RESEARCH ARTICLE

Construction data mining methods in the prediction of death in hemodialysis patients using support vector machine, neural network, logistic regression and decision tree

SALMAN KHAZAEI[1], SOMAYEH NAJAFI-GHOBADI[2], VAJIHE RAMEZANI-DOROH[3,4]

[1] Research Center for Health Sciences, Hamadan University of Medical Sciences, Hamadan, Iran; [2] Department of Industrial Engineering, Faculty of Engineering, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran; [3] Department of Health Management and Economics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran; [4] Modeling of Non-communicable diseases research center, Hamadan University of Medical Sciences, Hamadan, Iran

Keywords

Hemodialysis, • Kidney failure • Survival • Data mining • Decision tree • Neural network • Support vector machine • Logistic regression

Summary

Objectives. *Chronic kidney disease (CKD) is one of the main causes of morbidity and mortality worldwide. Detecting survival modifiable factors could help in prioritizing the clinical care and offers a treatment decision-making for hemodialysis patients. The aim of this study was to develop the best predictive model to explain the predictors of death in Hemodialysis patients by data mining techniques.*

Methods. *In this study, we used a dataset included records of 857 dialysis patients. Thirty-one potential risk factors, that might be associated with death in dialysis patients, were selected. The performances of four classifiers of support vector machine, neural network, logistic regression and decision tree were compared in terms of sensitivity, specificity, total accuracy, positive likelihood ratio and negative likelihood ratio.*

Results. *The average total accuracy of all methods was over 61%; the greatest total accuracy belonged to logistic regression (0.71). Also, logistic regression produced the greatest specificity (0.72), sensitivity (0.69), positive likelihood ratio (2.48) and the lowest negative likelihood ratio (0.43).*

Conclusions. *Logistic regression had the best performance in comparison to other methods for predicting death among hemodialysis patients. According to this model female gender, increasing age at diagnosis, addiction, low Iron level, C-reactive protein positive and low urea reduction ratio (URR) were the main predictors of death in these patients.*

Introduction

Chronic kidney disease (CKD) is one of the main causes of morbidity and mortality worldwide [1]. Death from CKD has increased dramatically in recent years, with a 134% increase in CKD deaths in 2013 compared to 1990 [1]. It is estimated that more than 1.5 million end-stage renal disease (ESRD) cases receive renal replacement therapy through dialysis or transplantation worldwide annually [2]. On the other hand, the increase in the number of chronic diseases, such as diabetes and hypertension, as the risk factors of CKD, enhances the prevalence of CKD among countries. Therefore, CKD should be considered as a health problem priority, especially in developing countries, and a large proportion of health resources should be devoted to this problem [3].

In Iran, hemodialysis (HD) is the main way of renal replacement therapy in ESRD patients. Accumulating evidence suggests that increasing age, malnutrition, cardiovascular disease, higher glomerular filtration rate (GFR) at the time of dialysis initiation, overhydration, use of catheter as a vascular access as well as hemoglobin, ferritin, C-reactive protein (CRP), serum albumin, and creatinine, among baseline laboratory markers, are the potential predictors of death in hemodialysis patients [4-7]. Apart from these accepted prognostic factors, the role of some factors is controversial on morbidity and mortality in hemodialysis patients and requires more investigations. Moreover, depending on different medical facilities of the hospitals as well as the cultural factors associated with the patients, the predictors for patient's death in different communities may be different. On the other hand, because of the seriousness of CKD and its complications, early screening of high-risk patients for mortality, using an accurate and efficient model, especially based on demographic characteristics, plays an important role in the prediction of death in hemodialysis patients. To this, it is possible to identify subjects who are at risk for mortality based on common risk factors, such as age and gender, through predictive statistical models.

Construction of a predictive model and identifying important risk factors of a dichotomous variable like dead/alive status for a patient is usually conducted through classical logistic regression (LR). Recently, machine learning techniques including support vector machine (SVM), neural networks (NN), and decision tree (DT) have been shown to have promising performance in classification problems [8-14].

https://doi.org/10.15167/2421-4248/jpmh2021.62.1.1837

Ideally, it would be interesting to construct a model with an increased predictive power through machine learning techniques for classification that require no distributional assumptions. These classifiers also consider nonlinear and complex relationships between the response variable and predictors. Therefore, they can produce accurate predictions for the response variable. Nevertheless, their performance may vary in different conditions and there are inconsistencies between studies about their superiority over classical models. So, they need to be investigated over different datasets.

Improved ability to identify those patients at an increased risk of death could help in prioritizing the clinical care and offers treatment decision-making for hemodialysis patients. Several studies have investigated the prediction performance of different outcomes in CKD patients [11, 14-17]. For example, Lacson used the SVM to test the hypothesis that "appropriately transformed sequential blood pressure measurements would significantly improve the prediction of mortality in hemodialysis patients" [18]. However, to the best of our knowledge, there was found no study that compared the predictive performance of SVM, NN, DT, and logistic regression in predicting hemodialysis mortality. Moreover, in each country and even in each setting, the predictors of mortality vary according to the available treatment facilities and the conditions of the patients. Therefore, the results of studies in other countries cannot be generalized to all communities. In general, the availability of a system for predicting death in hemodialysis patients can help improve the quality of the care proportionate to the patient's demographic and clinical features and can reduce costs. In this study, we selected the most widely used machine learning methods of the SVM, NN, and DT and compared their performance with that of the logistic regression. Therefore, the purposes of this research were to develop predictive models to explain the relationships between examined independent variables and the Hemodialysis patient's survival/mortality and choosing the model with better performance by comparison their sensitivity, specificity, accuracy, positive and negative likelihood ratio. We also compared the performance of these models with that of the classical logistic regression. Finally, we identified important variables that were associated with the outcome of interest with the model that provides better predictive power over the dataset used here.

