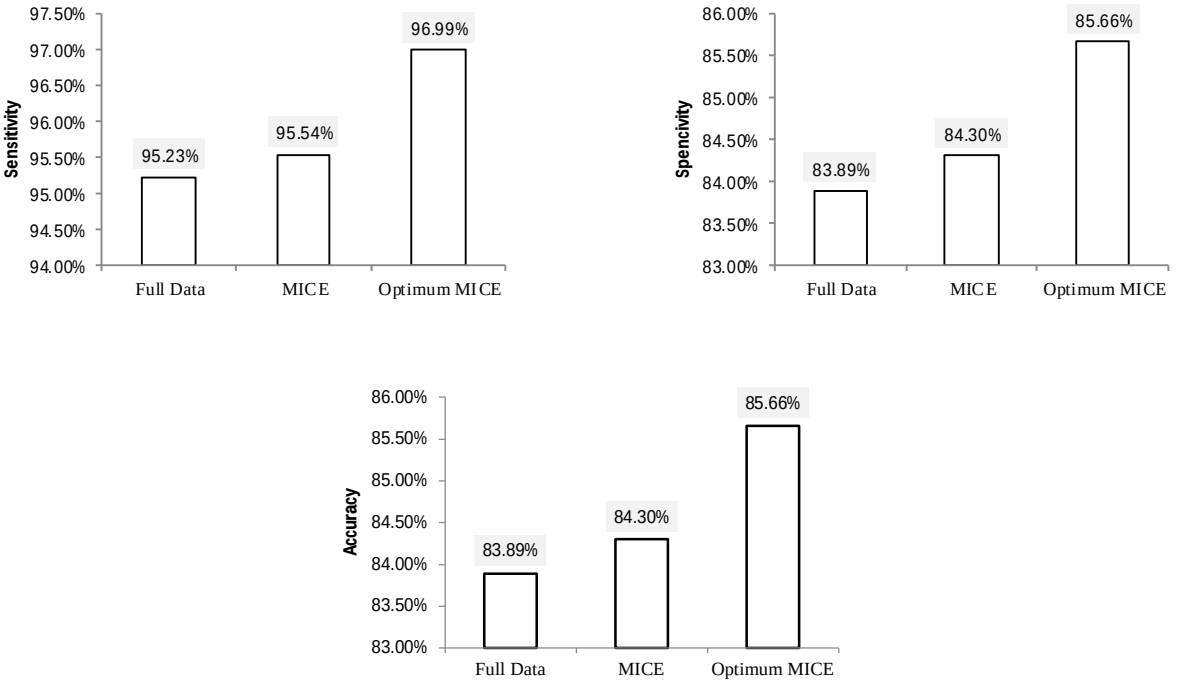

Figure 2. Comparison of the three methods in terms of the three parameters of sensitivity, specificity and accuracy

Discussion

This study was carried out in order to create a predictive model for the patients with thalassemia death by using the proposed algorithm for the missing data imputation. The created model was also able to predict the death occurrence with 94.35% accuracy, which was a high level of accuracy in comparison with the other models. Moreover, the proposed algorithm for the missing data imputation increased the accuracy in classification algorithm.

Yaman et al. conducted a cohort study on 67 patients with beta thalassemia from 2004 to 2009, it was demonstrated that disease complications increased by the age increasing. In this regard, endocrine (38.8%), cardiovascular (22.4%), digestive (19.4%), allergic (9%), inflectional (1.5%) and thrombocytes (1.5%) were the most common complications.[11] Dimensionality reduction results in this research indicated a significant relationship between the mortality and patient age, as well as age with the osteoporosis and hypogonadism diagnosis.