Methods

Study design

We performed a retrospective cohort study to develop predictive models to explain the relationships between examined independent variables and the Hemodialysis patient's dead/alive status.

Settings and patients

We examined 758 patients who were undergoing hemodialysis treatment in Hamadan province from March 2007 to March 2017 to investigate predictors of mortality among hemodialysis patients. Hamadan province, with an area of 19,493 square kilometers in extent, is located in the west of Iran and has a population of 1,738,234 people; according to the national census held in 2016 by the Statistical Center of Iran. We used information of patients from eight hospitals in the province with dialysis ward including Alimoradian, Besat, Vali-asr, Ghaem, Imam Hossein, Valiasr, Imam Reza, and Shahid-Beheshti in Nahavand, Hamadan, Tuyserkan, Asadabad, Malayer, Razan, Kabudarahang, and Hamadan city.

Patients with acute renal failure or under treatment with peritoneal dialysis, patients on transient hemodialysis, and patients with incomplete medical records were excluded from this study. This study was approved by the Ethics Committee of Hamadan University of Medical Sciences (IR.UMSHA.REC.1399.029).

Output and input variables

Here, we used all the medical information of the patients, including demographic, clinical, and laboratory information, to model building. Therefore, we provided a researcher-made checklist according to the hospital records of HD patients for collecting data. The checklist included features related to demographic profiles (age at diagnosis (year), gender (male, female), marital status (married, single, divorced, widow), body mass index (BMI, kg/m^2), residence area (urban, rural), educational level (illiterate, primary, guidance, high school, academic), the history of tobacco use (yes, no) and substance abuse (yes, no); and clinical information (including hemoglobin (g/dl), blood urea nitrogen (BUN) (mg/dL), creatinine (mg/dL), CRP status (positive, negative), sodium (mEq/L), calcium (mg/dL), phosphor (mg/dL), iPTH (pg/ml), albumin (g/dl) and ESRD cause (Hypertension, Diabetes, Urologic& obstructive diseases, Polycystic Kidney, Glomerulonephritis, Un-Known). Clinical and laboratory data related to the time before the first dialysis session were used for each patient. To minimize measurement variability, all these clinical data were measured twice and two measures for each patient were averaged. Time since diagnosis to mortality/follow up (year) was also used as a confounder and all analyses were adjusted for it. These records were collected by reviewing patients' medical records and, if necessary, asking patients or instructor of the sector.

Data pre-processing and dealing with missing values

Before model building and any analysis, the data were checked with regard to spelling errors and other irregularities/irrelevancies. Missing values were imputed. Therefore, for the variables with missingness greater than 1%, including uric acid (3.73%), hemoglobin (1.4%), Alk (1.28%), and Iron (1.05%), we used CART regression trees for imputing missing values. For the variables with missingness less than 1%, including albumin (0.47%), Plt (0.23%), hematocrit levels (0.12%) and urea reduction ratio (0.12%), we applied simple imputation using their median. The mode was used for imputing the missing data of the blood group (0.12%) and Rh (0.12%).

Anomaly detection was used for finding the outlier records to improve the precision of modeling. This technique utilizes clustering methods to recognize anomalous data. Anomaly detection provides very significant and critical information for outlier detection in various applications [19]. One record with an anomaly index greater than 2 [20] was eliminated from further investigations. To this, data mine was attached to anomaly node in the modeler. Then, the created model was added to the project. This technique uses the clustering methods. After running the model, data were divided into four clusters and then by adjusting the model, anomalous data in the model building process were deleted.

Statistical methods

Decision tree

Decision tree is assumed as a machine learning method that utilizes recursive partitioning of potential predictors space (each partitioning happens at a node) [21] and creates a hierarchical partitioning tree and then predicts the response variable using final nodes for both categorical and continuous responses. Optimal cut-off point for a continuous predictor is the one that produces lower prediction error. In this study we used this method in identifying individuals who experienced death through an easily interpretable rule induced by binary splitting of covariates according to the predictors. Here, we used C5.0 decision tree that can automatically inspect the variables before constructing the tree which keeps only relevant variables in the model. A rule is induced by a binary split on inputs with questions such as "Is the patient female or male?" or "Is the subject a smoker or nonsmoker?". For continuous variables, the algorithm automatically searches for the best split, using some criteria and the data are partitioned accordingly. The procedure continues until the data set is split into a number of mutually exclusive groups.

Neural network

Neural network is of the most widely used machine learning methods that works based on the human brain structure. This method uses an input layer, a hidden layer, and an output layer connected with some associated weights. These weights are adjusted during learning process, to provide a better prediction performance for the response variable. Although there are several structures for the neural network, in this study, we used the most commonly used NN called Multilayer Perceptron (MLP). In this study, to find the best performance of the NN, complex nonlinear mapping between input and output layers is conducted using different number of nodes and the NN approach with one hidden layer, using a hyperbolic tangent activation function *($f(x) = 2 / (1+exp(-2x)) -1$)*, was utilized (the performance of the method did not improve with a greater number of hidden layers). In the output layer, the "SoftMax" activation function ($f_i(x) = exp(x_j) / \Sigma\, exp(x_i)$, $j = 1, ..., p$, with x as the input vector) was used.

Support vector machine

Support vector machine is a commonly used machine learning method that recruits a kernel function to project the predictor space into a higher dimensional space where a linear hyper plane instead of the non-linear separator in the lower dimension [22]. The linear hyperplane is fitted in such a way that the training data have the maximal distance from its margin. In this study, our binary response was the dead/alive status of the patients. Therefore, considering:

$$y_i \in \{-1,1\}$$

as the binary response for:

$$i \in \{1,2,...,n\}$$

and x_i as the input vector, the equation of the hyperplane classification takes the following form:

$$\sum^{n} \alpha_i \gamma_i (x_i) + b = 0$$

Where γ is a function of x, α the original input vector of x, is the regression coefficients vector and b is the bias term (or intercept). These coefficients are obtained through a quadratic optimization problem. For more detail see Tapak et al. 2017 [11].

Stepwise logistic regression

Logistic regression is a parametric regression model that is the most commonly used model in modeling the relationship between some inputs and a binary response/output. The model is written as follows:

$$\log(\pi / (1-\pi)) = \alpha_0 + \sum^{p} \alpha_i x_i$$

In this equation, α stands for the regression coefficients vector and π is the probability of the response variable takes the value 1 for the event of interest. Stepwise logistic regression is a method of fitting logistic regression model in which the choice of predictive variables is carried out automatically and in each step, a variable is considered for including or excluding from a set of inputs based on some pre-specified criterion like Bayesian information criterion (BIC).

Model building

In all analysis, we used the variable of mortality status (dead/alive) as the output variable and all other variables in the checklist were considered as inputs. For comparing the models, we used 10-fold cross-validation: one with 90% subjects for training and the other with 10% subjects for validation. This process repeated 10 times. Then, sensitivity, specificity, total accuracy, positive likelihood ratio and negative likelihood ratio were computed to compare the models. The calculations were based on the following formulas:

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

$$Total\ Acuraccy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Positive\ likelihood\ ratio = \frac{Sensitivity}{1-Specificity}$$

$$Negetive\ likelihood\ ratio = \frac{1-Sensitivity}{Specificity}$$

Where *TP*, *FP*, *TN*, and *FN* represent the number of true positives, false positives, true negatives, and false negatives, respectively. The IBM SPSS modeler 15 was applied for data analysis.

Variable importance

The importance of a variable in each input, for the classifiers that were used in this study, was calculated according to the percentage increase in the prediction error when the variable/input was removed from the analysis. So, the input that led to the most increase in the prediction error by a classifier was selected as the most important. After scoring the importance of the variables, they were ranked based on their scores.

Results

Summary statistics of the variables included in the analysis for the patients were shown in Tables I and II. According to the results, the majority of the alive patients were male (57%), married (82.0%), non-smoker (82.0%), non-substance abuser (0.89%), illiterate (43.0%) and lived in rural area (64.0%). The main causes of ESRD for 28.0% and 20.0% of them were hypertension and diabetes alone, respectively. Permanent catheter in the 48% of cases was the dominant way of vascular access. Summary statistics of continuous variables also were provided in Table II.

Different settings of the parameters were tested, and the best result was obtained by expert mode with pruning severity 75 and minimum records per child branch 2. The most informative variables, according to the values of the variable importance, estimated by the DT model were shown in Figure 1. According to the results, the variables of “Age at diagnosis”, “CRP” and “Hemoglobin” were

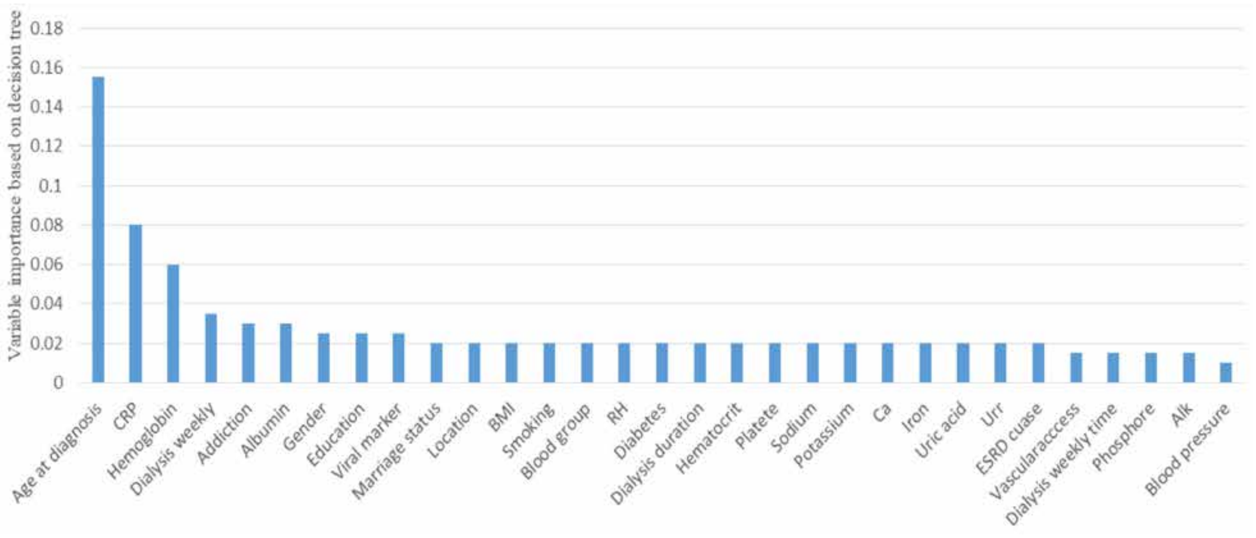
Tab. I. Descriptive statistics of discrete features related to participants.