In another study done by Modell et al. entitled as "survival in thalassemia major treated with desferal", it was concluded that patients, who received desferal consistently had higher longevity, in comparison with patients that were not using desferal, or those who received it with delay.[12]

In another study accomplished by Ansari et al. a survival analysis model was used from 2005 to2006 entitled as "evaluation of factors associated with the occurrence of hypogonadism in patients with thalassemia major". That study stablished the relationship between the age at desferal injection and age at hypogonadism complications, and survival without hypogonadism complications in patients with thalassemia. They reported that blood transfusion timely initiation could increase the complications and their pre-term onset incidence rate in case of delayed or lack of desferal injection.[4]

In another cross-sectional research by Roodbari et al. entitled as "survival of patients with thalassemia major in southeast of Iran in 2007", it was indicated that elevated blood hemoglobin level would lead to have reduction in mortality rate. Furthermore, mortality could decrease by 18% with only one time increase in the blood transfusions number.[13]

One of the major drawbacks of the study was considering the fact that the applied data belonged to only one center. Due to the fact that use of the proposed algorithm for the missing data imputation can lead to the models creation with high levels of accuracy, it is recommended that this algorithm be used for the missing data imputation in other researches. Moreover, it is recommended that the presented model in this study be applied in order to predict the patient mortality rate. However, this model clinical application should be validated using patients with other diseases data.

In this section, the results of research were explained and at the same time, the comprehensive discussion is presenting. Results can be presented in figures, graphs, tables, etc. which make the reader understand it easily.[2,5] This discussion can be made in several sub-chapters.

Acknowledgement

This work was a part of the last author's PhD dissertation in supported by a grant [grant #931034] from Mashhad University

of Medical Sciences Research Council. The funder was involved in the manuscript preparation and publication process.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Allehaiby AH, Alluheibi SM, Alnassar SM, Bayydih MA, Almohammadi MM, Alnashry LM, et al. Assessment of Patients with Beta-thalassemia. The Egyptian Journal of Hospital Medicine. 2017;69:2814-9. doi:10.12816/0042571

2. Shamsi A, Amiri F, Ebadi A, Ghaderi M. The effect of partnership care model on mental health of patients with thalassemia major. Depress Res Treat 2017;2017:3685402. doi:10.1155/2017/3685402

3. Kuo KH, Mrkobrada M. A systematic review and meta-analysis of deferiprone monotherapy and in combination with deferoxamine for reduction of iron overload in chronically transfused patients with β-thalassemia. Hemoglobin 2014;38:409-21. doi:10.3109/03630269.2014.965781

4. Sayehmiri K, Tardeh Z, Mansouri A, Borji M, Azami M. The prevalence of hypogonadism in patients with thalassemia major in Iran–a systematic review and meta-analysis study. J Shahrekord Univ Med Sci 2016;18:140-51.

5. Garcia-Santos D, Hamdi A, Zidova Z, Horvathova M, Ponka P. Heme Oxygenase 1 Inhibition Reverses Anemia in β-Thalassemia Mice. Am Soc Hematology 2016;128:2462.

6. Sivanandam SN, Deepa SN. Introduction to Genetic Algorithms: Springer Shop. Berlin Heidelberg; 2007.

7. Royston P, White IR. Multiple imputation by chained equations (MICE): implementation in Stata. J Stat Softw 2011;45:1-20.

8. Bahadir B, İnci H, Karadavut U. Determination of outlier in live-weight performance data of Japanese quails (Coturnix coturnix japonica) by Dfbeta and Dfbetas techniques. Italian Journal of Animal Science 2014;13:3113. doi:10.4081/ijas.2014.3113

9. Ranga Suri NNR, Narasimha Murty M, Athithan G. Outlier Detection: Techniques and Applications. A Data Mining Perspective. Springer, 2019.

10. David W, Hosmer Jr, Lemeshow S, Rodney X, Sturdivan. Applied Logistic Regression. 3 rd ed. Hoboken, New Jersey: Wiley, 2013.

11. Amoozgar H, Zeighami S, Haghpanah S, Karimi M. A comparison of heart function and arrhythmia in clinically asymptomatic patients with beta thalassemia intermedia and beta thalassemia major. Hematology 2017;22:25-9. doi:10.1080/10245332.2016.1226699

12. Modell B, Letsky EA, Flynn DM, Peto R, Weatherall DJ. Survival and desferrioxamine in thalassaemia major. Br Med J Clin Res Ed 1982;284:1081-4. doi:10.1136/bmj.284.6322.1081

13. Roudbari M, Soltani-Rad M, Roudbari S. The survival analysis of beta thalassemia major patients in South East of Iran. Saudi Med J 2008;29:1031-5.