Variables	Alive		Death	
	N	%	N	%
Gender:				
male	256	57	208	51
female	192	43	200	49
Marriage status:				
single	76	17	6	1.5
married	366	82	396	97.5
divorce	6	1	6	1
Education:				
illiterate	194	43.3	260	63.7
primary and middle	145	32.4	104	25.5
high school	87	19.3	36	8.8
university	22	5	8	2.0
Location:				
urban	287	64	248	61
rural	161	36	160	39
Smoking:				
no	366	82	296	72
yes	82	18	112	28
Addiction:				
no	399	89	331	81
yes	49	11	77	19
Blood group:				
A	153	34	126	31
B	103	23	97	24
AB	30	7	34	8
O	162	36	151	37
Rh:				
+	407	91	361	89
-	41	9	41	11
Viral marker:				
HIV	2	0.5	3	0.7
HBV	7	1.6	7	1.7
HVC	11	2.4	3	0.7
no	428	95.5	395	96.9
Diabetes:				
no	312	70	246	60
yes	136	30	162	40
Hypertension:				
no	269	60	239	59
yes	179	40	169	41
ESRD cause:				
diabetes	91	20.3	116	28.4
diabetes & bp	45	10.0	46	11.3
blood pressure	127	28.3	120	29.4
urological & obstructive diseases	38	8.5	40	9.8
polycystic kidney	18	4.1	18	4.4
glomerulonephritis	47	10.5	13	3.2
other	82	18.3	55	13.5
Vascular access:				
temporary catheter	69	15.4	66	16
permanent catheter	151	33.7	135	33
fistula	217	48.4	192	47
graft	11	2.5	15	4
Dialysis weekly:				
1	8	2	7	1.7
2	95	21	74	18.1
3	340	76	317	77.7
4	5	1	10	2.5
CRP:				
negative	310	69	242	59
positive	138	31	166	41

Tab. II. Summary of continues variables related to participants.

Variables	Alive					Death				
	N	Min	Max	Mean	Std. Dev	N	Min	Max	Mean	Std. Dev
Age at diagnosis (yr)	448	8	83	50.29	15.73	408	18	88	61.97	10.92
BMI (kg/m^2)	448	12.33	41.26	23.09	4.21	408	14.69	38.09	23.31	3.79
Dialysis duration (hour)	448	2	4	3.65	0.46	408	2	4	3.63	0.46
Dialysis weekly time (hour)	448	2	16	10.15	2.33	408	2	16	10.24	2.22
Hemoglobin (gr/dlit)	448	1.3	17.4	10.48	2.06	408	4	15.7	10.47	1.75
Hematocrit levels (%)	448	16.1	60	192.52	72.085	408	16.1	60	32.46	6.05
Plt (1,000/mm^3)	448	21	463	192.52	72.08	408	27	670	187.73	68.68
Sodium (mg/dlit)	448	105	198	138.78	6.8	408	106	193	138.89	7.56
Potassium (mg/dlit)	448	3	9.6	4.9	0.94	408	3	9.6	4.96	0.95
Calcium (mg/dlit)	448	5.1	12	8.9	0.90	408	5.9	11.3	8.7	0.84
Phosphor (mg/dlit)	448	2.3	12.3	5.11	1.55	408	1.7	12	5.13	1.6
Iron (ug/ dlit)	448	2	1028	111.90	115.48	408	3	520	98.3	73.49
Uric acid (mg/dlit)	448	1	14.6	6.77	1.54	408	3.2	13.7	6.70	1.42
Albumin (g/dlit)	448	1	5.6	3.74	0.73	408	1	6.4	3.59	0.72
Alk (U/L)	448	4.6	2612	316.14	262.65	408	4.2	2349	301.58	213.35
Urea reduction ratio (%)	448	1	0.96	0.643	0.124	408	0	0.89	0.62	0.13
Time to mortality/ follow-up (yr)	448	0.08	10.70	2.35	2.38	408	0.08	10.30	2.23	2.18

Fig 1. Importance of variables estimated by the decision tree.



the first topmost important variables in the prediction of death. In this study, the NN was trained with all inputs (one for each predictor) in the input layer and one hidden layer with 10 neurons. Figure 2 showed the importance of variables associated with death of dialysis patients by the NN model. According to the results, the variables of "Age at diagnosis", "Alkaline phosphatase" and "Iron" were the three topmost important variables in the prediction of death. Since linear function had better results than other functions, it was used as kernel function for the SVM model. Regularization (C) was optimized by trying different values, and the best-obtained value was 10. We used expert mode and the stopping criterion was set 0.001. The SVM model ranked all of the variables the final results were shown in Figure 3. According to the results, the variables of "Age at diagnosis", "Education" and "CRP" were the three topmost important variables in the prediction of death. Based on $p < 0.05$, the stepwise LR model indicated gender, age at diagnosis, marriage status, addiction, onset year, iron, URR, and CRP as significant variables (Tab. III). For the dependent variable, we considered death as the reference group.

Table III shows the total accuracy, sensitivity, specificity, positive likelihood ratio, and negative likelihood ratio (Mean and standard deviation) estimated by the cross-validation method over the testing set for each model in 100 repetitions. According to the results shown in Table IV, LR showed a better performance compared to

Fig. 2. Importance of variables estimated by the natural network.

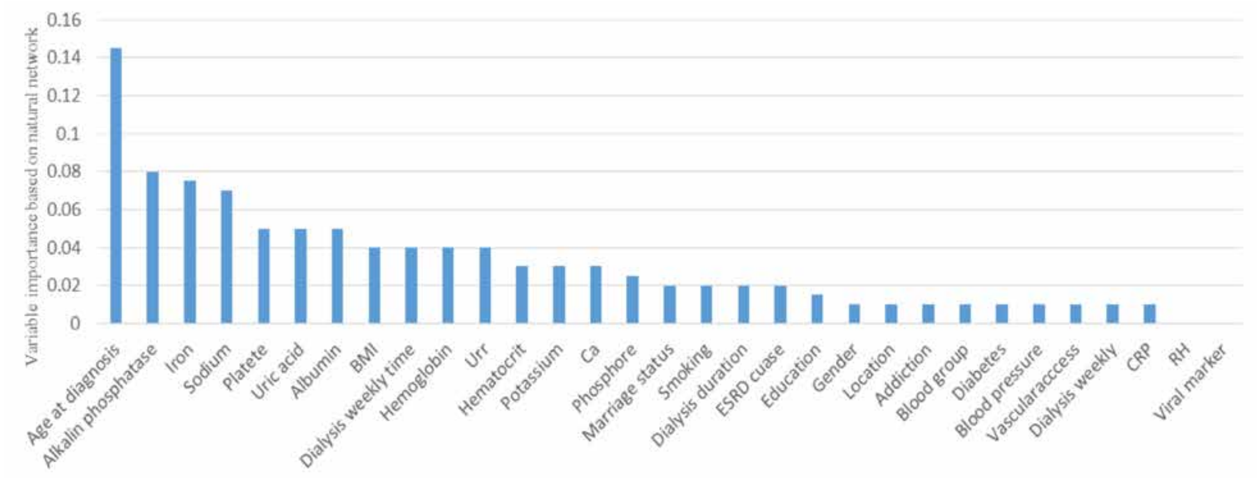
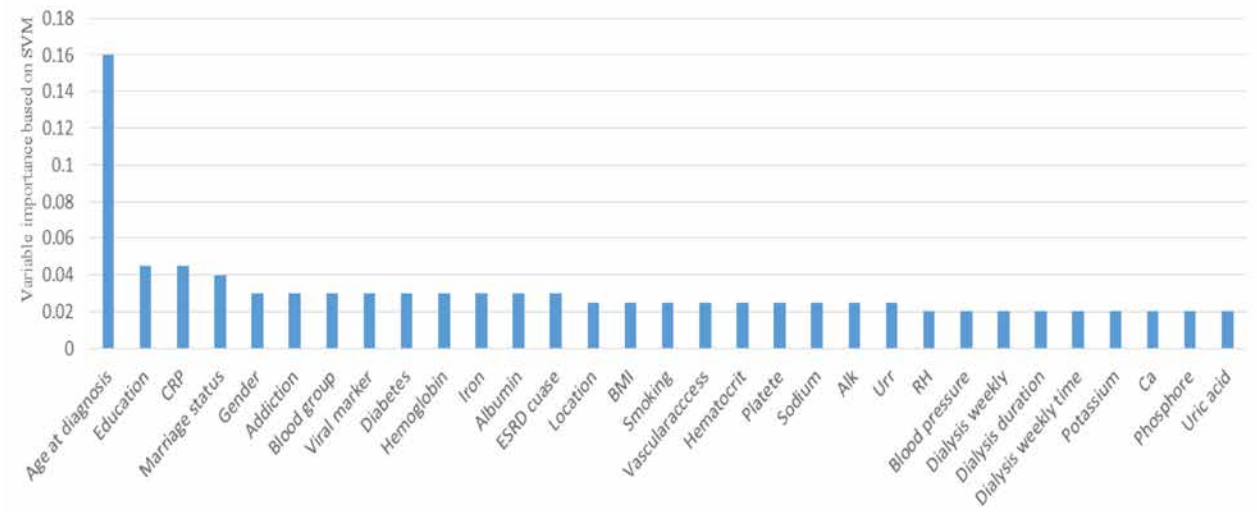


Fig. 3. Importance of variables estimated by the support vector machine.



the machine learning models in terms of total accuracy (0.71), sensitivity (0.69), positive likelihood ratio (2.48), and negative likelihood ratio (0.43). The specificity of LR and logistic regression was the same (0.72) over 100 repetitions. However, among the three machine learning methods, the performance of the SVM (sensitivity compared to specificity) was closest to the LR (positive likelihood ratio (2.25) and negative likelihood ratio (0.45) with a specificity of 0.70. Therefore, the SVM and LR were compatible and showed almost similar performance in predicting death in hemodialysis patients.

Discussion

In this study, we investigated the performance of the three most widely used machine learning methods of decision tree, neural network, and support vector machine in the prediction of death in hemodialysis patients in an Iranian population and compared it with that of the classical logistic regression. Our results showed that the prediction accuracy of the SVM was closest to the logistic regression. Our results showed that the discriminative performances of the used machine learning methods were equivalent to that of the stepwise LR method which is commonly applied for this purpose. So, these methods could be used successfully in detecting death in hemodialysis patients with clinical measurements, and laboratory tests. Other studies showed that the machine learning methods performed as good as the LR [10], which is in agreement with our results. However, other studies indicated machine learning methods of SVM, NN and DT outperformed logistic regression in terms of prediction performance [10, 11, 14, 23]. In those studies, Tab. III. Logistic regression model.

Variables	B	OR*	Wald	P-value
Gender:				
female	-	-	-	-
male	0.428	1.534	6.459	0.011
Age at diagnosis	-0.087	0.917	135.731	0.000
Addiction:				
yes	-	-	-	-
no	0.703	2.019	8.792	0.003
Iron	0.002	1.002	7.040	0.008
Urea reduction ratio	1.562	4.767	6.120	0.013
CRP:				
yes	-	-	-	-
no	1.595	4.930	62.731	0.000

*OR: Odds Ratio which is calculated as Exp (β).

Tab. IV. Mean and standard deviation of total accuracy, sensitivity, specificity, positive likelihood ratio and negative likelihood ratio for DT, NN, SVM and LR.

Models	Total accuracy		Sensitivity		Specificity		Positive likelihood ratio		Negative likelihood ratio	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Decision tree	0.66	0.05	0.65	0.06	0.67	0.05	2.03	0.47	0.53	0.11
Neural network	0.61	0.05	0.58	0.09	0.72	0.07	1.7	0.38	0.65	0.17
Support vector machine	0.68	0.04	0.66	0.07	0.70	0.04	2.25	0.45	0.48	0.10
Logistic regression	0.71	0.05	0.69	0.07	0.72	0.04	2.48	0.53	0.43	0.11

there were imbalanced data such that a large proportion of the samples did not experience the event of interest. But in our study the number of observed deaths was large, and the data were almost balanced. This might indicate that machine learning methods might be more useful in the case of imbalanced data. It is also possible that the effect of the covariates on the binary outcome in this study be linear as the traditional LR assumed the linear effects of covariates and a limited number of nonlinear effects. LR also performs well in large sample sizes. However, the SVM can handle complex relationships between variables and all interactions simultaneously even with a small sample size data set.

In this study, we also investigated the predictors of death in hemodialysis patients according to the model with a better performance. Therefore, the LR with simple interpretations was used to interpret the results. In spite of the considerable improvements in medical and technical support, the rate of mortality in these patients is high. We found that female gender, addiction, increasing age at diagnosis, low iron level, CRP positive and low URR were the most important predictors of death in HD patients in the present study. These novel findings may have important clinical and public health implications since they can be used for designing preventive interventions to reduce death among these patients.

In the present study female gender was independently associated with an increase of mortality. This finding was in line with Ratna Prabha et al. [24] and in contrast to those of Depner et al. [25] and USRDS data [14]. More access of males to health resources and a lower rate of CKD risk factors including diabetes and hypertension in males may be in connection with their better survival. Also, we found that addiction is another predictor of death in these patients. Side effects of drug addiction in different aspects of health are widely accepted [26, 27].

Increasing age at diagnosis in the present study was associated with rising odds of death in HD patients. Coric et al. [8] achieved a similar result in their study. Fatal infections are usually more common in elderly patients due to the weakness of their immune system. These patients have a higher rate of mortality due to infections and cardiovascular disease [9].

Among clinical predictors, being CRP positive was strongly associated with mortality. This positive relation has been previously identified [28, 29]. Lseki et al. in their study showed that regardless of serum albumin and other possible confounders, CRP is a significant predictor of death in HD patients [30]. Inflammation usually is in relation to insulin resistance, oxidative stress, wasting, infections, and endothelial dysfunction [25]. In the present study, the increase in iron supplement was associated with better survival. Some factors including inadequate intake of iron due to loss of appetite, continuous blood sampling for biochemical testing, and chronic iron loss through intestinal hemorrhage induced by uremic platelet dysfunction are associated with iron deficiency in the HD patients [31]. Motonishi et al. in their study showed that iron deficiency is a critical risk factor for deterioration of physical and mental conditions in maintenance HD patients [32]. However, studies in regard to the relationship between iron supplementation and death in HD patients are rare and conflicting [33, 34].

Regarding the role of adequate URR on the lower mortality rates in HD patients, consistent with our finding many studies have shown that dose of dialysis is a strong predictor of patient mortality through the highest ranges of URR recorded (i.e., URR > 75) [35, 36].

This study had several limitations. First, because of the retrospective design of the study, verifying quality control of the data was not possible. Second, the influence of risk factors on patients' death is time-varying over time, while we assessed only the influence of baseline patient features, our effect estimates may underestimate the association between mortality and investigated variables. Third, that the best performance of the LR in our study is not enough to say that LR is the best in the prediction of death among the patients with CKD, because the performance of machine learning methods of SVMs, NNs, DT, and logistic regression are bound to change depending on the situation of the dataset. Another potential limitation of the present study is that addiction and smoking status of the patients were based on their self-report and therefore, this information was prone to information bias and finally the quality of the services and technology may vary over time, and also the quality of service provision in the dialysis wards of hospitals is not the same, which could not be considered in this study.

Results of the study indicated that logistic regression had the best performance in comparison to other methods for predicting death among HD patients. According to this model female gender, increasing age at diagnosis, low iron level, CRP positive and low URR were the main predictors of death in hemodialysis patients.

Ethical statement

Ethic code: IR.UMSHA.REC.1399.029), research ID: 9911218196.

Acknowledgements

Funding sources: the Deputy of Research and Technology, Hamadan University of Medical Sciences funded this study.

The Authors would like to thank all contributors and participants who make this study project possible as well as the personnel of the hemodialysis wards in hospitals of Hamadan province for their kind collaboration.

Conflicts of interest statement

The authors declare no conflict of interest.

Authors' contributions

SKH: Study conception and Design; Analysis and Interpretation of the data; Drafting of the manuscript.

SNGH: Study conception and Design; Data extraction; Drafting of the manuscript.

VRD: Study conception and Design; Interpretation of the data; Drafting of the manuscript.

All authors approved the final version of the manuscript for publication.

References

[1] Abubakar I, Tillmann T, Banerjee A. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet 2015;385:117-71. https://doi.org/10.1016/S0140-6736(14)61682-2

[2] Bello AK, Levin A, Tonelli M, Okpechi IG, Feehally J, Harris D, Jindal K, Salako BL, Rateb A, Osman MA, Qarni B. Assessment of global kidney health care status. Jama 2017;317:1864-81. https://doi.org/10.1001/jama.2017.4046

[3] Trillini M, Perico N, Remuzzi G. Epidemiology of end-stage renal failure: the burden of kidney diseases to global health. In: Kidney Transplantation, Bioengineering and Regeneration: Elsevier 2017, pp. 5-11. https://doi.org/10.1038/ki.2014.419

[4] Floege J, Gillespie IA, Kronenberg F, Anker SD, Gioni I, Richards S, Pisoni RL, Robinson BM, Marcelli D, Froissart M, Eckardt KU. Development and validation of a predictive mortality risk score from a European hemodialysis cohort. Kidney Int 2015;87:996-1008. https://doi.org/10.1038/ki.2014.419

[5] Onofriescu M, Siriopol D, Voroneanu L, Hogas S, Nistor I, Apetrii M, Florea L, Veisa G, Mititiuc I, Kanbay M, Sascau R. Overhydration, cardiac function and survival in hemodialysis patients. PLoS One 2015;10:e0135691. https://doi.org/10.1371/journal.pone.0135691

[6] Dekker M, Marcelli D, Canaud B, Konings CJ, Leunissen K, Levin NW, Carioni P, Maheshwari V, Raimann JG, van der Sande FM, Usvyat LA. Unraveling the relationship between mortality, hyponatremia, inflammation and malnutrition in hemodialysis patients: results from the international MONDO initiative. Eur J Clin Nutr 2016;70:779. https://doi.org/10.1038/ejcn.2016.49

[7] Cooper BA, Branley P, Bulfone L, Collins JF, Craig JC, Fraenkel MB, Harris A, Johnson DW, Kesselhut J, Li JJ, Luxton G. A randomized, controlled trial of early versus late initiation of dialysis. N Engl J Med Overseas Ed 2010;363:609-19. https://doi.org/10.1128/JCM.03584-13

[8] Coric A, Resic H, Celik D, Masnic F, Ajanovic S, Prohic N, Beciragic A, Grosa E, Smajlovic A, Mujakovic A. Mortality in hemodialysis patients over 65 years of age. Mater Sociomed 2015;27:91. https://doi.org/10.5455/msm.2015.27.91-94

[9] Xue JL, Ma JZ, Louis TA, Collins AJ. Forecast of the number of patients with end-stage renal disease in the United States to the year 2010. J Am Soc Nephrol 2001;12:2753-8.

[10] Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clin Epidemiol Glob Health 2019;7:293-9. https://doi.org/10.1016/j.cegh.2018.10.003

[11] Tapak L, Hamidi O, Amini P, Poorolajal J. Prediction of kidney graft rejection using artificial neural network. Healthc Inform Res 2017;23:277-84. https://doi.org/10.4258/hir.2017.23.4.277

[12] Tapak L, Sheikh V, Jenabi E, Khazaei S. Predictors of mortality among hemodialysis patients in Hamadan province using random survival forests. J Prev Med Hyg 2020;61:E482. https://doi.org/10.15167/2421-4248/jpmh2020.61.3.1421

[13] Dekamin A, Shaibatalhamdi A. Real-data comparison of data mining methods in prediction of coronary artery disease in Iran. Journal of Health Management & Informatics 2017;4:87-94.

[14] Hamidi O, Poorolajal J, Farhadian M, Tapak L. Identifying im-

portant risk factors for survival in kidney graft failure patients using random survival forests. Iran J Public Health 2016;45:27.

[15] Chiang P-Y, Chao PC-P, Tu T-Y, Kao Y-H, Yang C-Y, Tarng D-C, Wey CL. Machine learning classification for assessing the degree of stenosis and blood flow volume at arteriovenous fistulas of hemodialysis patients using a new photoplethysmography sensor device. Sensors (Basel) 2019;19:3422. https://doi.org/10.3390/s19153422

[16] Tsai Y-T, Yang F-J, Lin H-M, Yeh J-C, Cheng B-W. Constructing a prediction model for physiological parameters for malnutrition in hemodialysis patients. Scientific Reports 2019;9:1-6. https://doi.org/10.1038/s41598-019-47130-7

[17] Martínez-Martínez JM, Escandell-Montero P, Barbieri C, Soria-Olivas E, Mari F, Martínez-Sober M, Amato C, López AJ, Bassi M, Magdalena-Benedito R, Stopper A. Prediction of the hemoglobin level in hemodialysis patients using machine learning techniques. Comput Methods Programs Biomed 2014;117:208-17. https://doi.org/10.1016/j.cmpb.2014.07.001

[18] Lacson R. Predicting hemodialysis mortality utilizing blood pressure trends. AMIA Annu Symp Proc 2008;2008:369-73.

[19] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Comput Surv 2009;41:15. https://doi.org/10.1145/1541880.1541882

[20] IBM. IBM Knowledge Center. Available from: https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/anomalydetectionnode_general.htm

[21] Han J, Kamber M, Pei J. Data mining. Concepts and techniques. Third edition. The Morgan Kaufmann Series in Data Management Systems 2011, pp. 83-124.

[22] Duda RO, Hart PE, Stork DG. Pattern classification: John Wiley & Sons 2012.

[23] Najafi-Ghobadi S, Najafi-Ghobadi K, Tapak L, Aghaei A. Application of data mining techniques and logistic regression to model drug use transition to injection: a case study in drug use treatment centers in Kermanshah Province, Iran. Subst Abuse Treat Prev Policy 2019;14:55. https://doi.org/10.1186/s13011-019-0242-1

[24] Prabha DR, Prasad G. Predictors of mortality among patients on maintenance hemodialysis. J Dr NTR Univ Health Sci 2016;5:255. https://doi.org/10.4103/2277-8632.196558

[25] Depner T, Daugirdas J, Greene T, Allon M, Beck G, Chumlea C, Delmez J, Gotch F, Kusek J, Levin N, Macon E. Dialysis dose and the effect of gender and body size on outcome in the HEMO Study Kidney Int 2004;65:1386-94. https://doi.org/10.1111/j.1523-1755.2004.00519.x

[26] Shedler J, Block J. Adolescent drug use and psychological health: a longitudinal inquiry. American Psychologist 1990;45:612. https://doi.org/10.1037/0003-066X.45.5.612

[27] Chen C-Y, Lin K-M. Health consequences of illegal drug use. Curr Opin Psychiatry 2009;22:287-92. https://doi.org/10.1097/YCO.0b013e32832a2349

[28] Yeun JY, Levine RA, Mantadilok V, Kaysen GA. C-reactive protein predicts all-cause and cardiovascular mortality in hemodialysis patients. Am J Kidney Dis 2000;35:469-76. https://doi.org/doi:10.1016/S0272-6386(00)70200-9

[29] DeFilippi C, Wasserman S, Rosanio S, Tiblier E, Sperger H, Tocchi M, Christenson R, Uretsky B, Smiley M, Gold J, Muniz H. Cardiac troponin T and C-reactive protein for predicting prognosis, coronary atherosclerosis, and cardiomyopathy in patients undergoing long-term hemodialysis. Jama 2003;290:353-9. https://doi.org/10.1001/jama.290.3.353

[30] Iseki K, Tozawa M, Yoshi S, Fukiyama K. Serum C-reactive protein (CRP) and risk of death in chronic dialysis patients. Nephrol Dial Transplant 1999;14:1956-60. https://doi.org/10.1093/ndt/14.8.1956

[31] Macdougall IC, Dahl NV, Bernard K, Li Z, Batycky A, Strauss WE. The Ferumoxytol for Anemia of CKD Trial (FACT) - a randomized controlled trial of repeated doses of ferumoxytol or iron sucrose in patients on hemodialysis: background and rationale. BMC Nephrol 2017;18:117. https://doi.org/10.1186/s12882-017-0523-8

[32] Motonishi S, Tanaka K, Ozawa T. Iron deficiency associates with deterioration in several symptoms independently from hemoglobin level among chronic hemodialysis patients. PloS One 2018;13(8). https://doi.org/10.1371/journal.pone.0201662

[33] Zitt E, Sturm G, Kronenberg F, Neyer U, Knoll F, Lhotta K, Weiss G. Iron supplementation and mortality in incident dialysis patients: an observational study. PLoS One 2014;9:e114144. https://doi.org/10.1371/journal.pone.0114144

[34] Feldman HI, Santanna J, Guo W, Furst H, Franklin E, Joffe M, Marcus S, Faich G. Iron administration and clinical outcomes in hemodialysis patients. J Am Soc Nephrol 2002;13:734-44.

[35] Wolfe RA, Ashby VB, Daugirdas JT, Agodoa LY, Jones CA, Port FK. Body size, dose of hemodialysis, and mortality. Am J Kidney Dis 2000;35:80-8. https://doi.org/10.1016/S0272-6386(00)70305-2

[36] Port FK, Ashby VB, Dhingra RK, Roys EC, Wolfe RA. Dialysis dose and body mass index are strongly associated with survival in hemodialysis patients. J Am Soc Nephrol 2002;13:1061-6.

Received on October 31, 2020. Accepted on December 21, 2021.

Correspondence: Vajihe Ramezani-Doroh, Hamadan University of Medical Sciences, Shahid Fahmide St., Pazhuhesh Square., Hamadan, Iran - Tel.: +98 9175375707 - E-mail: vajihe.r@gmail.com

How to cite this article: Khazaei S, Najafi-Ghobadi S, Ramezani-Doroh V. Construction data mining methods in the prediction of death in hemodialysis patients using support vector machine, neural network, logistic regression and decision tree. J Prev Med Hyg 2021;62:E222-E230. https://doi.org/10.15167/2421-4248/jpmh2021.62.1.1837